

$$U_{it} = \beta' \mathbf{x}_{it} + \varepsilon_{it},$$

EDITED BY

BADI H.

**BALTAGI**

$$Cov(\varepsilon_{it}, \varepsilon_{js}) = \mathbf{1}[i=j]\sigma_{ts}.$$

$$y_{it} = \mathbf{1}[U_{it} > 0].$$

$$\Delta_x(\mathbf{x}) = E_u \left[ \frac{\partial B(\beta' \mathbf{x} + \sigma u)}{\partial \mathbf{x}} \right] = \left[ \frac{\partial E_u[B(\beta' \mathbf{x} + \sigma u)]}{\partial \mathbf{x}} \right]$$

$$\Delta_x(\mathbf{x}) = \frac{\partial \Phi \left( \frac{\beta' \mathbf{x}}{1 + \sigma^2} \right)}{\partial \mathbf{x}} = \frac{1}{\sqrt{1 + \sigma^2}} \beta \phi \left( \frac{\beta' \mathbf{x}}{\sqrt{1 + \sigma^2}} \right)$$

≡ The Oxford Handbook of  
PANEL DATA

THE OXFORD HANDBOOK OF

# PANEL DATA

## **CONSULTING EDITORS**

Michael Szenberg  
*Lubin School of Business, Pace University*

Lall Ramrattan  
*University of California, Berkeley Extension*

THE OXFORD HANDBOOK OF

---

# PANEL DATA

---

*Edited by*

BADI H. BALTAGI

OXFORD  
UNIVERSITY PRESS



Oxford University Press is a department of the University of Oxford. It furthers the University's objective of excellence in research, scholarship, and education by publishing worldwide.

Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi  
Kuala Lumpur Madrid Melbourne Mexico City Nairobi  
New Delhi Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece  
Guatemala Hungary Italy Japan Poland Portugal Singapore  
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trade mark of Oxford University Press  
in the UK and in certain other countries

Published in the United States of America by  
Oxford University Press  
198 Madison Avenue, New York, NY 10016,

© Oxford University Press 2015

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, without the prior permission in writing of Oxford University Press, or as expressly permitted by law, by license, or under terms agreed with the appropriate reproduction rights organization. Inquiries concerning reproduction outside the scope of the above should be sent to the Rights Department, Oxford University Press, at the address above.

You must not circulate this work in any other form  
and you must impose this same condition on any acquirer.

Library of Congress Cataloging-in-Publication Data  
The Oxford handbook of panel data / edited by Badi H. Baltagi.  
pages cm  
Includes bibliographical references and index.  
ISBN 978-0-19-994004-2 (alk. paper)  
1. Panel analysis. 2. Econometrics. I. Baltagi, Badi H. (Badi Hani)  
H61.26.O94 2015  
330.01'5195—dc23  
2014024763

1 3 5 7 9 8 6 4 2

Printed in the United States of America  
on acid-free paper

# CONTENTS

---

<i>List of Contributors</i>	vii
<i>Introduction</i>	xi
BADI H. BALTAGI	

## PART I PANEL DATA MODELS AND METHODS

1. Large Panel Data Models with Cross-Sectional Dependence: A Survey ALEXANDER CHUDIK AND M. HASHEM PESARAN	3
2. Panel Cointegration IN CHOI	46
3. Dynamic Panel Data Models MAURICE J.G. BUN AND VASILIS SARAFIDIS	76
4. Incidental Parameters and Dynamic Panel Modeling HYUNGSIK ROGER MOON, BENOIT PERRON, AND PETER C.B. PHILLIPS	111
5. Unbalanced Panel Data Models with Interactive Effects JUSHAN BAI, YUAN LIAO, AND JISHENG YANG	149
6. Panel Data Models for Discrete Choice WILLIAM GREENE	171
7. Panel Conditional and Multinomial Logit Estimators MYOUNG-JAE LEE	202
8. Count Panel Data A. COLIN CAMERON AND PRAVIN K. TRIVEDI	233
9. Treatment Effects and Panel Data MICHAEL LECHNER	257

10.	Nonparametric Panel Data Regression Models YIGUO SUN, YU YVETTE ZHANG, AND QI LI	285
11.	Measurement Error in Panel Data ERIK MEIJER, LAURA SPIERDIJK, AND TOM WANSBEEK	325
12.	Spatial Panel Data Models LUNG-FEI LEE AND JIHAI YU	363
13.	Random Coefficients Models In Panels CHENG HSIAO	402
14.	Robust Panel Data Methods and Influential Observations BADI H. BALTAGI AND GEORGES BRESSON	418

## PART II PANEL DATA APPLICATIONS

15	The Analysis Of Macroeconomic Panel Data JÖRG BREITUNG	453
16.	Cohort Data in Health Economics STEPHANIE VON HINKE KESSLER SCHOLDER AND ANDREW M. JONES	493
17.	Panel Data And Productivity Measurement ROBIN C. SICKLES, JIAQI HAO AND CHENJUN SHANG	517
18.	Panel Data Discrete Choice Models of Consumer Demand MICHAEL P. KEANE	548
19.	Panel Econometrics of Labor Market Outcomes THOMAS J. KNIESNER AND JAMES P. ZILIAK	583
20.	Panel Data Gravity Models of International Trade BADI H. BALTAGI, PETER EGGER, AND MICHAEL PFAFFERMAYR	608
	<i>Author Index</i>	643
	<i>Subject Index</i>	661

## LIST OF CONTRIBUTORS

---

**Jushan Bai** is a Professor in the Department of Economics at Columbia University.

**Badi H. Baltagi** is Distinguished Professor in the Department of Economics and Senior Research Fellow in the Center for Policy Research at Syracuse University.

**Jörg Breitung** is Professor in the Department of Economics at the University of Bonn, Germany.

**Georges Bresson** is a Professor in the Department of Economics at Université Paris II, Sorbonne Universités, France.

**Maurice J. G. Bun** is Associate Professor in the Faculty of Economics and Business at Amsterdam School of Economics of University of Amsterdam and Research Fellow at the Tinbergen Institute, Netherlands.

**A. Colin Cameron** is Professor in the Department of Economics at University of California, Davis.

**In Choi** is Professor in the Department of Economics at Sogang University, South Korea.

**Alexander Chudik** is Senior Research Economist at the Federal Reserve Bank of Dallas.

**Peter Egger** is Professor in the Department of Applied Economics at the Swiss Federal Institute of Technology (ETH) Zurich, Switzerland.

**William Greene** is Robert Stansky Professor in the Department of Economics at Stern School of Business, New York University.

**Jiaqi Hao**, US Card Decision Science Partnership, Capital One

**Stephanie von Hinke Kessler Scholder** is Lecturer in the Department of Economics and Related Studies at University of York, UK.

**Cheng Hsiao** is Professor in the Department of Economics at the University of Southern California.

**Andrew M. Jones** is Professor in the Department of Economics and Related Studies at the University of York.

**Michael P. Keane** is the Nuffield Professor of Economics at the University of Oxford, UK.

**Thomas J. Kniesner** is Krisher Professor of Economics in the Maxwell School of Citizenship and Public Affairs and Senior Research Associate at Center for Policy Research at Syracuse University. Also, University Professor of Economics, Claremont Graduate University.

**Michael Lechner** is Professor of Econometrics in the Swiss Institute for Empirical Economic Research (SEW) at University of St. Gallen, Switzerland.

**Lung-fei Lee** is Professor in the Department of Economics at Ohio State University.

**Myoung-jae Lee** is Professor in the Department of Economics at Korea University.

**Qi Li** is Hugh Roy Cullen Professor in Liberal Arts in the Department of Economics at Texas A&M University.

**Yuan Liao** is Assistant Professor in the Department of Mathematics at University of Maryland.

**Erik Meijer** is Senior Economist at Center for Economic and Social Research at the University of Southern California and an Economist at the RAND Corporation.

**Hyungsik Roger Moon** is Professor in the Department of Economics at University of Southern California and at Yonsei University, South Korea.

**Benoit Perron** is Professor in the Department of Economics at University of Montreal, Canada.

**M. Hashem Pesaran** is John Elliott Chair in Economics and Professor of Economics at University of Southern California.

**Michael Pfaffermayr** is Professor in the Department of Economic Theory, Economic Policy and Economic History at University of Innsbruck.

**Peter C. B. Phillips** is Sterling Professor of Economics and Professor of Statistics at Cowles Foundation for Research in Economics at Yale University.

**Vasilis Sarafidis** is Professor in the Department of Econometrics & Business Statistics at Monash University, Australia.

**Chenjun Shang** is a PhD student in the Department of Economics at Rice University.

**Robin C. Sickles** is Reginald Henry Hargrove Chair of Economics, Department of Economics at Rice University.

**Laura Spierdijk** is Professor in the Faculty of Economics and Business at University of Groningen, Netherlands.

**Yiguo Sun** is Professor in the Department of Economics at University of Guelph, Canada.

**Pravin K. Trivedi** is Professor in the School of Economics at Queensland University, Australia.

**Tom Wansbeek** is Professor in the Faculty of Economics and Business at the University of Groningen, Netherlands.

**Jisheng Yang** is a faculty member of the School of Economics at the Huazhong University of Science and Technology, China.

**Jihai Yu** is Associate Professor in the Guanghua School of Management at Peking University, China.

**Yu Yvette Zhang** is Visiting Assistant Professor in the Department of Agricultural Economics at Texas A&M University.

**James P. Ziliak** is Carol Martin Gatton Chair in Microeconomics at University of Kentucky and Founding Director of University of Kentucky Center for Poverty Research.



---

# INTRODUCTION

---

BADI H. BALTAGI

Panel data econometrics has evolved rapidly over the last three decades. Micro and macro panels are increasing in numbers and availability and methods to deal with these data are in high demand from practitioners. Despite the availability of two chapters on panel data in the *Handbook of Econometrics* dating to Chamberlain (1984) and Arellano and Honore (2001) and a recent handbook by Matyas and Sevestre (2008) on panel data, as well as several textbooks including Hsiao (2003), Wooldridge (2010), and Baltagi (2013), there is still need to study the important new contributions in this field, highlighting them to researchers and practitioners. Dynamic panel data estimation, nonlinear panel data methods, and the phenomenal growth in non-stationary panel data econometrics make this an exciting area of research. Applications in finance, development, trade, and marketing and micro as well as macroeconomics applications including health, labor, and consumer economics attest to the usefulness of these methods in applied economics.

This Oxford handbook examines new developments in theory and applications in panel data. It includes basic topics like non-stationary panels, cointegration in panels, multifactor panel models, panel unit roots, measurement error in panels, incidental parameters and dynamic panels, spatial panels, nonparametric panel data, random coefficients, treatment effects, sample selection, count panel data, limited dependent variable panel models, unbalanced panel models with interactive effects, and influential observations in panel data. Part II of this handbook targets applications of panel data in economics, including health, labor, marketing, trade, productivity, and macro applications in panels.

This handbook should be of interest both to those who are relatively new to the area and to those wishing to extend their knowledge to the frontier. Below is the list of contents and contributors, as well as a summary of each chapter.

Alexander Chudik and M. Hashem Pesaran provide a review of the recent literature on estimation and inference in large panel data models with cross-sectional dependence. The chapter reviews the concepts of weak and strong cross-sectional dependence

and discusses the exponent of cross-sectional dependence that characterizes the different degrees of cross-sectional dependence. It considers a number of alternative estimators for static and dynamic panel data models, distinguishing between factor and spatial models of cross-sectional dependence. The chapter also provides an overview of tests of independence and weak cross-sectional dependence.

In Choi surveys the literature on panel cointegration, complementing earlier review papers by Baltagi and Kao (2000), Choi (2006), and Breitung and Pesaran (2008). First, cointegrating panel regressions for cross-sectionally independent and correlated panels are discussed. Next, tests for panel cointegration are introduced. Three groups of tests are examined: residual-based tests for the null of no cointegration, residual-based tests for the null of cointegration, and tests based on vector auto-regression.

Maurice J. G. Bun and Vasilis Sarafidis review the recent literature on dynamic panel data models with a short time span and a large cross-section dimension. First, they give a broad overview of available inference methods placing emphasis on Generalized Method of Moments (GMM). Then, they discuss the assumption of mean stationarity underlying the system GMM estimator. The consequence of deviations from mean stationarity and how to test for it.

Hyungsik Roger Moon, Benoit Perron, and Peter C. B. Phillips study incidental parameters and dynamic panel modeling. The challenges presented by incidental parameters are particularly acute in dynamic panels where behavioral effects over time are being measured in conjunction with individual effects. Maximum likelihood estimation of such models leads to inconsistent estimates of the parameters that govern the dynamics. The problems are even more acute in non-stationary panels and panels with incidental trends. This chapter reviews some of the established methodology in the field, the ground that has been won in developing a theory of inference for dynamic panel modeling, as well as some exciting ongoing research that seeks to address some of the many remaining challenges.

Jushan Bai, Yuan Liao, and Jisheng Yang focus on unbalanced panel data models with interactive effects. They propose new algorithms to estimate the model when missing data are present, allowing for various types of missing patterns such as the block missing, regular missing, and random missing. They adapt the expectation maximization (EM) algorithm. In particular, when common factors are deterministic (smooth in  $t$ ), the functional principal components method can be applied. Their proposed algorithms also work for stochastic (nonsmooth) common factors, and therefore are applicable to a broad class of panel data models. Extensions to the dynamic model with instrumental variables are discussed.

William Greene surveys panel data models for discrete choice and focuses on modeling cross-sectional heterogeneity in three foundational discrete choice settings: binary, ordered multinomial, and unordered multinomial models. While, Myoung-jae Lee studies panel conditional logit estimators (PCLEs) mainly for binary responses. This is then generalized for ordered discrete responses and multinomial responses. The review covers the underlying theories leading to the PCLEs as well as their practical implementations.

---

A. Colin Cameron and Pravin K. Trivedi survey panel data methods when the dependent variable is a count taking on nonnegative integer values. The focus is on panels with a short time span and a large cross-section dimension. The survey covers both static and dynamic models with random and fixed effects. It surveys quasi-maximum likelihood methods based on Poisson, as well as negative binomial models, finite mixture models, and hurdle models.

Michael Lechner studies treatment effects and panel data. The first part of this chapter focuses on the static treatment model. Lechner shows how panel data can be used to improve the credibility of matching and instrumental variable estimators. In addition to improving the credibility of static causal models, panel data may allow one to credibly estimate dynamic causal models. Lechner considers three approaches that figure prominently in the applied literature. Starting with matching and regression type methods, differences-in-differences methods, and instrumental variable estimation.

Yiguo Sun, Yu Yvette Zhang, and Qi Li selectively review some recent developments using nonparametric panel data regression models. This includes different estimation methods developed for nonparametric panel data mean regression models, some introduction on nonparametric panel data quantile regression models, nonseparable nonparametric panel data models, nonparametric poolability tests, and cross-sectional independence tests.

Erik Meijer, Laura Spierdijk, and Tom Wansbeek explore the basics of measurement error in the simplest possible panel data model with measurement error. This chapter shows that Ordinary Least Squares is inconsistent, in very much the same way as with a single cross-section, but one can exploit the panel character of the data to decrease the inconsistency. There are various ways to eliminate correlated (or “fixed”) effects, all leading to estimators that are inconsistent in the presence of measurement error, but in different ways. Short versus long differences are considered. The inconsistency decreases when differences are taken far apart in time. This chapter also discusses the random effects model and concludes with a discussion of the identification of this model. Restrictions on the various parameter matrices can be helpful in achieving identification and thus the construction of consistent estimators.

Lung-fei Lee and Jihai Yu survey recent developments in spatial panel data models. They first investigate various model specifications for both static and dynamic models with spatial interactions. Detailed estimation procedures such as maximum likelihood estimation and generalized method of moments are studied for these static and dynamic models.

Cheng Hsiao focuses on random coefficients models in panels. This chapter provides an example demonstrating the importance of taking into account parameter heterogeneity in empirical analysis. It also provides a test for heterogeneity and whether we should treat unobserved differences as fixed and different or random.

It is well known (see Huber 1981) that only 3% of atypical values in the sample is sufficient to reveal the weakness of classical consistent estimators. Panel data has a large number of observations but if 3% is enough to spoil the soup, this large number of good points is far from drowning a few bad points. Despite their relevance, studies of

robust methods using panel data are a few. Badi H. Baltagi and Georges Bresson survey the robust panel data methods and influential observations literature. They present robust estimators of linear static panel data models, followed by robust instrumental variables (IV) panel data estimators, robust Hausman and Taylor (1981) estimation method, dynamic panel data GMM methods, as well as, robust nonlinear panel data methods. Finally they consider detection of influential observations and outliers in the context of panel data.

Jörg Breitung reviews a wide range of recently developed econometric tools for analyzing macroeconomic panel data. He argues that the analysis of macroeconomic panels is still in its infancy. Modeling observable and unobservable interactions across sectors, regions, and economies is important for valid statistical inference. He reviews the econometric panel tools used in dynamic panels, GMM when  $T$  is large, the problem of many instruments, bias-corrected estimators, maximum likelihood estimators, Vector Auto Regression (VAR) models, heterogeneous panel models, cross-section dependence, robust standard errors, SUR-GLS, common factors, and global VAR to mention a few. Breitung recommends the global VAR approach as a promising direction for future research, as it allows one to model typical features of macroeconomic data like non-stationarity, parameter heterogeneity, and cross-section dependence.

Stephanie von Hinke Kessler Scholder and Andrew M. Jones study longitudinal birth cohorts in the United Kingdom. In particular, they consider the scientific rationale for studying birth cohorts and refer to some key papers in economics that use these data. They compare the use of birth cohorts to other longitudinal research designs and review some of the econometric methods that have been applied to these cohort studies.

Robin C. Sickles, Jiaqi Hao, and Chenjun Shang discuss panel data and productivity analysis in applied economic modeling. They formulate methods to decompose productivity growth based on a Solow-type residual into innovation and catch-up, the latter referred to as technical efficiency change in the stochastic frontier literature. They point out why panel data are needed to identify and measure productive efficiency and innovation. The focus of this chapter is on aggregate productivity, which emphasizes the parallels between efficiency, economic growth and development, and the panel data econometrics literature.

Michael P. Keane focuses on panel data discrete choice models of consumer demand, more specifically, scanner panel data, which shows substantial persistence by consumers in terms of brand loyalty. Uncovering whether state dependence exists is of great importance in both marketing and industrial organization. Keane discusses the econometric methods needed to estimate such models. More specifically, the theoretical issues involved in distinguishing state dependence from heterogeneity. Keane also discusses the empirical work on state dependence and/or choice dynamics.

Thomas J. Kniesner and James P. Ziliak demonstrate how the panel data techniques labor economists use manifest themselves in an applied setting, specifically the canonical hedonic labor-market equilibrium model of the wage-fatal risk trade-off.

Badi H. Baltagi, Peter Egger, and Michael Pfaffermayr focus on the estimation of gravity models of bilateral trade of goods (or services) and other bilateral international outcomes such as foreign direct investment or migration stocks or flows. Stochastic versions of this model have become the empirical workhorse to study gravity models since the nineteenth century. The estimates obtained (especially on bilateral geographical distance) reflect some of the most robust relationships in empirical economics. This chapter discusses the application of panel econometric methods to gravity modeling. It also discusses single cross-section, as well as repeated cross-sections, of country pairs over time. The nature of the data calls for panel econometric methods due to their inherent double and even triple indexation.

## REFERENCES

---

- Arellano, M., and B. Honoré. 2001. "Panel data models: Some recent developments," Chapter 53 in J. Heckman and E. Leamer (eds.), *Handbook of Econometrics*, 3229–3296. Amsterdam: North-Holland.
- Baltagi, Badi H. 2013. *Econometric Analysis of Panel Data*, 5th edn. Chichester, England: John Wiley & Sons.
- Baltagi, B. H., and C. Kao. 2000. Nonstationary panels, cointegration in panels and dynamic panels: A survey. *Advances in Econometrics*, 15, 7–51.
- Breitung, J., and M. H. Pesaran. 2008. "Unit roots and cointegration in panels," Chapter 9 in L. Matyas and P. Sevestre (eds.), *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*, 279–322. Berlin: Springer.
- Chamberlain, G. 1984. "Panel data," Chapter 22 in Z. Griliches and M. Intriligator (eds.), *Handbook of Econometrics*, 1247–1318. Amsterdam: North-Holland.
- Choi, I. 2006. "Nonstationary panels," Chapter 13 in T.C. Mills and K. Patterson (eds.), *Palgrave Handbooks of Econometrics*, Vol. 1, pp. 511–539. Palgrave, Macmillan.
- Hsiao, C. 2003. *Analysis of Panel Data*. Cambridge: Cambridge University Press.
- Huber, P. J. 1981. *Robust Statistics, Series in Probability and Mathematical Statistics*, 1st edn. New York: John Wiley.
- Mátyás, L., and Sevestre, P. eds. 2008. *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*. Berlin: Springer.
- Wooldridge, J. M. 2010. *Econometric Analysis of Cross-Section and Panel Data*. Massachusetts: MIT Press.



P A R T I

---

**PANEL DATA MODELS  
AND METHODS**

---



## CHAPTER 1

---

# LARGE PANEL DATA MODELS WITH CROSS-SECTIONAL DEPENDENCE *A SURVEY*

---

ALEXANDER CHUDIK AND M. HASHEM PESARAN

## 1.1 INTRODUCTION

---

THIS chapter reviews econometric methods for large linear panel data models subject to error cross-sectional dependence. Early panel data literature assumed cross-sectionally independent errors and homogeneous slopes. Heterogeneity across units was confined to unit-specific intercepts, treated as fixed or random (see, e.g. the survey by Chamberlain 1984). Dependence of errors was only considered in spatial models, but not in standard panels. However, with an increasing availability of data (across countries, regions, or industries), the panel literature moved from predominantly micro panels, where the cross dimension ( $N$ ) is large and the time series dimension ( $T$ ) is small, to models with both  $N$  and  $T$  large, and it has been recognized that, even after conditioning on unit-specific regressors, individual units, in general, need not be cross-sectionally independent.

Ignoring cross-sectional dependence of errors can have serious consequences, and the presence of some form of cross-sectional correlation of errors in panel data applications in economics is likely to be the rule rather than the exception. Cross correlations of errors could be due to omitted common effects, spatial effects, or could arise as a result of interactions within socioeconomic networks. Conventional panel estimators, such as fixed or random effects, can result in misleading inference and even in inconsistent estimators, depending on the extent of the cross-sectional dependence and on whether the source generating the cross-sectional dependence (such as an unobserved common shock) is correlated with regressors (Phillips and Sul 2003; Andrews 2005; Phillips and Sul 2007; Sarafidis and Robertson 2009). Correlation across units

in panels may also have serious drawbacks on commonly used panel unit root tests, since several of the existing tests assume independence. As a result, when applied to cross-sectionally dependent panels, such unit root tests can have substantial size distortions (O'Connell 1998). If, however, the extent of cross-sectional dependence of errors is sufficiently weak, or limited to a sufficiently small number of cross-sectional units, then its consequences on conventional estimators will be negligible. Furthermore, the consistency of conventional estimators can be affected only when the source of cross-sectional dependence is correlated with regressors. The problems of testing for the extent of cross-sectional correlation of panel residuals and modelling the cross-sectional dependence of errors are therefore important issues.

In the case of panel data models where the cross-section dimension is short and the time series dimension is long, the standard approach to cross-sectional dependence is to consider the equations from different cross-sectional units as a system of seemingly unrelated regression equations (SURE), and then estimate it by Generalized Least Squares techniques (see Zellner 1962). This approach assumes that the source generating cross-sectional dependence is not correlated with regressors and this assumption is required for the consistency of the SURE estimator. If the time series dimension is not sufficiently large, and in particular if  $N > T$ , the SURE approach is not feasible.

Currently, there are two main strands in the literature for dealing with error cross-sectional dependence in panels where  $N$  is large, namely the spatial econometric and the residual multifactor approaches. The spatial econometric approach assumes that the structure of cross-sectional correlation is related to location and distance among units, defined according to a pre-specified metric given by a ‘connection or spatial’ matrix that characterizes the pattern of spatial dependence according to pre-specified rules. Hence, cross-sectional correlation is represented by means of a spatial process, which explicitly relates each unit to its neighbors (see Whittle (1954), Moran (1948), Cliff and Ord (1973 and 1981), Anselin (1988 and 2001), Haining (2003, Chapter 7), and the recent survey by Lee and Yu (2013)). This approach, however, typically does not allow for slope heterogeneity across the units and requires a priori knowledge of the weight matrix.

The residual multifactor approach assumes that the cross dependence can be characterized by a small number of unobserved common factors, possibly due to economy-wide shocks that affect all units albeit with different intensities. Geweke (1977) and Sargent and Sims (1977) introduced dynamic factor models, which have more recently been generalized to allow for weak cross-sectional dependence by Forni and Lippi (2001), Forni et al. (2000), and Forni et al. (2004). This approach does not require any prior knowledge regarding the ordering of individual cross-section units.

The main focus of this chapter is on estimation and inference in the case of large  $N$  and  $T$  panel data models with a common factor error structure. We provide a synthesis of the alternative approaches proposed in the literature (such as principal components and common correlated effects approaches), with a particular focus on key assumptions and their consequences from the practitioners' view point. In particular, we discuss robustness of estimators to cross-sectional dependence of errors,

consequences of coefficient heterogeneity, panels with strictly or weakly exogenous regressors, including panels with a lagged dependent variable, and we highlight how to test for residual cross-sectional dependence.

The outline of the chapter is as follows: an overview of the different types of cross-sectional dependence is provided in Section 1.2. The analysis of cross-sectional dependence using a factor error structure is presented in Section 1.3. A review of estimation and inference in the case of large panels with a multifactor error structure and strictly exogenous regressors is provided in Section 1.4, and its extension to models with lagged dependent variables and/or weakly exogenous regressors is given in Section 1.5. A review of the tests of error cross-sectional dependence in static and dynamic panels is presented in Section 1.6. Section 1.7 discusses the application of common correlated effects estimators and the tests of error cross-sectional dependence to unbalanced panels, and the final section concludes.

## 1.2 TYPES OF CROSS-SECTIONAL DEPENDENCE

---

A better understanding of the extent and nature of cross-sectional dependence of errors is an important issue in the analysis of large panels. This section introduces the notions of weak and strong cross-sectional dependence and the notion of the exponent of cross-sectional dependence to characterize the correlation structure of  $\{z_{it}\}$  over the cross-sectional dimension,  $i$ , at a given point in time,  $t$ . Consider the double index process  $\{z_{it}, i \in \mathbb{N}, t \in \mathbb{Z}\}$ , where  $z_{it}$  is defined on a suitable probability space, the index  $t$  refers to an ordered set such as time, and  $i$  refers to units of an unordered population. We make the following assumption:

**ASSUMPTION CSD.1:** For each  $t \in T \subseteq \mathbb{Z}$ ,  $\mathbf{z}_t = (z_{1t}, \dots, z_{Nt})'$  has mean  $E(\mathbf{z}_t) = \mathbf{0}$ , and variance  $Var(\mathbf{z}_t) = \boldsymbol{\Sigma}_t$ , where  $\boldsymbol{\Sigma}_t$  is an  $N \times N$  symmetric, nonnegative definite matrix. The  $(i,j)$ -th element of  $\boldsymbol{\Sigma}_t$ , denoted by  $\sigma_{ij,t}$ , is bounded such that  $0 < \sigma_{ii,t} \leq K$ , for  $i = 1, 2, \dots, N$ , where  $K$  is a finite constant independent of  $N$ .

Instead of assuming unconditional mean and variances, one could consider conditioning on a given information set,  $\Omega_{t-1}$ , for  $t = 1, 2, \dots, T$ , as done in Chudik et al. (2011). The assumption of zero means can also be relaxed to  $E(\mathbf{z}_t) = \boldsymbol{\mu}$  (or  $E(\mathbf{z}_t | \Omega_{t-1}) = \boldsymbol{\mu}_{t-1}$ ). The covariance matrix,  $\boldsymbol{\Sigma}_t$ , fully characterizes cross-sectional correlations of the double index process  $\{z_{it}\}$ , and this section discusses summary measures based on the elements of  $\boldsymbol{\Sigma}_t$  that can be used to characterize the extent of the cross-sectional dependence in  $\mathbf{z}_t$ .

Summary measures of cross-sectional dependence based on  $\boldsymbol{\Sigma}_t$  can be constructed in a number of different ways. One possible measure, that has received a great deal of attention in the literature, is the largest eigenvalue of  $\boldsymbol{\Sigma}_t$ , denoted by  $\lambda_1(\boldsymbol{\Sigma}_t)$ . See, for example, Bai and Silverstein (1998), Hachem et al. (2005) and Yin et al. (1988). However, the existing work in this area suggests that the estimates of  $\lambda_1(\boldsymbol{\Sigma}_t)$

based on sample estimates of  $\Sigma_t$  could be very poor when  $N$  is large relative to  $T$ , and consequently using estimates of  $\lambda_1(\Sigma_t)$  for the analysis of cross-sectional dependence might be problematic in cases where  $T$  is not sufficiently large relative to  $N$ . Accordingly, other measures based on matrix norms of  $\Sigma_t$  have also been used in the literature. One prominent choice is the absolute column sum matrix norm, defined by  $\|\Sigma_t\|_1 = \max_{j \in \{1, 2, \dots, N\}} \sum_{i=1}^N |\sigma_{ij,t}|$ , which is equal to the absolute row sum matrix norm of  $\Sigma_t$ , defined by  $\|\Sigma_t\|_\infty = \max_{i \in \{1, 2, \dots, N\}} \sum_{j=1}^N |\sigma_{ij,t}|$ , due to the symmetry of  $\Sigma_t$ . It is easily seen that  $|\lambda_1(\Sigma_t)| \leq \sqrt{\|\Sigma_t\|_1 \|\Sigma_t\|_\infty} = \|\Sigma_t\|_1$ . See Chudik et al. (2011). Another possible measure of cross-sectional dependence can be based on the behavior of (weighted) cross-sectional averages, which is often of interest in panel data econometrics, as well as in macroeconomics and finance where the object of the analysis is often the study of aggregates or portfolios of asset returns. In view of this, Bailey et al. (2012) and Chudik et al. (2011) suggest the extent of cross-sectional dependence based on the behavior of cross-sectional averages  $\bar{z}_{wt} = \sum_{i=1}^N w_{it} z_{it} = \mathbf{w}'_t \mathbf{z}_t$ , at a point in time  $t$ , for  $t \in \mathcal{T}$ , where  $\mathbf{z}_t$  satisfies Assumption CSD.1 and the sequence of weight vectors  $\mathbf{w}_t$  satisfies the following assumption.

**ASSUMPTION CSD.2:** Let  $\mathbf{w}_t = (w_{1t}, \dots, w_{Nt})'$ , for  $t \in \mathcal{T} \subseteq \mathbb{Z}$  and  $N \in \mathbb{N}$ , be a vector of non-stochastic weights. For any  $t \in \mathcal{T}$ , the sequence of weight vectors  $\{\mathbf{w}_t\}$  of growing dimension ( $N \rightarrow \infty$ ) satisfies the ‘granularity’ conditions:

$$\|\mathbf{w}_t\| = \sqrt{\mathbf{w}'_t \mathbf{w}_t} = O\left(N^{-\frac{1}{2}}\right), \quad (1)$$

$$\frac{w_{jt}}{\|\mathbf{w}_t\|} = O\left(N^{-\frac{1}{2}}\right) \text{ uniformly in } j \in \mathbb{N}. \quad (2)$$

Assumption CSD.2, known in finance as the granularity condition, ensures that the weights  $\{w_{it}\}$  are not dominated by a few of the cross-section units.<sup>1</sup> Although we have assumed the weights to be non-stochastic, this is done for expositional convenience and can be relaxed by allowing the weights,  $w_t$ , to be random but distributed independently of  $\mathbf{z}_t$ . Chudik et al. (2011) define the concepts of weak and strong cross-sectional dependence based on the limiting behavior of  $\bar{z}_{wt}$  at a given point in time  $t \in \mathcal{T}$ , as  $N \rightarrow \infty$ .

**Definition 1 Weak and strong cross-sectional dependence.** *The process  $\{z_{it}\}$  is said to be cross-sectionally weakly dependent (CWD) at a given point in time  $t \in \mathcal{T}$ , if for any sequence of weight vectors  $\{\mathbf{w}_t\}$  satisfying the granularity conditions (1)-(2) we have*

$$\lim_{N \rightarrow \infty} \text{Var}(\mathbf{w}'_t \mathbf{z}_t) = 0. \quad (3)$$

*$\{z_{it}\}$  is said to be cross-sectionally strongly dependent (CSD) at a given point in time  $t \in \mathcal{T}$ , if there exists a sequence of weight vectors  $\{\mathbf{w}_t\}$  satisfying (1)-(2) and a constant  $K$  independent of  $N$  such that for any  $N$  sufficiently large (and as  $N \rightarrow \infty$ )*

$$\text{Var}(\mathbf{w}'_t \mathbf{z}_t) \geq K > 0. \quad (4)$$

The above concepts can also be defined conditional on a given information set,  $\Omega_{t-1}$ , see Chudik et al. (2011). The choice of the conditioning set largely depends on the nature of the underlying processes and the purpose of the analysis. For example, in the case of dynamic stationary models, the information set could contain all lagged realizations of the process  $\{z_{it}\}$ , that is  $\Omega_{t-1} = \{\mathbf{z}_{t-1}, \mathbf{z}_{t-2}, \dots\}$ , whilst for dynamic non-stationary models, such as unit root processes, the information included in  $\Omega_{t-1}$ , could start from a finite past. The conditioning information set could also contain contemporaneous realizations, which might be useful in applications where a particular unit has a dominant influence on the rest of the units in the system. For further details, see Chudik and Pesaran (2013b).

The following proposition establishes the relationship between weak cross-sectional dependence and the asymptotic behavior of the largest eigenvalue of  $\Sigma_t$ .

**Proposition 1.** *The following statements hold:*

- (i) *The process  $\{z_{it}\}$  is CWD at a point in time  $t \in \mathcal{T}$ , if  $\lambda_1(\Sigma_t)$  is bounded in  $N$  or increases at rate slower than  $N$ .*
- (ii) *The process  $\{z_{it}\}$  is CSD at a point in time  $t \in \mathcal{T}$ , if and only if for any  $N$  sufficiently large (and as  $N \rightarrow \infty$ ),  $N^{-1}\lambda_1(\Sigma_t) \geq K > 0$ .*

*Proof.* First, suppose  $\lambda_1(\Sigma_t)$  is bounded in  $N$  or increases at rate slower than  $N$ . We have

$$\text{Var}(\mathbf{w}'_t \mathbf{z}_t) = \mathbf{w}'_t \Sigma_t \mathbf{w}_t \leq (\mathbf{w}'_t \mathbf{w}_t) \lambda_1(\Sigma_t), \quad (5)$$

and under the granularity conditions (1)-(2) it follows that

$$\lim_{N \rightarrow \infty} \text{Var}(\mathbf{w}'_t \mathbf{z}_t) = 0,$$

namely that  $\{z_{it}\}$  is CWD, which proves (i). Proof of (ii) is provided in Chudik et al. (2011) ■

It is often of interest to know not only whether  $\bar{z}_{wt}$  converges to its mean, but also the rate at which this convergence (if at all) takes place. To this end, Bailey et al. (2012) propose to characterize the degree of cross-sectional dependence by an exponent of cross-sectional dependence defined by the rate of change of  $\text{Var}(\bar{z}_{wt})$  in terms of  $N$ . Note that in the case where  $z_{it}$  are independently distributed across  $i$ , we have  $\text{Var}(\bar{z}_{wt}) = O(N^{-1})$ , whereas in the case of strong cross-sectional dependence  $\text{Var}(\bar{z}_{wt}) \geq K > 0$ . There is, however, a range of possibilities in-between, where  $\text{Var}(\bar{z}_{wt})$  decays but at a rate slower than  $N^{-1}$ . In particular, using a factor framework, Bailey et al. (2012) show that in general

$$\text{Var}(\bar{z}_{wt}) = \kappa_0 N^{2(\alpha-1)} + \kappa_1 N^{-1} + O(N^{\alpha-2}), \quad (6)$$

where  $\kappa_i > 0$  for  $i = 0$  and  $1$ , are bounded in  $N$ , which will be time invariant in the case of stationary processes. Since the rate at which  $\text{Var}(\bar{z}_{wt})$  tends to zero with  $N$

cannot be faster than  $N^{-1}$ , the range of  $\alpha$  identified by  $Var(\bar{z}_{wt})$  lies in the restricted interval  $-1 < 2\alpha - 2 \leq 0$  or  $1/2 < \alpha \leq 1$ . Note that (3) holds for all values of  $\alpha < 1$ , whereas (4) holds only for  $\alpha = 1$ . Hence the process with  $\alpha < 1$  is CWD, and a CSD process has the exponent  $\alpha = 1$ . Bailey et al. (2012) show that under certain conditions on the underlying factor model,  $\alpha$  is identified in the range  $1/2 < \alpha \leq 1$ , and can be consistently estimated. Alternative bias-adjusted estimators of  $\alpha$  are proposed and shown by Monte Carlo experiments to have satisfactory small sample properties.

A particular form of a CWD process arises when pair-wise correlations take nonzero values only across finite subsets of units that do not spread widely as the sample size increases. A similar situation arises in the case of spatial processes, where direct dependence exists only amongst adjacent observations, and the indirect dependence is assumed to decay with distance. For further details see Pesaran and Tosetti (2011).

Since  $\lambda_1(\Sigma_t) \leq \|\Sigma_t\|_1$ , it follows from (5) that both the spectral radius and the column norm of the covariance matrix of a CSD process will be increasing at the rate  $N$ . Similar situations also arise in the case of time series processes with long memory or strong temporal dependence where autocorrelation coefficients are not absolutely summable. Along the cross-section dimension, common factor models represent examples of strong cross-sectional dependence.

### 1.3 COMMON FACTOR MODELS

---

Consider the  $m$  factor model for  $\{z_{it}\}$

$$z_{it} = \gamma_{i1}f_{1t} + \gamma_{i2}f_{2t} + \dots + \gamma_{im}f_{mt} + e_{it}, \quad i = 1, 2, \dots, N, \quad (7)$$

which can be written more compactly as

$$\mathbf{z}_t = \boldsymbol{\Gamma} f_t + \mathbf{e}_t, \quad (8)$$

where  $\mathbf{f}_t = (f_{1t}, f_{2t}, \dots, f_{mt})'$ ,  $\mathbf{e}_t = (e_{1t}, e_{2t}, \dots, e_{Nt})'$ , and  $\boldsymbol{\Gamma} = (\gamma_{ij})$ , for  $i = 1, 2, \dots, N$ ,  $j = 1, 2, \dots, m$ , is an  $N \times m$  matrix of fixed coefficients, known as factor loadings. The common factors,  $\mathbf{f}_t$ , simultaneously affect all cross-sectional units, albeit with different degrees as measured by  $\boldsymbol{\gamma}_i = (\gamma_{i1}, \gamma_{i2}, \dots, \gamma_{im})'$ . Examples of observed common factors that tend to affect all households' and firms' consumption and investment decisions include interest rates and oil prices. Aggregate demand and supply shocks represent examples of common unobserved factors. In multifactor models, interdependence arises from the reaction of units to some external events. Further, according to this representation, correlation between any pair of units does not depend on how far these observations are apart, and violates the distance decay effect that underlies the spatial interaction model.

The following assumptions are typically made regarding the common factors,  $f_{it}$ , and the idiosyncratic errors,  $e_{it}$ .

ASSUMPTION CF.1: The  $m \times 1$  vector  $\mathbf{f}_t$  is a zero mean covariance stationary process, with absolutely summable autocovariances, distributed independently of  $e_{it'}$  for all  $i, t, t'$ , such that  $E(f_{\ell t}^2) = 1$  and  $E(f_{\ell t} f_{\ell' t'}) = 0$ , for  $\ell \neq \ell' = 1, 2, \dots, m$ .

ASSUMPTION CF.2:  $\text{Var}(e_{it}) = \sigma_i^2 < K < \infty$ ,  $e_{it}$  and  $e_{jt}$  are independently distributed for all  $i \neq j$  and for all  $t$ . Specifically,  $\max_i(\sigma_i^2) = \sigma_{\max}^2 < K < \infty$ .

Assumption CF.1 is an identification condition, since it is not possible to separately identify  $\mathbf{f}_t$  and  $\boldsymbol{\Gamma}$ . The above factor model with a fixed number of factors and cross-sectionally independent idiosyncratic errors is often referred to as an exact factor model. Under the above assumptions, the covariance of  $\mathbf{z}_t$  is given by

$$E(\mathbf{z}_t \mathbf{z}'_t) = \boldsymbol{\Gamma} \boldsymbol{\Gamma}' + \mathbf{V},$$

where  $\mathbf{V}$  is a diagonal matrix with elements  $\sigma_i^2$  on the main diagonal.

The assumption that the idiosyncratic errors,  $e_{it}$ , are cross-sectionally independent is not necessary and can be relaxed. The factor model that allows the idiosyncratic shocks,  $e_{it}$ , to be cross-sectionally weakly correlated is known as the approximate factor model. See Chamberlain (1983). In general, the correlation patterns of the idiosyncratic errors can be characterized by

$$\mathbf{e}_t = \mathbf{R} \boldsymbol{\varepsilon}_t, \quad (9)$$

where  $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \varepsilon_{2t}, \dots, \varepsilon_{Nt})' \sim (\mathbf{0}, \mathbf{I}_N)$ . In the case of this formulation  $\mathbf{V} = \mathbf{R} \mathbf{R}'$ , which is no longer diagonal when  $\mathbf{R}$  is not diagonal, and further identification restrictions are needed so that the factor specification can be distinguished from the cross-sectional dependence assumed for the idiosyncratic errors. To this end it is typically assumed that  $\mathbf{R}$  has bounded row and column sum matrix norms (so that the cross-sectional dependence of  $\mathbf{e}_t$  is sufficiently weak) and the factor loadings are such that  $\lim_{N \rightarrow \infty} (N^{-1} \boldsymbol{\Gamma}' \boldsymbol{\Gamma})$  is a full rank matrix.

A leading example of  $\mathbf{R}$  arises in the context of the first-order spatial autoregressive, SAR(1), model, defined by

$$\mathbf{e}_t = \rho \mathbf{W} \mathbf{e}_t + \boldsymbol{\Lambda} \boldsymbol{\varepsilon}_t, \quad (10)$$

where  $\boldsymbol{\Lambda}$  is a diagonal matrix with strictly positive and bounded elements,  $0 < \sigma_i < \infty$ ,  $\rho$  is a spatial autoregressive coefficient, and the matrix  $\mathbf{W}$  is the ‘connection or spatial’ weight matrix which is taken as given. Assuming that  $(\mathbf{I}_N - \rho \mathbf{W})$  is invertible, we then have  $\mathbf{R} = (\mathbf{I}_N - \rho \mathbf{W})^{-1} \boldsymbol{\Lambda}$ . In the spatial literature,  $\mathbf{W}$  is assumed to have non-negative elements and is typically row-standardized so that  $\|\mathbf{W}\|_\infty = 1$ . Under these assumptions,  $|\rho| < 1$  ensures that  $|\rho| \|\mathbf{W}\|_\infty < 1$ , and we have

$$\begin{aligned} \|\mathbf{R}\|_\infty &= \|\boldsymbol{\Lambda}\|_\infty \|\mathbf{I}_N + \rho \mathbf{W} + \rho^2 \mathbf{W}^2 + \dots\|_\infty \\ &\leq \|\boldsymbol{\Lambda}\|_\infty [1 + |\rho| \|\mathbf{W}\|_\infty + |\rho|^2 \|\mathbf{W}\|_\infty^2 + \dots] = \frac{\|\boldsymbol{\Lambda}\|_\infty}{1 - |\rho| \|\mathbf{W}\|_\infty} < K < \infty, \end{aligned}$$

where  $\|\boldsymbol{\Lambda}\|_\infty = \max_i(\sigma_i) < \infty$ . Similarly,  $\|\mathbf{R}\|_1 < K < \infty$ , if it is further assumed that  $|\rho| \|\mathbf{W}\|_1 < 1$ . In general,  $\mathbf{R} = (\mathbf{I}_N - \rho \mathbf{W})^{-1} \boldsymbol{\Lambda}$  has bounded row and column

sum matrix norms if  $|\rho| < \min(1/\|W\|_1, 1/\|W\|_\infty)$ . In the case where  $W$  is a row and column stochastic matrix (often assumed in the spatial literature), this sufficient condition reduces to  $|\rho| < 1$ , which also ensures the invertibility of  $(I_N - \rho W)$ . Note that for a doubly stochastic matrix  $\rho(W) = \|W\|_1 = \|W\|_\infty = 1$ , where  $\rho(W)$  is the spectral radius of  $W$ . It turns out that almost all spatial models analyzed in the spatial econometrics literature characterize weak forms of cross-sectional dependence. See Sarafidis and Wansbeek (2012) for further discussion.

Turning now to the factor representation, to ensure that the factor component of (8) represents strong cross-sectional dependence (so that it can be distinguished from the idiosyncratic errors) it is sufficient that the absolute column sum matrix norm of  $\|\Gamma\|_1 = \max_{j \in \{1, 2, \dots, N\}} \sum_{i=1}^N |\gamma_{ij}|$  rises with  $N$  at the rate  $N$ , which implies that  $\lim_{N \rightarrow \infty} (N^{-1} \Gamma' \Gamma)$  is a full rank matrix, as required earlier.

The distinction between weak and strong cross-sectional dependence in terms of factor loadings is formalized in the following definition.

**Definition 2 Strong and weak factors.** *The factor  $f_{\ell t}$  is said to be strong if*

$$\lim_{N \rightarrow \infty} N^{-1} \sum_{i=1}^N |\gamma_{i\ell}| = K > 0. \quad (11)$$

*The factor  $f_{\ell t}$  is said to be weak if*

$$\lim_{N \rightarrow \infty} \sum_{i=1}^N |\gamma_{i\ell}| = K < \infty. \quad (12)$$

It is also possible to consider intermediate cases of semi-weak or semi-strong factors. In general, let  $\alpha_\ell$  be a positive constant in the range  $0 \leq \alpha_\ell \leq 1$  and consider the condition

$$\lim_{N \rightarrow \infty} N^{-\alpha_\ell} \sum_{i=1}^N |\gamma_{i\ell}| = K < \infty. \quad (13)$$

Strong and weak factors correspond to the two values of  $\alpha_\ell = 1$  and  $\alpha_\ell = 0$ , respectively. For any other values of  $\alpha_\ell \in (0, 1)$  the factor  $f_{\ell t}$  can be said to be semi-strong or semi-weak. It will prove useful to associate the semi-weak factors with values of  $0 < \alpha_\ell < 1/2$ , and the semi-strong factors with values of  $1/2 \leq \alpha_\ell < 1$ . In a multi-factor set up, the overall exponent can be defined by  $\alpha = \max(\alpha_1, \alpha_2, \dots, \alpha_m)$ .

**Example 1.** Suppose that  $z_{it}$  are generated according to the simple factor model,  $z_{it} = \gamma_i f_t + e_{it}$ , where  $f_t$  is independently distributed of  $\gamma_i$ , and  $e_{it} \sim \text{IID}(0, \sigma_i^2)$ , for all  $i$  and  $t$ ,  $\sigma_i^2$  is non-stochastic for expositional simplicity and bounded,  $E(f_t^2) = \sigma_f^2 < \infty$ ,  $E(f_t) = 0$  and  $f_t$  is independently distributed of  $e_{it'}$  for all  $i, t$  and  $t'$ . The factor loadings are given by

$$\gamma_i = \mu + \nu_i, \text{ for } i = 1, 2, \dots, [N^{\alpha_\gamma}] \quad (14)$$

$$\gamma_i = 0 \text{ for } i = [N^{\alpha_\gamma}] + 1, [N^{\alpha_\gamma}] + 2, \dots, N, \quad (15)$$

for some constant  $\alpha_\gamma \in [0, 1]$ , where  $[N^{\alpha_\gamma}]$  is the integer part of  $N^{\alpha_\gamma}$ ,  $\mu \neq 0$ , and  $v_i$  are IID with mean 0 and the finite variance,  $\sigma_v^2$ .<sup>2</sup> Note that  $\sum_{i=1}^N |\gamma_i| = O_p([N^{\alpha_\gamma}])$  and the factor  $f_t$  with loadings  $\gamma_i$  is strong for  $\alpha_\gamma = 1$ , weak for  $\alpha_\gamma = 0$ , and semi-weak or semi-strong for  $0 < \alpha_\gamma < 1$ . Consider the variance of the (simple) cross-sectional averages  $\bar{z}_t = N^{-1} \sum_{i=1}^N z_{it}$

$$\text{Var}_N(\bar{z}_t) = \text{Var}\left(\bar{z}_t \mid \{\gamma_i\}_{i=1}^N\right) = \bar{\gamma}_N^2 \sigma_f^2 + N^{-1} \bar{\sigma}_N^2, \quad (16)$$

where (dropping the integer part sign,  $[.]$ , for further clarity)

$$\begin{aligned} \bar{\gamma}_N &= N^{-1} \sum_{i=1}^N \gamma_i = N^{-1} \sum_{i=1}^{N^{\alpha_\gamma}} \gamma_i = \mu N^{\alpha_\gamma - 1} + N^{\alpha_\gamma - 1} \left( \frac{1}{N^{\alpha_\gamma}} \sum_{i=1}^{N^{\alpha_\gamma}} v_i \right) \\ \bar{\sigma}_N^2 &= N^{-1} \sum_{i=1}^N \sigma_i^2 > 0. \end{aligned}$$

But, noting that

$$E(\bar{\gamma}_N) = \mu N^{\alpha_\gamma - 1}, \quad \text{Var}(\bar{\gamma}_N) = N^{\alpha_\gamma - 2} \sigma_v^2,$$

we have

$$E(\bar{\gamma}_N^2) = [E(\bar{\gamma}_N)]^2 + \text{Var}(\bar{\gamma}_N) = \mu^2 N^{2(\alpha_\gamma - 1)} + N^{\alpha_\gamma - 2} \sigma_v^2.$$

Therefore, using this result in (16), we now have

$$\text{Var}(\bar{z}_t) = E[\text{Var}_N(\bar{z}_t)] = \sigma_f^2 \mu^2 N^{2(\alpha_\gamma - 1)} + \bar{\sigma}_N^2 N^{-1} + \sigma_v^2 \sigma_f^2 N^{\alpha_\gamma - 2} \quad (17)$$

$$= \sigma_f^2 \mu^2 N^{2(\alpha_\gamma - 1)} + \bar{\sigma}_N^2 N^{-1} + O(N^{\alpha_\gamma - 2}). \quad (18)$$

Thus the exponent of cross-sectional dependence of  $z_{it}$ , denoted as  $\alpha_z$ , and the exponent  $\alpha_\gamma$  coincide in this example, so long as  $\alpha_\gamma > 1/2$ . When  $\alpha_\gamma = 1/2$ , one cannot use  $\text{Var}(\bar{z}_t)$  to distinguish the factor effects from those of the idiosyncratic terms. Of course, this does not necessarily mean that other more powerful techniques cannot be found to distinguish such weak factor effects from the effects of the idiosyncratic terms. Finally, note also that in this example  $\sum_{i=1}^N \gamma_i^2 = O_p(N^{\alpha_\gamma})$ , and the largest eigenvalue of the  $N \times N$  covariance matrix,  $\text{Var}(\mathbf{z}_t)$ , also rises at the rate of  $N^{\alpha_\gamma}$ .

The relationship between the notions of CSD and CWD and the definitions of weak and strong factors are explored in the following theorem.

**Theorem 2.** Consider the factor model (8) and suppose that Assumptions CF.1-CF.2 hold, and there exists a positive constant  $\alpha = \max(\alpha_1, \alpha_2, \dots, \alpha_m)$  in the range  $0 \leq \alpha \leq 1$ , such that condition (13) is met for any  $\ell = 1, 2, \dots, m$ . Then the following statements hold:

- (i) The process  $\{z_{it}\}$  is cross-sectionally weakly dependent at a given point in time  $t \in \mathcal{T}$  if  $\alpha < 1$ , which includes cases of weak, semi-weak or semi-strong factors,  $f_{\ell t}$ , for  $\ell = 1, 2, \dots, m$ .

- (ii) The process  $\{z_{it}\}$  is cross-sectionally strongly dependent at a given point in time  $t \in \mathcal{T}$  if and only if there exists at least one strong factor.

Proof is provided in Chudik et al. (2011).

Since a factor structure can lead to strong as well as weak forms of cross-sectional dependence, cross-sectional dependence can also be characterized more generally by the following  $N$  factor representation:

$$z_{it} = \sum_{j=1}^N \gamma_{ij} f_{jt} + \varepsilon_{it}, \text{ for } i = 1, 2, \dots, N,$$

where  $\varepsilon_{it}$  is independently distributed across  $i$ . Under this formulation, to ensure that the variance of  $z_{it}$  is bounded in  $N$ , we also require that

$$\sum_{\ell=1}^N |\gamma_{i\ell}| \leq K < \infty, \text{ for } i = 1, 2, \dots, N. \quad (19)$$

$z_{it}$  can now be decomposed as

$$z_{it} = z_{it}^s + z_{it}^w, \quad (20)$$

where

$$z_{it}^s = \sum_{\ell=1}^m \gamma_{i\ell} f_{\ell t}; \quad z_{it}^w = \sum_{\ell=m+1}^N \gamma_{i\ell} f_{\ell t} + \varepsilon_{it}, \quad (21)$$

and  $\gamma_{i\ell}$  satisfy conditions (11) for  $\ell = 1, \dots, m$ , where  $m$  must be finite in view of the absolute summability condition (19) that ensures finite variances. Remaining loadings  $\gamma_{i\ell}$  for  $\ell = m+1, m+2, \dots, N$  must satisfy either (12) or (13) for some  $\alpha < 1$ .<sup>3</sup> In light of Theorem 2, it can be shown that  $z_{it}^s$  is CSD and  $z_{it}^w$  is CWD. Also, notice that when  $z_{it}$  is CWD, we have a model with no strong factors and potentially an infinite number of weak or semi-strong factors. Seen from this perspective, spatial models considered in the literature can be viewed as an  $N$  weak factor model.

Consistent estimation of factor models with weak or semi-strong factors may be problematic, as evident from the following example.

**Example 2.** Consider the single factor model with known factor loadings

$$z_{it} = \gamma_i f_t + \varepsilon_{it}, \quad \varepsilon_{it} \sim IID(0, \sigma^2).$$

The least squares estimator of  $f_t$ , which is the best linear unbiased estimator, is given by

$$\hat{f}_t = \frac{\sum_{i=1}^N \gamma_i z_{it}}{\sum_{i=1}^N \gamma_i^2}, \quad \text{Var}(\hat{f}_t) = \frac{\sigma^2}{\sum_{i=1}^N \gamma_i^2}.$$

In the weak factor case where  $\sum_{i=1}^N \gamma_i^2$  is bounded in  $N$ , then  $\text{Var}(\hat{f}_t)$  does not vanish as  $N \rightarrow \infty$ , and  $\hat{f}_t$  need not be a consistent estimator of  $f_t$ . See also Onatski (2012).

The presence of weak or semi-strong factors in errors does not affect consistency of conventional panel data estimators, but affects inference, as is evident from the following example.

**Example 3.** Consider the following panel data model

$$y_{it} = \beta x_{it} + u_{it}, \quad u_{it} = \gamma_i f_t + \varepsilon_{it},$$

where

$$x_{it} = \delta_i f_t + v_{it}.$$

To simplify the exposition, we assume that,  $\varepsilon_{it}$ ,  $v_{js}$  and  $f_t$  are independently, and identically distributed across all  $i, j, t, s$ , and  $t'$ , as  $\varepsilon_{it} \sim \text{IID}(0, \sigma_\varepsilon^2)$ ,  $v_{it} \sim \text{IID}(0, \sigma_v^2)$ , and  $f_t \sim \text{IID}(0, 1)$ . The pooled estimator of  $\beta$  satisfies

$$\sqrt{NT} (\hat{\beta}_P - \beta) = \frac{\frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T x_{it} u_{it}}{\frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T x_{it}^2}, \quad (22)$$

where the denominator converges in probability to  $\sigma_v^2 + \lim_{N \rightarrow \infty} N^{-1} \sum_{i=1}^N \delta_i^2 > 0$ , while the numerator can be expressed, after substituting for  $x_{it}$  and  $u_{it}$ , as

$$\frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T x_{it} u_{it} = \frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T \gamma_i \delta_i f_t^2 + \frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T (\delta_i f_t \varepsilon_{it} + \gamma_i v_{it} f_t + v_{it} \varepsilon_{it}). \quad (23)$$

Under the above assumptions it is now easily seen that the second term in the above expression is  $O_p(1)$ , but the first term can be written as

$$\begin{aligned} \frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T \gamma_i \delta_i f_t^2 &= \frac{1}{\sqrt{N}} \sum_{i=1}^N \gamma_i \delta_i \cdot \frac{1}{\sqrt{T}} \sum_{t=1}^T f_t^2 \\ &= \frac{1}{\sqrt{N}} \sum_{i=1}^N \gamma_i \delta_i \cdot O_p(T^{1/2}). \end{aligned}$$

Suppose now that  $f_t$  is a factor such that loadings  $\gamma_i$  and  $\delta_i$  are given by (14)-(15) with the exponents  $\alpha_\gamma$  and  $\alpha_\delta$  ( $0 \leq \alpha_\gamma, \alpha_\delta \leq 1$ ), respectively, and let  $\alpha = \min(\alpha_\gamma, \alpha_\delta)$ . It then follows that  $\sum_{i=1}^N \gamma_i \delta_i = O_p(N^\alpha)$ , and

$$\frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T \gamma_i \delta_i f_t^2 = O_p(N^{\alpha-1/2} T^{1/2}).$$

Therefore, even if  $\alpha < 1$  the first term in (23) diverges, and overall we have  $\hat{\beta}_P - \beta = O_p(N^{\alpha-1}) + O_p(T^{-1/2} N^{-1/2})$ . It is now clear that even if  $f_t$  is not a strong factor, the rate of convergence of  $\hat{\beta}_P$  and its asymptotic variance will still be affected by the factor structure of the error term. In the case where  $\alpha = 0$ , and the errors are spatially dependent, the

variance matrix of the pooled estimator also depends on the nature of the spatial dependence, which must be taken into account when carrying out inference on  $\beta$ . See Pesaran and Tosetti (2011) for further results and discussions.

Weak, strong, and semi-strong common factors may be used to represent very general forms of cross-sectional dependence. For example, a factor process with an infinite number of weak factors, and no idiosyncratic errors, can be used to represent spatial processes. In particular, the spatial model (9) can be represented by  $e_{it} = \sum_{j=1}^N \gamma_{ij} f_{jt}$ , where  $\gamma_{ij} = r_{ij}$  and  $f_{jt} = \varepsilon_{jt}$ . Strong factors can be used to represent the effect of the cross-section units that are “dominant” or pervasive, in the sense that they impact all the other units in the sample and their effect does not vanish as  $N$  tends to infinity (Chudik and Pesaran 2013b). As argued in Holly et al. (2011), a large city may play a dominant role in determining house prices nationally. Semi-strong factors may exist if there is a cross-section unit or an unobserved common factor that affects only a subset of the units and the number of affected units rises more slowly than the total number of units. Estimates of the exponent of cross-sectional dependence reported by Bailey et al. (2012) (2012, Tables 1 and 2) suggest that for typical large macroeconomic data sets the estimates of  $\alpha$  fall in the range of 0.77–0.92, which fall short of 1 assumed in the factor literature. For cross-country quarterly real GDP growth, inflation and real equity prices the estimates of  $\alpha$  are much closer to unity and tend to be around 0.97.

## 1.4 LARGE PANELS WITH STRICTLY EXOGENOUS REGRESSORS AND A FACTOR ERROR STRUCTURE

---

Consider the following heterogeneous panel data model:

$$y_{it} = \boldsymbol{\alpha}'_i \mathbf{d}_t + \boldsymbol{\beta}'_i \mathbf{x}_{it} + u_{it}, \quad (24)$$

where  $\mathbf{d}_t$  is a  $n \times 1$  vector of observed common effects (including deterministics such as intercepts or seasonal dummies),  $\mathbf{x}_{it}$  is a  $k \times 1$  vector of observed individual-specific regressors on the  $i$ th cross-section unit at time  $t$ , and disturbances,  $u_{it}$ , have the following common factor structure

$$u_{it} = \gamma_{i1} f_{1t} + \gamma_{i2} f_{2t} + \dots + \gamma_{im} f_{mt} + e_{it} = \boldsymbol{\gamma}'_i \mathbf{f}_t + e_{it}, \quad (25)$$

in which  $\mathbf{f}_t = (f_{1t}, f_{2t}, \dots, f_{mt})'$  is an  $m$ -dimensional vector of unobservable common factors, and  $\boldsymbol{\gamma}_i = (\gamma_{i1}, \gamma_{i2}, \dots, \gamma_{im})'$  is the associated  $m \times 1$  vector of factor loadings. The number of factors,  $m$ , is assumed to be fixed relative to  $N$ , and in particular  $m \ll N$ . The idiosyncratic errors,  $e_{it}$ , could be CWD, for example, being generated by a spatial process, or, more generally, by a weak factor structure. For estimation purposes, as in

the case of panels with group effects, the factor loadings,  $\gamma_i$ , could be either random or fixed unknown coefficients. We distinguish between the homogeneous coefficient case where  $\beta_i = \beta$  for all  $i$ , and the heterogeneous case where  $\beta_i$  are random draws from a given distribution. In the latter case, we assume that the object of interest is the mean coefficients,  $\beta = E(\beta_i)$ , for all  $i$ . When the regressors,  $x_{it}$ , are strictly exogenous and the deviations  $v_i = \beta_i - \beta$  are distributed independently of the errors and the regressors, the mean coefficients,  $\beta$ , can be consistently estimated using pooled as well as mean group estimation procedures. But mean group estimation will only be consistent if the regressors are weakly exogenous and/or if the deviations are correlated with the regressors/errors.<sup>4</sup>

The assumption of slope homogeneity is also crucially important for the derivation of the asymptotic distribution of the pooled or the mean group estimators of  $\beta$ . Under slope homogeneity, the asymptotic distribution of the estimator of  $\beta$  typically converges at the rate of  $\sqrt{NT}$ , whilst under slope heterogeneity the rate is  $\sqrt{N}$ . In view of the uncertainty regarding the assumption of slope heterogeneity, non-parametric estimators of the variance matrix of the pooled and mean group estimators are proposed.<sup>5</sup> In the following sub-sections we review a number of different estimators of  $\beta$  proposed in the literature.

### 1.4.1 PC estimators

The principal components (PC) approach proposed by Coakley et al. (2002) and Bai (2009), by requiring that  $N^{-1}\Gamma'\Gamma$  tends to a positive definite matrix, implicitly assumes that all the unobserved common factors in (25) are strong. Coakley et al. (2002) consider the panel data model with strictly exogenous regressors and homogeneous slopes (i.e.,  $\beta_i = \beta$ ), and propose a two-stage estimation procedure. In the first stage, PCs are extracted from the OLS residuals as proxies for the unobserved variables, and in the second step the estimated factors are treated as observable and the following augmented regression is estimated

$$y_{it} = \alpha'_i d_t + \beta' x_{it} + \gamma'_i \hat{f}_t + \varepsilon_{it}, \text{ for } i = 1, 2, \dots, N; t = 1, 2, \dots, T, \quad (26)$$

where  $\hat{f}_t$  is an  $m \times 1$  vector of principal components of the residuals computed in the first stage. The resultant estimator of  $\beta$  is consistent for large  $N$  and  $T$ , so long as  $f_t$  and the regressors,  $x_{it}$ , are uncorrelated. However, if the factors and the regressors are correlated, which is likely to be the case in practice, the two-stage estimator becomes inconsistent (Pesaran 2006).

Building on Coakley et al. (2002), Bai (2009) has proposed an iterative method which consists of alternating the PC method applied to OLS residuals and the least squares estimation of (26), until convergence. In particular, to simplify the exposition suppose  $\alpha_i = 0$ . Then the least squares estimator of  $\beta$  and  $F$  is the solution of the

following set of non-linear equations:

$$\hat{\beta}_{PC} = \left( \sum_{i=1}^N \mathbf{X}'_i \mathbf{M}_{\hat{F}} \mathbf{X}_i \right)^{-1} \sum_{i=1}^N \mathbf{X}'_i \mathbf{M}_{\hat{F}} \mathbf{y}_i,$$

$$\frac{1}{NT} \sum_{i=1}^N (\mathbf{y}_i - \mathbf{X}_i \hat{\beta}_{PC}) (\mathbf{y}_i - \mathbf{X}_i \hat{\beta}_{PC})' \hat{F} = \hat{F} \hat{V},$$

where  $\mathbf{X}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{iT})'$  is the matrix of observations on  $\mathbf{x}_{it}$ ,  $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{iT})'$  is the vector of observations on  $y_{it}$ ,  $\mathbf{M}_{\hat{F}} = \mathbf{I}_T - \hat{F}(\hat{F}'\hat{F})^{-1}\hat{F}'$ ,  $\hat{F} = (\hat{f}_1, \hat{f}_2, \dots, \hat{f}_T)'$ , and  $\hat{V}$  is a diagonal matrix with the  $m$  largest eigenvalues of the matrix  $\frac{1}{NT} \sum_{i=1}^N (\mathbf{y}_i - \mathbf{X}_i \hat{\beta}_{PC})(\mathbf{y}_i - \mathbf{X}_i \hat{\beta}_{PC})'$  arranged in a decreasing order. The solution  $\hat{\beta}_{PC}$ ,  $\hat{F}$ , and  $\hat{\gamma}_i = (\hat{F}'\hat{F})^{-1}\hat{F}'(\mathbf{y}_i - \mathbf{X}_i \hat{\beta}_{PC})$  minimizes the sum of squared residuals function,

$$SSR_{NT} (\boldsymbol{\beta}, \{\boldsymbol{\gamma}_i\}_{i=1}^N, \{\mathbf{f}_t\}_{t=1}^T) = \sum_{i=1}^N (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta} - \mathbf{F} \boldsymbol{\gamma}_i)' (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta} - \mathbf{F} \boldsymbol{\gamma}_i),$$

where  $\mathbf{F} = (\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_T)'$ . This function is a Gaussian quasi maximum likelihood function of the model and in this respect, Bai's iterative principal components estimator can also be seen as a quasi maximum likelihood estimator, since it minimizes the quasi likelihood function.

Bai (2009) shows that such an estimator is consistent even if common factors are correlated with the explanatory variables. Specifically, the least square estimator of  $\boldsymbol{\beta}$  obtained from the above procedure,  $\hat{\beta}_{PC}$ , is consistent if both  $N$  and  $T$  tend to infinity, without any restrictions on the ratio  $T/N$ . When in addition  $T/N \rightarrow K > 0$ ,  $\hat{\beta}_{PC}$  converges at the rate  $\sqrt{NT}$ , but the limiting distribution of  $\sqrt{NT}(\hat{\beta}_{PC} - \boldsymbol{\beta})$  does not necessarily have a zero mean. Nevertheless, Bai shows that the asymptotic bias can be consistently estimated and proposes a bias corrected estimator.

But it is important to bear in mind that PC-based estimators generally require the determination of the unknown number of strong factors (PCs),  $m$ , to be included in the second stage of estimation, and this can introduce some degree of sampling uncertainty into the analysis. There is now a large literature that considers the estimation of  $m$ , assuming all the  $m$  factors to be strong. See, for example, Bai and Ng (2002) (2002, 2007), Kapetanios (2004) (2004, 2010), Amengual and Watson (2007), Hallin and Liska (2007), Onatski (2009) (2009, 2010), Ahn and Horenstein (2013), Breitung and Pigorsch (2013), Choi and Jeong (2013) and Harding (2013). There are also a number of useful surveys by Bai and Ng (2008), Stock and Watson (2011) and Breitung and Choi (2013), amongst others, that can be consulted for detailed discussions of these methods and additional references. An extensive Monte Carlo investigation into the small sample performance of different selection/estimation methods is provided in Choi and Jeong (2013).

### 1.4.2 CCE Estimators

Pesaran (2006) suggests the Common Correlated Effects (CCE) estimation procedure that consists of approximating the linear combinations of the unobserved factors by cross-sectional averages of the dependent and explanatory variables, and then running standard panel regressions augmented with these cross-sectional averages. Both pooled and mean group versions are proposed, depending on the assumption regarding the slope homogeneity.

Under slope heterogeneity, the CCE approach assumes that  $\beta'_i$ 's follow the random coefficient model

$$\beta_i = \beta + v_i, \quad v_i \sim IID(\mathbf{0}, \Omega_v) \quad \text{for } i = 1, 2, \dots, N,$$

where the deviations,  $v_i$ , are distributed independently of  $e_{jt}$ ,  $x_{jt}$ , and  $d_t$ , for all  $i, j$  and  $t$ . Since in many empirical applications where cross-sectional dependence is caused by unobservable factors, these factors are correlated with the regressors, and the following model for the individual-specific regressors in (24) is adopted

$$x_{it} = A'_i d_t + \Gamma'_i f_t + v_{it}, \quad (27)$$

where  $A_i$  and  $\Gamma_i$  are  $n \times k$  and  $m \times k$  factor loading matrices with fixed components,  $v_{it}$  is the idiosyncratic component of  $x_{it}$  distributed independently of the common effects  $f_{t'}$  and errors  $e_{jt'}$  for all  $i, j, t$  and  $t'$ . However,  $v_{it}$  is allowed to be serially correlated, and cross-sectionally weakly correlated.

Equations (24), (25), and (27) can be combined into the following system of equations

$$z_{it} = \begin{pmatrix} y_{it} \\ x_{it} \end{pmatrix} = B'_i d_t + C'_i f_t + \xi_{it}, \quad (28)$$

where

$$\begin{aligned} \xi_{it} &= \begin{pmatrix} e_{it} + \beta'_i v_{it} \\ v_{it} \end{pmatrix}, \\ B_i &= (\alpha_i \quad A_i) \begin{pmatrix} 1 & 0 \\ \beta_i & I_k \end{pmatrix}, \quad C_i = (\gamma_i \quad \Gamma_i) \begin{pmatrix} 1 & 0 \\ \beta_i & I_k \end{pmatrix}. \end{aligned}$$

Consider the weighted average of  $z_{it}$  using the weights  $w_i$  satisfying the granularity conditions (1)–(2):

$$\bar{z}_{wt} = \bar{B}'_w d_t + \bar{C}'_w f_t + \bar{\xi}_{wt},$$

where

$$\bar{z}_{wt} = \sum_{i=1}^N w_i z_{it},$$

$$\bar{B}_w = \sum_{i=1}^N w_i B_i, \quad \bar{C}_w = \sum_{i=1}^N w_i C_i, \quad \text{and} \quad \bar{\xi}_{wt} = \sum_{i=1}^N w_i \xi_{it}.$$

Assume that<sup>6</sup>

$$\text{Rank}(\bar{\mathbf{C}}_w) = m \leq k + 1, \quad (29)$$

we have

$$\mathbf{f}_t = (\bar{\mathbf{C}}_w \bar{\mathbf{C}}'_w)^{-1} \bar{\mathbf{C}}_w (\bar{\mathbf{z}}_{wt} - \bar{\mathbf{B}}'_w \mathbf{d}_t - \bar{\boldsymbol{\xi}}_{wt}). \quad (30)$$

Under the assumption that  $e_{it}$ 's and  $v_{it}$ 's are CWD processes, it is possible to show that (see Pesaran and Tosetti (2011))

$$\bar{\boldsymbol{\xi}}_{wt} \xrightarrow{q.m.} \mathbf{0}, \quad (31)$$

which implies

$$\mathbf{f}_t - (\bar{\mathbf{C}}_w \bar{\mathbf{C}}'_w)^{-1} \bar{\mathbf{C}}_w (\bar{\mathbf{z}}_{wt} - \bar{\mathbf{B}}'_w \mathbf{d}_t) \xrightarrow{q.m.} \mathbf{0}, \text{ as } N \rightarrow \infty, \quad (32)$$

where

$$\mathbf{C} = \lim_{N \rightarrow \infty} (\bar{\mathbf{C}}_w) = \tilde{\boldsymbol{\Gamma}} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \boldsymbol{\beta} & \mathbf{I}_k \end{pmatrix}, \quad (33)$$

$\tilde{\boldsymbol{\Gamma}} = (E(\boldsymbol{\gamma}_i), E(\boldsymbol{\Gamma}_i))$ , and  $\boldsymbol{\beta} = E(\boldsymbol{\beta}_i)$ . Therefore, the unobservable common factors,  $\mathbf{f}_t$ , can be well approximated by a linear combination of observed effects,  $\mathbf{d}_t$ , the cross-sectional averages of the dependent variable,  $\bar{y}_{wt}$ , and those of the individual-specific regressors,  $\bar{\mathbf{x}}_{wt}$ .

When the parameters of interest are the cross-sectional means of the slope coefficients,  $\boldsymbol{\beta}$ , we can consider two alternative estimators, the CCE Mean Group (CCEMG) estimator, originally proposed by Pesaran and Smith (1995), and the CCE Pooled (CCEP) estimator. Let  $\bar{\mathbf{M}}_w$  be defined by

$$\bar{\mathbf{M}}_w = \mathbf{I}_T - \bar{\mathbf{H}}_w (\bar{\mathbf{H}}'_w \bar{\mathbf{H}}_w)^+ \bar{\mathbf{H}}'_w, \quad (34)$$

where  $\mathbf{A}^+$  denotes the Moore-Penrose inverse of matrix  $\mathbf{A}$ ,  $\bar{\mathbf{H}}_w = (\mathbf{D}, \bar{\mathbf{Z}}_w)$ , and  $\mathbf{D}$  and  $\bar{\mathbf{Z}}_w$  are, respectively, the matrices of the observations on  $\mathbf{d}_t$  and  $\bar{\mathbf{z}}_{wt} = (\bar{y}_{wt}, \bar{\mathbf{x}}'_{wt})'$ .

The CCEMG is a simple average of the estimators of the individual slope coefficients<sup>7</sup>

$$\hat{\boldsymbol{\beta}}_{\text{CCEMG}} = N^{-1} \sum_{i=1}^N \hat{\boldsymbol{\beta}}_{\text{CCE},i}, \quad (35)$$

where

$$\hat{\boldsymbol{\beta}}_{\text{CCE},i} = (\mathbf{X}'_i \bar{\mathbf{M}}_w \mathbf{X}_i)^{-1} \mathbf{X}'_i \bar{\mathbf{M}}_w \mathbf{y}_i. \quad (36)$$

Pesaran (2006) shows that, under some general conditions,  $\hat{\boldsymbol{\beta}}_{\text{CCEMG}}$  is asymptotically unbiased for  $\boldsymbol{\beta}$ , and as  $(N, T) \rightarrow \infty$ ,

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{\text{CCEMG}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_{\text{CCEMG}}), \quad (37)$$

where  $\Sigma_{CCEMG} = \Omega_v$ . A consistent estimator of the variance of  $\hat{\beta}_{CCEMG}$ , denoted by  $Var(\hat{\beta}_{CCEMG})$ , can be obtained by adopting the non-parametric estimator

$$\widehat{Var}(\hat{\beta}_{CCEMG}) = N^{-1} \hat{\Sigma}_{CCEMG} = \frac{1}{N(N-1)} \sum_{i=1}^N (\hat{\beta}_{CCE,i} - \hat{\beta}_{CCEMG})(\hat{\beta}_{CCE,i} - \hat{\beta}_{CCEMG})' . \quad (38)$$

The CCEP estimator is given by

$$\hat{\beta}_{CCEP} = \left( \sum_{i=1}^N w_i \mathbf{X}'_i \bar{\mathbf{M}}_w \mathbf{X}_i \right)^{-1} \sum_{i=1}^N w_i \mathbf{X}'_i \bar{\mathbf{M}}_w \mathbf{y}_i . \quad (39)$$

Under some general conditions, Pesaran (2006) proves that  $\hat{\beta}_{CCEP}$  is asymptotically unbiased for  $\beta$ , and, as  $(N, T) \rightarrow \infty$ ,

$$\left( \sum_{i=1}^N w_i^2 \right)^{-1/2} (\hat{\beta}_{CCEP} - \beta) \xrightarrow{d} N(0, \Sigma_{CCEP}),$$

where

$$\begin{aligned} \Sigma_{CCEP} &= \Psi^{*-1} \mathbf{R}^* \Psi^{*-1}, \\ \Psi^* &= \lim_{N \rightarrow \infty} \left( \sum_{i=1}^N w_i \Sigma_i \right), \quad \mathbf{R}^* = \lim_{N \rightarrow \infty} \left[ N^{-1} \sum_{i=1}^N \tilde{w}_i^2 (\Sigma_i \Omega_v \Sigma_i) \right], \\ \Sigma_i &= p \lim_{T \rightarrow \infty} (\mathbf{X}'_i \bar{\mathbf{M}}_w \mathbf{X}_i), \text{ and } \tilde{w}_i = \frac{w_i}{\sqrt{N^{-1} \sum_{i=1}^N w_i^2}}. \end{aligned}$$

A consistent estimator of  $Var(\hat{\beta}_{CCEP})$ , denoted by  $\widehat{Var}(\hat{\beta}_{CCEP})$ , is given by

$$\widehat{Var}(\hat{\beta}_{CCEP}) = \left( \sum_{i=1}^N w_i^2 \right) \hat{\Sigma}_{CCEP} = \left( \sum_{i=1}^N w_i^2 \right) \hat{\Psi}^{*-1} \hat{\mathbf{R}}^* \hat{\Psi}^{*-1}, \quad (40)$$

where

$$\begin{aligned} \hat{\Psi}^* &= \sum_{i=1}^N w_i \left( \frac{\mathbf{X}'_i \bar{\mathbf{M}}_w \mathbf{X}_i}{T} \right), \\ \hat{\mathbf{R}}^* &= \frac{1}{N-1} \sum_{i=1}^N \tilde{w}_i^2 \left( \frac{\mathbf{X}'_i \bar{\mathbf{M}}_w \mathbf{X}_i}{T} \right) (\hat{\beta}_{CCE,i} - \hat{\beta}_{CCEMG})(\hat{\beta}_{CCE,i} - \hat{\beta}_{CCEMG})' \left( \frac{\mathbf{X}'_i \bar{\mathbf{M}}_w \mathbf{X}_i}{T} \right). \end{aligned}$$

The rate of convergence of  $\hat{\beta}_{CCEMG}$  and  $\hat{\beta}_{CCEP}$  is  $\sqrt{N}$  when  $\Omega_v \neq 0$ . Note that even if  $\beta_i$  were observed for all  $i$ , the estimate of  $\beta = E(\beta_i)$  cannot converge at a faster rate than  $\sqrt{N}$ . If the individual slope coefficients  $\beta_i$  are homogeneous (namely if  $\Omega_v = 0$ ),  $\hat{\beta}_{CCEMG}$  and  $\hat{\beta}_{CCEP}$  are still consistent and converge at the rate  $\sqrt{NT}$  rather than  $\sqrt{N}$ .

An advantage of the nonparametric estimators  $\hat{\Sigma}_{CCEMG}$  and  $\hat{\Sigma}_{CCEP}$  is that they do not require knowledge of the weak cross-sectional dependence of  $e_{it}$  (provided it is sufficiently weak) nor the knowledge of serial correlation of  $e_{it}$ . An important question is whether the non-parametric variance estimators  $\widehat{Var}(\hat{\beta}_{CCEMG})$  and  $\widehat{Var}(\hat{\beta}_{CCEP})$  can be used in both cases of homogeneous and heterogeneous slopes. As established in Pesaran and Tosetti (2011), the asymptotic distribution of  $\hat{\beta}_{CCEMG}$  and  $\hat{\beta}_{CCEP}$  depends on nuisance parameters when slopes are homogeneous ( $\Omega_v = 0$ ), including the extent of cross-sectional correlations of  $e_{it}$  and their serial correlation structure. However, it can be shown that the robust non-parametric estimators  $\widehat{Var}(\hat{\beta}_{CCEMG})$  and  $\widehat{Var}(\hat{\beta}_{CCEP})$  are consistent when the regressor-specific components,  $v_{it}$ , are independently distributed across  $i$ .

The CCE continues to be applicable even if the rank condition (29) is not satisfied. Failure of the rank condition can occur if there is an unobserved factor for which the average of the loadings in the  $y_{it}$  and  $x_{it}$  equations tends to a zero vector. This could happen if, for example, the factor in question is weak, in the sense defined above. Another possible reason for failure of the rank condition is if the number of unobservable factors,  $m$ , is larger than  $k + 1$ , where  $k$  is the number of unit-specific regressors included in the model. In such cases, common factors cannot be estimated from cross-sectional averages. However, it is possible to show that the cross-sectional means of the slope coefficients,  $\beta_i$ , can still be consistently estimated, under the additional assumption that the unobserved factor loadings,  $\gamma_i$ , in equation (25) are independently and identically distributed across  $i$ , and of  $e_{jt}$ ,  $v_{jt}$ , and  $\mathbf{g}_t = (\mathbf{d}'_t, \mathbf{f}'_t)'$  for all  $i, j$ , and  $t$ , and uncorrelated with the loadings attached to the regressors,  $\Gamma_i$ . The consequences of the correlation between loadings  $\gamma_i$  and  $\Gamma_i$  for the performance of CCE estimators in the rank deficient case are documented in Sarafidis and Wansbeek (2012).

An advantage of the CCE approach is that it yields consistent estimates under a variety of situations. Kapetanios et al. (2011) consider the case where the unobservable common factors follow unit root processes and could be cointegrated. They show that the asymptotic distribution of panel estimators in the case of I(1) factors is similar to that in the stationary case. Pesaran and Tosetti (2011) prove consistency and asymptotic normality for CCE estimators when  $\{e_{it}\}$  are generated by a spatial process. Chudik et al. (2011) prove consistency and asymptotic normality of the CCE estimators when errors are subject to a finite number of unobserved strong factors and an infinite number of weak and/or semi-strong unobserved common factors as in (20)–(21), provided that certain conditions on the loadings of the infinite factor structure are satisfied. A further advantage of the CCE approach is that it does not require an a priori knowledge of the number of unobserved common factors.

In a Monte Carlo (MC) study, Coakley et al. (2006) compare ten alternative estimators for the mean slope coefficient in a linear heterogeneous panel regression with strictly exogenous regressors and unobserved common (correlated) factors. Their results show that, overall, the mean group version of the CCE estimator stands out as the most efficient and robust. These conclusions are in line with those in Kapetanios

and Pesaran (2007) and Chudik et al. (2011), who investigate the small sample properties of CCE estimators and the estimators based on principal components. The MC results show that PC augmented methods do not perform as well as the CCE approach, and can lead to substantial size distortions, due, in part, to the small sample errors in the number of factors selection procedure. In a recent theoretical study, Westerlund and Urbain (2011) investigate the merits of the CCE and PC estimators in the case of homogeneous slopes and a known number of unobserved common factors and find that, although the PC estimates of factors are more efficient than the cross-sectional averages, the CCE estimators of slope coefficients generally perform the best.

## 1.5 DYNAMIC PANEL DATA MODELS WITH A FACTOR ERROR STRUCTURE

---

The problem of the estimation of panels subject to cross-sectional error dependence becomes much more complicated once the assumption of strict exogeneity of the unit-specific regressors is relaxed. One important example is the panel data model with lagged dependent variables and unobserved common factors (possibly correlated with the regressors):<sup>8</sup>

$$y_{it} = \lambda_i y_{i,t-1} + \beta_i' x_{it} + u_{it}, \quad (41)$$

$$u_{it} = \gamma_i' f_t + e_{it}, \quad (42)$$

for  $i = 1, 2, \dots, N$ ;  $t = 1, 2, \dots, T$ . It is assumed that  $|\lambda_i| < 1$  and the dynamic processes have started a long time in the past. As in the previous section, we distinguish between the case of homogeneous coefficients, where  $\lambda_i = \lambda$  and  $\beta_i = \beta$  for all  $i$ , and the heterogeneous case, where  $\lambda_i$  and  $\beta_i$  are randomly distributed across units and the object of interest are the mean coefficients  $\lambda = E(\lambda_i)$  and  $\beta = E(\beta_i)$ . This distinction is more important for dynamic panels since not only the rate of convergence is affected by the presence of coefficient heterogeneity, but, as shown by Pesaran and Smith (1995), pooled least squares estimators are no longer consistent in the case of dynamic panel data models with heterogeneous coefficients.

It is convenient to define the vector of regressors  $\zeta_{it} = (y_{i,t-1}, x_{it}')'$  and the corresponding parameter vector  $\pi_i = (\lambda_i, \beta_i')'$  so that (41) can be written as

$$y_{it} = \pi_i' \zeta_{it} + u_{it}. \quad (43)$$

### 1.5.1 Quasi Maximum Likelihood Estimator

Moon and Weidner (2010) assume  $\pi_i = \pi$  for all  $i$  and develop a Gaussian quasi maximum likelihood estimator (QMLE) of the homogeneous coefficient vector  $\pi$ .<sup>9</sup> The

QMLE of  $\boldsymbol{\pi}$  is

$$\hat{\boldsymbol{\pi}}_{QMLE} = \arg \min_{\boldsymbol{\pi} \in \mathbb{B}} L_{NT}(\boldsymbol{\pi}),$$

where  $\mathbb{B}$  is a compact parameter set assumed to contain the true parameter values, and the objective function is the profile likelihood function:

$$L_{NT}(\boldsymbol{\pi}) = \min_{\{\boldsymbol{\gamma}_i\}_{i=1}^N, \{\mathbf{f}_t\}_{t=1}^T} \frac{1}{NT} \sum_{i=1}^N (\mathbf{y}_i - \boldsymbol{\Xi}_i \boldsymbol{\pi} - \mathbf{F}\boldsymbol{\gamma}_i)' (\mathbf{y}_i - \boldsymbol{\Xi}_i \boldsymbol{\pi} - \mathbf{F}\boldsymbol{\gamma}_i),$$

where

$$\boldsymbol{\Xi}_i = \begin{pmatrix} y_{i1} & \mathbf{x}'_{i,2} \\ y_{i2} & \mathbf{x}'_{i,3} \\ \vdots & \vdots \\ y_{iT-1} & \mathbf{x}'_{iT} \end{pmatrix}.$$

Both  $\hat{\boldsymbol{\pi}}_{QMLE}$  and  $\hat{\boldsymbol{\beta}}_{PC}$  minimize the same objective function and therefore, when the same set of regressors is considered, these two estimators are numerically the same, but there are important differences in their bias-corrected versions and in other aspects of the analysis of Bai (2009) and the analysis of Moon and Weidner (2010). The latter paper allows for more general assumptions on regressors, including the possibility of weak exogeneity, and adopts a quadratic approximation of the profile likelihood function, which allows the authors to work out the asymptotic distribution and to conduct inference on the coefficients.

Moon and Weidner (MW) show that  $\hat{\boldsymbol{\pi}}_{QMLE}$  is a consistent estimator of  $\boldsymbol{\pi}$ , as  $(N, T) \rightarrow \infty$  without any restrictions on the ratio  $T/N$ . To derive the asymptotic distribution of  $\hat{\boldsymbol{\pi}}_{QMLE}$ , MW require  $T/N \rightarrow \varkappa$ ,  $0 < \varkappa < \infty$ , as  $(N, T) \rightarrow \infty$ , and assume that the idiosyncratic errors,  $e_{it}$ , are cross-sectionally independent. Under certain high level assumptions, they show that  $\sqrt{NT}(\hat{\boldsymbol{\pi}}_{QMLE} - \boldsymbol{\pi})$  converges to a normal distribution with a non-zero mean, which is due to two types of asymptotic bias. The first follows from the heteroskedasticity of the error terms, as in Bai (2009), and the second is due to the presence of weakly exogenous regressors. The authors provide consistent estimators of these two components, and propose a bias-corrected QMLE.

There are, however, two important considerations that should be borne in mind when using the QMLE proposed by MW. First, it is developed for the case of full slope homogeneity, namely under  $\boldsymbol{\pi}_i = \boldsymbol{\pi}$  for all  $i$ . This assumption, for example, rules out the inclusion of fixed effects into the model which can be quite restrictive in practice. Although, the unobserved factor component,  $\boldsymbol{\gamma}'_i \mathbf{f}_t$ , does in principle allow for fixed effects if the first element of  $\mathbf{f}_t$  can be constrained to be unity at the estimation stage. A second consideration is the small sample properties of QMLE in the case of models with fixed effects, which are primarily of interest in empirical applications. Simulations reported in Chudik and Pesaran (2013a) suggest that the bias correction does not go far enough and the QMLE procedure could yield tests which are grossly over-sized. To check the robustness of the QMLE to the presence of fixed effects, we carried out a

small Monte Carlo experiment in the case of a homogeneous AR(1) panel data model with fixed effects,  $\lambda_i = 0.70$ , and  $N = T = 100$ . Using  $R = 2,000$  replications, the bias of the bias-corrected QMLE,  $\hat{\lambda}_{QMLE}$ , turned out to be  $-0.024$ , and tests based on  $\hat{\lambda}_{QMLE}$  were grossly oversized with the size exceeding 60%.

### 1.5.2 PC Estimators for Dynamic Panels

Song (2013) extends Bai (2009)'s approach to dynamic panels with heterogeneous coefficients. The focus of Song's analysis is on the estimation of unit-specific coefficients  $\boldsymbol{\pi}_i = (\lambda_i, \boldsymbol{\beta}'_i)'$ . In particular, Song proposes an iterated least squares estimator of  $\boldsymbol{\pi}_i$ , and shows as in Bai (2009) that the solution can be obtained by alternating the PC method applied to the least squares residuals and the least squares estimation of (41) until convergence. In particular, the least squares estimators of  $\boldsymbol{\pi}_i$  and  $\mathbf{F}$  are the solution to the following set of non-linear equations

$$\hat{\boldsymbol{\pi}}_{i,PC} = (\mathbf{\Xi}'_i \mathbf{M}_{\hat{\mathbf{F}}} \mathbf{\Xi}_i)^{-1} \mathbf{\Xi}'_i \mathbf{M}_{\hat{\mathbf{F}}} \mathbf{y}_i, \text{ for } i = 1, 2, \dots, N, \quad (44)$$

$$\frac{1}{NT} \sum_{i=1}^N (\mathbf{y}_i - \mathbf{\Xi}_i \hat{\boldsymbol{\pi}}_{i,PC}) (\mathbf{y}_i - \mathbf{\Xi}_i \hat{\boldsymbol{\pi}}_{i,PC})' \hat{\mathbf{F}} = \hat{\mathbf{F}} \hat{\mathbf{V}}. \quad (45)$$

Song (2013) establishes consistency of  $\hat{\boldsymbol{\pi}}_{i,PC}$  when  $(N, T) \rightarrow \infty$  without any restrictions on  $T/N$ . If in addition  $T/N^2 \rightarrow 0$ , Song (2013) shows that  $\hat{\boldsymbol{\pi}}_{i,PC}$  is  $\sqrt{T}$  consistent, and derives the asymptotic distribution under some additional requirements including the cross-sectional independence of  $e_{it}$ . Song (2013) does not provide theoretical results on the estimation of the mean coefficients  $\boldsymbol{\pi} = E(\boldsymbol{\pi}_i)$ , but he considers the following mean group estimator based on the individual estimates  $\hat{\boldsymbol{\pi}}_{i,PC}$ ,

$$\hat{\boldsymbol{\pi}}_{PCMG}^s = \frac{1}{N} \sum_{i=1}^N \hat{\boldsymbol{\pi}}_{i,PC},$$

in a Monte Carlo study and finds that  $\hat{\boldsymbol{\pi}}_{PCMG}^s$  has satisfactory small sample properties in terms of bias and root mean squared error. But he does not provide any results on the asymptotic distribution of  $\hat{\boldsymbol{\pi}}_{PCMG}^s$ . However, results of a Monte Carlo study presented in Chudik and Pesaran (2013a) suggest that  $\sqrt{N}(\hat{\boldsymbol{\pi}}_{PCMG}^s - \boldsymbol{\pi})$  is asymptotically normally distributed with mean zero and a covariance matrix that can be estimated by (as in the case of the CCEMG estimator),

$$\widehat{Var}(\hat{\boldsymbol{\pi}}_{PCMG}^s) = \frac{1}{N(N-1)} \sum_{i=1}^N (\hat{\boldsymbol{\pi}}_i^s - \hat{\boldsymbol{\pi}}_{MG}^s) (\hat{\boldsymbol{\pi}}_i^s - \hat{\boldsymbol{\pi}}_{MG}^s)'.$$

The test results based on this conjecture tend to perform well so long as  $T$  is sufficiently large. However, as with the other PC based estimators, knowledge of the number of

factors and the assumption that the factors under consideration are strong continue to play an important role in the small sample properties of the tests based on  $\hat{\pi}_{MGPC}^s$ .

### 1.5.3 Dynamic CCE Estimators

The CCE approach as it was originally proposed in Pesaran (2006) does not cover the case where the panel includes a lagged dependent variable or weakly exogenous regressors.<sup>10</sup> The extension of the CCE approach to dynamic panels with heterogeneous coefficients and weakly exogenous regressors is proposed by Chudik and Pesaran (2013a). In what follows we refer to this extension as dynamic CCE.

The inclusion of a lagged dependent variable amongst the regressors has three main consequences for the estimation of the mean coefficients. The first is the well known time series bias,<sup>11</sup> which affects the individual specific estimates and is of order  $O(T^{-1})$ . The second consequence is that the full rank condition becomes necessary for consistent estimation of the mean coefficients unless the  $\mathbf{f}_t$  is serially uncorrelated. The third complication arises from the interaction of dynamics and coefficient heterogeneity, which leads to infinite lag order relationships between unobserved common factors and cross-sectional averages of the observables when  $N$  is large. This issue also arises in cross-sectional aggregation of heterogeneous dynamic models. See Granger (1980) and Chudik and Pesaran (2014).

To illustrate these complications, using (41) and recalling assumption  $|\lambda_i| < 1$ , for all  $i$ , then we have

$$y_{it} = \sum_{\ell=0}^{\infty} \lambda_i^\ell \boldsymbol{\beta}'_i \mathbf{x}_{i,t-\ell} + \sum_{\ell=0}^{\infty} \lambda_i^\ell \boldsymbol{\gamma}'_i \mathbf{f}_{t-\ell} + \sum_{\ell=0}^{\infty} \lambda_i^\ell e_{i,t-\ell}. \quad (46)$$

Taking weighted cross-sectional averages, and assuming independence of  $\lambda_i$ ,  $\boldsymbol{\beta}_i$ , and  $\boldsymbol{\gamma}_i$ , strict exogeneity of  $\mathbf{x}_{it}$ , and weak cross-sectional dependence of  $\{e_{it}\}$ , we obtain (following the arguments in Chudik and Pesaran (2014))

$$\bar{y}_{wt} = a(L) \boldsymbol{\gamma}' \mathbf{f}_t + a(L) \boldsymbol{\beta}' \bar{\mathbf{x}}_{wt} + \xi_{wt}, \quad (47)$$

where  $a(L) = \sum_{\ell=0}^{\infty} a_\ell L^\ell$ , with  $a_\ell = E(\lambda_i^\ell)$ ,  $\boldsymbol{\beta} = E(\boldsymbol{\beta}_i)$ , and  $\boldsymbol{\gamma} = E(\boldsymbol{\gamma}_i)$ . Under the assumption that the idiosyncratic errors are cross-sectionally weakly dependent, we have  $\xi_{wt} \xrightarrow{P} 0$ , as  $N \rightarrow \infty$ , with the rate of convergence depending on the degree of cross-sectional dependence of  $\{e_{it}\}$  and the granularity of  $w$ . In the case where  $w$  satisfies the usual granularity conditions (1)-(2), and the exponent of cross-sectional dependence of  $e_{it}$  is  $\alpha_e \leq 1/2$ , we have  $\xi_{wt} = O_p(N^{-1/2})$ . In the special case where  $\boldsymbol{\beta} = 0$  and  $m = 1$ , (47) reduces to

$$\bar{y}_{wt} = \gamma a(L) f_t + O_p(N^{-1/2}).$$

The extent to which  $f_t$  can be accurately approximated by  $\bar{y}_{wt}$  and its lagged values depends on the rate at which  $a_\ell = E(\lambda_i^\ell)$ , the coefficients in the polynomial lag operator,  $a(L)$ , decay with  $\ell$ , and the size of the cross-section dimension,  $N$ . The coefficients in  $a(L)$  are given by the moments of  $\lambda_i$ , and therefore these coefficients need not be absolutely summable if the support of  $\lambda_i$  is not sufficiently restricted in the neighborhood of the unit circle (see Granger 1980 and Chudik and Pesaran 2014). Assuming that for all  $i$  the support of  $\lambda_i$  lies strictly within the unit circle, it is then easily seen that  $a_\ell$  will then decay exponentially and for  $N$  sufficiently large,  $f_t$  can be well approximated by  $\bar{y}_{wt}$  and a number of its lagged values.<sup>11</sup> The number of lagged values of  $\bar{y}_{wt}$  needed to approximate  $f_t$  rises with  $T$  but at a slower rate.<sup>12</sup>

In the general case where  $\beta$  is nonzero,  $x_{it}$  are weakly exogenous, and  $m \geq 1$ , Chudik and Pesaran (2013a) show that there exists the following large  $N$  distributed lag relationship between the unobserved common factors and cross-sectional averages of the dependent variable and the regressors,  $\bar{z}_{wt} = (\bar{y}_{wt}, \bar{x}'_{wt})'$ ,

$$\Lambda(L)\tilde{\Gamma}'f_t = \bar{z}_{wt} + O_p(N^{-1/2}),$$

where as before  $\tilde{\Gamma} = E(\gamma_i, \Gamma_i)$  and the decay rate of the matrix coefficients in  $\Lambda(L)$  depends on the heterogeneity of  $\lambda_i$  and  $\beta_i$  and other related distributional assumptions. The existence of a large  $N$  relationship between the unobserved common factors and cross-sectional averages of variables is not surprising considering that only the components with the largest exponents of cross-sectional dependence can survive cross-sectional aggregation with granular weights. Assuming  $\tilde{\Gamma}$  has full row rank, i.e.,  $\text{rank}(\tilde{\Gamma}) = m$ , and the distributions of coefficients are such that  $\Lambda^{-1}(L)$  exists and has exponentially decaying coefficients yields the following unit-specific dynamic CCE regressions,

$$y_{it} = \lambda_i y_{i,t-1} + \beta'_i x_{it} + \sum_{\ell=0}^{p_T} \delta'_{i\ell} \bar{z}_{w,t-\ell} + e_{yit}, \quad (48)$$

where  $\bar{z}_{wt}$  and its lagged values are used to approximate  $f_t$ . The error term  $e_{yit}$  consists of three parts: an idiosyncratic term,  $e_{it}$ , an error component due to the truncation of a possibly infinite distributed lag function, and an  $O_p(N^{-1/2})$  error component due to the approximation of unobserved common factors based on large  $N$  relationships:

Chudik and Pesaran (2013a) consider the least squares estimates of  $\pi_i = (\lambda_i, \beta'_i)'$  based on the above dynamic CCE regressions, denoted as  $\hat{\pi}_i = (\hat{\lambda}_i, \hat{\beta}'_i)'$ , and the mean group estimate of  $\pi = E(\pi_i)$  based on  $\hat{\pi}_i$ . To define these estimators, we introduce the following data matrices

$$\tilde{\Xi}_i = \begin{pmatrix} y_{ip_T} & x'_{i,p_T+1} \\ y_{i,p_T+1} & x'_{i,p_T+2} \\ \vdots & \vdots \\ y_{i,T-1} & x'_{iT} \end{pmatrix}, \quad \bar{Q}_w = \begin{pmatrix} \bar{z}'_{w,p_T+1} & \bar{z}'_{w,p_T} & \cdots & \bar{z}'_{w,1} \\ \bar{z}'_{w,p_T+2} & \bar{z}'_{w,p_T+1} & \cdots & \bar{z}'_{w,2} \\ \vdots & \vdots & & \vdots \\ \bar{z}'_{w,T} & \bar{z}'_{w,T-1} & \cdots & \bar{z}'_{w,T-p_T} \end{pmatrix}, \quad (49)$$

and the projection matrix  $\bar{\mathbf{M}}_q = \mathbf{I}_{T-p_T} - \bar{\mathbf{Q}}_w(\bar{\mathbf{Q}}'_w\bar{\mathbf{Q}}_w)^+\bar{\mathbf{Q}}'_w$ , where  $\mathbf{I}_{T-p_T}$  is a  $(T-p_T) \times (T-p_T)$  dimensional identity matrix.<sup>13</sup>  $p_T$  should be set such that  $p_T^2/T$  tends to zero as  $p_T$  and  $T$  both tend to infinity. Monte Carlo experiments reported in Chudik and Pesaran (2013a) suggest that setting  $p_T = T^{1/3}$  could be a good choice in practice.

The individual estimates,  $\hat{\boldsymbol{\pi}}_i$ , can now be written as

$$\hat{\boldsymbol{\pi}}_i = \left( \tilde{\boldsymbol{\Xi}}'_i \bar{\mathbf{M}}_q \tilde{\boldsymbol{\Xi}}_i \right)^{-1} \tilde{\boldsymbol{\Xi}}'_i \bar{\mathbf{M}}_q \tilde{\mathbf{y}}_i, \quad (50)$$

where  $\tilde{\mathbf{y}}_i = (y_{i,p_T+1}, y_{i,p_T+2}, \dots, y_{i,T})'$ . The mean group estimator of  $\boldsymbol{\pi} = E(\boldsymbol{\pi}_i) = (\lambda, \boldsymbol{\beta}')'$  is given by

$$\hat{\boldsymbol{\pi}}_{MG} = \frac{1}{N} \sum_{i=1}^N \hat{\boldsymbol{\pi}}_i. \quad (51)$$

Chudik and Pesaran (2013a) show that  $\hat{\boldsymbol{\pi}}_i$  and  $\hat{\boldsymbol{\pi}}_{MG}$  are consistent estimators of  $\boldsymbol{\pi}_i$  and  $\boldsymbol{\pi}$ , respectively, assuming that the rank condition is satisfied and  $(N, T, p_T) \rightarrow \infty$  such that  $p_T^3/T \rightarrow \varkappa$ ,  $0 < \varkappa < \infty$ , but without any restrictions on the ratio  $N/T$ . The rank condition is necessary for the consistency of  $\hat{\boldsymbol{\pi}}_i$  because the unobserved factors are allowed to be correlated with the regressors. If the unobserved common factors were serially uncorrelated (but still correlated with  $x_{it}$ ), then  $\hat{\boldsymbol{\pi}}_{MG}$  is consistent also in the rank deficient case, despite the inconsistency of  $\hat{\boldsymbol{\pi}}_i$ , so long as factor loadings are independently and identically distributed across  $i$ . The convergence rate of  $\hat{\boldsymbol{\pi}}_{MG}$  is  $\sqrt{N}$  due to the heterogeneity of the slope coefficients. Chudik and Pesaran (2013a) show that  $\hat{\boldsymbol{\pi}}_{MG}$  converges to a normal distribution as  $(N, T, p_T) \rightarrow \infty$  such that  $p_T^3/T \rightarrow \varkappa_1$  and  $T/N \rightarrow \varkappa_2$ ,  $0 < \varkappa_1, \varkappa_2 < \infty$ . The ratio  $N/T$  needs to be restricted for conducting inference, due to the presence of small time series bias. In the full rank case, the asymptotic variance of  $\hat{\boldsymbol{\pi}}_{MG}$  is given by the variance of  $\boldsymbol{\pi}_i$  alone. When the rank condition does not hold, but factors are serially uncorrelated, then the asymptotic variance depends also on other parameters, including the variance of factor loadings. In both cases the asymptotic variance can be consistently estimated non-parametrically, as in (38).

Monte Carlo experiments in Chudik and Pesaran (2013a) show that the dynamic CCE approach performs reasonably well (in terms of bias, RMSE, size and power). This is particularly the case when the parameter of interest is the average slope of the regressors ( $\boldsymbol{\beta}$ ), where the small sample results are quite satisfactory even if  $N$  and  $T$  are relatively small (around 40). But the situation is different if the parameter of interest is the mean coefficient of the lagged dependent variable ( $\lambda$ ). In the case of  $\lambda$ , the CCEMG estimator suffers from the well known time series bias and tests based on it tend to be over-sized, unless  $T$  is sufficiently large. To mitigate the consequences of this bias, Chudik and Pesaran (2013a) consider application of a half-panel jackknife procedure (Dhaene and Jochmans 2012), and the recursive mean adjustment procedure (So and Shin 1999), both of which are easy to implement. The proposed jackknife bias-corrected CCEMG estimator is found to be more effective in mitigating the time series

bias, but it cannot fully deal with the size distortion when  $T$  is relatively small. Improving the small  $T$  sample properties of the CCEMG estimator of  $\lambda$  in the heterogeneous panel data models still remains a challenge to be taken on in the future.

The application of the CCE approach to static panels with weakly exogenous regressors (namely without lagged dependent variables) has not yet been investigated in the literature. In order to investigate whether the standard CCE mean group and pooled estimators could be applied in this setting, we conducted Monte Carlo experiments. We used the following data generating process

$$y_{it} = c_{yi} + \beta_{0i}x_{it} + \beta_{1i}x_{i,t-1} + u_{it}, \quad u_{it} = \boldsymbol{\gamma}'_t \mathbf{f}_t + \varepsilon_{it}, \quad (52)$$

and

$$x_{it} = c_{xi} + \alpha_{xi}y_{i,t-1} + \boldsymbol{\gamma}'_{xi} \mathbf{f}_t + v_{it}, \quad (53)$$

for  $i = 1, 2, \dots, N$ , and  $t = -99, \dots, 0, 1, 2, \dots, T$  with the starting values  $y_{i,-100} = x_{i,-100} = 0$ . This set up allows for feedbacks from  $y_{i,t-1}$  to the regressors, thus rendering  $x_{it}$  weakly exogenous. The size of the feedback is measured by  $\alpha_{xi}$ . The unobserved common factors in  $\mathbf{f}_t$  and the unit-specific components  $v_{it}$  are generated as independent stationary AR(1) processes:

$$\begin{aligned} f_{\ell t} &= \rho_{f\ell} f_{\ell-1,t} + \varsigma_{f\ell t}, \quad \varsigma_{f\ell t} \sim IIDN\left(0, 1 - \rho_{f\ell}^2\right), \\ v_{it} &= \rho_{xi} v_{i,t-1} + \varsigma_{it}, \quad \varsigma_{it} \sim IIDN\left(0, \sigma_{vi}^2\right), \end{aligned} \quad (54)$$

for  $i = 1, 2, \dots, N$ ,  $\ell = 1, 2, \dots, m$ , and for  $t = -99, \dots, 0, 1, 2, \dots, T$  with the starting values  $f_{\ell,-100} = 0$  and  $v_{i,-100} = 0$ . The first 100 time observations ( $t = -99, -98, \dots, 0$ ) are discarded. We generate  $\rho_{xi}$ , for  $i = 1, 2, \dots, N$  as  $IIDU[0, 0.95]$ , and set  $\rho_{f\ell} = 0.6$ , for  $\ell = 1, 2, \dots, m$ . We also set  $\sigma_{vi} = \sqrt{1 - [E(\rho_{xi})]^2}$  for all  $i$ .

The fixed effects are generated as  $c_{yi} \sim IIDN(1, 1)$ ,  $c_{xi} = c_{yi} + \varsigma_{cxi}$ , where  $\varsigma_{cxi} \sim IIDN(0, 1)$ , thus allowing for dependence between  $x_{it}$  and  $c_{yi}$ . We set  $\beta_{1i} = -0.5$  for all  $i$ , and generate  $\beta_{0i}$  as  $IIDU(0.5, 1)$ . We consider two possibilities for the feedback coefficients  $\alpha_{xi}$ : weakly exogenous regressors where we generate  $\alpha_{xi}$  as draws from  $IIDU(0, 1)$  (in which case  $E(\alpha_{xi}) = 0.5$ ), and strictly exogenous regressors where we set  $\alpha_{xi} = 0$  for all  $i$ . We consider  $m = 3$  unobserved common factors, with all factor loadings generated independently in the same way as in Chudik and Pesaran (2013a). Similarly, the idiosyncratic errors,  $\varepsilon_{it}$ , are generated as in Chudik and Pesaran (2013a) to be heteroskedastic and weakly cross-sectionally dependent. We consider the following combinations of sample sizes:  $N \in \{40, 50, 100, 150, 200\}$ ,  $T \in \{20, 50, 100, 150, 200\}$ , and set the number of replications to  $R = 2000$ .

The small sample results for the CCE mean group and pooled estimators (with lagged augmentations) in the case of these experiments with weakly exogenous regressors are presented on the upper panel of Table 1.1. The rank condition in these experiment does not hold, but this does not seem to cause any major problems for the CCE mean group estimator, which performs very well (in terms of bias and RMSE)

for  $T > 50$  and for all values of  $N$ . Also tests based on this estimator are correctly sized and have good power properties. When  $T \leq 50$ , we observe a negative bias and the tests are oversized (the rejection rates are in the range of 9 to 75 percent, depending on the sample size). The CCE pooled estimator, however, is no longer consistent in the case of weakly exogenous regressors with heterogeneous coefficients, due to the bias caused by the correlation between the slope coefficients and the regressors. For comparison, we also provide, at the bottom panel of Table 1.1, the results of the same experiments but with strictly exogenous regressors ( $\alpha_{xi} = 0$ ), where the bias is negligible and all tests are correctly sized.

## 1.6 TESTS OF ERROR CROSS-SECTIONAL DEPENDENCE

---

In this section we provide an overview of alternative approaches to testing the cross-sectional independence or weak dependence of the errors in the following panel data model

$$y_{it} = a_i + \boldsymbol{\beta}'_i \mathbf{x}_{it} + u_{it}, \quad (55)$$

where  $a_i$  and  $\boldsymbol{\beta}_i$  for  $i = 1, 2, \dots, N$  are assumed to be fixed unknown coefficients, and  $\mathbf{x}_{it}$  is a  $k$ -dimensional vector of regressors. We consider both cases where the regressors are strictly and weakly exogenous, as well as when they include lagged values of  $y_{it}$ .

The literature on testing for error cross-sectional dependence in large panels follows two separate strands, depending on whether the cross-section units are ordered or not. In the case of ordered data sets (which could arise when observations are spatial or belong to given economic or social networks) tests of cross-sectional independence that have high power with respect to such ordered alternatives have been proposed in the spatial econometrics literature. A prominent example of such tests is Moran's I test. See Moran (1948) with further developments by Anselin (1988), Anselin and Bera (1998), Haining (2003), and Baltagi et al. (2003).

In the case of cross-section observations that do not admit an ordering, tests of cross-sectional dependence are typically based on estimates of pair-wise error correlations ( $\rho_{ij}$ ) and are applicable when  $T$  is sufficiently large so that relatively reliable estimates of  $\rho_{ij}$  can be obtained. An early test of this type is the Lagrange multiplier (LM) test of Breusch and Pagan (1980, pp. 247–248) which tests the null hypothesis that *all* pair-wise correlations are zero, namely that  $\rho_{ij} = 0$  for all  $i \neq j$ . This test is based on the average of the *squared* estimates of pair-wise correlations, and under standard regularity conditions it is shown to be asymptotically (as  $T \rightarrow \infty$ ) distributed as  $\chi^2$  with  $N(N - 1)/2$  degrees of freedom. The LM test tends to be highly over-sized in the case of panels with relatively large  $N$ .

**Table 1.1** Small sample properties of CCEMG and CCEP estimators of mean slope coefficients in panel data models with weakly and strictly exogenous regressors

(N,T)	Bias (x100)					RMSE (x100)					Size (x100)					Power (x100)				
	20	50	100	150	200	20	50	100	150	200	20	50	100	150	200	20	50	100	150	200
Experiments with weakly exogenous regressors																				
CCEMG																				
40	-5.70	-1.46	-0.29	0.00	0.11	7.82	3.65	2.80	2.67	2.61	23.70	9.35	6.20	6.05	6.25	86.80	94.05	96.00	96.30	96.95
50	-5.84	-1.56	-0.39	0.04	0.11	7.56	3.43	2.56	2.41	2.33	29.50	9.30	7.00	6.70	6.20	93.40	96.75	98.75	98.70	99.20
100	-5.88	-1.50	-0.41	-0.05	0.07	6.82	2.63	1.83	1.70	1.64	46.70	13.10	6.00	5.75	5.25	99.75	99.95	100.00	100.00	100.00
150	-6.11	-1.59	-0.45	-0.11	0.08	6.73	2.36	1.53	1.35	1.30	66.05	16.15	6.60	4.75	4.80	100.00	100.00	100.00	100.00	100.00
200	-6.04	-1.55	-0.43	-0.12	0.01	6.54	2.17	1.37	1.18	1.18	74.65	19.70	7.35	4.50	6.10	100.00	100.00	100.00	100.00	100.00
CCEP																				
40	-3.50	-0.09	0.76	0.98	1.23	6.58	3.71	3.33	3.24	3.35	14.80	6.75	7.50	7.55	9.85	72.30	78.45	80.55	82.70	82.55
50	-3.55	-0.27	0.70	1.08	1.19	6.07	3.31	2.96	3.00	2.96	14.00	5.70	6.20	8.65	8.80	79.70	86.90	88.55	88.70	90.90
100	-3.56	-0.10	0.76	1.08	1.17	5.11	2.42	2.22	2.27	2.26	21.75	5.50	6.75	9.10	10.45	96.05	97.80	98.80	98.95	99.30
150	-3.78	-0.10	0.74	1.10	1.16	4.86	1.98	1.87	1.99	1.98	30.45	5.85	7.60	11.45	12.60	99.15	99.75	99.95	99.95	100.00
200	-3.66	-0.19	0.80	1.08	1.13	4.56	1.77	1.67	1.78	1.77	35.65	6.25	8.35	12.50	12.45	100.00	100.00	100.00	100.00	100.00
Experiments with strictly exogenous regressors																				
CCEMG																				
40	0.19	-0.05	0.02	0.07	0.04	6.43	3.91	3.06	2.91	2.75	6.20	6.40	4.60	6.40	5.55	36.20	74.40	89.95	93.90	95.60
50	-0.02	0.08	0.11	-0.05	-0.02	5.72	3.48	2.83	2.68	2.46	5.25	6.10	5.90	6.75	5.75	43.90	82.20	93.70	96.80	98.05
100	-0.06	0.01	0.02	-0.05	-0.01	4.13	2.42	2.02	1.79	1.78	5.55	6.45	4.90	4.95	6.20	69.95	97.60	99.75	99.95	100.00
150	0.06	0.03	0.00	0.02	0.01	3.29	2.03	1.62	1.50	1.42	5.40	6.00	5.50	5.05	5.30	85.65	99.95	100.00	100.00	100.00
200	-0.06	0.03	-0.02	-0.03	-0.01	2.87	1.75	1.39	1.33	1.23	4.50	5.30	4.85	6.50	5.15	94.10	100.00	100.00	100.00	100.00
CCEP																				
40	0.21	0.17	0.02	-0.01	-0.02	5.78	3.85	3.16	3.08	2.85	6.40	6.45	5.95	7.10	6.35	74.55	72.90	88.10	92.15	93.50
50	0.03	-0.01	-0.13	0.02	-0.02	5.20	3.48	2.84	2.59	2.54	5.60	6.25	6.25	6.00	5.95	83.35	83.30	94.80	96.30	97.30
100	-0.01	-0.06	0.05	-0.04	0.07	3.67	2.56	2.03	1.89	1.76	5.60	6.15	5.00	5.35	5.65	98.50	97.75	99.85	100.00	100.00
150	0.05	0.02	0.02	0.01	0.01	2.95	2.02	1.65	1.52	1.49	4.50	5.20	5.50	4.95	5.60	99.80	99.95	100.00	100.00	100.00
200	-0.09	-0.04	-0.06	0.03	0.02	2.57	1.74	1.43	1.38	1.28	6.05	5.75	5.15	5.75	4.95	100.00	100.00	100.00	100.00	100.00

Notes: Observations are generated as  $y_{it} = c_{yi} + \beta_{0i}x_{it} + \beta_{1i}x_{i,t-1} + u_{it}$ ,  $u_{it} = \gamma_i'f_t + \varepsilon_{it}$ , and  $x_{it} = c_{xi} + \alpha_{xi}y_{i,t-1} + \gamma_{xi}'f_t + v_{it}$ , (see (52)-(53)), where  $\beta_{0i} \sim IIDU(0.5, 1)$ ,  $\beta_{1i} = -0.5$  for all  $i$ , and  $m = 3$  (number of unobserved common factors). Fixed effects are generated as  $c_{yi} \sim IIDN(1, 1)$ , and  $c_{xi} = c_{yi} + IIDN(0, 1)$ . In the case of weakly exogenous regressors,  $\alpha_{xi} \sim IIDU(0, 1)$  (with  $E(\alpha_{xi}) = 0.5$ ), and under the case of strictly exogenous regressors  $\alpha_{xi} = 0$  for all  $i$ . The errors are generated to be heteroskedastic and weakly cross-sectionally dependent. See Section 1.5.3 for a more detailed description of the MC design.

In what follows we review the various attempts made in the literature to develop tests of cross-sectional dependence when  $N$  is large and the cross-section units are unordered. But before proceeding further, we first need to consider the appropriateness of the null hypothesis of cross-sectional "independence" or "uncorrelatedness", that underlie the LM test of Breusch and Pagan (1980), namely that all  $\rho_{ij}$  are zero for all  $i \neq j$ , when  $N$  is large. The null that underlies the LM test is sensible when  $N$  is small and fixed as  $T \rightarrow \infty$ . But when  $N$  is relatively large and rising with  $T$ , it is unlikely to matter if out of the total  $N(N - 1)/2$  pair-wise correlations only a few are non-zero. Accordingly, Pesaran (2014) argues that the null of cross-sectionally uncorrelated errors, defined by

$$H_0 : E(u_{it} u_{jt}) = 0, \text{ for all } t \text{ and } i \neq j, \quad (56)$$

is restrictive for large panels and the null of a sufficiently weak cross-sectional dependence could be more appropriate since mere incidence of isolated error dependencies are of little consequence for estimation or inference about the parameters of interest, such as the individual slope coefficients,  $\beta_i$ , or their average value,  $E(\beta_i) = \beta$ .

Consider the panel data model (55), and let  $\hat{u}_{it}$  be the OLS estimator of  $u_{it}$  defined by

$$\hat{u}_{it} = y_{it} - \hat{a}_i - \hat{\beta}'_i x_{it}, \quad (57)$$

with  $\hat{a}_i$ , and  $\hat{\beta}_i$  being the OLS estimates of  $a_i$  and  $\beta_i$ , based on the  $T$  sample observations,  $y_t, x_{it}$ , for  $t = 1, 2, \dots, T$ . Consider the sample estimate of the pair-wise correlation of the residuals,  $\hat{u}_{it}$  and  $\hat{u}_{jt}$ , for  $i \neq j$

$$\hat{\rho}_{ij} = \hat{\rho}_{ji} = \frac{\sum_{t=1}^T \hat{u}_{it} \hat{u}_{jt}}{\left( \sum_{t=1}^T \hat{u}_{it}^2 \right)^{1/2} \left( \sum_{t=1}^T \hat{u}_{jt}^2 \right)^{1/2}}.$$

In the case where the  $u_{it}$  is symmetrically distributed and the regressors are strictly exogenous, then under the null hypothesis of no cross-sectional dependence,  $\hat{\rho}_{ij}$  and  $\hat{\rho}_{is}$  are cross-sectionally uncorrelated for all  $i, j$  and  $s$  such that  $i \neq j \neq s$ . This follows since

$$E(\hat{\rho}_{ij} \hat{\rho}_{is}) = \sum_{t=1}^T \sum_{t'=1}^T E(\hat{\eta}_{it} \hat{\eta}_{it'} \hat{\eta}_{jt} \hat{\eta}_{st'}) = \sum_{t=1}^T \sum_{t'=1}^T E(\hat{\eta}_{it} \hat{\eta}_{it'}) E(\hat{\eta}_{jt}) E(\hat{\eta}_{st'}) = 0. \quad (58)$$

where  $\hat{\eta}_{it} = \hat{u}_{it} / (\sum_{t=1}^T \hat{u}_{it}^2)^{1/2}$ . Note when  $x_{it}$  is strictly exogenous for each  $i$ ,  $\hat{u}_{it}$ , being a linear function of  $u_{it}$ , for  $t = 1, 2, \dots, T$ , will also be symmetrically distributed with zero means, which ensures that  $\eta_{it}$  is also symmetrically distributed around its mean which is zero. Further, under (56) and when  $N$  is finite, it is known that (see Pesaran 2004)

$$\sqrt{T} \hat{\rho}_{ij} \xrightarrow{a} N(0, 1), \quad (59)$$

for a given  $i$  and  $j$ , as  $T \rightarrow \infty$ . The above result has been widely used for constructing tests based on the sample correlation coefficient or its transformations. Noting

that, from (59),  $T\hat{\rho}_{ij}^2$  is asymptotically distributed as a  $\chi_1^2$ , it is possible to consider the following statistic

$$CD_{LM} = \sqrt{\frac{1}{N(N-1)}} \sum_{i=1}^{N-1} \sum_{j=i+1}^N (T\hat{\rho}_{ij}^2 - 1). \quad (60)$$

Based on the Euclidean norm of the matrix of sample correlation coefficients, (60) is a version of the Lagrange Multiplier test statistic due to Breusch and Pagan (1980). Frees (1995) first explored the finite sample properties of the LM statistic, calculating its moments for fixed values of  $T$  and  $N$ , under the normality assumption. He advanced a non-parametric version of the LM statistic based on the Spearman rank correlation coefficient. Dufour and Khalaf (2002) have suggested applying Monte Carlo exact tests to correct the size distortions of  $CD_{LM}$  in finite samples. However, these tests, being based on the bootstrap method applied to the  $CD_{LM}$ , are computationally intensive, especially when  $N$  is large.

An alternative adjustment to the LM test is proposed by Pesaran et al. (2008), where the LM test is centered to have a zero mean for a fixed  $T$ . These authors also propose a correction to the variance of the LM test. The basic idea is generally applicable, but analytical bias corrections can be obtained only under the assumption that the regressors,  $\mathbf{x}_{it}$ , are strictly exogenous and the errors,  $u_{it}$  are normally distributed. Under these assumptions, Pesaran et al. (2008) show that the exact mean and variance of  $(N-k)\hat{\rho}_{ij}^2$  are given by:

$$\begin{aligned} \mu_{Tij} &= E[(N-k)\hat{\rho}_{ij}^2] = \frac{1}{T-k} Tr[E(\mathbf{M}_i \mathbf{M}_j)], \\ v_{Tij}^2 &= Var[(N-k)\hat{\rho}_{ij}^2] = \{Tr[E(\mathbf{M}_i \mathbf{M}_j)]\}^2 a_{1T} + 2 \left\{ Tr[E(\mathbf{M}_i \mathbf{M}_j)^2] \right\} a_{2T}, \end{aligned}$$

where  $a_{1T} = a_{2T} - (\frac{1}{T-k})^2$ , and  $a_{2T} = 3[\frac{(T-k-8)(T-k+2)+24}{(T-k+2)(T-k-2)(T-k-4)}]^2$ ,  $\mathbf{M}_i = \mathbf{I}_T - \tilde{\mathbf{X}}_i(\tilde{\mathbf{X}}_i'\tilde{\mathbf{X}}_i)^{-1}\tilde{\mathbf{X}}_i'$  and  $\tilde{\mathbf{X}}_i$  is a  $T \times (k+1)$  matrix of observations on  $(1, \mathbf{x}'_{it})'$ . The adjusted LM statistic is now given by

$$LM_{Adj} = \sqrt{\frac{2}{N(N-1)}} \sum_{i=1}^{N-1} \sum_{j=i+1}^N \frac{(T-k)\hat{\rho}_{ij}^2 - \mu_{Tij}}{v_{Tij}}, \quad (61)$$

which is asymptotically  $N(0, 1)$  under  $H_0$ ,  $T \rightarrow \infty$  followed by  $N \rightarrow \infty$ . The asymptotic distribution of  $LM_{Adj}$  is derived under sequential asymptotics, but it might be possible to establish it under the joint asymptotics following the method of proof in Schott (2005) or Pesaran (2014).

The application of the  $LM_{Adj}$  test to dynamic panels or panels with weakly exogenous regressors is further complicated by the fact that the bias corrections depend on the true values of the unknown parameters and will be difficult to implement. The implicit null of LM tests when  $T$  and  $N \rightarrow \infty$  jointly rather than sequentially could also differ from the null of uncorrelatedness of all pair-wise correlations. To overcome some of

these difficulties, Pesaran (2004) has proposed a test that has exactly mean zero for fixed values of  $T$  and  $N$ . This test is based on the average of pair-wise correlation coefficients

$$CD_P = \sqrt{\frac{2T}{N(N-1)}} \left( \sum_{i=1}^{N-1} \sum_{j=i+1}^N \hat{\rho}_{ij} \right). \quad (62)$$

As it is established in (58), under the null hypothesis  $\hat{\rho}_{ij}$  and  $\hat{\rho}_{is}$  are uncorrelated for all  $i \neq j \neq s$ , but they need not be independently distributed when  $T$  is finite. Therefore, the standard central limit theorems cannot be applied to the elements of the double sum in (62) when  $(N, T) \rightarrow \infty$  jointly, and as shown in Pesaran (2014, Theorem 2) the derivation of the limiting distribution of the  $CD_P$  statistic involves a number of complications. It is also important to bear in mind that the implicit null of the test in the case of large  $N$  depends on the rate at which  $T$  expands with  $N$ . Indeed, as argued in Pesaran (2004), under the null hypothesis of  $\rho_{ij} = 0$  for all  $i \neq j$ , we continue to have  $E(\hat{\rho}_{ij}) = 0$ , even when  $T$  is fixed, so long as  $u_{it}$  are symmetrically distributed around zero, and the  $CD_P$  test continues to hold.

Pesaran (2014) extends the analysis of the  $CD_P$  test and shows that the implicit null of the test is weak cross-sectional dependence. In particular, the implicit null hypothesis of the test depends on the relative expansion rates of  $N$  and  $T$ .<sup>14</sup> Using the exponent of cross-sectional dependence,  $\alpha$ , developed in Bailey et al. (2012) and discussed above, Pesaran (2014) shows that when  $T = O(N^\epsilon)$  for some  $0 < \epsilon \leq 1$ , the implicit null of the  $CD_P$  test is given by  $0 \leq \alpha < (2 - \epsilon)/4$ . This yields the range  $0 \leq \alpha < 1/4$  when  $(N, T) \rightarrow \infty$  at the same rate such that  $T/N \rightarrow \varkappa$  for some finite positive constant  $\varkappa$ , and the range  $0 \leq \alpha < 1/2$  when  $T$  is small relative to  $N$ . For larger values of  $\alpha$ , as shown by Bailey et al. (2012),  $\alpha$  can be estimated consistently using the variance of the cross-sectional averages.

Monte Carlo experiments reported in Pesaran (2014) show that the  $CD_P$  test has good small sample properties for values of  $\alpha$  in the range  $0 \leq \alpha \leq 1/4$ , even in cases where  $T$  is small relative to  $N$ , as well as when the test is applied to residuals from pure autoregressive panels so long as there are no major asymmetries in the error distribution.

Other statistics have also been proposed in the literature to test for zero contemporaneous correlation in the errors of panel data model (55).<sup>16</sup> Using results from the literature on *spacing* discussed in (Pyke 1965), Ng (2006) considers a statistic based on the  $q^{th}$  differences of the cumulative normal distribution associated to the  $N(N-1)/2$  pair-wise correlation coefficients ordered from the smallest to the largest, in absolute value. Building on the work of John (1971), and under the assumption of normal disturbances, strictly exogenous regressors, and homogeneous slopes, Baltagi et al. (2011) propose a test of the null hypothesis of sphericity, defined by

$$H_0^{BKF} : \mathbf{u}_t \sim IIDN(\mathbf{0}, \sigma_u^2 \mathbf{I}_N),$$

based on the statistic

$$J_{BFK} = \frac{T \left( \text{tr}(\hat{\mathbf{S}})/N \right)^{-2} \text{tr}(\hat{\mathbf{S}}^2)/N - T - N}{2} - \frac{1}{2} - \frac{N}{2(T-1)}, \quad (63)$$

where  $\hat{\mathbf{S}}$  is the  $N \times N$  sample covariance matrix, computed using the fixed effects residuals under the assumption of slope homogeneity,  $\beta_i = \beta$ . Under  $H_0^{BFK}$ , errors  $u_{it}$  are cross-sectionally independent and homoskedastic and the  $J_{BFK}$  statistic converges to a standardized normal distribution as  $(N, T) \rightarrow \infty$  such that  $N/T \rightarrow \varkappa$  for some finite positive constant  $\varkappa$ . The rejection of  $H_0^{BFK}$  could be caused by cross-sectional dependence, heteroskedasticity, slope heterogeneity, and/or non-normal errors. Simulation results reported in Baltagi et al. (2011) show that this test performs well in the case of homoskedastic, normal errors, strictly exogenous regressors, and homogeneous slopes, although it is oversized for panels with large  $N$  and small  $T$ , and is sensitive to non-normality of disturbances. Joint assumption of homoskedastic errors and homogeneous slopes is quite restrictive in applied work and therefore the use of the  $J_{BFK}$  statistics as a test of cross-sectional dependence should be approached with care.

A slightly modified version of the  $CD_{LM}$  statistic, given by

$$LM_S = \sqrt{\frac{T+1}{N(N-1)(T-2)}} \sum_{i=1}^{N-1} \sum_{j=i+1}^N \left[ (T-1) \hat{\rho}_{ij}^2 - 1 \right] \quad (64)$$

has also been considered by Schott (2005), who shows that when the  $LM_S$  statistic is computed based on normally distributed observations, as opposed to panel residuals, it converges to  $N(0, 1)$  under  $\rho_{ij} = 0$  for all  $i \neq j$  as  $(N, T) \rightarrow \infty$  such that  $N/T \rightarrow \varkappa$  for some  $0 < \varkappa < \infty$ . Monte Carlo simulations reported in Jensen and Schmidt (2011) suggests that the  $LM_S$  test has good size properties for various sample sizes when applied to panel residuals in the case when slopes are homogeneous and estimated using the fixed effects approach. However, the  $LM_S$  test can lead to severe over-rejection when the slopes are in fact heterogeneous and the fixed effects estimators are used. The over-rejection of the  $LM_S$  test could persist even if mean group estimates are used in the computation of the residuals to take care of slope heterogeneity. This is because for relatively small values of  $T$ , unlike the  $LM_{Adj}$  statistic defined by (61), the  $LM_S$  statistic defined by (64) is not guaranteed to have a zero mean exactly.

The problem of testing for cross-sectional dependence in limited dependent variable panel data models with strictly exogenous covariates has also been investigated by Hsiao et al. (2012). In this paper the authors derive a LM test and show that in terms of the generalized residuals of Gourieroux et al. (1987), the test reduces to the LM test of Breusch and Pagan (1980). However, not surprisingly as with the linear panel data models, the LM test based on generalized residuals tends to over-reject in panels with large  $N$ . They then develop a CD type test based on a number of different residuals, and using Monte Carlo experiments they find that the CD test performs well for most combinations of  $N$  and  $T$ .

Sarafidis, Yagamata, and Robertson (2009) propose a test for the null hypothesis of homogeneous cross-sectional dependence

$$H_0 : \text{Var}(\boldsymbol{\gamma}_i) = 0, \quad (65)$$

in a lagged dependent variable model with regressors and residual factor structure (41)-(42) with cross-sectionally uncorrelated idiosyncratic innovations  $e_{it}$  against the alternative of heterogeneous cross-sectional dependence

$$H_1 : \text{Var}(\boldsymbol{\gamma}_i) \neq 0. \quad (66)$$

Following Sargan (1988) and exploring two different sets of moment conditions, one valid only under the null and the other valid under both hypotheses, Sarafidis, Yagamata, and Robertson (2009) derive Sargan's difference test based on the first-differenced as well as system based GMM estimators in a large  $N$  and fixed  $T$  setting. The null hypothesis (65) does not imply that the errors are cross-sectionally uncorrelated, and it allows examination of whether any cross-section dependence of errors remains after including time dummies, or after the data is transformed in terms of deviations from time-specific averages. In such cases the  $CD_P$  test lacks power and the test by Sarafidis, Yagamata, and Robertson (2009) could have some merits.

The existing literature on testing for error cross-sectional dependence, with the exception of Sarafidis, Yagamata, and Robertson (2009), has mostly focused on the case of strictly exogenous regressors. This assumption is required for both  $LM_{Adj}$  and  $J_{BFR}$  tests, while Pesaran (2004) shows that the  $CD_P$  test is also applicable to autoregressive panel data models so long as the errors are symmetrically distributed. The properties of the  $CD_P$  test for dynamic panels that include weakly or strictly exogenous regressors have not yet been investigated.

We conduct Monte Carlo experiments to investigate the performance of these tests in the case of dynamic panels and to shed light also on the performance of the  $LM_S$  test in the case of heterogeneous slopes. We generate the dependent variable and the regressors in the same way as described in Section 1.5.3 with the following two exceptions. First, we introduce lags of the dependent variable in (60):

$$y_{it} = c_{yi} + \lambda_i y_{i,t-1} + \beta_{0i} x_{it} + \beta_{1i} x_{i,t-1} + u_{it}. \quad (67)$$

and generate  $\lambda_i$  as  $IIDU(0, 0.8)$ . As discussed in Chudik and Pesaran (2013a), the lagged dependent variable coefficients,  $\lambda_i$ , and the feedback coefficients,  $\alpha_{xi}$ , in (53) need to be chosen such as to ensure the variances of  $y_{it}$  remain bounded. We generate  $\alpha_{xi}$  as  $IIDU(0, 0.35)$ , which ensures that this condition is met and  $E(\alpha_{xi}) = 0.35/2$ . For comparison purposes, we also consider the case of strictly exogenous regressors where we set  $\lambda_i = \alpha_{xi} = 0$  for all  $i$ . The second exception is the generation of the reduced form errors. In order to consider different options for cross-sectional dependence, we use the following residual factor model to generate the errors  $u_{it}$ .

$$u_{it} = \gamma_i g_t + \varepsilon_{it}, \quad (68)$$

where  $\varepsilon_{it} \sim IIDN(0, \frac{1}{2}\sigma_i^2)$  with  $\sigma_i^2 \sim \chi^2(2)$ ,  $g_t \sim IIDN(0, 1)$  and the factor loadings are generated as

$$\begin{aligned}\gamma_i &= v_{\gamma i}, \text{ for } i = 1, 2, \dots, M_\alpha, \\ \gamma_i &= 0, \text{ for } i = M_\alpha + 1, M_\alpha + 2, \dots, N,\end{aligned}$$

where  $M_\alpha = [N^\alpha]$ ,  $v_{\gamma i} \sim IIDU[\mu_\nu - 0.5, \mu_\nu + 0.5]$ . We set  $\mu_\nu = 1$ , and consider four values of the exponent of the cross-sectional dependence for the errors, namely  $\alpha = 0, 0.25, 0.5$ , and  $0.75$ . We also consider the following combinations of  $N \in \{40, 50, 100, 150, 200\}$ , and  $T \in \{20, 50, 100, 150, 200\}$ , and use 2000 replications for all experiments.

Table 1.2 presents the findings for the  $CD_P$ ,  $LM_{Adj}$ , and  $LM_S$  tests. The rejection rates for  $J_{BFK}$  in all cases, including the cross-sectionally independent case of  $\alpha = 0$ , were all close to 100%, in part due to the error variance heteroskedasticity, and are not included in Table 1.2. The top panel of Table 1.2 reports the test results for the case of strictly exogenous regressors, and the bottom part gives the results for the panel data models with weakly exogenous regressors. We see that the  $CD_P$  test continues to perform well even when the panel data model contains a lagged dependent variable and other weakly exogenous regressors, for the combination of  $N$  and  $T$  samples considered. The results also confirm the theoretical finding discussed above that shows the implicit null of the  $CD_P$  test is  $0 \leq \alpha \leq 0.25$ . In contrast, the  $LM_{Adj}$  test tends to over-reject when the panel includes dynamics and  $T$  is small compared to  $N$ . The reported rejection rate when  $N = 200$  and  $T = 20$  is 14.25 percent.<sup>17</sup> Furthermore, the findings also suggest that the  $LM_{Adj}$  test has power when the cross-sectional dependence is very weak, namely in the case when the exponent of cross-sectional dependence is  $\alpha = 0.25$ .  $LM_S$  also over-rejects when  $T$  is small relative to  $N$ , but the over-rejection is much more severe as compared to the  $LM_{adj}$  test since in the weakly exogenous regressor case it is not centered at zero for a fixed  $T$ .

The over-rejection of the  $J_{BFK}$  test in these experiments is caused by a combination of several factors, including heteroskedastic errors and heterogeneous coefficients. In order to distinguish between these effects, we also conducted experiments with homoskedastic errors where we set  $Var(\varepsilon_{it}) = \sigma_i^2 = 1$ , for all  $i$ , and strictly exogenous regressors (by setting  $\alpha_{xi} = 0$  for all  $i$ ), and consider two cases for the coefficients: heterogeneous and homogeneous (we set  $\beta_{i0} = E(\beta_{i0}) = 0.75$ , for all  $i$ ). The results under homoskedastic errors and homogeneous slopes are summarized in the upper part of Table 1.3. As to be expected, the  $J_{BFK}$  test has good size and power when  $T > 20$  and  $\alpha = 0$ . But the test tends to over-reject when  $T = 20$  and  $N$  relatively large even under these restrictions. The bottom part of Table 1.3 presents findings for the experiments with slope heterogeneity, whilst maintaining the assumptions of homoskedastic errors and strictly exogenous regressors. We see that even a small degree of slope heterogeneity can cause the  $J_{BFK}$  test to over-reject badly.

Finally, it is important to bear in mind that even the  $CD_P$  test is likely to over-reject in the case of models with weakly exogenous regressors if  $N$  is much larger than  $T$ .

**Table 1.2 Size and power of  $CD_P$ ,  $LM_{Adj}$ , and  $LM_S$  tests in the case of panels with weakly and strictly exogenous regressors (nominal size is set at 5%)**

(N,T)	$\alpha = 0$					$\alpha = 0.25$					$\alpha = 0.5$					$\alpha = 0.75$				
	20	50	100	150	200	20	50	100	150	200	20	50	100	150	200	20	50	100	150	200
Experiments with strictly exogenous regressors																				
$CD_P$ test																				
40	5.65	6.25	5.15	5.15	4.95	6.20	6.50	6.20	6.25	6.75	24.60	46.70	74.60	86.35	91.95	99.50	100.00	100.00	100.00	100.00
50	5.40	4.80	5.30	5.15	5.40	5.05	5.15	6.85	7.25	6.50	28.55	52.60	78.85	90.45	96.10	99.70	100.00	100.00	100.00	100.00
100	5.45	5.45	5.40	4.40	5.45	5.10	6.15	6.75	6.60	8.20	32.50	60.15	82.55	92.95	97.45	99.95	100.00	100.00	100.00	100.00
150	4.80	4.75	4.65	4.95	5.05	5.05	5.70	5.85	5.15	6.10	31.90	56.45	83.45	92.80	97.50	100.00	100.00	100.00	100.00	100.00
200	5.85	4.70	5.25	6.60	4.50	6.00	5.80	5.30	5.55	6.40	30.00	57.60	83.65	94.15	97.95	100.00	100.00	100.00	100.00	100.00
$LM_{Adj}$ test																				
40	4.75	5.25	5.50	4.30	5.20	6.80	7.65	15.95	28.30	36.55	43.05	93.35	99.80	100.00	100.00	99.70	100.00	100.00	100.00	100.00
50	6.05	5.25	4.00	4.95	4.95	6.05	6.45	12.40	19.70	31.50	47.85	95.85	100.00	100.00	99.70	100.00	100.00	100.00	100.00	
100	7.00	5.10	4.75	4.70	4.80	7.35	8.80	18.40	34.30	46.25	53.75	98.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	
150	6.55	4.85	5.10	5.15	5.40	7.45	6.70	11.95	18.55	28.65	49.85	98.55	99.95	100.00	100.00	100.00	100.00	100.00	100.00	100.00
200	7.75	4.95	5.15	3.90	5.10	8.75	6.45	8.50	13.25	19.00	52.05	98.70	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
$LM_S$ test																				
40	11.35	5.30	4.70	5.70	5.30	15.75	10.85	20.10	28.15	41.50	63.35	95.65	99.70	99.95	100.00	99.80	100.00	100.00	100.00	100.00
50	17.65	6.70	5.90	5.40	4.90	18.90	11.40	15.65	21.05	31.45	73.85	97.05	99.95	100.00	99.95	100.00	100.00	100.00	100.00	100.00
100	44.70	9.40	5.80	6.15	6.15	49.70	19.80	24.80	40.00	51.05	88.65	99.20	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
150	67.25	14.30	7.25	7.30	5.45	70.40	25.85	19.35	25.20	35.40	94.15	99.70	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
200	85.10	24.45	9.45	6.55	6.60	85.75	31.10	19.85	22.05	26.05	98.20	99.85	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Experiments with lagged dependent variable and other weakly exogenous regressors																				
$CD_P$ test																				
40	5.90	4.75	5.55	5.05	4.90	6.00	6.10	5.70	6.40	5.90	26.00	49.75	72.55	84.80	93.10	99.80	100.00	100.00	100.00	100.00
50	6.35	5.00	5.15	5.60	4.40	5.90	6.75	6.00	5.95	6.45	28.90	55.00	77.90	90.50	96.00	99.75	100.00	100.00	100.00	100.00
100	6.55	5.50	5.05	5.10	3.95	6.85	6.85	6.80	6.75	8.15	31.75	59.15	81.30	93.65	97.10	100.00	100.00	100.00	100.00	100.00
150	7.55	5.90	4.50	5.60	4.35	8.30	5.75	6.80	5.25	6.45	34.30	56.15	80.95	94.15	97.00	100.00	100.00	100.00	100.00	100.00
200	8.10	4.75	5.05	5.65	4.60	10.30	6.00	7.25	6.40	6.45	35.75	61.10	83.20	94.05	98.45	100.00	100.00	100.00	100.00	100.00
$LM_{Adj}$ test																				
40	5.05	4.70	5.25	4.10	5.85	6.40	6.85	15.25	26.70	38.65	32.60	92.00	99.80	99.95	100.00	99.40	100.00	100.00	100.00	100.00
50	5.80	4.90	4.70	4.90	4.90	5.30	6.15	12.10	20.80	30.85	35.65	95.55	99.85	100.00	100.00	99.65	100.00	100.00	100.00	100.00
100	6.45	5.60	5.05	4.70	4.80	7.85	7.50	18.45	31.05	47.00	36.40	98.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
150	11.55	5.95	5.30	5.20	3.85	10.35	6.50	10.35	19.65	28.60	31.60	97.65	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
200	14.25	5.50	5.35	4.90	5.25	12.85	6.00	8.20	11.55	19.25	31.55	98.80	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
$LM_S$ test																				
40	15.40	6.10	5.40	4.50	5.10	17.00	10.90	19.45	28.35	41.70	61.30	94.85	99.85	99.85	100.00	99.65	100.00	100.00	100.00	100.00
50	18.60	6.55	5.60	4.25	5.05	22.25	10.25	14.25	22.80	31.70	72.80	97.35	99.95	100.00	100.00	99.80	100.00	100.00	100.00	100.00
100	50.25	10.55	6.60	5.10	5.55	55.60	21.65	26.35	37.95	54.40	88.20	99.25	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
150	77.65	18.25	7.70	6.25	6.95	77.70	28.20	18.75	27.70	34.95	95.40	99.55	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
200	89.60	29.55	11.90	6.90	6.30	87.95	36.65	19.95	23.40	25.85	98.70	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Notes: Observations are generated using the equations  $y_{it} = c_{yi} + \lambda_i y_{i,t-1} + \beta_{0i} x_{it} + \beta_{1i} x_{i,t-1} + u_{it}$ ,  $x_{it} = c_{xi} + \alpha_{xi} y_{i,t-1} + \gamma'_{xi} \mathbf{f}_t + v_{it}$ , (see (67) and (53), respectively), and  $u_{it} = \gamma_1 \tilde{f}_t + \varepsilon_{it}$ , (see (68)). Four values of  $\alpha = 0, 0.25, 0.5$  and  $0.75$  are considered. Null of weak cross-sectional dependence is characterized by  $\alpha = 0$  and  $\alpha = 0.25$ . In the case of panels with strictly exogenous regressors  $\lambda_i = \alpha_{xi} = 0$ , for all  $i$ . For a more detailed account of the MC design see Section 1.6.  $LM_S$  test statistic is computed using the fixed effects estimators.

**Table 1.3 Size and power of the  $J_{BFK}$  test in the case of panel data models with strictly exogenous regressors and homoskedastic idiosyncratic shocks  $\epsilon_{it}$  (nominal size is set to 5%)**

(N,T)	$\alpha = 0$					$\alpha = 0.25$					$\alpha = 0.5$					$\alpha = 0.75$					
	20	50	100	150	200	20	50	100	150	200	20	50	100	150	200	20	50	100	150	200	
Experiments with homogeneous slopes																					
40	7.85	5.60	5.60	5.20	5.90	21.85	53.80	79.40	86.65	92.50	82.70	99.30	100.00	100.00	100.00	99.70	100.00	100.00	100.00	100.00	
50	8.90	5.90	6.00	6.10	4.20	17.85	44.90	73.75	83.75	89.10	84.35	99.90	100.00	100.00	100.00	99.85	100.00	100.00	100.00	100.00	
100	9.70	6.10	5.65	5.30	5.50	19.35	52.30	81.30	91.90	95.55	88.25	100.00	100.00	100.00	100.00	99.95	100.00	100.00	100.00	100.00	
150	15.00	5.90	5.30	5.10	5.60	14.65	39.60	69.80	83.95	91.00	87.95	99.95	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	
200	21.30	6.60	5.30	4.60	5.60	15.90	27.45	58.70	75.45	84.55	87.10	99.95	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	
Experiments with heterogeneous slopes																					
40	7.30	9.10	13.70	22.10	31.80	22.15	55.15	83.90	93.30	96.45	81.40	99.60	100.00	100.00	100.00	99.75	100.00	100.00	100.00	100.00	
50	7.60	8.80	18.20	30.90	40.45	18.65	53.25	80.95	92.45	96.95	85.45	99.85	100.00	100.00	100.00	99.85	100.00	100.00	100.00	100.00	
100	9.40	16.85	42.05	65.20	83.10	21.40	65.65	94.80	99.20	99.90	88.75	100.00	100.00	100.00	100.00	99.90	100.00	100.00	100.00	100.00	
150	12.65	24.70	60.80	86.25	96.25	17.35	62.30	94.10	99.60	99.95	88.45	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	
200	15.20	36.80	78.90	95.65	99.00	16.05	64.85	96.15	99.75	99.90	87.75	99.95	100.00	100.00	100.00	99.95	100.00	100.00	100.00	100.00	

Notes: The data generating process is the same as the one used to generate the results in Table 1.2 with strictly exogenous regressors, but with two exceptions: error variances are assumed homoskedastic ( $Var(\epsilon_{it}) = \sigma_i^2 = 1$ , for all  $i$ ) and two possibilities are considered for the slope coefficients: heterogeneous and homogeneous (in the latter case  $\beta_{i0} = E(\beta_{i0}) = 0.75$ , for all  $i$ ). Null of weak cross-sectional dependence is characterized by  $\alpha = 0$  and  $\alpha = 0.25$ . See also the notes to Table 1.2. The  $J_{BFK}$  test statistic is computed using the fixed effects estimates.

**Table 1.4 Size and power of the  $CD_P$  test for large  $N$  and short  $T$  panels with strictly and weakly exogenous regressors (nominal size is set to 5%)**

(N,T)	$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$
	10	10	10	10
Panel with strictly exogenous regressors				
1000	5.10	6.30	20.50	99.90
Pure AR(1) panel				
1000	5.50	6.05	22.10	100.00
Dynamic panel with weakly exogenous regressors				
1000	69.45	70.70	73.95	100.00

Notes: See the notes to Tables 1.1 and 1.2, and Section 1.6 for further details. In particular, note that the null of weak cross-sectional dependence is characterized by  $\alpha = 0$  and  $\alpha = 0.25$ , with alternatives of semi-strong and strong cross-sectional dependence given by values of  $\alpha \geq 1/2$ .

Only in the case of models with strictly exogenous regressors, and pure autoregressive models with symmetrically distributed disturbances, we would expect the  $CD_P$  test to perform well even if  $N$  is much larger than  $T$ . To illustrate this property, we provide empirical size and power results when  $N = 1,000$  and  $T = 10$  in Table 1.4. As can be seen the  $CD_P$  test has the correct size when we consider panel data models with strictly exogenous regressors or in the case of pure AR(1) models, which is in contrast to the case of panels with weakly exogenous regressors where the size of the  $CD_P$  test is close to 70 percent. It is clear that the small sample properties of the  $CD_P$  test for very large  $N$  and small  $T$  panels very much depends on whether the panel includes weakly exogenous regressors.

## 1.7 APPLICATION OF CCE ESTIMATORS AND CD TESTS TO UNBALANCED PANELS

---

CCE estimators can be readily extended to unbalanced panels, a situation which frequently arises in practice. Denote the set of cross-section units with the available data on  $y_{it}$  and  $\mathbf{x}_{it}$  in period  $t$  as  $\mathcal{N}_t$  and the number of elements in the set by  $\#\mathcal{N}_t$ . Initially, we suppose that data coverage for the dependent variables and regressors is the same and later we relax this assumption. The main complication of applying a CCE estimator to the case of unbalanced panels is the inclusion of cross-sectional averages in the individual regressions. There are two possibilities regarding the units to include in the

computation of cross-sectional averages, either based on the same number of units or based on a varying number of units. In both cases, cross-sectional averages should be constructed using at least a minimum number of units, say  $N_{\min}$ , which based on the current Monte Carlo evidence suggests the value of  $N_{\min} = 20$ . If the same units are used, we have

$$\bar{y}_t = \frac{1}{\#\mathcal{N}} \sum_{i \in \mathcal{N}} y_{it}, \text{ and similarly } \bar{x}_t = \frac{1}{\#\mathcal{N}} \sum_{i \in \mathcal{N}} x_{it},$$

for  $t = \underline{t}, \underline{t} + 1, \dots, \bar{t}$  where  $\mathcal{N} = \bigcap_{t=\underline{t}}^{\bar{t}} \mathcal{N}_t$  and the starting and ending points of the sample  $\underline{t}$  and  $\bar{t}$  are chosen to maximize the use of data subject to the constraint  $\#\mathcal{N} \geq N_{\min}$ . The second possibility utilizes data in a more efficient way:

$$\bar{y}_t = \frac{1}{\#\mathcal{N}_t} \sum_{i \in \mathcal{N}_t} y_{it}, \text{ and } \bar{x}_t = \frac{1}{\#\mathcal{N}_t} \sum_{i \in \mathcal{N}_t} x_{it},$$

for  $t = \underline{t}, \underline{t} + 1, \dots, \bar{t}$ , where  $\underline{t}$  and  $\bar{t}$  are chosen such that  $\#\mathcal{N}_t \geq N_{\min}$  for all  $t = \underline{t}, \underline{t} + 1, \dots, \bar{t}$ . Both procedures are likely to perform similarly when  $\#\mathcal{N}$  is reasonably large, and the occurrence of missing observations is random. In cases where new cross-section units are added to the panel over time and such additions can have systematic influences on the estimation outcomes, it might be advisable to de-mean or de-trend the observations for individual cross-section units before computing the cross-section averages to be used in the CCE regressions.

Now suppose that the cross-section coverage differs for each variable. For example, the dependent variable can be available only for OECD countries, whereas some of the regressors could be available for a larger set of countries. Then, it is preferable to also utilize data on non-OECD countries to maximize the number of units for the computation of CS averages for each of the individual variables.

The *CD* and *LM* tests can also be readily extended to unbalanced panels. Denote by  $\mathcal{T}_i$ , the set of dates over which time series observations on  $y_{it}$  and  $x_{it}$  are available for the  $i^{th}$  individual, and the number of the elements in the set by  $\#\mathcal{T}_i$ . For each  $i$ , compute the OLS residuals based on the full set of time series observations for that individual. As before, denote these residuals by  $\hat{u}_{it}$ , for  $t \in \mathcal{T}_i$ , and compute the pairwise correlations of  $\hat{u}_{it}$  and  $\hat{u}_{jt}$  using the common set of data points in  $\mathcal{T}_i \cap \mathcal{T}_j$ . Since, the estimated residuals need not sum to zero over the common sample period  $\rho_{ij}$  could be estimated by

$$\hat{\rho}_{ij} = \frac{\sum_{t \in \mathcal{T}_i \cap \mathcal{T}_j} (\hat{u}_{it} - \bar{\hat{u}}_i)(\hat{u}_{jt} - \bar{\hat{u}}_j)}{\left[ \sum_{t \in \mathcal{T}_i \cap \mathcal{T}_j} (\hat{u}_{it} - \bar{\hat{u}}_i)^2 \right]^{1/2} \left[ \sum_{t \in \mathcal{T}_i \cap \mathcal{T}_j} (\hat{u}_{jt} - \bar{\hat{u}}_j)^2 \right]^{1/2}},$$

where

$$\bar{\hat{u}}_i = \frac{\sum_{t \in \mathcal{T}_i \cap \mathcal{T}_j} \hat{u}_{it}}{\#(\mathcal{T}_i \cap \mathcal{T}_j)}.$$

The CD (similarly the LM type) statistics for the unbalanced panel can then be computed as usual by

$$CD_P = \sqrt{\frac{2}{N(N-1)}} \left( \sum_{i=1}^{N-1} \sum_{j=i+1}^N \sqrt{T_{ij}} \hat{\rho}_{ij} \right), \quad (69)$$

where  $T_{ij} = \#(\mathcal{T}_i \cap \mathcal{T}_j)$ . Under the null hypothesis  $CD_P \sim N(0, 1)$  for  $T_i > k + 1$ ,  $T_{ij} > 3$ , and sufficiently large  $N$ .

## 1.8 CONCLUDING REMARKS

---

This chapter provides a review of the literature on large panel data models with cross-sectional error dependence. The survey focuses on large  $N$  and  $T$  panel data models where a natural ordering across the cross-section dimension is not available. This excludes the literature on spatial panel econometrics, which is recently reviewed by Lee and Yu (2010 and 2013). We provide a brief account of the concepts of weak and strong cross-sectional dependence, and discuss the exponent of cross-sectional dependence that characterizes the different degrees of cross-sectional dependence. We then attempt a synthesis of the literature on estimation and inference in large  $N$  and  $T$  panel data models with a common factor error structure. We distinguish between strictly and weakly exogenous regressors and panels with homogeneous and heterogeneous slope coefficients. We also provide an overview of tests of error cross-sectional dependence in static and dynamic panel data models.

## ACKNOWLEDGMENTS

---

We are grateful to an anonymous referee, Cheng Chou, Ron Smith, Vanessa Smith, Wei Xie, Takashi Yamagata and Qiankun Zhou for helpful comments. Pesaran acknowledges financial support from ESRC grant no. ES/I031626/1. The views expressed in this chapter are those of the authors and do not necessarily reflect those of the Federal Reserve Bank of Dallas or the Federal Reserve System.

## NOTES

---

1. Conditions (1)–(2) imply existence of a finite constant  $K$  (which does not depend on  $i$  or  $N$ ) such that  $|w_{it}| < KN^{-1}$  for any  $i = 1, 2, \dots, N$  and any  $N \in \mathbb{N}$ .

2. The assumption of zero loadings for  $i > [N^{\alpha_\gamma}]$  could be relaxed so long as  $\sum_{i=[N^{\alpha_\gamma}]+1}^N |\gamma_i| = O_p(1)$ . But for expositional simplicity we maintain  $\gamma_i = 0$  for  $i = [N^{\alpha_\gamma}] + 1, [N^{\alpha_\gamma}] + 2, \dots, N$ .
3. Note that the number of factors with  $\alpha_\ell > 0$  is limited by the absolute summability condition (19).
4. Pooled estimation is carried out assuming that  $\beta_i = \beta$  for all  $i$ , whilst mean group estimation allows for slope heterogeneity and estimates  $\beta$  by the average of the individual estimates of  $\beta_i$  (Pesaran and Smith 1995).
5. Tests of the slope homogeneity hypothesis in static and dynamic panels are discussed in Pesaran and Yamagata (2008).
6. This assumption can be relaxed. See Pesaran (2006).
7. Pesaran (2006) also considered a weighted average of individual  $\hat{b}_i$ , with weights inversely proportional to the individual variances.
8. Fixed effects and observed common factors (denoted by  $d_t$  previously) can also be included in the model. They are excluded to simplify the notations.
9. See also Lee et al. (2012) for an extension of this framework to panels with measurement errors.
10. See Everaert and Groote (2012) who derive the asymptotic bias of the CCE pooled estimator in the case of dynamic *homogeneous* panels.
11. This bias was first quantified in the case of a simple AR(1) model by Hurwicz (1950).
12. For example if  $\lambda_i$  is distributed uniformly over the range  $(0, b)$  where  $0 < b < 1$ , we have  $\alpha_\ell = E(\lambda_i^\ell) = b^\ell / (1 + \ell)$ , which decays exponentially with  $\ell$ .
13. The number of lags cannot increase too fast, otherwise there will not be a sufficient number of observations to accurately estimate the parameters, whilst at the same time a sufficient number of lags are needed to ensure that the factors are well approximated. Setting the number of lags equal to  $T^{1/3}$  seems to be a good choice, balancing the effects of the above two opposing considerations. See Berk (1974), Said and Dickey (1984), and Chudik and Pesaran (2013b) for a related discussion on the choice of lag truncation for estimation of infinite order autoregressive models.
14. Matrices  $\Xi_i$ ,  $\bar{Q}_w$ , and  $\bar{M}_q$  depend also on  $p_T$ ,  $N$  and  $T$ , but we omit these subscripts to simplify notations.
15. Pesaran (2014) also derives the exact variance of the  $CD_P$  test under the null of cross-sectional independence and proposes a slightly modified version of the  $CD_P$  test distributed exactly with mean zero and a unit variance.
16. A recent review is provided by Moscone and Tosetti (2009).
17. The rejection rates based on the  $LM_{Adj}$  test were above 90 percent for the sample size  $N = 500, 1000$  and  $T = 10$ .

## REFERENCES

- Ahn, S. C. and A. R. Horenstein (2013). Eigenvalue ratio test for the number of factors. *Econometrica* 81(3), 1203–1207.
- Amengual, D. and M. W. Watson (2007). Consistent estimation of the number of dynamic factors in a large  $N$  and  $T$  panel. *Journal of Business and Economic Statistics* 25(1), 91–96.
- Andrews, D. (2005). Cross section regression with common shocks. *Econometrica* 73, 1551–1585.

- Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Anselin, L. (2001). Spatial econometrics. In B. H. Baltagi (Ed.), *A Companion to Theoretical Econometrics*. Oxford: Blackwell.
- 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.
- Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica* 77, 1229–1279.
- Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica* 70, 191–221.
- Bai, J. and S. Ng (2007). Determining the number of primitive shocks in factor models. *Journal of Business and Economic Statistics* 25(1), 52–60.
- Bai, J. and S. Ng (2008). Large dimensional factor analysis. *Foundations and Trends in Econometrics* 3(2), 89–168.
- Bai, Z. D. and J. W. Silverstein (1998). No eigenvalues outside the support of the limiting spectral distribution of large dimensional sample covariance matrices. *Annals of Probability* 26(1), 316–345.
- Bailey, N., G. Kapetanios, and M. H. Pesaran (2012). Exponents of cross-sectional dependence: Estimation and inference. CESifo Working Paper No. 3722, revised July 2013.
- Baltagi, B. H., Q. Feng, and C. Kao (2011). Testing for sphericity in a fixed effects panel data model. *The Econometrics Journal* 14, 25–47.
- Baltagi, B. H., S. Song, and W. Koh (2003). Testing panel data regression models with spatial error correlation. *Journal of Econometrics* 117, 123–150.
- Berk, K. N. (1974). Consistent autoregressive spectral estimates. *The Annals of Statistics* 2, 489–502.
- Breitung, J. and I. Choi (2013). Factor models. In N. Hashimzade and M. A. Thornton (Eds.), *Handbook Of Research Methods and Applications in Empirical Macroeconomics*, Chapter 11. Edward Elgar. Cheltenham: UK.
- Breitung, J. and U. Pigorsch (2013). A canonical correlation approach for selecting the number of dynamic factors. *Oxford Bulletin of Economics and Statistics* 75(1), 23–36. Cheltenham: UK.
- Breusch, T. S. and A. R. Pagan (1980). The Lagrange Multiplier test and its application to model specifications in econometrics. *Review of Economic Studies* 47, 239–253.
- Chamberlain, G. (1983). Funds, factors and diversification in arbitrage pricing models. *Econometrica* 51, 1305–1324.
- Chamberlain, G. (1984). Panel data. In Z. Griliches and M. Intrilligator (Eds.), *Handbook of Econometrics*, Volume 2, Chapter 22, pp. 1247–1318. Amsterdam: North-Holland.
- Choi, I. and H. Jeong (2013). Model selection for factor analysis: Some new criteria and performance comparisons. Research Institute for Market Economy (RIME) Working Paper No.1209, Sogang University.
- Chudik, A. and M. H. Pesaran (2014). Aggregation in large dynamic panels. *Journal of Econometrics*. 178(2), 273–285.
- Chudik, A. and M. H. Pesaran (2013a). Common correlated effects estimation of heterogeneous dynamic panel data models with weakly exogenous regressors. CESifo Working Paper No. 4232.
- Chudik, A. and M. H. Pesaran (2013b). Econometric analysis of high dimensional VARs featuring a dominant unit. *Econometric Reviews* 32, 592–649.

- Chudik, A., M. H. Pesaran, and E. Tosetti (2011). Weak and strong cross section dependence and estimation of large panels. *The Econometrics Journal* 14, C45–C90.
- 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.
- Coakley, J., A. M. Fuertes, and R. Smith (2002). A principal components approach to cross-section dependence in panels. Birkbeck College Discussion Paper 01/2002.
- Coakley, J., A. M. Fuertes, and R. Smith (2006). Unobserved heterogeneity in panel time series. *Computational Statistics and Data Analysis* 50, 2361–2380.
- Dhaene, G. and K. Jochmans (2012). Split-panel jackknife estimation of fixed-effect models. Mimeo, 21 July 2012.
- Dufour, J. M. and L. Khalaf (2002). Exact tests for contemporaneous correlation of disturbances in seemingly unrelated regressions. *Journal of Econometrics* 106, 143–170.
- Everaert, G. and T. D. Groote (2012). Common correlated effects estimation of dynamic panels with cross-sectional dependence. Mimeo, 9 November 2012.
- Forni, M. and M. Lippi (2001). The generalized factor model: Representation theory. *Econometric Theory* 17, 1113–1141.
- Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000). The generalized dynamic factor model: Identification and estimation. *Review of Economics and Statistics* 82, 540–554.
- Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2004). The generalized dynamic factor model: Consistency and rates. *Journal of Econometrics* 119, 231–235.
- Frees, E. W. (1995). Assessing cross sectional correlation in panel data. *Journal of Econometrics* 69, 393–414.
- Geweke, J. (1977). The dynamic factor analysis of economic time series. In D. Aigner and A. Goldberger (Eds.), *Latent variables in socio-economic models*. Amsterdam: North-Holland.
- Gourieroux, C., A. Monfort, E. Renault, and A. Trognon (1987). Generalised residuals. *Journal of Econometrics* 34, 5–32.
- Granger, C. W. J. (1980). Long memory relationships and the aggregation of dynamic models. *Journal of Econometrics* 14, 227–238.
- Hachem, W., P. Loubaton, and J. Najim (2005). The empirical eigenvalue distribution of a gram matrix: From independence to stationarity. *Markov Processes and Related Fields* 11(4), 629–648.
- Haining, R. P. (2003). *Spatial Data Analysis: Theory and Practice*. Cambridge: Cambridge University Press.
- Hallin, M. and R. Liska (2007). The generalized dynamic factor model: Determining the number of factors. *Journal of the American Statistical Association* 102, 603–617.
- Harding, M. (2013). Estimating the number of factors in large dimensional factor models. Mimeo, April 2013.
- Holly, S., M. H. Pesaran, and T. Yagamata (2011). Spatial and temporal diffusion of house prices in the UK. *Journal of Urban Economics* 69, 2–23.
- Hsiao, C., M. H. Pesaran, and A. Pick (2012). Diagnostic tests of cross-section independence for limited dependent variable panel data models. *Oxford Bulletin of Economics and Statistics* 74, 253–277.
- Hurwicz, L. (1950). Least squares bias in time series. In T. C. Koopman (Ed.), *Statistical Inference in Dynamic Economic Models*. New York: Wiley.
- Jensen, P. S. and T. D. Schmidt (2011). Testing cross-sectional dependence in regional panel data. *Spatial Economic Analysis* 6(4), 423–450.

- John, S. (1971). Some optimal multivariate tests. *Biometrika* 58, 123–127.
- Kapetanios, G. (2004). A new method for determining the number of factors in factor models with large datasets. Queen Mary University of London, Working Paper No. 525.
- Kapetanios, G. (2010). A testing procedure for determining the number of factors in approximate factor models with large datasets. *Journal of Business and Economic Statistics* 28(3), 397–409.
- Kapetanios, G. and M. H. Pesaran (2007). Alternative approaches to estimation and inference in large multifactor panels: Small sample results with an application to modelling of asset returns. In G. Phillips and E. Tzavalis (Eds.), *The Refinement of Econometric Estimation and Test Procedures: Finite Sample and Asymptotic Analysis*. Cambridge: Cambridge University Press.
- Kapetanios, G., M. H. Pesaran, and T. Yagamata (2011). Panels with nonstationary multifactor error structures. *Journal of Econometrics* 160, 326–348.
- Lee, L.-F. and J. Yu (2010). Some recent developments in spatial panel data model. *Regional Science and Urban Economics* 40, 255–271.
- Lee, L.-F. and J. Yu (2013). Spatial panel data models. Mimeo, April, 2013.
- Lee, N., H. R. Moon, and M. Weidner (2012). Analysis of interactive fixed effects dynamic linear panel regression with measurement error. *Economics Letters* 117(1), 239–242.
- Moon, H. R. and M. Weidner (2010). Dynamic linear panel regression models with interactive fixed effects. Mimeo, July 2010.
- Moran, P. A. P. (1948). The interpretation of statistical maps. *Biometrika* 35, 255–260.
- Moscone, F. and E. Tosetti (2009). A review and comparison of tests of cross section independence in panels. *Journal of Economic Surveys* 23, 528–561.
- Ng, S. (2006). Testing cross section correlation in panel data using spacings. *Journal of Business and Economic Statistics* 24, 12–23.
- O'Connell, P. G. J. (1998). The overvaluation of purchasing power parity. *Journal of International Economics* 44, 1–19.
- Onatski, A. (2009). Testing hypotheses about the number of factors in large factor models. *Econometrica* 77, 1447–1479.
- Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *Review of Economics and Statistics* 92, 1004–1016.
- Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics* 168, 244–258.
- Pesaran, M. H. (2004). General diagnostic tests for cross section dependence in panels. CESifo Working Paper No. 1229.
- Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with multi-factor error structure. *Econometrica* 74, 967–1012.
- Pesaran, M. H. (2014). Testing weak cross-sectional dependence in large panels. *Econometric Reviews*.
- Pesaran, M. H. and R. Smith (1995). Estimation of long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics* 68, 79–113.
- Pesaran, M. H. and E. Tosetti (2011). Large panels with common factors and spatial correlation. *Journal of Econometrics* 161(2), 182–202.
- Pesaran, M. H., A. Ullah, and T. Yamagata (2008). A bias-adjusted LM test of error cross section independence. *The Econometrics Journal* 11, 105–127.
- Pesaran, M. H. and T. Yamagata (2008). Testing slope homogeneity in large panels. *Journal of Econometrics* 142, 50–93.

- Phillips, P. C. B. and D. Sul (2003). Dynamic panel estimation and homogeneity testing under cross section dependence. *The Econometrics Journal* 6, 217–259.
- Phillips, P. C. B. and D. Sul (2007). Bias in dynamic panel estimation with fixed effects, incidental trends and cross section dependence. *Journal of Econometrics* 137, 162–188.
- Pyke (1965). Spacings. *Journal of the Royal Statistical Society, Series B* 27, 395–449.
- Said, E. and D. A. Dickey (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika* 71, 599–607.
- Sarafidis, V. and D. Robertson (2009). On the impact of error cross-sectional dependence in short dynamic panel estimation. *The Econometrics Journal* 12, 62–81.
- Sarafidis, V. and T. Wansbeek (2012). Cross-sectional dependence in panel data analysis. *Econometric Reviews* 31, 483–531.
- Sarafidis, V., T. Yagamata, and D. Robertson (2009). A test of cross section dependence for a linear dynamic panel model with regressors. *Journal of Econometrics* 148, 149–161.
- Sargan, J. (1988). Testing for misspecification after estimation using instrumental variables. In E. Maasoumi (Ed.), *Contributions to Econometrics: John Denis Sargan*, Volume 1. Cambridge: University Press.
- Sargent, T. J. and C. A. Sims (1977). Business cycle modeling without pretending to have too much a-priori economic theory. In C. Sims (Ed.), *New Methods in Business Cycle Research*. Minneapolis: Federal Reserve Bank of Minneapolis.
- Schott, J. R. (2005). Testing for complete independence in high dimensions. *Biometrika* 92, 951–956.
- So, B. S. and D. W. Shin (1999). Recursive mean adjustment in time series inferences. *Statistics & Probability Letters* 43, 65–73.
- Song, M. (2013). Asymptotic theory for dynamic heterogeneous panels with cross-sectional dependence and its applications. Mimeo, 30 January 2013.
- Stock, J. H. and M. W. Watson (2011). Dynamic factor models. In M. P. Clements and D. F. Hendry (Eds.), *The Oxford Handbook of Economic Forecasting*. New York: Oxford University Press.
- Westerlund, J. and J. Urbain (2011). Cross-sectional averages or principal components? Research Memoranda 053, Maastricht: METEOR, Maastricht Research School of Economics of Technology and Organization.
- Whittle, P. (1954). On stationary processes on the plane. *Biometrika* 41, 434–449.
- Yin, Y. Q., Z. D. Bai, and P. R. Krishnainiah (1988). On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. *Probability Theory and Related Fields* 78(4), 509–521.
- Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association* 57, 348–368.

## CHAPTER 2

---

# PANEL COINTEGRATION

---

IN CHOI

### 2.1 INTRODUCTION

---

THE concept of cointegration has been studied intensely since it was first introduced in Engle and Granger (1987). In more recent years, cointegration has been studied in the setup of panel data. Panel data offer higher power for cointegration tests and better precision for the estimation of slope coefficients when they are homogeneous across cross-sectional units. However, panel data also pose challenges since they should not be assumed to be independent for macroeconomic and financial applications. Methods for modelling cross-section correlations have been proposed in the literature for this reason, which has in turn enriched the literature on panel data.

This chapter surveys the literature on panel cointegration, complementing earlier review papers by Choi (2006) and Breitung and Pesaran (2008). First, cointegrating panel regressions for cross-sectionally independent and correlated panels are discussed. Next, tests for panel cointegration are introduced. Three groups of the tests are examined: residual-based tests for the null of noncointegration, residual-based tests for the null of cointegration, and tests based on vector autoregression.

This chapter is structured as follows. Section 2.2 introduces panel regressions for cross-sectionally independent and cross-sectionally correlated panels. Poolability tests are also discussed in this section. Section 2.3 discusses tests for panel cointegration. Residual-based tests for the null of noncointegration, residual-based tests for the null of cointegration, and panel VAR cointegration tests are treated separately. Section 2.4 provides a summary and further remarks.

A few words on our notation may be helpful. The integer part of  $x$  is denoted by  $[x]$ , and weak convergence by  $\Rightarrow$ . The index  $i$  runs from 1 to  $N$ ;  $t$  runs from 1 to  $T$ .

## 2.2 PANEL REGRESSIONS

---

This section introduces the literature on panel regressions involving  $I(1)$  variables. In the early stage of this literature, cross-sectional units were assumed to be independent, which makes related analysis easier than when they are not. However, the assumption of independence is not suitable for most macropans, which prompted various methods that allow cross-sectional correlation. This section will review these two strands of the literature.

### 2.2.1 Regressions for Cross-Sectionally Independent Panels

This subsection introduces properties of the regression estimators for independent, cointegrated panels. Cointegrating panel regressions are studied in Kao and Chiang (2000), Pedroni (2000), Phillips and Moon (1999), and Mark and Sul (2003). The contrasting case of spurious panel regressions involving  $I(1)$  regressors and errors are studied in Entorf (1997), Kao (1999), Phillips and Moon (1999), and Urbain and Westerlund (2011). It is shown in Kao (1999) and Phillips and Moon (1999) that the estimators they study converge to what Phillips and Moon call the “long-run average regression coefficient” and that  $\sqrt{N}$ -asymptotics holds for the spurious panel regression.

Kao and Chiang (2000) consider the model

$$\begin{aligned} y_{it} &= \alpha_i + \beta' x_{it} + u_{it}, \\ x_{it} &= x_{i,t-1} + \varepsilon_{it}, \end{aligned} \tag{1}$$

for which  $\{y_{it}, x_{it}\}$  are assumed to be independent across  $i$ . For this model, letting  $w_{it} = (u_{it}, \varepsilon'_{it})'$ , Kao and Chiang (2000) assume  $\frac{1}{\sqrt{T}} \sum_{t=1}^{[Tr]} w_{it} \Rightarrow B_i(r)$ , where  $B_i(r)$  is a vector Brownian motion with the homogeneous long-run covariance matrix

$$\Omega = \sum_{j=-\infty}^{\infty} E(w_{ij} w_{i0}') = \Sigma + \Gamma + \Gamma' = \begin{bmatrix} \Omega_{uu} & \Omega_{ue} \\ \Omega_{eu} & \Omega_{ee} \end{bmatrix},$$

with  $\Gamma = \sum_{j=1}^{\infty} E(w_{ij} w_{i0}')$  and  $\Sigma = E(w_{i0} w_{i0}')$ . In this assumption, each  $w_{it}$  has the same long-run covariance matrix across  $i$ . We also let  $\Delta = \Sigma + \Gamma$ . The matrices  $\Sigma$ ,  $\Gamma$ , and  $\Delta$  are partitioned conformable to  $\Omega$  as

$$\Sigma = \begin{bmatrix} \Sigma_{uu} & \Sigma_{ue} \\ \Sigma_{eu} & \Sigma_{ee} \end{bmatrix}, \quad \Gamma = \begin{bmatrix} \Gamma_{uu} & \Gamma_{ue} \\ \Gamma_{eu} & \Gamma_{ee} \end{bmatrix}, \quad \text{and} \quad \Delta = \begin{bmatrix} \Delta_{uu} & \Delta_{ue} \\ \Delta_{eu} & \Delta_{ee} \end{bmatrix}.$$

Kao and Chiang (2000) report that the Within-OLS estimator of  $\beta$ , defined by

$$\hat{\beta} = \left( \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_{i\cdot})(x_{it} - \bar{x}_{i\cdot})' \right)^{-1} \left( \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_{i\cdot})(y_{it} - \bar{y}_{i\cdot}) \right),$$

has the limiting distribution

$$\sqrt{NT}(\hat{\beta} - \beta) - \sqrt{N}\delta_{NT} \Rightarrow N(0, 6\Omega_{\varepsilon\varepsilon}^{-1}\Omega_{u\varepsilon}) \text{ as } T \rightarrow \infty \text{ and } N \rightarrow \infty,$$

where  $z_i = \frac{1}{T} \sum_{t=1}^T z_{it}$  ( $z = x, y$ ),

$$\begin{aligned} \delta_{NT} &= \left( \frac{1}{NT^2} \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_{i\cdot})(x_{it} - \bar{x}_{i\cdot})' \right)^{-1} \\ &\times \left( \frac{1}{N} \sum_{i=1}^N \Omega_{\varepsilon\varepsilon}^{1/2} \left( \int_0^1 \bar{W}_i(r) dW_i'(r) \right) \Omega_{\varepsilon\varepsilon}^{-1/2} \Omega_{u\varepsilon}' + \Delta_{u\varepsilon}' \right), \end{aligned}$$

$\bar{W}_i(r) = W_i(r) - \int_0^1 W_i(s) ds$ , and  $W_i(r)$  is a vector standard Brownian motion for each  $i$ . This result shows that the Within-OLS estimator is inconsistent when the regressors and errors are correlated, which is in contrast to the consistency property of the time series OLS estimator under the same circumstance.

In order to eliminate the bias of the Within-OLS estimator, Kao and Chiang (2000) and Phillips and Moon (1999) consider the fully modified OLS (FM-OLS) estimator of  $\beta$ , defined by

$$\hat{\beta}_{\text{FM}} = \left( \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_{i\cdot})(x_{it} - \bar{x}_{i\cdot})' \right)^{-1} \left( \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_{i\cdot})\hat{y}_{it}^+ - NT\hat{\Delta}_{\varepsilon u}^+ \right),$$

where  $\hat{y}_{it}^+ = y_{it} - \hat{\Omega}_{u\varepsilon}\hat{\Omega}_{\varepsilon\varepsilon}^{-1}\Delta x_{it}$ ,  $\hat{\Delta}_{\varepsilon u}^+ = \hat{\Delta}_{\varepsilon u} - \hat{\Delta}_{\varepsilon\varepsilon}\hat{\Omega}_{\varepsilon\varepsilon}^{-1}\hat{\Omega}_{\varepsilon u}$  and  $\hat{a}$  denotes the estimator of parameter  $a$  using the OLS residuals. This estimator extends Phillips and Hansen's (1990) FM-OLS estimator to panel regression. As  $T \rightarrow \infty$  and  $N \rightarrow \infty$  in sequence,

$$\sqrt{NT}(\hat{\beta}_{\text{FM}} - \beta) \Rightarrow N(0, 6\Omega_{\varepsilon\varepsilon}^{-1}\Omega_{u\varepsilon}),$$

where  $\Omega_{u\varepsilon} = \Omega_{uu} - \Omega_{u\varepsilon}\Omega_{\varepsilon\varepsilon}^{-1}\Omega_{\varepsilon u}$ . This result shows that the panel FM-OLS estimator is consistent, converges to the true parameter vector at a faster rate than the time series FM-OLS estimator, and is normally distributed in the limit.

Kao and Chiang (2000) also consider the panel regression model for dynamic OLS (DOLS),

$$y_{it} = \alpha_i + \beta' x_{it} + \sum_{j=-q}^q \zeta_{ij}' \Delta x_{i,t-j} + \dot{v}_{it}, \quad (2)$$

(see Phillips and Loretan 1991; Saikkonen 1991; and Stock and Watson 1993, for the time-series DOLS) and show that the Within-OLS estimator of the parameter vector  $\beta$  has the same distribution as the FM-OLS estimator.<sup>1</sup>

Kao and Chiang (2000) and Pedroni (2000) also study FM-OLS and DOLS for the heterogeneous panels where  $\{B_i(r)\}$  have different variance–covariance matrices  $\{\Omega_i\}$  across  $i$ . For these, we need to estimate  $\{\Omega_i\}$  and use transformations similar to those for the FM-OLS and DOLS applied to homogeneous panels. The interested reader is referred to the aforementioned articles for details.

Mark and Sul (2003) consider the same model as (2). They estimate the coefficients  $\beta$  and  $\{\zeta_{ij}\}$  by applying pooled OLS after subtracting time-series means from the regressors and regressand. They report their limiting distributions for the cases of fixed  $N$  and large  $T$  and of large  $N$  and  $T$ . Mark and Sul extend Model (2) to

$$y_{it} = \alpha_i + \lambda_i t + \beta' x_{it} + \sum_{j=-q}^q \zeta'_{ij} \Delta x_{i,t-j} + v_{it}, \quad (3)$$

which contains a heterogeneous linear time trend, reporting the limiting distribution of the Within-OLS estimator of  $\beta$ .

What would happen if we run a cross-section regression for Model (1)? This question is studied by Madsen (2005) using a model slightly more general than Model (1) in the sense that the regressors and the error terms contain variables for both individual and time effects. Madsen finds that the cross-sectional OLS regression provides a  $\sqrt{N}$ -consistent and asymptotically normally distributed estimator of the slope coefficient for each  $t$ . Since the convergence is slower than the estimator using the whole sample, the cross-sectional estimator is unlikely to be used in practice unless  $T$  is quite small.

Kauppi (2000) shows that the FM-OLS estimator does not, in general, provide asymptotically normal distributions when regressors are endogenous and nearly nonstationary in the sense that  $x_{it} - \exp(\frac{C}{T}) x_{i,t-1}$  is  $I(0)$  with  $C$  being a constant matrix. In response to this, Choi (2002) shows that the instrumental variables (IV) estimation method can be used for panel regression with endogenous, nearly nonstationary regressors. The panel Within-IV estimator of parameter  $\beta$ ,  $\tilde{\beta}$ , has the weak limit for a nearly nonstationary instrument  $z_{it}$   $T \rightarrow \infty$

$$\begin{aligned} \sqrt{NT}(\tilde{\beta} - \beta) &= \frac{\sum_{i=1}^N \sum_{t=1}^T (z_{it} - \bar{z}_{i.}) u_{it} / (T\sqrt{N})}{\sum_{i=1}^N \sum_{t=1}^T (z_{it} - \bar{z}_{i.})(x_{it} - \bar{x}_{i.}) / (T^2 N)} \\ &\Rightarrow \frac{\sum_{i=1}^N \int_0^1 \bar{K}_{z_i}(r) d B_{u_i}(r) / \sqrt{N}}{\sum_{i=1}^N \int_0^1 \bar{K}_{z_i}(r) \bar{K}_{x_i}(r) dr / N} \end{aligned} \quad (4)$$

The numerator of the weak limit in relation (4) is a standardized sum of zero-mean random variables when  $E(z_{it} u_{is}) = 0$ , for all  $t$  and  $s$ , and the denominator is a standardized sum of random variables. Thus, when proper conditions are given and  $N$  is large, we may apply the central limit theorem and law of large numbers for the numerator

and denominator, respectively, which leads to the asymptotic normality result for the panel IV estimator. This intuition forms the basis of the asymptotic normality results for more involved IV estimators such as within-IV-OLS, IV-GLS, and within-IV-GLS in Choi (2002).

Breitung (2005) proposes a two-step procedure for the estimation of a common cointegrating vector across individuals. The procedure is based on the vector error-correction model

$$\Delta Y_{it} = \alpha_i \beta' Y_{i,t-1} + \varepsilon_{it}, \quad (5)$$

where  $E(\varepsilon_{it}) = 0$ ,  $E(\varepsilon_{it}\varepsilon'_{it}) = \Sigma_i$ , and  $\varepsilon_{it}$  and  $\varepsilon_{jt}$  ( $i \neq j$ ) are independent. Note that all the individuals share the same cointegrating vector  $\beta$ . Premultiplying equation (5) by  $(\alpha'_i \Sigma_i^{-1} \alpha_i)^{-1} \alpha'_i \Sigma_i^{-1}$  yields

$$Z_{it} = \beta' Y_{i,t-1} + v_{it}, \quad (6)$$

where  $Z_{it} = (\alpha'_i \Sigma_i^{-1} \alpha_i)^{-1} \alpha'_i \Sigma_i^{-1} \Delta Y_{it}$  and  $v_{it} = (\alpha'_i \Sigma_i^{-1} \alpha_i)^{-1} \alpha'_i \Sigma_i^{-1} \varepsilon_{it}$ . When the normalization  $\beta = [I, \beta'_2]$  with the conformable partition  $Y_{it} = (Y_{it}^1, Y_{it}^2)'$  is used, equation (6) is rewritten as

$$Z_{it} - Y_{i,t-1}^1 = \beta'_2 Y_{i,t-1}^2 + v_{it}.$$

Breitung suggests estimating this model by pooled OLS using Johansen's (1988) estimates of  $\alpha_i$  and  $\Sigma_i$ . He shows that the resulting estimator of  $\beta_2$  is  $\sqrt{NT}$ -consistent and has a multivariate normal distribution in the limit. Breitung's simulation results show that the estimator has good finite-sample properties.

Baltagi, Kao, and Liu (2008) study efficiency of various estimators for the random effects model

$$\begin{aligned} y_{it} &= \alpha + \beta x_{it} + u_{it}, \\ u_{it} &= \mu_i + v_{it}, \quad \mu_i \sim i.i.d.(0, \sigma_\mu^2), \end{aligned}$$

where both  $\{x_{it}\}$  and  $\{v_{it}\}$  are univariate AR(1) processes which are either  $I(1)$  or  $I(0)$  and  $\{\mu_i\}$  and  $\{x_{it}\}$  are uncorrelated. When the regressor is endogenous and  $I(1)$ , they report that the Within-OLS, first-differenced, GLS and OLS estimators are all inconsistent regardless of the integration order of the error terms. When the regressor is totally exogenous and  $I(1)$ , they report that all those estimators become consistent for both  $I(0)$  and  $I(1)$  errors. But Within-OLS and GLS are most efficient for the  $I(0)$  errors and GLS is so for the  $I(1)$  errors. Thus, GLS is the preferred estimator in their study. Building on the results of Baltagi, Kao, and Liu (2008), Baltagi, Kao, and Na (2011) study the  $t$ -ratios for the slope coefficient  $\beta$  for the case of exogenous regressors and conclude that the  $t$ -ratio based on the feasible GLS estimator is most preferred.

## 2.2.2 Regressions for Cross-Sectionally Correlated Panels

The estimators considered in the previous subsection are based on the assumptions of cross-sectional independence and noncointegration. When these assumptions are violated, these estimators are subject to efficiency loss as shown in Wagner and Hlouskova's (2010) simulation study. This subsection introduces regression methods that drop such assumptions.

Mark and Sul (2003) extend Model (2) to

$$y_{it} = \alpha_i + \lambda_i t + \theta_t + \beta' x_{it} + \sum_{j=-q}^q \zeta'_{ij} \Delta x_{i,t-j} + v_{it}, \quad (7)$$

where  $\{\theta_t\}$  are variables for time effect that introduce cross-sectional dependence. For Model (7), individual components  $\{\alpha_i\}$ ,  $\{\lambda_i t\}$ , and  $\{\theta_t\}$  are partialled out in the first step and then pooled OLS is applied to estimate the slope coefficient vector  $\beta$ . Mark and Sul report the limiting distribution of the pooled OLS estimator of  $\beta$ . The cross-sectional correlation introduced in Model (7) is, however, less general than that of factor models as will be discussed below.

Mark, Ogaki, and Sul (2005) use the same model as (2) except that heterogeneity is allowed for the slope coefficient vector and that  $\{v_{it}\}$  are cross-sectionally correlated. For this extended model, they propose using the seemingly unrelated regression estimator, which requires estimating the inverse of the long-run variance-covariance matrix of  $\{v_{it}\}$ . For fixed  $N$ , Mark, Ogaki, and Sul show that the estimator is  $T$ -consistent and has a mixture normal distribution. They also show that the estimator is more efficient than Saikkonen's (1993) system DOLS. When  $N$  is large, however, the estimator is expected to not work well since it requires estimating  $N(N+1)/2$  elements of the long-run variance-covariance matrix of  $\{v_{it}\}$ . Westerlund's (2005a, Table 3) simulation study confirms this. In addition, it is well documented that estimating large covariance matrices invites serious inferential problems (cf. Johnstone 2001). Moreover, only the error terms are cross-sectionally correlated in Mark, Ogaki, and Sul's approach. A more general approach should consider cross-sectional correlation of the regressors too.

Moon and Perron (2005) adopt the same approach as in Mark, Ogaki, and Sul (2005) and show that the dynamic GLS estimator is most efficient. For fixed  $N$ , Choi and Chue's (2007) subsampling method can also be used for statistical inference on cross-sectionally correlated and cointegrated panels.

Bai and Kao (2006) extend the FM-OLS of Kao and Chiang (2000) to a model involving unobserved  $I(0)$  factors. They consider the model

$$\begin{aligned} y_{it} &= \alpha_i + \beta' x_{it} + u_{it}, \\ x_{it} &= x_{i,t-1} + \varepsilon_{it}, \\ u_{it} &= \lambda'_i f_t + \eta_{it}, \end{aligned} \quad (8)$$

where  $\{\lambda_i\}$  and  $\{f_t\}$  are the factor loadings and unobserved  $I(0)$  factors, respectively. Under the given factor structure of the error terms, the dependent variables  $\{y_{it}\}$  are cross-sectionally correlated. Furthermore, they assume that  $\{x_{it}\}$  are cross-sectionally independent, although they claim that the assumption can be relaxed by assuming a factor structure for  $\{\varepsilon_{it}\}$ . For this model, the spaces generated by the factor loadings and unobserved factors are estimated by the principal components method (cf. Chamberlain and Rothschild 1983; and Connor and Korajczyk 1986), which uses the OLS residuals  $\{\hat{u}_{it}\}$ .

Letting  $w_{it} = (f'_t, \eta_{it}, \varepsilon'_{it})'$ , Bai and Kao (2006) assume  $\frac{1}{\sqrt{T}} \sum_{t=1}^{[Tr]} w_{it} \Rightarrow B_i(r)$ , where  $B_i(r)$  is a vector Brownian motion with the homogeneous long-run covariance matrix

$$\Omega = \sum_{j=-\infty}^{\infty} E(w_{ij} w'_{i0}) = \Sigma + \Gamma + \Gamma' = \begin{bmatrix} \Omega_{ff} & \Omega_{f\eta} & \Omega_{f\varepsilon} \\ \Omega_{\eta f} & \Omega_{\eta\eta} & \Omega_{\eta\varepsilon} \\ \Omega_{\varepsilon f} & \Omega_{\varepsilon u} & \Omega_{\varepsilon\varepsilon} \end{bmatrix},$$

with  $\Gamma = \sum_{j=1}^{\infty} E(w_{ij} w'_{i0})$  and  $\Sigma = E(w_{i0} w'_{i0})$ . Letting  $\Delta = \Sigma + \Gamma$ , the matrices  $\Sigma$ ,  $\Gamma$ , and  $\Delta$  are partitioned conformable to  $\Omega$ .

Bai and Kao (2006) propose the FM-OLS estimator

$$\begin{aligned} \hat{\beta}_{\text{FM}} = & \left( \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_{i.})(x_{it} - \bar{x}_{i.})' \right)^{-1} \\ & \times \left( \sum_{i=1}^N \left( \sum_{t=1}^T (x_{it} - \bar{x}_{i.}) \hat{y}_{it}^+ - T(\hat{\Delta}_{\varepsilon u}^+ + \hat{\Delta}_{\varepsilon f}^+ \hat{\lambda}_i) \right) \right), \end{aligned}$$

where  $\hat{y}_{it}^+ = y_{it} - (\hat{\Omega}_{ue} + \hat{\lambda}'_i \hat{\Omega}_{fe}) \hat{\Omega}_{ee}^{-1} \Delta x_{it}$ ,  $(\hat{\Delta}_{\varepsilon f}^+ \quad \hat{\Delta}_{\varepsilon u}^+) = \hat{\Delta}_{\varepsilon b} - \hat{\Delta}_{\varepsilon e} \hat{\Omega}_{ee}^{-1} \hat{\Omega}_{eb}$  with  $z_{\varepsilon b} = (z_{\varepsilon f} \quad z_{\varepsilon u})$  ( $z = \Delta, \Omega$ ) and  $\hat{a}$  denotes the consistent estimator of parameter  $a$  using the OLS residuals. This estimator adapts Phillips and Hansen's (1990) FM-OLS estimator to account for the presence of factors. The FM-OLS estimator is  $\sqrt{NT}$ -consistent and has a multivariate normal distribution in the limit as  $N, T \rightarrow \infty$  simultaneously along with the condition  $\frac{\sqrt{N}}{T} \rightarrow 0$ . They also propose what they call a continuously updated fully modified (CUP-FM) estimator, which repeats the  $\beta$  estimation by FM-OLS using residuals from the FM-OLS of the previous stage until convergence. Bai and Kao show the CUP-FM estimator is more accurate than the FM-OLS estimator in finite samples.

Using Model (8) with heterogeneous long-run covariance matrix

$$\Omega_i = \sum_{j=-\infty}^{\infty} E(w_{ij} w'_{i0}) = \Sigma_i + \Gamma_i + \Gamma'_i = \begin{bmatrix} \Omega_{ff} & \Omega_{f\eta i} & \Omega_{f\varepsilon i} \\ \Omega_{\eta fi} & \Omega_{\eta\eta i} & \Omega_{\eta\varepsilon i} \\ \Omega_{\varepsilon fi} & \Omega_{\varepsilon ui} & \Omega_{\varepsilon\varepsilon i} \end{bmatrix},$$

Westerlund (2007) proposes a bias-adjusted estimator of  $\beta$ , which is defined as

$$\hat{\beta}^+ = \hat{\beta} - b_{NT}, \tag{9}$$

where

$$\begin{aligned}\hat{\beta} &= \left( \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_{i\cdot})(x_{it} - \bar{x}_{i\cdot})' \right)^{-1} \left( \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_{i\cdot})y_{it} \right), \\ b_{NT} &= \left( \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_{i\cdot})(x_{it} - \bar{x}_{i\cdot})' \right)^{-1} \sum_{i=1}^N (U_{\varepsilon f i} \lambda_i + U_{\varepsilon \eta i}), \\ U_{\varepsilon f i} &= \left( \sum_{t=1}^T (x_{it} - \bar{x}_{i\cdot}) \Delta x'_{it} - T \Delta_{\varepsilon \varepsilon i} \right) \Omega_{\varepsilon \varepsilon i}^{-1} \Omega_{\varepsilon f i} + T \Delta_{\varepsilon f i}, \\ U_{\varepsilon \eta i} &= \left( \sum_{t=1}^T (x_{it} - \bar{x}_{i\cdot}) \Delta x'_{it} - T \Delta_{\varepsilon \varepsilon i} \right) \Omega_{\varepsilon \varepsilon i}^{-1} \Omega_{\varepsilon \eta i} + T \Delta_{\varepsilon \eta i}.\end{aligned}\tag{10}$$

He shows that the bias-adjusted estimator is  $\sqrt{NT}$ -consistent and has a multivariate normal distribution in the limit as  $N, T \rightarrow \infty$  simultaneously along with the condition  $\frac{\sqrt{N}}{T} \rightarrow 0$ . The same applies to the feasible version of the bias-adjusted estimator. Westerlund reports that the bias-adjusted estimator performs slightly better than the FM-OLS estimator.

Bai, Kao, and Ng (2009) change Model (8) to

$$\begin{aligned}y_{it} &= \alpha_i + \beta' x_{it} + \lambda'_i f_t + u_{it}, \\ x_{it} &= x_{i,t-1} + \varepsilon_{it}, \\ f_t &= f_{t-1} + \eta_t,\end{aligned}\tag{11}$$

where  $\{f_t\}$  are observed or unobserved  $I(1)$  time series called the stochastic trends. The regressors  $\{x_{it}\}$  are assumed to be independent across  $i$ . Bai, Kao, and Ng propose a CUP-FM estimator for the coefficient  $\beta$ . When the stochastic trends are not observed, this estimator is calculated by an iterative method that requires initial estimates of either  $\beta$  or  $\{f_t\}$ . The CUP-FM estimator is shown to be  $\sqrt{NT}$ -consistent and has a multivariate normal distribution in the limit as  $N, T \rightarrow \infty$  simultaneously along with the condition  $\frac{\sqrt{N}}{T} \rightarrow 0$ .

Kapetanios, Pesaran, and Yamagata (2011; KPY hereafter) employ the model

$$y_{it} = \alpha'_i d_t + \beta'_i x_{it} + \lambda'_i f_t + u_{it},\tag{12}$$

where  $\{d_t\}$  are vectors of observed common effects partitioned as  $d_t = (d'_{1t}, d'_{2t})'$  with  $d_{it}$  denoting deterministic regressors and  $d_{2t}$   $I(1)$  regressors,  $\beta_i = \beta + \chi_i$  with  $\chi_i$  being an *i.i.d.* random variable with zero mean, and  $\{u_{it}\}$  are *i.i.d.* across  $i$ . The unobserved factor  $\{f_t\}$  is assumed to be  $I(1)$ . The regressor  $\{x_{it}\}$  is modelled as

$$x_{it} = A'_i d_t + \Lambda'_i f_t + v_{it}.\tag{13}$$

KPY's model assumes that both the regressors and regressands are cross-sectionally correlated. Moreover, the regressors may be cross-sectionally cointegrated. KPY's model is the most general thus far in the literature and is therefore worthy of serious considerations in empirical applications.

Since relations (12) and (13) imply

$$\begin{pmatrix} y_{it} \\ x_{it} \end{pmatrix} = B'_i d_t + C'_i f_t + \varepsilon_{it},$$

the cross-sectional mean of  $z_{it}$ ,  $\bar{z}_t$ , is written as

$$\bar{z}_t = \bar{B}' d_t + \bar{C}' f_t + \bar{\varepsilon}_t,$$

where  $\bar{A} = \frac{1}{N} \sum_{i=1}^N A_i$  ( $A = B, C$ ). If  $\bar{C}\bar{C}'$  has full rank, this relation yields

$$f_t \simeq (\bar{C}\bar{C}')^{-1} \bar{C} (\bar{z}_t - \bar{B}' d_t),$$

since  $\bar{\varepsilon}_t$  is negligible for large  $N$ . This shows that  $f_t$  can be estimated by combining vectors  $\bar{z}_t$  and  $d_t$ . Thus, if  $\bar{z}_t$  is used as an additional regressor for the regression equation (12), the cross-sectional correlation induced by the factors can be eliminated. This is KPY's basic idea, which was first proposed by Pesaran (2006).

After adding  $\bar{z}_t$  as an additional regressor to the regression equation (12), KPY study the group mean estimator of Pesaran and Smith (1995), which is the cross-sectional average of the time-series OLS estimator of  $\beta_i$ . In addition, they study the pooled estimator of  $\beta$ . Both of them are shown to be  $\sqrt{N}$ -consistent for  $\beta$  and asymptotically normal as  $N$  and  $T$  go to infinity simultaneously. But it seems hard to compare their efficiency. KPY also show that their theory works even when  $\bar{C}\bar{C}'$  is of deficient rank if the factor loadings are assumed to be independent of factors and individual-specific error processes  $\{u_{it}\}$  and  $\{v_{it}\}$ . KPY show through simulation that their estimators and test work well in finite samples.

Kao, Trapani, and Urga (2012) consider the model

$$y_{it} = \alpha_i + \beta' f_t + u_{it}, \quad (14)$$

where  $\{f_t\}$  are difference stationary and  $\{u_{it}\}$  are cross-sectionally correlated. In this model,  $\{y_{it}\}$  are cross-sectionally correlated due the presence of the common shocks  $\{f_t\}$  and cross-sectionally correlated error terms. Unlike in the previous studies that assume  $\{f_t\}$  to be unobserved factors, Kao, Trapani, and Urga assume that  $\{f_t\}$  are either observed or unobserved. Moreover, their main interest lies in the estimation of  $\beta$ . This is again in contrast to the previous studies, where  $\{f_t\}$  are nuisance variables that simply help modelling cross-sectional correlation. Kao, Trapani, and Urga's interest in their model arises from practical examples where each microunit is influenced by common macroeconomic factors  $\{f_t\}$ . When  $\{f_t\}$  are not observed, they are assumed to be factors that are extracted from panel data. Kao, Trapani, and Urga study asymptotic properties of the OLS estimators of the parameter  $\beta$  for Model (14) and its differenced

version. For the error terms, it is assumed that  $u_{it} \sim I(1)$  or  $u_{it} \sim I(0)$ . Extensions of their results to models with additional regressors that change across  $i$  and with slope coefficients varying over  $i$  are also made.

### 2.2.3 Poolability Tests

Using Westerlund's (2007) model discussed in the previous subsection (i.e., Model (8) with heterogeneous long-run covariance matrices), Westerlund and Hess (2011) propose a poolability test for cointegrated panels. The null hypothesis of their test is  $H_0 : \beta_i = \beta$  for all  $i$ , and the alternative is  $H_0 : \beta_i \neq \beta$  for at least one  $i$ . Their test statistic compares two estimators of the cointegration parameters – one individual and one pooled. The pooled estimator,  $\hat{\beta}^+$ , is defined in (9), and the individual one,  $\hat{\beta}_i^+$ , is the same as the pooled estimator if the summation signs over  $i$  (i.e.,  $\sum_{i=1}^N$ ) of  $\hat{\beta}$  and  $b_{NT}$  in (10) are eliminated. The test statistic is defined as  $\hat{H}_{\max} = \max_{1 \leq i \leq N} \hat{H}_i$ , where

$$\hat{H}_i = T^2(\hat{\beta}_i^+ - \hat{\beta}^+)' \left( M_i^{-1} \left( \frac{1}{6} \hat{\Omega}_{e,\varepsilon i} \hat{\Omega}_{\varepsilon e i} \right) M_i^{-1} \right)^{-1} (\hat{\beta}_i^+ - \hat{\beta}^+)$$

with  $M_i = \sum_{t=1}^T (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)'$ ,  $\hat{\Omega}_{e,\varepsilon i} = \hat{\lambda}_i' \hat{\Omega}_{f\varepsilon i} \hat{\lambda}_i + \hat{\Omega}_{\eta\varepsilon i}$  and  $\hat{a}$  denotes the consistent estimator of parameter  $a$  using the OLS residuals. Westerlund and Hess propose the normalized test statistic  $\hat{Z}_{\max} = \frac{1}{a_N} (\hat{H}_{\max} - b_N)$ , where  $a_N = 2$  and  $b_N = F^{-1}(1 - \frac{1}{N})$  with  $F(\cdot)$  being the chi-squared distribution function with degrees of freedom equal to the dimension of  $\alpha$ .  $\hat{Z}_{\max}$  has the limiting Gumbel distribution. That is,  $P(\hat{Z}_{\max} \leq z) \rightarrow \exp(-e^{-z})$  as  $N, T \rightarrow \infty$  with  $\frac{\sqrt{N}}{T} \rightarrow 0$ .

## 2.3 TESTS FOR PANEL COINTEGRATION

This section introduces tests for panel cointegration. There are three groups of tests: residual-based tests for the null of noncointegration, residual-based tests for the null of cointegration, and tests based on vector autoregression. The first group takes the null hypothesis as noncointegration and applies unit root tests to panel regression residuals. This idea originates from Engle and Granger (1987) and is developed further by Phillips and Ouliaris (1990). The second group's null hypothesis is cointegration and employs tests for the null of stationarity. The third group adapts Johansen's (1988, 1991) methods for panel data.

### 2.3.1 Residual-Based Tests for the Null of Noncointegration

Let  $\{y_{it}\}$  and  $\{x_{it}\}$  be noncointegrated,  $I(1)$  time series. In the Within-OLS panel regression

$$y_{it} - \bar{y}_{i\cdot} = \hat{\beta}'(x_{it} - \bar{x}_{i\cdot}) + \hat{u}_{it}, \quad (15)$$

it is expected that  $\{\hat{u}_{it}\}$  behave like an integrated process for large  $T$  and  $N$ . A similar result is reported in the literature on time series (see Phillips 1986). Thus, unit root tests applied to  $\{\hat{u}_{it}\}$  can test the null of noncointegration between  $\{y_{it}\}$  and  $\{x_{it}\}$ . This is the basic idea underlying Kao's (1999) procedure.

Kao (1999) considers his tests in a bivariate setting (i.e.,  $\{y_{it}\}$  and  $\{x_{it}\}$  are sequences of scalar random variables). He assumes that  $\{w_{it}\} = \{(\Delta y_{it} \quad \Delta x_{it})'\}$  are independent across  $i$  and that  $\{w_{it}\}$  satisfy the functional central limit theorem  $\frac{1}{\sqrt{T}} \sum_{t=1}^{[Tr]} w_{it} w'_{it} \Rightarrow \Omega^{1/2} W_i(r)$ , where

$$\Omega = E \left[ \sum_{t=1}^T \begin{pmatrix} \Delta y_{it} \\ \Delta x_{it} \end{pmatrix} \right] \left[ \sum_{t=1}^T \begin{pmatrix} \Delta y_{it} \\ \Delta x_{it} \end{pmatrix} \right]' = \begin{bmatrix} \Omega_{yy} & \Omega_{yx} \\ \Omega_{xy} & \Omega_{xx} \end{bmatrix} \text{ for every } i$$

and  $W_i(r)$  is a two-dimensional Wiener process. In addition, let

$$\Sigma = E \left( \begin{pmatrix} \Delta y_{it} \\ \Delta x_{it} \end{pmatrix} (\Delta y_{it} \quad \Delta x_{it})' \right) = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix} \text{ for every } i.$$

Note that we can estimate  $\Sigma$  and  $\Omega$  using  $\{\Delta y_{it}, \Delta x_{it}\}$ . For the former, we can use the usual formula,  $\frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T w_{it} w'_{it}$ . For the latter, we need use estimation methods for the long-run variance-covariance matrix (see, e.g., Newey and West 1987). These estimators are consistent and denoted as  $\hat{\Sigma}$  and  $\hat{\Omega}$ .

Under the null of noncointegration, Kao (1999) reports that  $\hat{\beta} \xrightarrow{p} \frac{\Omega_{yx}}{\Omega_{xx}}$  as  $T \rightarrow \infty$  followed by  $N \rightarrow \infty$ . Thus,  $\{\hat{u}_{it}\}$  behave like an integrated process for large  $T$  and  $N$  as in the case of time-series regression.

Letting the pooled AR(1) coefficient estimator using  $\{\hat{u}_{it}\}$  be  $\hat{\rho} = \frac{\sum_{i=1}^N \sum_{t=2}^T \hat{u}_{it} \hat{u}_{i,t-1}}{\sum_{i=1}^N \sum_{t=2}^T \hat{u}_{i,t-1}^2}$ , Kao (1999) shows that  $\sqrt{NT}(\hat{\rho} - 1)$  and the  $t$ -ratio for the unit root hypothesis  $t_\rho = \frac{\hat{\rho} - 1}{\sqrt{s_e^2 \left( \sum_{i=1}^N \sum_{t=2}^T \hat{u}_{i,t-1}^2 \right)^{-1}}}$  with  $s_e^2 = \frac{1}{NT} \sum_{i=1}^N \sum_{t=2}^T (\hat{u}_{it} - \hat{\rho} \hat{u}_{i,t-1})^2$  have asymptotic distributions that depend on nuisance parameters. Thus, the usual Dickey–Fuller test statistics cannot be used without modification. To overcome this difficulty, he proposes modified Dickey–Fuller coefficient and  $t$ -test statistics defined as, respectively,

$$DF_\rho^* = \frac{\sqrt{NT}(\hat{\rho} - 1) + \frac{3\sqrt{N}\hat{\sigma}_v}{\hat{\sigma}_{0v}^2}}{\sqrt{3 + \frac{36\hat{\sigma}_v^4}{5\hat{\sigma}_{0v}^4}}}$$

and

$$DF_t^* = \frac{t_\rho + \frac{\sqrt{6N}\hat{\sigma}_v}{2\hat{\sigma}_{0v}}}{\sqrt{\frac{\hat{\sigma}_{0v}^2}{2\hat{\sigma}_v^2} + \frac{3\hat{\sigma}_v^2}{10\hat{\sigma}_{0v}^2}}},$$

where  $\hat{\sigma}_v^2 = \hat{\Sigma}_{yy} - \hat{\Sigma}_{yx}\hat{\Sigma}_{xx}^{-1}$  and  $\hat{\sigma}_{0v}^2 = \hat{\Omega}_{yy} - \hat{\Omega}_{yx}\hat{\Omega}_{xx}^{-1}$ . These test statistics have a standard normal distribution in the limit as  $T \rightarrow \infty$  and  $N \rightarrow \infty$ .

Kao (1999) also studies the augmented Dickey–Fuller regression using  $\{\hat{u}_{it}\}$

$$\hat{u}_{it} = \tilde{\rho}\hat{u}_{i,t-1} + \sum_{j=0}^k \tilde{\omega}_j \Delta \hat{u}_{i,t-1-j} + \tilde{w}_{it} \quad (16)$$

and finds that the limiting distribution of the augmented Dickey–Fuller test statistic depends on nuisance parameters. As a remedy to the dependency on nuisance parameters, he modifies the augmented Dickey–Fuller test statistic as

$$ADF = \frac{t_{ADF} + \frac{\sqrt{6N}\hat{\sigma}_v}{2\hat{\sigma}_{0v}}}{\sqrt{\frac{\hat{\sigma}_{0v}^2}{2\hat{\sigma}_v^2} + \frac{3\hat{\sigma}_v^2}{10\hat{\sigma}_{0v}^2}}},$$

where  $t_{ADF}$  is the augmented Dickey–Fuller test statistic using Model (16), and reports that it has a standard normal as its limiting distribution.

Kao's (1999) test statistics are based on a bivariate regression. However, extending Kao's test statistics to the case of multiple regressors is straightforward: No changes in the test statistics are necessary except that  $\hat{\sigma}_v^2$  and  $\hat{\sigma}_{0v}^2$  now require estimating  $\Sigma_{yx}$ ,  $\Sigma_{xx}$ ,  $\Omega_{yx}$ , and  $\Omega_{xx}$ , each of which is either a vector or a matrix.

Simulation results in Kao (1999) show that the test statistics  $DF_\rho^*$ ,  $DF_t^*$ , and  $ADF$  have reasonable empirical size unless there is a moving average component in  $\{u_{it}\}$  that brings negative correlations in the errors. But when  $T$  is as small as 10 or 25, all the test statistics are subject to size distortions even with a large  $N$ . Comparing the three,  $DF_\rho^*$  and  $DF_t^*$  tend to keep the nominal size better than  $ADF$ .

Pedroni (1999, 2004) considers the panel regression model with heterogeneous coefficients

$$y_{it} = \alpha_i + \gamma_i t + \beta_i' x_{it} + u_{it}, \quad (17)$$

where  $\{y_{it}\}$  and  $\{x_{it}\}$  are scalar and  $k \times 1$   $I(1)$  time series, respectively. It is required that  $\{y_{it}\}$  and  $\{x_{it}\}$  satisfy conditions for the functional central limit theorem and that they are independent over  $i$ , as in Kao (1999). It is also required that  $\{x_{it}\}$  are not cointegrated. In addition, Pedroni's model allows the long-run variance of  $\{\Delta y_{it}, \Delta x_{it}\}$  to be heterogeneous.

Pedroni estimates Model (17) separately for each  $i$  by OLS and proposes seven test statistics that use the resulting regression residuals. Simulation results in Gutierrez

(2003) and Pedroni (2004) indicate that a variant of Im, Pesaran, and Shin's (2003) test statistic is most promising among the seven in the sense that it keeps nominal sizes within acceptable limits even under a small  $N$ . Similar to Im, Pesaran, and Shin, this test statistic is a normalized sum of Phillips and Ouliaris's (1990) coefficient test statistic for noncointegration using residuals from the  $i$ -th individual. It is defined as

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N Z_i, \quad (18)$$

where  $Z_i$  is the Phillips–Perron coefficient test statistic (cf. Phillips and Perron 1988) using the OLS residuals from the  $i$ -th equation of Model (17). As in Im, Pesaran, and Shin, this statistic requires mean and variance adjustments for it to have a standard normal distribution in the limit.<sup>2</sup> That is, we need to use  $\frac{1}{\sqrt{N}} \sum_{i=1}^N (Z_i - m_Z) / \sqrt{v_Z}$  in practice. Values of  $m_Z$  and  $v_Z$  are reported in Pedroni. We reject the null of noncointegration when the value of the test statistics is smaller than a critical value taken from the left-hand-side tail of a standard normal distribution.

Maddala and Wu (1999) and Choi (2001) have independently proposed panel unit root test statistics that use combinations of  $p$ -values. These can also be used for testing the null of noncointegration. Suppose that the  $p$ -value of a time-series cointegration test using the residuals from Model (17) is denoted as  $p_i$ . Maddala and Wu suggest using Fisher's test statistic defined by

$$P = -2 \sum_{i=1}^N \ln(p_i). \quad (19)$$

Choi considers the inverse normal test statistic defined by

$$Z = \frac{1}{\sqrt{N}} \sum_{i=1}^N \Phi^{-1}(p_i), \quad (20)$$

where  $\Phi(\cdot)$  is the standard normal cdf, a modified Fisher's test statistic

$$P_m = -\frac{1}{\sqrt{N}} \sum_{i=1}^N (\ln(p_i) + 1) \quad (21)$$

and some others. The modification for Fisher's test statistic is required since  $P \xrightarrow{P} \infty$  as  $T \rightarrow \infty$  and then  $N \rightarrow \infty$  (see Choi 2001, for details). These test statistics have a standard normal distribution in the limit when both  $T$  and  $N$  are large. Critical values for the  $P$  and  $P_m$  test statistics should be taken from the right-hand-side tail of the chi-square distribution and the left-hand-side tail of a standard normal distribution, respectively. In using  $Z$ , we reject the null hypothesis of noncointegration when its value is smaller than a critical value from the left-hand-side tail of a standard normal distribution.

Extending Breitung's (2002) variance-ratio test for a unit root to panel data, Westerlund (2005b) proposes test statistics for panel cointegration

$$VR_G = \sum_{i=1}^N \frac{\sum_{t=1}^T \hat{E}_{it}^2}{\hat{R}_i} \text{ and } VR_P = \frac{\sum_{i=1}^N \sum_{t=1}^T \hat{E}_{it}^2}{\sum_{i=1}^N \hat{R}_i},$$

where  $\hat{E}_{it}$  and  $\hat{R}_i$  are constructed using the OLS regression residuals from equation (17) and are defined as  $\hat{E}_{it} = \sum_{j=1}^T \hat{u}_{ij}$  and  $\hat{R}_i = \sum_{t=1}^T \hat{u}_{it}^2$ . An advantage of these test statistics is that they do not require estimating long-run variances, which often brings size distortions. The test statistic  $VR_G$  is the group mean of individual variance ratios and  $VR_P$  can be considered as a pooled variance ratio. Westerlund's test statistics allow serial correlation in the data, but assume cross-sectional independence as in Kao (1999) and Pedroni (1999). Furthermore, they require modifications as previous test statistics. That is,

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N (T^{-2} VR_G - m_G) / \sqrt{v_G} \text{ and } \sqrt{N} \sum_{i=1}^N (T^{-2} VR_P - m_P) / \sqrt{v_P}$$

have standard normal distributions in the limit as  $T \rightarrow \infty$  and then  $N \rightarrow \infty$ . The adjustment factors for the mean and variance are simulated and reported in Westerlund's Table 1. Westerlund shows that his test statistics keep nominal sizes better than those of Kao (1999) and Pedroni (1999), which confirms the advantage of his test statistics mentioned earlier.

Westerlund (2007) proposes a test for the null of noncointegration using the error-correction model. The idea for the test dates back to Banerjee, Dolado, and Mestre (1998) in the time series literature. Westerlund assumes the DGP

$$y_{it} = \alpha_i + \gamma_i t + z_{it}, \quad (22)$$

$$\phi_i(B) \Delta z_{it} = \delta_i(z_{i,t-1} - \beta_i' x_{i,t-1}) + \theta_i(B)' v_{it} + u_{it}, \quad (23)$$

where  $\{x_{it}\}$  is a noncointegrated,  $k \times 1$  integrated process represented by  $x_{it} = x_{i,t-1} + v_{it}$ ,  $\phi_i(z) = 1 - \sum_{j=1}^{p_i} \phi_{ij} z^j$ ,  $\theta_i(z) = 1 - \sum_{j=1}^{p_i} \theta_{ij} z^j$ , all the roots of the characteristic equation  $\phi_i(z) = 0$  lie outside the unit circle,  $E(u_{it} v_{js}) = 0$  for all  $i, t, j, s$ , and both  $v_{it}$  and  $u_{it}$  are  $I(0)$ . Equation (23) is the conditional model for  $\{z_{it}\}$  given  $\{x_{it}\}$  in a standard vector error-correction model. Westerlund's test requires that  $\{x_{it}\}$  not be error-correcting. That is, equation (23) with  $\Delta z_{it}$  being replaced by  $\Delta x_{it}$  is not allowed. One may find this assumption too restrictive. Moreover, if this assumption is violated, his test shows quite low power according to his simulation results.

Substituting (23) into (22), we obtain

$$\phi_i(B) \Delta y_{it} = \alpha_i^* + \gamma_i^* t + \delta_i(y_{i,t-1} - \beta_i' x_{i,t-1}) + \theta_i(B)' v_{it} + u_{it}. \quad (24)$$

If  $\delta_i = 0$ , there is no error-correction term and  $\{y_{it}\}$  is a unit root process that is not cointegrated with  $\{x_{it}\}$ . But if  $\delta_i < 0$ , there is error correction, implying that  $\{y_{it}\}$  and

$\{x_{it}\}$  are cointegrated. Thus, testing the null hypothesis  $H_0 : \delta_i = 0$  against the alternative  $H_0 : \delta_i < 0$  is tantamount to testing the null of noncointegration against the alternative of cointegration.

Rewriting equation (24) as

$$\Delta y_{it} = \alpha_i^* + \gamma_i^* t + \delta_i y_{i,t-1} + \lambda_i' x_{i,t-1} + \sum_{j=1}^{p_i} \alpha_{ij} \Delta y_{i,t-j} + \sum_{j=0}^{p_i} \theta_{ij} v_{i,t-j} + u_{it}$$

with  $\theta_{i0} = 1$ , Westerlund (2007) proposes four test statistics based on this equation. Among these, the group mean statistic defined as  $G_\tau = \frac{1}{N} \sum_{i=1}^N \frac{\hat{\delta}_i}{SE(\hat{\delta}_i)}$ , where  $\hat{\delta}_i$  is the OLS estimator of  $\delta_i$  and  $SE(\hat{\delta}_i)$  is its standard error, is simple to calculate and performs better than the others in terms of finite sample size. A modified version of this test statistic,  $\sqrt{N}(G_\tau - m_G)/\sqrt{v_G}$ , follows a standard normal distribution as  $T \rightarrow \infty$  and then  $N \rightarrow \infty$ . The values of  $m_G$  and  $v_G$  are reported in Westerlund's (2005b) Table 1. The test statistic works well in finite samples and is more powerful than some of Pedroni's (1999) residual-based test statistics when  $\{x_{it}\}$  is not error correcting.

Westerlund (2006a) proposes a test for the null of noncointegration in the presence of a structural change in level. That is, in Model (17), a structural change for  $\alpha_i$  occurs at  $T_i = [\lambda_i T]$  for all  $i$ . Structural changes for  $\beta_i$  are not considered. He adapts Pedroni's (1999) test statistics to accommodate a one-time structural change. For test statistic (18), the adapted test statistics are  $\frac{1}{\sqrt{N}} \sum_{i=1}^N \min_{\lambda_i} Z_i(\lambda_i)$  and  $\frac{1}{\sqrt{N}} \sum_{i=1}^N (\min_{\lambda_i} Z_i(\lambda_i) - m_Z) / \sqrt{v_Z}$ , where  $Z_i(\lambda_i)$  denotes the Phillips–Perron coefficient test statistic incorporating a structural change at  $T_i$ . These follow a standard normal distribution as  $T \rightarrow \infty$  and  $N \rightarrow \infty$  with appropriate mean and variance adjustments. The simulated mean and variance adjustments are reported in Westerlund's Table 1.

Cross-sectional independence has been assumed for all the tests discussed so far. When there exists cross-sectional correlation, tests introduced above become theoretically invalid and subject to size distortions as reported in Wagner and Hlouskova's (2010) simulation study. Gengenbach, Palm, and Urbain (2006; GPU hereafter) analyze the properties of Kao's (1999) and Pedroni's (1999) tests for noncointegration under cross-sectional correlation and cointegration using the factor model. GPU assume the DGP<sup>3</sup>

$$z_{it} = \Lambda_i f_t + U_{it},$$

where  $\{f_t\}$  denote noncointegrated,  $I(1)$ , unobserved, multidimensional factors, and  $\{U_{it}\}$  are either  $I(0)$  or  $I(1)$  and cross-sectionally independent. In this DGP, since  $\{f_t\}$  are common to all individuals,  $\{z_{it}\}$  are cross-sectionally correlated. In addition, letting

$$z_{it} = \begin{pmatrix} y_{it} \\ x_{it} \end{pmatrix} \begin{matrix} 1 \\ k \end{matrix} \text{ and } \Lambda_i = \begin{bmatrix} \Lambda_i^y & 0 \\ 0 & \Lambda_i^x \end{bmatrix} \begin{matrix} 1 \\ k \end{matrix}, \text{ we may write}$$

$$y_{it} = \Lambda_i^y f_t^y + U_{it}^y,$$

$$x_{it} = \Lambda_i^x f_t^x + U_{it}^x,$$

where  $\begin{pmatrix} a_{it}^y \\ a_{it}^x \end{pmatrix}$  is a partition of  $a_{it}$  conformable to  $z_{it}$ . Note that both  $\{y_{it}\}$  and  $\{x_{it}\}$  are cross-sectionally cointegrated. A linear combination of  $\{z_{it}\}$  is written as

$$y_{it} - \beta'_i x_{it} = \Lambda_i^y \left( f_t^y - \frac{\beta'_i \Lambda_i^x}{\Lambda_i^y} f_t^x \right) + (U_{it}^y - \beta'_i U_{it}^x). \quad (25)$$

Using this DGP, GPU study properties of Kao's (1999) and Pedroni's (1999) test statistics and report that their testing procedures become invalid whether or not  $\{U_{it}\}$  is  $I(1)$ .<sup>4</sup>

If we assume  $\frac{\beta'_i \Lambda_i^x}{\Lambda_i^y}$  is a constant  $\delta$  across  $i$ , Model (25) is written as

$$y_{it} - \beta'_i x_{it} = \Lambda_i^y G_t + U_{it}^*,$$

where  $G_t = f_t^y - \delta f_t^x$  and  $U_{it}^* = U_{it}^y - \beta'_i U_{it}^x$ . Using this representation and noting that  $\{y_{it}\}$  and  $\{x_{it}\}$  are noncointegrated when both or either of  $\{G_t\}$  and  $\{U_{it}^*\}$  are  $I(1)$ , GPU suggest the following sequential approach to testing for noncointegration.

Step 1: Extract the common factors from  $\{x_{it}\}$  and  $\{y_{it}\}$  using the principal components method as in Bai and Ng (2004) and test for a unit root in both the factors and the idiosyncratic components.

- Step 2: a. If Step 1 presents evidence for  $f_t \sim I(1)$  and  $U_{it}^* \sim I(0)$ , test the null of noncointegration between  $\{f_t^y\}$  and  $\{f_t^x\}$  using the estimated factors from Step 1.  
b. If Step 1 presents evidence for  $f_t \sim I(1)$  and  $U_{it}^* \sim I(1)$ , perform Step 2a. Next, defactor the series  $\{x_{it}\}$  and  $\{y_{it}\}$  separately to obtain

$$U_{it}^y = \sum_{s=1}^t (\Delta y_{it} - \hat{\Lambda}_i^y \Delta \hat{f}_t^y),$$

$$U_{it}^x = \sum_{s=1}^t (\Delta x_{it} - \hat{\Lambda}_i^x \Delta \hat{f}_t^x)$$

and test for noncointegration by applying, for example, Pedroni's (1999) test to  $\{U_{it}^y\}$  and  $\{U_{it}^x\}$ .

In Step 2b, we reject the null of noncointegration when both the tests reject. Note that GPU's procedure needs to assume the constancy of  $\frac{\beta'_i \Lambda_i^x}{\Lambda_i^y}$  across  $i$  for it to be theoretically valid. This assumption might be difficult to be satisfied in some applications. In addition, the method of how to test the null of noncointegration between  $\{f_t^y\}$  and  $\{f_t^x\}$  using the estimated factors needs to be spelled out in more detail. Using Johansen's (1988, 1991) test may not be suitable since some elements of  $\{f_t^y\}$  and  $\{f_t^x\}$  may overlap. The extant residual-based cointegration tests (e.g., Phillips and Ouliaris, 1991) are

based on the univariate regression where the dependent variable is a scalar. Extending this to the case of multivariate regression must be straightforward, but new critical values may need to be computed.

Westerlund (2008) assumes the DGP

$$\begin{aligned} y_{it} &= \alpha_i + \beta_i x_{it} + u_{it}, \\ x_{it} &= \delta_i x_{i,t-1} + \varepsilon_{it}, \end{aligned}$$

where  $\delta_i = 1$  or  $|\delta_i| < 1$ . He considers this DGP with application to an examination of the Fisher effect in mind, where the regressors, inflation rates, are not necessarily unit root processes. He also assumes a single regressor, although it does not seem to be difficult to extend his theory to the case of multiple regressors. For the error terms  $\{u_{it}\}$ , a factor structure is assumed such that

$$\begin{aligned} u_{it} &= \lambda'_i f_t + e_{it}, \\ f_t &= A f_{t-1} + w_t, \\ e_{it} &= \rho_i e_{i,t-1} + \eta_{it}, \end{aligned}$$

where  $A$  is a diagonal matrix whose diagonal elements take values less than one such that  $f_t \sim I(0)$ . The null hypothesis is  $H_0 : \rho_i = 1$  for all  $i$ , and the alternatives are  $H_1^p : \rho_i = \rho$  and  $\rho < 1$  for all  $i$  and  $H_1^g : \rho_i < 1$  for at least some  $i$ . Under the null hypothesis, the idiosyncratic terms  $\{e_{it}\}$  are  $I(1)$ . Following Bai and Ng (2004), he suggests estimating the differenced model

$$\Delta u_{it} = \lambda'_i \Delta f_t + \Delta e_{it}$$

using the OLS residuals  $\{\hat{u}_{it}\}$  and the principal components method. Letting the estimates of  $\Delta e_{it}$  be denoted as  $\Delta \hat{e}_{it}$ , Westerlund estimates  $\{e_{it}\}$  as  $\hat{e}_{it} = \sum_{j=2}^t \Delta \hat{e}_{it}$ . Then he applies Choi's (1994) Durbin–Hausman test to  $\{\hat{e}_{it}\}$ . The Durbin–Hausman test is based on the difference between the OLS estimator of  $\rho_i$  and the IV estimator using  $\{\hat{e}_{it}\}$  as an instrument. The OLS and IV estimators are both consistent under the null, but they have different probability limits under the alternative. Thus, a test statistic using the difference of the OLS and IV estimators can discriminate the null from the alternatives.

Let the OLS estimator of  $\rho_i$  from each time series be  $\hat{\rho}_i$  and the corresponding IV estimator be  $\tilde{\rho}_i$ . The pooled counterparts are denoted as  $\hat{\rho}$  and  $\tilde{\rho}$ . The Durbin–Hausman test statistics are defined as

$$\begin{aligned} DH^g &= \sum_{i=1}^N \hat{S}_i (\tilde{\rho}_i - \hat{\rho}_i)^2 \sum_{t=2}^T \hat{e}_{i,t-1}^2, \\ DH^p &= \hat{S}_N (\tilde{\rho} - \hat{\rho})^2 \sum_{i=1}^N \sum_{t=2}^T \hat{e}_{i,t-1}^2, \end{aligned}$$

where  $\hat{S}_i = \hat{\omega}_i^2 / \hat{\sigma}_i^4$ ,  $\hat{S}_N = \hat{\omega}_N^2 / \hat{\sigma}_N^4$ , and  $\hat{\omega}_i^2$  and  $\hat{\sigma}_i^2$  are the short- and long-run variance estimators using the OLS residuals from regressing  $\{\hat{e}_{it}\}$  on  $\{\hat{e}_{i,t-1}\}$ ,  $\hat{\omega}_N^2 = \frac{1}{N} \sum_{i=1}^N \hat{\omega}_i^2$ , and  $\hat{\sigma}_N^2 = \frac{1}{N} \sum_{i=1}^N \hat{\sigma}_i^2$ . The test statistic  $DH^g$  is for the alternative hypothesis  $H_1^g$ , and  $DH^p$  for  $H_1^p$ . Westerlund shows that asymptotic distributions of  $DH^g$  and  $DH^p$  are represented as

$$N^{-1/2} DH^g - \sqrt{N} E(B_i) \Rightarrow N(0, \text{Var}(B_i)),$$

$$N^{-1/2} DH^p - \sqrt{N} E(C_i)^{-1} \Rightarrow N(0, E(C_i)^{-4} \text{Var}(C_i)),$$

where  $B_i = \left(\int_0^1 W_i(r)^2 dr\right)^{-1}$  and  $C_i = 1/B_i$ . The simulated values of  $E(B_i)$ ,  $\text{Var}(B_i)$ ,  $E(C_i)$  and  $\text{Var}(C_i)$  are, respectively, 5.5464, 36.7673, 0.5005 and 0.3348. These values can be used to make the test statistics operational in practice. Westerlund shows by simulation that the tests statistics have good size and higher power than some of Pedroni's (1999) test statistics.

Gengenbach, Urbain, and Westerlund (2008) consider the triangular system<sup>5</sup>

$$y_{it} = b'_i x_{it} + \lambda'_{1i} f_t + u_{1it},$$

$$\Delta x_{it} = \lambda'_{2i} \Delta f_t + u_{2it},$$

$$\Delta f_t = \eta_t,$$

where  $\{u_{1it}, u_{2it}, \eta_t\}$  is a stationary linear process and  $\{u_{1it}, u_{2it}\}$  are independent of  $\{\eta_t\}$ . Let  $\beta_i = (1, b'_i, \lambda'_{1i})$  and assume  $\{x_{it}\}$  are weakly exogenous with respect to  $\beta_i$  and  $\alpha_{1i}$  below. Then, Gengenbach, Urbain, and Westerlund show that the triangular system can be written as a vector error-correction model

$$\Delta y_{it} = \alpha_{1i} \beta'_i z_{i,t-1} + B_i(L) \Delta y_{i,t-1} + C_i(L) \Delta x_{i,t-1} + D_i(L) \Delta f_t + \varepsilon_{it}, \quad (26)$$

where  $z_{it} = (y_{it}, x'_{it}, f'_t)'$ , and  $B_i(L)$ ,  $C_i(L)$ , and  $D_i(L)$  are lag polynomials. As in Westerlund (2007), if  $\alpha_{1i} = 0$ , there is no error-correction term and  $\{y_{it}\}$  is a unit root process and is not cointegrated with  $\{x_{it}, f_t\}$ . But if  $\alpha_{1i} < 0$ , there is error correction, implying that  $\{y_{it}\}$  and  $\{x_{it}, f_t\}$  are cointegrated. Thus, a test for the null of noncointegration against the alternative of cointegration can be implemented by testing the null hypothesis  $H_0 : \alpha_{1i} = 0$  against the alternative  $H_0 : \alpha_{1i} < 0$ .

Model (26) can be reparametrized as

$$\Delta y_{it} = \alpha_{1i} y_{i,t-1} + \gamma'_{1i} x_{i,t-1} + \gamma'_{2i} \beta'_i f_{t-1} + B_i(L) \Delta y_{i,t-1} \\ + C_i(L) \Delta x_{i,t-1} + D_i(L) \Delta f_t + \varepsilon_{it},$$

where  $\gamma_{1i} = -\alpha_{1i} b'_i$  and  $\gamma_{2i} = -\alpha_{1i} \lambda'_{1i}$ . Since  $\{f_t\}$  are not observed, following Pesaran (2007), Gengenbach, Urbain, and Westerlund (2008) replace  $\{f_t\}$  with the cross-sectional averages of  $(y_{it}, x'_{it})'$  and devise a test for the null hypothesis  $H_0 : \alpha_{1i} = 0$  and a Wald test for the null hypothesis  $H_0 : \alpha_{1i} = 0, \gamma_{1i} = 0, \gamma_{2i} = 0$ . The asymptotic distribution of the  $t$ -test statistics depends on nuisance parameters, while that of the

Wald test statistic does not. However, these tests should be performed for each individual separately, and the method of pooling these tests for the case of unknown factors is not provided in Gengenbach, Urbain, and Westerlund.

Bai and Carrion-i-Silvestre (2013) consider the model<sup>6</sup>

$$y_{it} = \beta'_i x_{it} + \lambda'_i f_t + u_{it},$$

where  $\{x_{it}\}$  are observed  $I(1)$  regressors,  $\{f_t\}$  are unobserved factors, and  $\{u_{it}\}$  are idiosyncratic errors. The factors are assumed to be  $I(1)$ ,  $I(0)$ , or a combination of both. The idiosyncratic errors are allowed to be serially and cross-sectionally correlated. For the regressors, it is assumed that

$$x_{it} = A_t \lambda_i + B_i f_t + \sum_{j=1}^r C_{i,j} (f_{t,j} \lambda_{i,j}) + \varepsilon_{it},$$

where  $\{A_t\}$ ,  $\{B_i\}$ , and  $\{C_{i,j}\}$  are matrices,  $f_{t,j}$  and  $\lambda_{i,j}$  are the  $j$ -th elements of  $f_t$  and  $\lambda_i$ , respectively, and  $\Delta \varepsilon_{it} \sim I(0)$ . The regressors are assumed to be correlated with  $\lambda_i$  or  $f_t$  or both. Obviously, they are cross-sectionally correlated. When both  $\{f_t\}$  and  $\{u_{it}\}$  are  $I(0)$ ,  $\{y_{it}\}$  and  $\{x_{it}\}$  are cointegrated. When  $\{u_{it}\}$  are  $I(0)$ ,  $\{y_{it}\}$ ,  $\{x_{it}\}$ , and  $\{f_t\}$  are cointegrated, a case studied by Bai, Kao, and Ng (2009) and KPY using slightly different models.

In order to control endogeneity of the regressors, Bai and Carrion-i-Silvestre (2013) use the model augmented by  $\{\Delta x_{i,t-j}\}_{j=-q, \dots, q}$

$$y_{it} = \beta'_i x_{it} + \sum_{j=-q}^q \zeta'_{ij} \Delta x_{i,t-j} + \lambda'_i f_t + v_{it}, \quad (27)$$

where the regressors are strictly exogenous relative to  $\{v_{it}\}$ . Note that  $\{u_{it}\}$  and  $\{v_{it}\}$  share the same order of integration. Differencing Model (27) once yields

$$\Delta y_{it} = \beta'_i \Delta x_{it} + \sum_{j=-q}^q \zeta'_{ij} \Delta^2 x_{i,t-j} + \lambda'_i \Delta f_t + \Delta v_{it}.$$

Using this model and an iterative least squares method, Bai and Carrion-i-Silvestre estimate  $\{\Delta v_{it}\}$  and  $\{\Delta f_t\}$ . Accumulating these estimates over  $t$ , estimates of  $\{v_{it}\}$  and  $\{f_t\}$ , denoted  $\{\hat{v}_{it}\}$  and  $\{\hat{f}_t\}$ , are obtained. Bai and Carrion-i-Silvestre apply the modified Sargan-Bhargava statistic (cf. Stock, 1999) to  $\{\hat{v}_{it}\}$  in order to test the null of noncointegration for each  $i$ . Then, they pool the test results employing Im, Pesaran, and Shin's (2003) test static, (19) and (21). In order to test the null hypothesis of noncointegration for the factors, they apply Ploberger and Phillips's test statistic to  $\{\hat{f}_t\}$  when the dimension of  $\{f_t\}$  is one, and Stock and Watson's (1988) procedure when it is greater than one.

### 2.3.2 Residual-Based Tests for the Null of Cointegration

McCoskey and Kao (1998) devise a test for the null of cointegration for independent panels. They consider the model

$$\begin{aligned} y_{it} &= \alpha_i + \beta'_i x_{it} + u_{it}, \\ u_{it} &= \varepsilon_{it} + v_{it}, \\ \varepsilon_{it} &= \varepsilon_{i,t-1} + \theta e_{it}, \varepsilon_{i0} = 0, \end{aligned} \tag{28}$$

where  $\{x_{it}\}$  is an  $I(1)$  process and  $\{v_{it}\}$  is an  $I(0)$  process. The null hypothesis of cointegration is  $H_0 : \theta = 0$  since  $\{u_{it}\}$  becomes  $I(0)$  under the null. It is assumed that  $w_{it} = (v_{it}, \Delta x_{it})$  is a mixing process that satisfies the functional central limit theorem,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{[Tr]} w_{it} \Rightarrow B(r), \tag{29}$$

where  $B(r)$  denotes the vector Brownian motion with the covariance matrix  $\Omega = \begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \Omega_{22} \end{bmatrix}$ .

McCoskey and Kao (1998) adapt the Lagrange multiplier test for the null of stationarity (see Kwiatkowski, Phillips, Schmidt, and Shin, 1992) to their problem. In order to compute their test statistic, they run either Phillips and Hansen's (1990) FM-OLS or DOLS on each individual. The residuals from the regressions, denoted here as  $\{\hat{u}_{it}^+\}$ , are used to formulate the test statistics

$$LM_{NT} = \frac{1}{N} \sum_{i=1}^N \frac{1}{T^2 \hat{\sigma}_i^2} \sum_{t=1}^T S_{it}^2, \tag{30}$$

where  $S_{it} = \sum_{j=1}^t \hat{u}_{ij}^+$ , and  $\hat{\sigma}_i^2$  is a consistent estimator of the parameter  $\omega_{11} - \omega'_{12} \Omega_{22}^{-1} \omega_{21}$ , which use the OLS residuals for each individual. McCoskey and Kao show that

$$\sqrt{N} (LM_{NT} - m_{LM}) / \sqrt{v_{LM}} \Rightarrow N(0, 1)$$

as  $T \rightarrow \infty$  and then  $N \rightarrow \infty$ , where  $m_{LM}$  and  $v_{LM}$  are mean and variance adjustment factors, respectively. Thus, in practice, the value of  $\sqrt{N} (LM_{NT} - m_{LM}) / \sqrt{v_{LM}}$  is compared with a critical value from the upper tail of a standard normal distribution to test the null of cointegration. McCoskey and Kao report simulated mean and variance adjustment. Additionally, their simulation results favor FM-OLS in formulating their test statistic; it offers better size and power properties than DOLS.

Westerlund (2006b) extends McCoskey and Kao's (1998) test so that it can accommodate structural changes in the deterministic component of a cointegrated panel regression. Westerlund uses the same model as (28) except that his model includes structural changes,

$$y_{it} = z'_{it} \gamma_j + \beta'_i x_{it} + u_{it},$$

where  $\{z_{it}\}$  is a vector of deterministic components. The index  $j = 1, \dots, M_i + 1$  is used to denote structural changes at the dates  $T_{i1}, \dots, T_{iM_i}$ . For example, if  $z_{it} = 1$  and  $M_i = 1$ , there are two regimes with

$$\gamma_{ij} = \begin{cases} \gamma_1 & t = 1, \dots, T_{i1} - 1 \\ \gamma_2 & t = T_{i1}, \dots, T \end{cases}.$$

Assuming that the number and dates of structural changes are known, Westerlund (2006b) defines his test statistic as  $Z_{NT} = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^{M_i+1} \sum_{t=T_{ij-1}}^{T_{ij}} \frac{1}{(T_{ij} - T_{ij-1})^2 \hat{\sigma}_i^2} S_{it}^2$ , where  $\hat{\sigma}_i^2$  is defined as for the test statistic (30),  $S_{it} = \sum_{j=T_{ij-1}+1}^t \hat{u}_{ij}^+$ ,  $\{\hat{u}_{ij}^+\}$  are residuals from an efficient cointegrating regression (e.g., FM-OLS and DOLS) that uses subsample ranging from  $T_{ij-1}$  to  $T_{ij}$ . In other words, the efficient regression is run for each subsample from which the residuals  $\{\hat{u}_{ij}^+\}$  are taken. As  $T \rightarrow \infty$  and then  $N \rightarrow \infty$ ,

$$\sqrt{N}(Z_{NT} - m_Z) / \sqrt{v_Z} \Rightarrow N(0, 1),$$

where  $m_Z$  and  $v_Z$  are mean and variance adjustment factors, respectively. Westerlund simulates mean and variance adjustment for the cases of intercept under structural changes and of both intercept and trend under structural changes in.

When the number and dates of structural changes are not known, Westerlund (2006b) suggests following Bai and Perron's (1998) procedure. Namely, the dates of structural changes are estimated for each  $M_i$  by  $\hat{T}_i = \arg \min_{T_i} \sum_{j=1}^{M_i+1} \sum_{t=T_{ij-1}}^{T_{ij}} \hat{u}_{it}^{+2}$ . Then, the number of changes,  $M_i$ , is estimated by the information criterion. The estimated number and dates are used to formulate the test statistic above.

All the tests discussed thus far assume cross-sectional independence. Westerlund and Edgerton (2007) apply Bühlmann's (1997) sieve bootstrap method to McCoskey and Kao's (1998) test such that cross-sectional correlation is allowed. They assume that  $\{w_t\}$  of equation (29) follow a linear process and approximate it as an AR process in their bootstrapping. The bootstrapped version of McCoskey and Kao's test is shown to keep nominal sizes well when  $N$  is as small as 10.

### 2.3.3 Panel VAR Cointegration Tests

Larsson, Lyhagen, and Löthgren (2001) consider the heterogeneous vector error-correction model of order  $k_i$ ,

$$\Delta Y_{it} = \Pi_i Y_{i,t-1} + \sum_{j=1}^{k_i} \Gamma_{ij} \Delta Y_{i,t-j} + \varepsilon_{it}, \quad (31)$$

where the  $p$ -dimensional vector  $\varepsilon_{it} \stackrel{\text{iid}}{\sim} N_p(0, \Omega_i)$ . They assume that cross-sectional units are independent and consider the null hypothesis

$$H_0 : \text{rank}(\Pi_i) = r_i \leq r \text{ for all } i = 1, \dots, N$$

against the alternative

$$H_1 : \text{rank}(\Pi_i) = p \text{ for all } i = 1, \dots, N.$$

Denote Johansen's (1988) likelihood ratio test for this hypothesis for individual  $i$  by  $LR_i$ , and their cross-sectional average as  $LRB_{NT} = \frac{1}{N} \sum_{i=1}^N LR_i$ . Then, Larsson, Lyhagen, and Löthgren's (2001) test statistic is defined by  $\sqrt{N}(LRB_{NT} - m_{LR}) / \sqrt{v_{LR}}$ , where  $m_{LR}$  and  $v_{LR}$  are the mean and variance adjustment factors. Its limiting distribution is  $N(0, 1)$  as  $N$  and  $T$  go to infinity such that  $\frac{\sqrt{N}}{T} \rightarrow 0$ . The simulated adjustment factors are reported in their Table 1.

Assuming a common cointegrating vector across individuals, Breitung (2005) employs the following vector error-correction model augmented by  $\beta'_{\perp} Y_{i,t-1}$

$$\Delta Y_{it} = \alpha_i \beta' Y_{i,t-1} + \gamma_i \beta'_{\perp} Y_{i,t-1} + \varepsilon_{it}, \quad (32)$$

where  $\beta'_{\perp} \beta = 0$  in order to test the null hypothesis

$$H_0 : \text{rank}(\beta) = r$$

against the alternative

$$H_1 : \text{rank}(\beta) = p.$$

Under the null hypothesis,  $\gamma_i = 0$  for  $i = 1, \dots, N$ , as shown in Saikkonen (1999). Premultiplying equation (32) by  $\alpha'_{i\perp}$  ( $\alpha'_{i\perp} \alpha_i = 0$ ), we have

$$\alpha'_{i\perp} \Delta Y_{it} = \phi_i \beta'_{\perp} Y_{i,t-1} + \alpha'_{i\perp} \varepsilon_{it}, \quad (33)$$

where  $\phi_i = \alpha'_{i\perp} \gamma_i$ , and  $\nu_{it} = \alpha'_{i\perp} \varepsilon_{it}$ . Using estimates of  $\alpha_{i\perp}$  and  $\beta_{\perp}$  ( $\hat{\alpha}_{i\perp}$  and  $\hat{\beta}_{\perp}$ ), Breitung estimates equation (33) for each individual and constructs a Lagrange multiplier test statistic as

$$LM_i(r) = T \times tr \left[ \sum_{t=2}^T \hat{f}_{it} g'_{i,t-1} \left( \sum_{t=2}^T g_{i,t-1} g'_{i,t-1} \right)^{-1} \sum_{t=2}^T g_{i,t-1} \hat{f}'_{it} \left( \sum_{t=2}^T f_{i,t-1} f'_{i,t-1} \right)^{-1} \right],$$

where  $\hat{f}_{it} = \hat{\alpha}'_{i\perp} \Delta Y_{it}$  and  $g_{i,t-1} = \hat{\beta}'_{\perp} Y_{i,t-1}$ . Then, following the idea of Im, Pesaran, and Shin (2003), Breitung proposes a test statistic

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N (LM_i(r) - E(LM_i(r))) / \sqrt{\text{Var}(LM_i(r))},$$

which tends to a standard normal distribution in the limit. Breitung presents the adjustment factors for the mean and variance for Model (32) with an intercept only and with an intercept and a linear time trend.

Larsson and Lyhagen (2007) extend Larsson, Lyhagen, and Löthgren (2001) to the case where cross-sectional correlation is allowed. They consider the stacked vector error-correction model

$$\begin{bmatrix} \Delta Y_{1t} \\ \vdots \\ \Delta Y_{Nt} \end{bmatrix} = \begin{bmatrix} \Pi_{11} & \Pi_{12} & \cdots & \Pi_{1N} \\ \Pi_{21} & \Pi_{22} & & \\ \vdots & & \ddots & \\ \Pi_{N1} & \Pi_{N2} & & \Pi_{NN} \end{bmatrix} \begin{bmatrix} Y_{1,t-1} \\ \vdots \\ Y_{N,t-1} \end{bmatrix} + \dots + \sum_{j=1}^k \begin{bmatrix} \Gamma_{11,j} & \cdots & \Gamma_{1N,j} \\ \vdots & & \\ \Gamma_{N1,j} & \cdots & \Gamma_{NN,j} \end{bmatrix} \begin{bmatrix} \Delta Y_{1,t-j} \\ \vdots \\ \Delta Y_{N,t-j} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \vdots \\ \varepsilon_{Nt} \end{bmatrix},$$

where  $\begin{bmatrix} \varepsilon_{1t} \\ \vdots \\ \varepsilon_{Nt} \end{bmatrix} \stackrel{\text{iid}}{\sim} N(0, \Omega)$  and  $\Omega > 0$ . Letting  $\Pi = (\Pi_{ij})_{i,j=1,\dots,N}$  and  $\Pi = AB'$ , Larsson and Lyhagen assume

$$B = \begin{bmatrix} \beta_{11} & & 0 \\ & \ddots & \\ 0 & & \beta_{NN} \end{bmatrix}.$$

These restrictions,  $\beta_{ij} = 0$  ( $i \neq j$ ), imply that cointegrating relationships are allowed only within each of the  $N$  individuals in the panel. Because matrices  $A$  and  $\Gamma_j = (\Gamma_{ab})_{a,b=1,\dots,N}$  ( $j = 1, \dots, k$ ) are unrestricted, there are no restrictions on the short-run dynamics of the vector error-correction model. Larsson and Lyhagen consider the null hypothesis

$$H_0 : \text{rank}(\Pi) \leq Nr$$

against the alternative

$$H_1 : \text{rank}(\Pi) = Np.$$

The likelihood ratio test statistic is formulated as in Johansen (1988), and Larsson and Lyhagen report its limiting distribution for finite  $N$ . Simulation results in Larsson and Lyhagen for  $N = 3$  show that the convergence to the limiting distribution is quite slow. As a remedy to this, they suggest using the Bartlett's correction of Johansen (2000, 2002). But if  $N$  is large, Larsson, and Lyhagen's test may not work well even with the Bartlett's correction. This issue requires further study.

Jacobson et al. (2008) assume

$$B = \begin{bmatrix} \beta_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & & \beta_{NN} \\ \beta_{N+1,1} & \cdots & \beta_{N+1,N} \end{bmatrix}$$

and develop a cointegration test as in Larsson and Lyhagen (2007). The particular structure of the matrix  $B$  is suitable to study purchasing power parity (PPP) relation where a numeraire country is required.

Groen and Kleibergen (2003) consider the following model obtained by stacking Model (31) (with  $k_i = 0$ )

$$\begin{aligned} \Delta Y_t &= \begin{pmatrix} \Pi_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Pi_N \end{pmatrix} Y_{t-1} + \varepsilon_t \\ &= \Pi_A Y_{t-1} + \varepsilon_t, \end{aligned} \quad (34)$$

where

$$Y_t = \begin{pmatrix} Y_{1t} \\ \vdots \\ Y_{Nt} \end{pmatrix}, \quad \varepsilon_t \stackrel{\text{iid}}{\sim} N(0, \Omega) \text{ and } \Omega = \begin{pmatrix} \Omega_{11} & \cdots & \Omega_{1N} \\ \vdots & \ddots & \vdots \\ \Omega_{N1} & \cdots & \Omega_{NN} \end{pmatrix}.$$

Since there are no restrictions on the variance–covariance matrix  $\Omega$ , cross-sectional correlation is allowed in Model (34). Groen and Kleibergen assume that the number of cointegrating vectors is  $r$  for all individuals. Model (34) is written in a vector error-correction form as

$$\begin{aligned} \Delta Y_t &= \begin{pmatrix} \alpha_1 \beta'_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \alpha_N \beta'_N \end{pmatrix} Y_{t-1} + \varepsilon_t \\ &= \Pi_B Y_{t-1} + \varepsilon_t. \end{aligned} \quad (35)$$

If all individuals are assumed to have the common cointegrating vector  $\beta$ , Model (35) is written as

$$\begin{aligned} \Delta Y_t &= \begin{pmatrix} \alpha_1 \beta' & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \alpha_N \beta' \end{pmatrix} Y_{t-1} + \varepsilon_t \\ &= \Pi_C Y_{t-1} + \varepsilon_t. \end{aligned}$$

Using likelihood ratio test statistics for both fixed and infinite  $N$ , Groen and Kleibergen testing

$$H_0 : \Pi_B \text{ vs. } H_1 : \Pi_A \quad (36)$$

and

$$H_0 : \Pi_C \text{ vs. } H_1 : \Pi_A. \quad (37)$$

The matrix  $\Pi_B$  has reduced rank, while  $\Pi_A$  is of full rank. Thus, the null and alternative hypothesis in (36) test the null hypothesis  $\text{rank}(\Pi_B) = Nr$  against the alternative  $\text{rank}(\Pi_B) = Np$ . In (37), further restrictions that  $\beta_i = \beta$  for all  $i$  are imposed under the null hypothesis. Groen and Kleibergen's test for infinite  $N$  again uses essentially the same idea as in Im, Pesaran, and Shin (2003). They also consider extending these tests to the model of higher-order dynamics and the model with deterministic components. Groen and Kleibergen's simulation results for small values of  $N$  show that using panel data increases power. How their test works for large  $N$  remains to be studied.

Larsson and Lyhagen (2007) and Groen and Kleibergen (2003) improve on Larsson, Lyhagen, and Löthgren (2001) and Breitung (2005) in that they allow for cross-sectional correlation. However, the improvements come at a cost: Their tests require restrictive assumptions that cointegration is allowed only within each individual and that each individual has the same cointegration rank. If these assumptions are violated, their tests become invalid. In fact, Banerjee, Marcelliono, and Osbat (2004) show via simulation for the cases  $N = 2, 4, 8$  that the tests of Larsson and Lyhagen (1999) and Groen and Kleibergen (2003) are subject to size distortions when the assumptions are violated. Additionally, when  $N$  is large, the tests may be subject to considerable size distortions even when the assumptions are satisfied as reported in Wagner and Hlouskova (2010).

## 2.4 SUMMARY AND FURTHER REMARKS

---

This chapter has surveyed the literature on panel cointegration. In the early stages of the literature, cross-sectional independence was assumed for regression estimators and tests for cointegration. Results under cross-sectional independence seem to be straightforward extensions of those for time series regressions and tests, although there are two indices in the panel data case which makes related analysis more involved than in the time-series case. Since the assumption of cross-sectional independence is not deemed to be suitable for applications to macropanels, attention has shifted to the case of cross-sectional dependence in the literature and various regression methods and tests for cointegration have been devised. These developments enriched the literature on panel cointegration and made it more relevant to empirical applications.

Among the various methods discussed in this paper, which should be used in practice? If we confine our attention to the case of cross-sectional dependence, KPY's regression model is the most general in the literature so far, and their estimators are intuitive and seem to work well although their optimality is unconfirmed. Regarding the tests for the null of noncointegration, GPU's model and testing procedure are suitable for application. Bai and Carrion-i-Silvestre's (2013) method is also recommendable for application with good initial values for some structural parameters, which are required for their iterative least squares method. Larsson and Lyhagen's (2007) and Groen and Kleibergen's (2003) panel VAR cointegration tests work well for small  $N$ , but their properties need further examination for large  $N$ .

The literature on panel cointegration still leaves room for further developments and refinements. Regression methods and cointegration tests under cross-sectional dependence should be improved further. Nonlinear models are unconsidered in the literature on panel cointegration. These issues certainly require further studies.

## ACKNOWLEDGMENTS

---

The author thanks Professor Badi Baltagi and two reviewers for helpful comments. He also thanks Sungwook Cho and HanBat Jeong for research assistance. The research reported in this paper was supported by the National Research Foundation of Korea (project # NRF-2010-342-B00006), which is gratefully acknowledged.

## NOTES

---

1. Westerlund (2005a) suggest using BIC to select the numbers of leads and lags in panel DOLS regressions.
2. See Pedroni (2004) for details along with proofs.
3. We omit the nonstochastic terms for brevity.
4. Banerjee, Marcelliono, and Osbat (2004) show via simulation that Pedroni's (1999) test statistics are subject to size distortions when there is cross-sectional cointegration. GPU's analysis partially explains this.
5. We omit deterministic terms from the model for simplicity.
6. We abstain from deterministic terms for simplicity.

## REFERENCES

---

- Bai, J., and J.L. Carrion-i-Silvestre (2013). "Testing Panel Cointegration with Unobservable Dynamic Common Factors That Are Correlated with the Regressors," *Econometrics Journal*, 16, 222–249.
- Bai, J., and C. Kao. (2006). "On the Estimation and Inference of a Panel Cointegration Model with Cross-Sectional Dependence," in *Panel Data Econometrics: Theoretical Contributions*

- and Empirical Applications (Contributions to Economic Analysis, Volume 274), edited by B.H. Baltagi and E. Sadka, pp. 3–30. Amsterdam: Emerald Group.
- Bai, J., C. Kao, and S. Ng. (2009). “Panel Cointegration with Global Stochastic Trends,” *Journal of Econometrics*, 149, 82–99.
- Bai, J., and S. Ng. (2004). “A Panic Attack on Unit Roots and Cointegration,” *Econometrica*, 72, 1127–1177.
- Bai, J., and P. Perron. (1998). “Estimating and Testing Linear Models with Multiple Structural Changes,” *Econometrica*, 66, 47–78.
- Baltagi, B., C. Kao, and L. Liu. (2008), “Asymptotic Properties of Estimators for the Linear Panel Regression Model with Individual Effects and Serially Correlated Errors: The Case of Stationary and Non-Stationary Regressors and Residuals,” *Econometrics Journal*, 11, 554–572.
- Baltagi, B., C. Kao, and S. Na. (2011), “Test of Hypotheses in Panel Data Models When the Regressor and Disturbances are Possibly Nonstationary,” *Advances in Statistical Analysis*, 95, 329–350.
- 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, 267–283.
- Banerjee, A., M. Marcellino, and C. Osbat. (2004). “Some Cautions on the Use of Panel Methods for Integrated Series of Macroeconomic Data,” *Econometrics Journal*, 7, 322–340.
- Breitung, J. (2002). “Nonparametric Tests for Unit Roots and Cointegration,” *Journal of Econometrics*, 108, 343–363.
- Breitung, J. (2005). “A Parametric Approach to the Estimation of Cointegration Vectors in Panel Data,” *Econometric Reviews*, 24, 151–173.
- Breitung, J., and M.H. Pesaran. (2008). “Unit Roots and Cointegration in Panels,” in *The Econometrics of Panel Data*, edited by L. Mátyás and P. Sevestre, pp. 279–322. Berlin: Springer.
- Bühlmann, P. (1997). “Sieve Bootstrap for Time Series,” *Bernoulli*, 3, 123–148.
- Chamberlain, G., and M. Rothschild. (1983). “Arbitrage, Factor Structure and Mean-Variance Analysis in Large Asset Markets,” *Econometrica*, 51, 1305–1324.
- Choi, I. (1994). “Durbin–Hausman Tests for Cointegration,” *Journal of Economic Dynamics and Control*, 18, 467–480.
- Choi, I. (2001). “Unit Root Tests for Panel Data,” *Journal of International Money and Finance*, 20, 249–272.
- Choi, I. (2002). “Instrumental Variables Estimation of a Nearly Nonstationary, Heterogeneous Error Component Model,” *Journal of Econometrics*, 109, 1–32.
- Choi, I. (2006). “Nonstationary Panels,” in *Palgrave Handbook of Econometrics*, Vol. 1, edited by T.C. Mills and K. Patterson, pp. 511–539. New York: Palgrave Macmillan.
- Choi, I., and T. Chue. (2007). “Subsampling Hypothesis Tests for Nonstationary Panels with Applications to Exchange Rates and Stock Prices,” *Journal of Applied Econometrics*, 22, 233–264.
- Connor, G., and R. Korajczyk. (1986). “Performance Measurement with the Arbitrage Pricing Theory: A New Framework for Analysis,” *Journal of Financial Economics*, 15, 373–394.
- Engle, R.F., and C.W.J. Granger. (1987). “Co-integration and Error Correction: Representation, Estimation, and Testing,” *Econometrica*, 55, 251–276.

- Entorf, H. (1997). "Random Walks with Drifts: Nonsense Regression and Spurious Fixed-Effect Estimation," *Journal of Econometrics*, 80, 287–296.
- Gengenbach, C., F.C. Palm, and J.P. Urbain. (2006). "Cointegration Testing in Panels with Common Factors," *Oxford Bulletin of Economics & Statistics*, 68, 683–719.
- Gengenbach, C., J. Urbain, and J. Westerlund. (2008). "Panel Error Correction Testing with Global Stochastic Trends," Working Paper, Maastricht University.
- Groen, J.J.J., and F. Kleibergen. (2003). "Likelihood-Based Cointegration Analysis in Panels of Vector Error-Correction Models," *Journal of Business & Economic Statistics*, 21, 295–318.
- Gutierrez, L. (2003). "On the Power of Panel Cointegration Tests: A Monte Carlo Comparison," *Economics Letters*, 80, 105–111.
- Im, K.S., M.H. Pesaran, and Y. Shin. (2003). "Testing for Unit Roots in Heterogeneous Panels," *Journal of Econometrics*, 115, 53–74.
- Jacobson, T., J. Lyhagen, R. Larsson, and M. Nessen. (2008). "Inflation, Exchange Rates and PPP in a Multivariate Panel Cointegration Model," *Econometrics Journal*, 11, 58–79.
- Johansen, S. (1988). "Statistical Analysis of Cointegration Vectors," *Journal of Economic Dynamics and Control*, 12, 231–254.
- Johansen, S. (1991). "Estimation and Hypothesis Testing of Cointegrating Vectors in Gaussian Vector Autoregressive Models," *Econometrica*, 59, 1551–1580.
- Johansen, S. (2000). "A Bartlett Correction Factor for Tests on the Cointegrating Relations," *Econometric Theory*, 16, 740.
- Johansen, S. (2002). "A Small Sample Correction for the Test of Cointegrating Rank in the Vector Autoregressive Model," *Econometrica*, 70, 1929–1961.
- Johnstone, I. (2001). "On the Distribution of the Largest Eigenvalue in Principal Components Analysis," *Annals of Statistics*, 29, 295–327.
- Kao, C. (1999). "Spurious Regression and Residual-Based Tests for Cointegration in Panel Data," *Journal of Econometrics*, 90, 1–44.
- Kao, C., and M.H. Chiang. (2000). "On the Estimation and Inference of a Cointegrated Regression in Panel Data, in: Baltagi B. (Ed.), Nonstationary Panels, Panel Cointegration, and Dynamic Panels," *Advances in Econometrics*, 15, 161–178.
- Kao, C., L. Trapani, and G. Urga. (2012). "Asymptotics for Panel Models with Common Shocks," *Econometric Reviews*, 31, 390–439.
- Kapetanios, G., M.H. Pesaran, and T. Yamagata. (2011). "Panels with Non-stationary Multifactor Error Structures," *Journal of Econometrics*, 160, 326–348.
- Kauppi, H. (2000). "Panel Data Limit Theory and Asymptotic Analysis of a Panel Regression with Near Integrated Regressors," *Advances in Econometrics*, 15, 239–274.
- Kwiatkowski, D., P.C.B. Phillips, P. Schmidt, and Y. Shin. (1992). "Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root: How Sure Are We That Economic Time Series Have a Unit Root?" *Journal of Econometrics*, 54, 159–178.
- Larsson, R., and J. Lyhagen. (1999). "Likelihood-Based Inference in Multivariate Panel Cointegration Models," Working Paper Series in Economic and Finance, 331.
- Larsson, R., and J. Lyhagen. (2007). "Inference in Panel Cointegration Models with Long Panels," *Journal of Business & Economic Statistics*, 25, 473–483.
- Larsson, R., J. Lyhagen, and M. Löthgren. (2001). "Likelihood-Based Cointegration Tests in Heterogeneous Panels," *Econometric Journal*, 4, 109–142.
- Maddala, G.S., and S. Wu. (1999). "A Comparative Study of Unit Root Tests with Panel Data and a New Simple Test," *Oxford Bulletin of Economics & Statistics*, 61, 631.

- Madsen, E. (2005). "Estimating Cointegrating Relations from a Cross Section," *Econometrics Journal*, 8, 380–405.
- Mark, N.C., M. Ogaki, and D. Sul. (2005). "Dynamic Seemingly Unrelated Cointegrating Regressions," *Review of Economic Studies*, 72, 797–820.
- Mark, N.C., and D. Sul. (2003). "Cointegration Vector Estimation by Panel DOLS and Long-Run Money Demand," *Oxford Bulletin of Economics & Statistics*, 65, 655–680.
- McCoskey, S., and C. Kao. (1998). "A Residual-Based Test of the Null of Cointegration in Panel Data," *Econometric Reviews*, 17, 57–84.
- Moon, H.R., and B. Perron. (2005). "Efficient Estimation of the Seemingly Unrelated Regression Cointegration Model and Testing for Purchasing Power Parity," *Econometric Reviews*, 23, 293–323.
- Newey, W.K., and K.D. West. (1987). "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.
- Pedroni, P. (1999). "Critical Values for Cointegration Tests in Heterogeneous Panels with Multiple Regressors," *Oxford Bulletin of Economics & Statistics*, 61, 653.
- Pedroni, P. (2000). "Fully Modified OLS for Heterogeneous Cointegrated Panels, in: Baltagi B. (Ed.), Nonstationary Panels, Panel Cointegration, and Dynamic Panels," *Advances in Econometrics*, 15, 93–130.
- Pedroni, P. (2004). "Panel Cointegration: Asymptotic and Finite Sample Properties of Pooled Time Series Tests with an Application to the PPP Hypothesis," *Econometric Theory*, 20, 597–625.
- Pesaran, M.H. (2006). "Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure," *Econometrica*, 74, 967–1012.
- Pesaran, M.H. (2007). "A Simple Panel Unit Root Test in the Presence of Cross-Section Dependence," *Journal of Applied Econometrics*, 22, 265–312.
- Pesaran, M.H., and R. Smith. (1995). "Estimating Long-Run Relationships from Dynamic Heterogeneous Panels," *Journal of Econometrics*, 68, 79–113.
- Phillips, P.C.B. (1986). "Understanding Spurious Regressions in Econometrics," *Journal of Econometrics*, 33, 311–340.
- Phillips, P.C.B., and B.E. Hansen. (1990). "Statistical Inference in Instrumental Variable Regression with  $I(1)$  Processes," *Review of Economic Studies*, 57, 99–125.
- Phillips, P.C.B., and M. Loretan. (1991). "Estimating Long-Run Economic Equilibria," *Review of Economic Studies*, 58, 407–436.
- Phillips, P.C.B., and H.R. Moon. (1999). "Linear Regression Limit Theory for Nonstationary Panel Data," *Econometrica*, 67, 1057–1111.
- Phillips, P.C.B., and S. Ouliaris. (1990). "Asymptotic Properties of Residual Based Tests for Cointegration," *Econometrica*, 58, 165–193.
- Phillips, P.C.B., and P. Perron. (1988). "Testing for a Unit Root in Time Series Regression," *Biometrika*, 75, 335–346.
- Saikkonen, P. (1991). "Asymptotically Efficient Estimation of Cointegration Regressions," *Econometric Theory*, 7, 1–21.
- Saikkonen, P. (1993). "Estimation of Cointegration Vectors with Linear Restrictions," *Econometric Theory*, 9, 19–35.
- Saikkonen, P. (1999). "Testing Normalization and Overidentification of Cointegrating Vectors in Vector Autoregressive Processes," *Econometric Reviews*, 18, 235–257.

- Stock, J. H. (1999). "A Class of Tests for Integration and Cointegration," in edited by R. F. Engle and H. White, *Cointegration, causality and forecasting. A Festschrift in Honour of Clive W. F. Granger*. Oxford University Press, Oxford, 135–167.
- Stock, J.H., and M.W. Watson. (1988). "Testing for Common Trends," *Journal of the American Statistical Association*, 83, 1097–1107.
- Stock, J.H., and M.W. Watson. (1993). "A Simple Estimator of Cointegrating Vectors in Higher Order Integrated Systems," *Econometrica*, 61, 783–820.
- Urbain, J., and J. Westerlund. (2011). "Least Squares Asymptotics in Spurious and Cointegrated Panel Regressions with Common and Idiosyncratic Stochastic Trends," *Oxford Bulletin of Economics and Statistics*, 73, 119–139.
- Wagner, M., and J. Hlouskova. (2010). "The Performance of Panel Cointegration Methods: Results from a Large Scale Simulation Study," *Econometric Reviews*, 29, 182–223.
- Westerlund, J. (2005a). "Data Dependent Endogeneity Correction in Cointegrated Panels," *Oxford Bulletin of Economics & Statistics*, 67, 691–705.
- Westerlund, J. (2005b). "New Simple Tests for Panel Cointegration," *Econometric Reviews*, 24, 297–316.
- Westerlund, J. (2006a). "Testing for Panel Cointegration with a Level Break," *Economics Letters*, 91, 27–33.
- Westerlund, J. (2006b). "Testing for Panel Cointegration with Multiple Structural Breaks," *Oxford Bulletin of Economics & Statistics*, 68, 101–132.
- Westerlund, J. (2007). "Testing for Error Correction in Panel Data," *Oxford Bulletin of Economics & Statistics*, 69, 709–748.
- Westerlund, J. (2008). "Panel Cointegration Tests of the Fisher Effect," *Journal of Applied Econometrics*, 23, 193–233.
- Westerlund, J., and D.L. Edgerton. (2007). "A Panel Bootstrap Cointegration Test," *Economics Letters*, 97, 185–190.
- Westerlund, J., and W. Hess. (2011). "A New Poolability Test for Cointegrated Panels," *Journal of Applied Econometrics*, 26, 56–88.

## CHAPTER 3

---

# DYNAMIC PANEL DATA MODELS

---

MAURICE J.G. BUN AND VASILIS SARAFIDIS

### 3.1 INTRODUCTION

---

THIS chapter reviews the recent literature on dynamic panel data models. Economic relationships usually involve dynamic adjustment processes. In time series regression models, it is common practice to deal with these by including in the specification lagged values of the covariates, the dependent variable, or both. The inclusion of lags of the dependent variable seems to provide an adequate characterization of many economic dynamic adjustment processes. However, in panel data analysis with a small number of time periods there often appear to be inference problems, such as small sample bias in coefficient estimation and hypothesis testing.

We consider a class of linear dynamic panel data models allowing for endogenous covariates. Sometimes it can be argued that the covariates are exogenous, at least conditional on individual- and time-specific effects, for example, when these covariates reflect natural phenomena. However, in many areas of economic inquiry this is often not the case. For instance, in empirical analysis of policy interventions, policy variables are most likely not strictly exogenous but simultaneously determined with the outcome variable of interest (e.g., Besley and Case 2000). Even if one is willing to assume that the covariates are not simultaneously determined, they may still be influenced by past values of the outcome variable.

Due to the various endogeneity problems mentioned above, least squares based inference methods (i.e., fixed effects or random effects estimators), are biased and inconsistent. Hence, it has become standard practice nowadays to use Instrumental Variables (IV) methods or the Generalized Method of Moments (GMM), which produce consistent parameter estimates for a finite number of time periods,  $T$ , and a large cross-sectional dimension,  $N$  (see, e.g., Arellano and Bond 1991; Arellano and Bover 1995; Blundell and Bond 1998). Within this class of methods, the system GMM estimator (Blundell and Bond 1998) has become increasingly popular. We do

not intend to provide a detailed overview of specific applications, but in labor economics (minimum wage effects, labor supply, returns to schooling, job training), development economics (effectiveness of foreign aid, transition economics), health economics (health expenditures, organization of health care, aging, addiction, insurance), industrial organization (mergers & acquisitions, evaluation of competition policy), international economics (effects of trade policy and economic integration), macroeconomics (economic growth, optimal currency areas), and finance (banking regulation) GMM inference methods have been applied extensively.

One main reason for their popularity in empirical research is that the GMM estimation approach may provide asymptotically efficient inference employing a relatively minimal set of statistical assumptions. However, despite its optimal asymptotic properties, the finite sample behavior of the GMM estimator and corresponding test statistics can be rather poor due to weakness and/or abundance of moment conditions and dependence on crucial nuisance parameters. As a result, several alternative inference methods have been proposed, often requiring different and more stringent assumptions. Here we will survey some of the most recent contributions.

In addition, an issue that has recently attracted further attention is the mean stationarity assumption that underlies the system GMM estimator. Roodman (2009) points out that this assumption is not trivial, which seems to be underappreciated in applied research. The effect of deviations from mean stationarity are analyzed theoretically by Hayakawa (2009) and Hayakawa and Nagata (2013). Kiviet (2007), Everaert (2013), and Juodis (2013) also explore this issue by Monte Carlo simulation. Consequently, in this chapter we will focus on mean stationarity in more detail and analyze the main arguments.

This is not the first review study on linear dynamic models for panel data; Blundell, Bond, and Windmeijer (2001) and Roodman (2009) provide excellent summaries of the GMM methodology. Arellano and Honoré (2001) also provide a very comprehensive analysis, including results for nonlinear models. Specific chapters in books on panel data also pay ample attention to dynamic panel data modeling (Arellano 2003a; Hsiao, 2003; Mátyás and Sevestre 2008; Baltagi 2013).

There are of course many other interesting and related topics that we do n't cover in this chapter. We do not discuss: (1) slope parameter heterogeneity; (2) cross-sectional dependence; (3) nonlinear models. Also we mainly focus on GMM inference methods but we briefly mention likelihood-based alternatives in Section 3.2. A discussion of some of these topics, however, can be found in other chapters of this volume.

## 3.2 REVIEW OF THE LITERATURE

---

Suppose the relation between the dependent variable  $y_{it}$  and a single covariate  $x_{it}$  can be modeled by the following dynamic specification:

$$y_{it} = \alpha y_{i,t-1} + \beta x_{it} + \eta_i + \varepsilon_{it}; \quad i = 1, \dots, N, t = 1, \dots, T, \quad (2.1)$$

where  $\eta_i$  denotes unobserved time-invariant heterogeneity and  $\varepsilon_{it}$  is the idiosyncratic error component.<sup>1</sup> We assume that  $y_{i0}$  and  $x_{i0}$  are observed. The dynamic panel data model in (2.1) permits the distinction between the long run, or equilibrium, relationship and the short-run dynamics. Note that  $x_{it}$  could also be a vector, containing both contemporaneous and lagged values of explanatory variables. It can be seen in this case that the above specification encompasses several important other specifications (i.e., static models, distributed lag or first-differenced specifications).

Often the individual-specific effect,  $\eta_i$ , is thought to be correlated with  $x_{it}$ . Furthermore, by construction the lagged dependent variable is correlated with the individual specific effect, i.e.  $E(\eta_i|y_{i,t-1}) \neq 0$ . Additionally, the covariate may also exhibit a nonzero correlation with the contemporaneous or lagged idiosyncratic errors, such that  $E(\varepsilon_{it}|x_{is}) \neq 0$  for  $t \leq s$ . All these endogeneity issues imply that least squares based estimators may be inconsistent. To this end, several alternative estimators have been proposed. In this chapter we focus on GMM estimators, although at the end of the section we briefly describe relative merits of other procedures, especially likelihood-based inference methods.

We consider models where idiosyncratic errors obey the following conditional moment restriction:

$$E[\varepsilon_{it}|\mathbf{y}_i^{t-1}, \mathbf{x}_i^s, \eta_i] = 0; \quad t = 1, \dots, T, \quad (2.2)$$

where  $\mathbf{y}_i^{t-1} = (y_{i0}, y_{i1}, \dots, y_{i,t-1})'$  and  $\mathbf{x}_i^s = (x_{i0}, x_{i1}, \dots, x_{is})'$ . Assumption (2.2) rules out serial correlation in  $\varepsilon_{it}$ , which is a base for constructing unconditional moments. However, it does not restrict the relationship between  $\eta_i$  and  $\mathbf{x}_i^s$ . Regarding the regressor  $x_{it}$  we distinguish between (i) strict exogeneity,  $s = T$ ; (ii) predeterminedness,  $s = t$ ; and (iii) endogeneity,  $s < t$ . That is, depending on  $s$ , equation (2.2) permits instantaneous or lagged feedback from  $y$  to  $x$ .

Based on assumption (2.2), the model can be expressed in first differences as

$$\Delta y_{it} = \alpha \Delta y_{i,t-1} + \beta \Delta x_{it} + \Delta \varepsilon_{it}, \quad (2.3)$$

for which the following (DIF) unconditional moment conditions are available:

$$E[\mathbf{y}_i^{t-2} \Delta \varepsilon_{it}] = 0; \quad E[\mathbf{x}_i^{s-1} \Delta \varepsilon_{it}] = 0; \quad t = 2, \dots, T, \quad (2.4)$$

with  $s$  depending on the exogeneity status of  $x_{it}$ . Lagged levels of the endogenous variables can be used as instruments for current changes. Simple IV estimators of this type were first proposed by Anderson and Hsiao (1981, 1982) for the first order autoregressive AR(1) model and in a multivariate setting and GMM framework by Holtz-Eakin, Newey, and Rosen (1988) and Arellano and Bond (1991).

Assumption (2.2) also rules out any correlations between  $\varepsilon_{it}$  and  $\eta_i$ .<sup>2</sup> This provides an additional set of  $T-2$  nonlinear moment conditions available for the model in first differences, as suggested by Ahn and Schmidt (1995):

$$E[(\eta_i + \varepsilon_{it}) \Delta \varepsilon_{i,t-1}] = 0, \quad t = 3, \dots, T. \quad (2.5)$$

Thus, under assumption (2.2) efficiency gains may occur by using (2.5) in addition to (2.4). Ahn and Schmidt (1995) show that the GMM estimator (labeled AS hereinafter) that makes use of (2.4) and (2.5) is efficient in the class of estimators that make use of second moment information. They also report substantial efficiency gains when comparing asymptotic variances for the AR(1) model. Especially when the series is highly persistent, the additional quadratic moment conditions become relatively informative compared with the moment conditions in (2.4) as can be seen from the calculations in Ahn and Schmidt (1995).

It is well known (see, e.g., Blundell and Bond, 1998) that the GMM estimator of the first-differenced model can have poor finite sample properties in terms of bias and precision when the series are persistent. One reason for this is that in this case lagged levels are weak predictors of the first differences. Blundell and Bond (1998) advocated the use of extra moment conditions that rely on certain stationarity restrictions on the time series properties of the data, as suggested by Arellano and Bover (1995). For the multivariate model in (2.1) these amount to assuming

$$E[\Delta y_{it}|\eta_i] = 0, \quad E[\Delta x_{it}|\eta_i] = 0, \quad (2.6)$$

which imply that the original series in levels have *constant correlation* over time with the individual-specific effects.<sup>3</sup> Assumption (2.6) leads to the following additional moment conditions for the model in levels (2.1) (hereinafter, LEV):

$$E[\Delta y_i^{t-1}(\eta_i + \varepsilon_{it})] = 0, \quad E[\Delta x_i^s(\eta_i + \varepsilon_{it})] = 0, \quad (2.7)$$

for  $t = 2, \dots, T$ , where  $\Delta y_i^{t-1} = (\Delta y_{i1}, \Delta y_{i2}, \dots, \Delta y_{i,t-1})'$  and so on. In words, with regards to endogenous variables, lagged changes can be used as instruments for current levels.

Notice that a subset of the moment conditions in (2.7) is redundant because it can be expressed as a linear combination of the moments in (2.4) (see, e.g., Kiviet, Pleus, and Poldermans 2013, for a proof of this result). Therefore, the complete set of non redundant linear moment conditions in levels can be specified as

$$E[\Delta y_{i,t-1}(\eta_i + \varepsilon_{it})] = 0, \quad t = 2, \dots, T, \quad (2.8)$$

together with

$$E[\Delta x_{i,t-1}(\eta_i + \varepsilon_{it})] = 0, \quad t = 2, \dots, T, \quad (2.9)$$

in case of endogenous  $x_{it}$ , or

$$E[\Delta x_{it}(\eta_i + \varepsilon_{it})] = 0, \quad t = 1, \dots, T, \quad (2.10)$$

in case of predetermined or strictly exogenous  $x_{it}$ . Combining (2.4) with (2.8) and either (2.9) or (2.10) leads to the system GMM estimator (labeled SYS). It should also be noted that (2.7) render the nonlinear moment conditions in (2.5) redundant. Hence under assumption (2.6) SYS is asymptotically efficient. Blundell and Bond (1998)

argued that SYS performs better than the DIF GMM estimator because the instruments in the LEV model remain good predictors for the endogenous variables even when the series are highly persistent.

Notwithstanding the popularity of the GMM methodology in applied economic research, producing accurate statistical inferences for panel data models using instrumental variables has not been a straightforward exercise. In particular, the desirable asymptotic properties of the estimators do not safeguard their performance in finite samples. In what follows, we summarize some of the issues that may arise in finite samples.

### 3.2.1 Asymptotic Standard Errors

As it has already been shown in the Monte Carlo study in Arellano and Bond (1991), estimated asymptotic standard errors of two step GMM estimators can be severely downward biased, suggesting more precision than is actually justified. Windmeijer (2005) showed that this is due to the fact that the weight matrix used in the second stage is based on initial parameter estimates, which themselves are subject to sampling variability that is not accounted for. Using asymptotic expansion techniques the author proposed a variance correction, leading to improved inference using the Wald test. In a rather extensive Monte Carlo study Bond and Windmeijer (2005) confirm the poor performance of the standard Wald test based on two step GMM. They find that using Wald statistics based on either one step GMM or the variance-corrected two step GMM or exploiting the LM statistic produces reliable inferences when identification is not too weak.

### 3.2.2 Many Instruments

Since dynamic panels are often largely overidentified, another important practical issue is how many moment conditions to use. Again, traditional first order asymptotics are not very helpful in answering this question as they imply ‘the more the merrier’. In practice, however, it is well documented that numerous instruments can overfit endogenous variables in finite samples (see, e.g., Bekker 1994), resulting in a trade-off between bias and efficiency. To gain some insight, consider a standard IV regression with one endogenous covariate; the  $R^2$  coefficient of the first stage regression takes the value of one when the number of instruments is equal to the number of observations. Thus, the instrumental variable is perfectly correlated with the endogenous variable and the IV estimator is numerically identical to the (biased) OLS estimator.

There is substantial theoretical work on the overfitting bias of GMM estimators in panel data models. For example, Koenker and Machado (1999) establish that a sufficient condition for the usual limiting distribution of the GMM estimator to remain

valid under instrument proliferation is  $m = o(N^{1/3})$ , where  $m$  denotes the number of instruments. Arellano (2003b) shows that in models with predetermined variables, such as a pure AR model, the bias as a result of overfitting is of order  $O(m/N)$ , while for models with endogenous variables the bias is of order  $O(mT/N)$ . Similarly, Alvarez and Arellano (2003) analyze a panel autoregressive model of order one, and show that although GMM remains consistent for  $T/N \rightarrow c$ , so long as  $0 \leq c \leq 2$ , for  $c > 0$  the estimator exhibits a bias in its asymptotic distribution that is of order  $1/N$ . Bun and Kiviet (2006) show that in comparison with GMM estimators that employ all available instruments, reducing the set of instruments by order T also decreases the bias by an order smaller in magnitude by a factor T. Ziliak (1997) examines the bias/efficiency trade-off issue using bootstrapping in an empirical application to life cycle labor supply under uncertainty. He shows that the bias of 2SLS and GMM estimators becomes larger as the number of instruments increases, and furthermore that GMM is biased downwards relative to 2SLS, arguably due to the nonzero correlation between the estimated weight matrix and the sample moments. Results from Monte Carlo simulation experiments vary, depending on the simulation design, the degree of overidentification in conjunction with the techniques employed for reducing the number of instruments, and finally the method employed in estimation. Windmeijer (2005) reported that for the two step DIF GMM, using only two lags of the dependent variable as instruments appeared to decrease the average bias by 40% relative to the estimator that made use of the full set of instruments, although the standard deviation of the estimator increased by about 7.5%. Roodman (2009) compared two popular approaches for limiting the number of instruments: (i) the use of (up to) certain lags instead of all available lags and (ii) combining instruments into smaller sets. His results show that the bias in SYS GMM based on the first approach is similar to the bias when using the full set of instruments. However, there is clear bias reduction under the second approach. On the other hand, Hayakawa (2009) shows that in panels with large unobserved heterogeneity the bias in DIF GMM can actually be larger when using a smaller set of instruments.

### 3.2.3 Dependence on Nuisance Parameters

Various studies (e.g., Binder, Hsiao, and Pesaran 2005; Bun and Kiviet 2006; Bun and Windmeijer, 2010) show that the finite sample properties of GMM estimators depend heavily on crucial nuisance parameters, especially the ratio of the variances of the individual-specific effects and the idiosyncratic errors ( $\sigma_\eta^2/\sigma_\varepsilon^2$ ). Binder, Hsiao, and Pesaran (2005) show that the asymptotic variance of the DIF GMM estimator increases with the variance of the individual-specific effects. Using asymptotic expansion techniques Bun and Kiviet (2006) approximate the bias of various one step GMM estimators. The asymptotic expansions provide analytic evidence on how the bias of the various GMM estimators depends on, among other things, the size of the

variance of the individual effects and the correlation between regressors and individual effects. Bun and Windmeijer (2010) analyze the bias of DIF, LEV, and SYS 2SLS estimators relative to bias in corresponding OLS estimators. They conclude that, although absolute bias of the LEV, and SYS 2SLS estimators tends to be small for persistent panel data, this bias is an increasing function of  $\sigma_\eta^2/\sigma_\varepsilon^2$ . Furthermore, relative biases of LEV and SYS 2SLS estimators are smaller and the associated Wald tests perform better than those of DIF when  $\sigma_\eta^2 < \sigma_\varepsilon^2$ . The reverse is the case when  $\sigma_\eta^2$  is larger than  $\sigma_\varepsilon^2$ . By Monte Carlo simulation these results are shown to extend to the panel data setting when estimating the model by GMM.

### 3.2.4 Weak Instruments

When instruments are weak (i.e., only lowly correlated with the endogenous variables), IV and GMM estimators can perform poorly in finite samples, see Bound, Jaeger, and Baker (1995), Staiger and Stock (1997), Stock and Wright, (2000) and Stock, Wright, and Yogo (2002). With weak instruments, IV or GMM estimators for panel data models are biased in the direction of the least squares estimator, and their distributions are non-normal (Wansbeek and Knaap 1999; Hahn, Hausman, and Kuersteiner 2007; Kruiniger 2009; Bun and Kleibergen 2013), affecting inference using standard  $t$  or Wald testing procedures.

To illustrate the weak instrument problem in dynamic panel data models, consider the the case of an AR(1) model, that is, impose  $\beta = 0$  in (2.1), and  $T = 2$ . The DIF and LEV models, (2.3) and (2.1), are now the following cross-sectional models:

$$\text{DIF: } \Delta y_{i2} = \alpha \Delta y_{i1} + \Delta \varepsilon_{i2}, \quad (2.11)$$

$$\text{LEV: } y_{i2} = \alpha y_{i1} + \eta_i + \varepsilon_{i2}. \quad (2.12)$$

The moment conditions for both models are:

$$\text{DIF: } E[y_{i0}(\Delta y_{i2} - \alpha \Delta y_{i1})] = 0, \quad (2.13)$$

$$\text{LEV: } E[\Delta y_{i1}(y_{i2} - \alpha y_{i1})] = 0, \quad (2.14)$$

hence for  $T = 2$  simple IV estimators result:

$$\hat{\alpha}_{\text{DIF}} = \frac{\sum_{i=1}^N y_{i0} \Delta y_{i2}}{\sum_{i=1}^N y_{i0} \Delta y_{i1}}, \quad \hat{\alpha}_{\text{LEV}} = \frac{\sum_{i=1}^N y_{i2} \Delta y_{i1}}{\sum_{i=1}^N y_{i1} \Delta y_{i1}}. \quad (2.15)$$

Assuming mean stationarity, i.e.  $y_{i0} = \frac{\eta_i}{1-\alpha} + \varepsilon_{i0}$ , the resulting covariance between regressor and instrument is:

$$\text{DIF: } E[y_{i0} \Delta y_{i1}] = (\alpha - 1) E[\varepsilon_{i0}^2], \quad (2.16)$$

$$\text{LEV: } E[y_{i1} \Delta y_{i1}] = \alpha (\alpha - 1) E[\varepsilon_{i0}^2] + E[\varepsilon_{i1}^2]. \quad (2.17)$$

Because  $E[\varepsilon_{i1}^2] \neq 0$  the LEV moment condition always seems to identify  $\alpha$  even for true values close to one, see Arellano and Bover (1995) and Blundell and Bond (1998). There is a caveat, however, because identification using LEV moment conditions is affected by the model for the initial observations.

Bond, Nauges, and Windmeijer (2005) show how identification of  $\alpha$  depends on the variance of the initial observations. The LEV first stage regression is:

$$y_{i1} = \pi_l \Delta y_{i1} + l_i, \quad (2.18)$$

with  $l_i$  being the reduced form error. When  $\alpha = 1$  we have  $\pi_l = 1$  and  $l_i = y_{i0}$ . Therefore, weak identification does not originate from  $\pi_l \rightarrow 0$ , but from  $Var(l_i) = Var(y_{i0})$  being large. When the number of time periods that the process has been in existence before the sample is drawn is fixed, then  $Var(y_{i0}) < \infty$ . In this case the LEV (and hence SYS) moment conditions identify  $\alpha$  even when its true value is one. For many DGPs, however,  $Var(y_{i0}) \rightarrow \infty$  when  $\alpha$  approaches one leading to identification failure. An example of such a DGP is that of covariance stationarity.

Kruiniger (2009) also shows that weakness of DIF and LEV moment conditions can manifest itself in different ways depending on the model for the initial observations. Following Han and Phillips (2006) sample moment conditions can be decomposed in “signal” and “noise”. Conventional asymptotics assume a strong signal, while noise is eliminated asymptotically. For the dynamic panel data model Kruiniger (2009) shows that, depending on the initial conditions, in some cases the signal becomes weak, while in other situations noise is dominating. For example, assuming covariance stationarity we have that  $E[\varepsilon_{i0}^2] = \frac{\sigma_\varepsilon^2}{1-\alpha^2}$  and  $E[\varepsilon_{i1}^2] = \sigma_\varepsilon^2$ , hence (2.16) and (2.17) become:

$$\text{DIF: } E[y_{i0} \Delta y_{i1}] = -\frac{\sigma_\varepsilon^2}{1+\alpha}, \quad (2.19)$$

$$\text{LEV: } E[y_{i1} \Delta y_{i1}] = \frac{\sigma_\varepsilon^2}{1+\alpha}. \quad (2.20)$$

These expressions suggest a strong “signal” for both DIF and LEV moment conditions, even when  $\alpha$  is (close to) one. However, at the same time the variance of the DIF and LEV moment conditions is proportional to  $\frac{1}{1-\alpha}$  implying explosive behavior when  $\alpha$  goes to one. In this case the noise in the moment equation dominates the signal and weak identification results for both DIF and LEV moment conditions.

Bun and Windmeijer (2010) show the weakness of DIF and LEV moment conditions in yet another way by calculating concentration parameters for DIF and LEV models assuming covariance stationarity. For a simple cross-sectional linear IV model, the concentration parameter is a measure of the information content of the instruments. When  $T = 2$  and assuming covariance stationarity they are equal for both models:

$$\sigma_\varepsilon^2 \frac{(1-\alpha)^2}{1-\alpha^2 + 2(1+\alpha)\frac{\sigma_\eta^2}{\sigma_\varepsilon^2}}. \quad (2.21)$$

This suggests a weak identification problem in the LEV model too when  $\alpha \rightarrow 1$  (and/or  $\frac{\sigma_\eta^2}{\sigma_\varepsilon^2} \rightarrow \infty$ ).

Bun and Kleibergen (2013) emphasize the arbitrariness of identification by the LEV moment condition by considering a joint limit process where both  $\alpha$  converges to one and  $N$  goes to infinity. Specifying the function  $h(\alpha)$  such that  $h(\alpha)^{-2} \propto \text{Var}(y_{i0})$  they show that when  $h(\alpha)\sqrt{N} \xrightarrow[N \rightarrow \infty, \alpha \uparrow 1]{} \infty$  the derivative of the LEV moment condition converges to a nonzero constant. However, when  $h(\alpha)\sqrt{N} \xrightarrow[N \rightarrow \infty, \alpha \uparrow 1]{} 0$ , it is the case that

$$h(\alpha) \frac{1}{\sqrt{N}} \sum_{i=1}^N y_{i1} \Delta y_{i1} \xrightarrow[N \rightarrow \infty, \theta_0 \uparrow 1]{d} N(0, \text{Var}(\varepsilon_{i1})). \quad (2.22)$$

This result shows identification failure since the derivative of the LEV moment condition converges to a random limit with mean zero.<sup>4</sup> Since any assumption on convergence rates of  $\alpha$  and  $N$  is arbitrary, identification by LEV moment conditions is arbitrary. Assuming  $h(\alpha)\sqrt{N} \xrightarrow[N \rightarrow \infty, \alpha \uparrow 1]{} 0$  Bun and Kleibergen (2013) show that 2-step DIF, LEV, and SYS GMM estimators and associated Wald statistics have non-standard large sample distributions, which results are qualitatively similar to those in Kruiniger (2009). They also show, however, that for  $T > 2$  it is possible to achieve identification of  $\alpha$  even when  $h(\alpha)\sqrt{N} \xrightarrow[N \rightarrow \infty, \alpha \uparrow 1]{} 0$  by combining SYS or AS moment conditions with the Lagrange multiplier GMM statistic proposed by Newey and West (1987), or with identification robust GMM statistics proposed by Stock and Wright (2000) and Kleibergen (2005).

Summarizing, whether the various sets of moment conditions identify the parameters of dynamic panel data models with persistent data depends on what seems reasonable to assume for the initial observations. In many microeconometric panel data a finite number of start-up periods may be a realistic scenario. In those cases identification issues are less severe, but this is not known on beforehand. Note that all above studies exploit mean stationarity and hence validity of LEV moment conditions. Strength of identification by the DIF (and also AS) moment conditions, however, may change substantially when we deviate from mean stationarity as we will discuss in Section 3.3.3 below.

### 3.2.5 Alternative Procedures

The dependence of finite sample distributions on the number and type of moment conditions as well as important nuisance parameters can be detrimental to the use of conventional GMM estimators in applied work. Hence, recent contributions propose to exploit alternative and possibly nonlinear moment conditions derived from inconsistent least squares procedures or likelihood based methods.

A central theme in linear dynamic panel data analysis is the fact that the fixed effects maximum likelihood (ML) estimator is inconsistent for a fixed number of time periods, as the number of cross-sectional units tends to infinity. This inconsistency is referred to as ‘Nickell bias’, due to Nickell (1981), and is an example of the incidental parameters problem (the number of parameters increasing with the sample size), analyzed first by Neyman and Scott (1948). This has led to an interest in likelihood-based methods that correct for the incidental parameters problem. Some of these methods are based on modifications of the profile likelihood, see Lancaster (2002) and Dhaene and Jochmans (2012). Other methods start from the likelihood function of the first differences, see Hsiao, Pesaran, and Tacmisioglu (2002), Binder, Hsiao, and Pesaran (2005) and Hayakawa and Pesaran (2014).

Well known transformations to remove individual-specific effects in panel data models are the within transformation and first differences. Kiviet (1995) and Bun and Carree (2005) exploit the possibility to correct the inconsistency of the fixed effects estimator, while Han and Phillips (2010) and Han, Phillips, and Sul (2014) recently developed efficient GMM methods based on alternative moment conditions arising from the model in first differences. However, the models considered in these studies are mainly autoregressive of nature (possibly with additional exogenous regressors) which currently limits their practical use.

A common advantage of these alternative likelihood based inference procedures is that they are largely invariant to the model parameters because unobserved heterogeneity is *a priori* transformed away. In comparison with the GMM approach, a limitation is that they impose exogeneity restrictions on the covariates and time series homoskedasticity, which may be violated in practice. Especially endogeneity with respect to the idiosyncratic errors is a common scenario in many applied studies. As mentioned by Hayakawa and Pesaran (2014), in principle it is feasible to exploit likelihood-based estimators in case of endogeneity too, however this requires supplementing the structural dynamic equation (2.1) with a reduced form equation for the endogenous regressors. Estimates of the parameters of interest could be retrieved from the resulting panel VAR coefficients. This is still a matter of future research.

### 3.3 REVISITING THE ISSUE OF INITIAL CONDITIONS

---

The popular system GMM estimator depends on (2.6), which is certainly satisfied if all variables are assumed to be mean stationary. A number of authors (see, e.g., Roodman 2009) have critically assessed the credibility of mean stationarity in applied economic research. In this section we discuss this issue in more detail. Furthermore, we describe consequences of departures from this assumption and statistical procedures to test it. Throughout the discussion the focus is on GMM inference methods.

### 3.3.1 Constant-Correlated Effects

The issue of initial conditions in models with fixed  $T$  has attracted considerable attention in the dynamic panel data literature since its infancy. For instance, Anderson and Hsiao (1982) and Bhargava and Sargan (1983) analyze the asymptotic properties of various maximum likelihood and instrumental variable type procedures under a large variety of assumptions about the initial conditions of the processes being studied.<sup>5</sup> One possibility is to assume that the initial condition is such that the process is mean stationary. The growing concern about the properties of dynamic panel estimators in finite samples may have contributed to placing large emphasis on this assumption, both in terms of theoretical developments, as well as in empirical applications.

In particular, mean stationarity has been employed for deriving additional moment conditions and developing new estimators (e.g., Arellano and Bover 1995; Blundell and Bond 1998). Given its mathematical convenience, it has also become a standard assumption in the many/weak instruments literature (e.g., Alvarez and Arellano 2003; Bun and Windmeijer 2010). Moreover, it is fair to say that in a large part of the literature during the last fifteen years or so, which reports results on the performance of GMM estimators based on Monte Carlo experiments, mean stationarity is either assumed from the outset, or it is effectively imposed as a byproduct of the simulation design. In the former case this is achieved by drawing the initial observations from a covariance stationary distribution (e.g., Blundell, Bond, and Windmeijer 2001). In the latter case the design entails generating  $T + S$  time series observations, with  $S$  equal to 50 or more, but using only  $T$  observations for estimation purposes. The first  $S$  observations are not considered in estimation, in order to ‘minimize the effect of initial conditions’ (e.g., Bun and Kiviet 2006). Although this practice is rather innocuous in panels with  $T$  large, it can have important consequences in panels with small  $T$ .

Another point that is easily discernible on selective reading of a huge empirical literature utilising panel data, is that the GMM estimator proposed by Ahn and Schmidt (1995), utilising (2.4) and (2.5), is rarely used in practice. This is despite the fact that this is the efficient estimator under a relatively minimal set of assumptions, excluding mean stationarity, and that using these moment conditions identification is achieved even for persistent panel data (Bun and Kleibergen 2013). Instead, in a substantial body of applied work the estimation strategy appears to involve the use of either DIF (which is not efficient under mean nonstationarity) or SYS (for which a sufficient condition for consistency is mean stationarity), or often both, without providing much theoretical justification for the implications of the underlying assumptions that validate the use of SYS specifically. The tendency to bypass AS is not surprising perhaps, given that both DIF and SYS are easy to compute and are readily available in several econometric packages of widespread use. On the contrary, so far as we know, AS is not yet part of a standard routine.

From a statistical perspective, and since most dynamic panel data models are typically overidentified, violations from mean stationarity are in principle detectable based

on Sargan's or Hansen's test of overidentifying restrictions. However, it is now well known that these tests can have very low power, especially when the number of instruments used is relatively large (see, e.g., Bowsher 2002; Roodman 2009). This could be partially mitigated by computing an incremental test based on AS and SYS, which involves a smaller number of degrees of freedom compared to an incremental test based on DIF and SYS. This is rarely implemented in practice.

In what follows, we revisit the conditions under which the LEV moment conditions hold true. We elaborate on what we call the “constant-correlated effects” assumption, which, for a limited lifespan of the time series, is a necessary and sufficient condition for the consistency of LEV and SYS GMM estimators. Since it is a rather intuitive concept to grasp, it has the benefit that, once one is prepared to motivate what unobserved heterogeneity is likely to capture in one's model, it becomes relatively straightforward to form an idea about how restrictive the condition appears to be on a specific application. If it does, the efficient estimator is AS and more effort should be made to apply it. Furthermore, we summarize some of the (limited) results existing and attempt to provide some guidance.

Recall that the LEV moment conditions (2.7) imply that the first difference of  $y_{it}$  and  $x_{it}$  are both uncorrelated with  $\eta_i$ , that is,

$$E(\Delta y_{it} \eta_i) = 0, \quad (3.1)$$

$$E(\Delta x_{it} \eta_i) = 0, \quad (3.2)$$

for  $t = 1, \dots, T$ . Thus, the moment conditions above imply that the first-differenced variables are free from the individual effects, which requires that the correlation between  $y_{it}$  ( $x_{it}$ ) and  $\eta_i$  is *constant over time*. We phrase this high level condition as the “constant-correlated effects” (cce) assumption, which can be expressed as

$$E(y_{it} \eta_i) = c_y, \quad (3.3)$$

$$E(x_{it} \eta_i) = c_x, \quad (3.4)$$

for all  $t$ . The issue of whether the variables of the model exhibit a constant correlation over time with unobserved time-invariant heterogeneity depends on the application in mind. Below we consider a few applications where GMM estimators have been popular. While the discussion should not be interpreted as indicative of a general pattern, it does suggest that the cce assumption is often taken too lightly by empirical researchers.

Suppose that (2.1) represents an earnings determination equation (see also Hause 1980; Arellano 2003a) with wage on the left-hand side and experience on the right-hand side (along with lagged wage and other variables, such as education and tenure). It is commonly viewed in this case that  $\eta_i$  captures, among other things, the effect of innate ability, or skills, which are unobserved to the econometrician and in any case hard to quantify. Consider the following scenario: the sample includes workers at different phases of their career; some of them are close to retirement and some are

new starters, having entered the labor market only recently for the first time, or having made a career change soon prior to the beginning of the sampling period. An argument could be made that the subgroup of new starters who are highly skilled, and therefore are employed in knowledge-intensive jobs, is likely to accumulate proportionally more experience as time progresses, and indeed receive higher salaries for this reason, relative to those individuals within the same group who have lower skills. This systematic relationship over time between unobserved skills and experience, or wage, is ruled out by the cce assumption.

Alternatively, one can draw from the literature of the estimation of production functions, in which  $\eta_i$  may capture the effect of technical inefficiency and unobserved managerial practices. Additionally, short-run dynamics may originate from autoregressive productivity shocks (Blundell and Bond 2000). One might argue that within new firms, or at least new entrants in a particular market, those which are more efficient are likely to be able to produce proportionally more output towards the end of the sampling period compared with inefficient firms, as the former group is able to learn better from past practices. Again, this scenario is ruled out by the cce assumption.

To obtain some insight about what the cce condition entails in our model, consider model (2.1) again, which is replicated below for ease of exposition

$$y_{it} = \alpha y_{i,t-1} + \beta x_{it} + \eta_i + \varepsilon_{it}. \quad (3.5)$$

We can express  $y_{it}$  recursively as follows:

$$y_{it} = \alpha^t \left( y_{i0} - \frac{\eta_i}{1-\alpha} \right) + \beta \sum_{s=0}^{t-1} \alpha^s x_{i,t-s} + \frac{\eta_i}{1-\alpha} + \sum_{s=0}^{t-1} \alpha^s \varepsilon_{i,t-s}. \quad (3.6)$$

It is immediately clear from the above expression that it is very unlikely that the correlation between  $y_{it}$  and  $\eta_i$  is constant over time, i.e.  $E(y_{it}\eta_i) = c_y$ , when the correlation between  $x_{it}$  and  $\eta_i$  is not. This is because  $y_{it}$  depends not only on the current but also on all lagged values of  $x_{it}$ , albeit their impact is declining with distance. To make further progress, let  $x_{it}$  form an AR(1) process such that

$$x_{it} = \rho x_{i,t-1} + \tau \eta_i + \nu_{it} = \rho^t \left( x_{i0} - \frac{\tau \eta_i}{1-\rho} \right) + \frac{\tau \eta_i}{1-\rho} + \sum_{s=0}^{t-1} \rho^s \nu_{i,t-s}, \quad (3.7)$$

where we assume that  $-1 < \rho < 1$ . As a result, (3.6) becomes

$$\begin{aligned} y_{it} = & \alpha^t \left( y_{i0} - \frac{\eta_i}{1-\alpha} \right) + \beta \sum_{s=0}^{t-1} \alpha^s \left[ \rho^{t-s} \left( x_{i0} - \frac{\tau \eta_i}{1-\rho} \right) + \frac{\tau \eta_i}{1-\rho} + \sum_{j=0}^{t-1-s} \rho^j \nu_{i,t-s-j} \right] \\ & + \frac{\eta_i}{1-\alpha} + \sum_{s=0}^{t-1} \alpha^s \varepsilon_{i,t-s} \end{aligned}$$

$$\begin{aligned}
&= \alpha^t \left( y_{i0} - \frac{1-\rho+\beta\tau}{(1-\alpha)(1-\rho)} \eta_i \right) + \beta \sum_{s=0}^{t-1} \alpha^s \rho^{t-s} \left( x_{i0} - \frac{\tau \eta_i}{1-\rho} \right) + \frac{1-\rho+\beta\tau}{(1-\alpha)(1-\rho)} \eta_i \\
&\quad + \beta \sum_{s=0}^{t-1} \alpha^s \sum_{j=0}^{t-1-s} \rho^j v_{i,t-s-j} + \sum_{s=0}^{t-1} \alpha^s \varepsilon_{i,t-s}.
\end{aligned} \tag{3.8}$$

The first (second) right-hand-side term within the brackets is the deviation of the initial in-sample observation on  $y$  ( $x$ ) from its steady state path, or its long run mean conditional on  $\eta_i$ . Eventually, assuming that the process for  $y$  and  $x$  is not altered, these deviations will die out because  $|\alpha| < 1$  and  $|\rho| < 1$ . However, in series with a limited lifespan, which is typically the case in microeconomics, these quantities are non-negligible, especially when the autoregressive coefficients are close to the value of one. Thus, the cce assumption suggests that any deviations from steady state behaviour need to be uncorrelated with  $\eta_i$ . We may also express this in an alternative form, as follows:

$$E \left[ \left( x_{i0} - \frac{\tau \eta_i}{1-\rho} \right) \frac{\tau \eta_i}{1-\rho} \right] = 0, \tag{3.9}$$

and

$$E \left[ \left( y_{i0} - \frac{1-\rho+\beta\tau}{(1-\alpha)(1-\rho)} \eta_i \right) \frac{1-\rho+\beta\tau}{(1-\alpha)(1-\rho)} \eta_i \right] = 0. \tag{3.10}$$

Both equations state effectively that *deviations* of the initial conditions from the steady state behavior are not systematically related to the *level* of the steady state itself. Under our hypothesised scenario in the earnings determination example, the expectation in (3.9) is likely to be negative as workers with higher innate ability (i.e., whose  $\eta_i$  value is relatively large) accumulate proportionately more experience, and thereby deviate to a greater extent from their steady state path of experience in the beginning of the sample, than workers with a small  $\eta_i$  value. Likewise, high skilled workers will systematically have lower wage in the beginning of the sample relative to their steady state earnings, in comparison with low skilled workers. It is clear that in order for the LEV moment conditions in (2.7) to be valid, one typically requires that *all* distinct covariates in a particular model satisfy a condition like (3.9) and the dependent variable satisfies (3.10).

The cce assumption is less strong than assuming that the series have a stationary mean. In other words, one can think of initial condition processes where the latter is not true but deviations from the steady state path remain uncorrelated with unobserved heterogeneity. It is useful to illustrate an example of such process with an application. Consider the empirics of growth models using country level data. The GMM methodology has been a popular estimation approach in this field. The Solow model takes the following form:

$$y_{it} - y_{i,t-1} = (\alpha - 1) y_{i,t-1} + \beta' x_{it} + \eta_i + \lambda_t + \varepsilon_{it}, \tag{3.11}$$

where  $y_{it} - y_{i,t-1}$  is the log difference in per capita GDP over a five year interval ( $t$ ),  $y_{i,t-1}$  denotes the logarithm of per capita GDP at the start of that period, and  $\mathbf{x}$  is a vector that contains variables such as the logarithm of the investment rate and the population growth rate, while in its augmented form various measures of human capital are included. Among other things,  $\eta_i$  reflects differences in the level of initial endowment of physical capital and natural resources across countries, as well as geographical location and topography, while  $\lambda_t$  reflects changes in productivity that are common to all countries. An equivalent representation of (3.11) arises by adding  $y_{i,t-1}$  on both sides, which resembles the standard dynamic panel data formulation as in (2.1). As we have already discussed, a sufficient condition for the first difference of per capita GDP,  $\Delta y_{it}$ , to be uncorrelated with  $\eta_i$  is mean stationarity of the level of per capita GDP,  $y_{it}$ , which also requires mean stationarity of the covariates used in the model. However, as Bond, Hoeffler, and Temple (2001) point out, while the Solow model is consistent with stationary conditional means of investment rates and population growth rates, this is clearly not the case for the per capita GDP series. One possibility is to assume that the conditional mean of  $y_{it}$  shifts intertemporally in some arbitrary way due to common technological progress. This is in fact what is already implied in equation (3.11) by the inclusion of common time effects,  $\lambda_t$ . Because this procedure is equivalent to transforming the series in terms of deviations from time-specific averages, we may consider instead the transformed model

$$\underline{y}_{it} = \alpha \underline{y}_{i,t-1} + \beta' \underline{\mathbf{x}}_{it} + \underline{\eta}_i + \underline{\varepsilon}_{it}, \quad (3.12)$$

where  $\underline{y}_{it} = y_{it} - \bar{y}_t$ ,  $\bar{y}_t = N^{-1} \sum_{i=1}^N y_{it}$ , and so on. The effect of common technological progress has been eliminated. Thus, any arbitrary pattern in the conditional mean of per capita GDP over time that is due to technological progress would be consistent with the cce assumption, provided that this is satisfied for the transformed series.

Nevertheless, the discussion above hinges on the assumption that the two-way error components formulation is adequate in explaining deviations from steady state behavior. One might object that what drives changes in the conditional mean of per capita GDP over time is the extent to which countries manage to absorb advances in technology available. Since this is likely to be different across  $i$ , depending on existing constraints and the production capacity that each country faces, among other considerations, a factor structure in the error term may be more appropriate to deal with this problem. It is worth emphasizing that a factor structure implies that changes in productivity are not common to all countries, and thereby deviations from steady state behavior are not identical across  $i$ , which can be an empirically relevant scenario. GMM type methods for estimating dynamic panel data models with a factor structure in the residuals and short  $T$ , have been developed by Ahn, Lee, and Schmidt (2013) and Robertson and Sarafidis (2013). Sarafidis and Wansbeek (2012) provide a recent overview of these methods. Panel data models with a factor structure are also discussed in chapter 2 of this volume.

### 3.3.2 Deviations of Initial Conditions from Steady State Behavior

Consider the following initial condition processes for  $x$  and  $y$  respectively:

$$x_{i0} = \delta_x \frac{\tau \eta_i}{1 - \rho} + w_{i0}, \quad (3.13)$$

$$y_{i0} = \delta_y \frac{(1 - \rho + \beta\tau) \eta_i}{(1 - \alpha)(1 - \rho)} + e_{i0}, \quad (3.14)$$

which can be motivated by (3.7) and (3.8). The conditional mean of  $x$  at the in-sample start-up period is  $E(x_{i0}|\eta_i) = \delta_x \frac{\tau \eta_i}{1 - \rho}$  and the conditional mean of  $y$  is  $E(y_{i0}|\eta_i) = \delta_y \frac{(1 - \rho + \beta\tau) \eta_i}{(1 - \alpha)(1 - \rho)}$ . Both  $\delta_x$  and  $\delta_y$  are meaningful in economic terms.<sup>6</sup> In particular  $0 < \delta_x < 1$  implies that the conditional mean of the initial in-sample observation is closer to zero than its steady state path. Therefore, assuming  $\tau > 0$ , if  $\eta_i > 0$  the series approaches its steady state from below and if  $\eta_i < 0$  then it approaches from above, with the rate of convergence depending on  $\rho$ . Similarly,  $\delta_x > 1$  implies that the value of the initial observation lies further away from zero than the series' long run conditional mean. Thus, if  $\eta_i > 0$  ( $\eta_i < 0$ ) the series converges from above (below). When  $\delta_x = 1$  the series is mean stationary, i.e. its conditional mean is constant over time throughout the sampling period. In this case one can readily check that (3.9) is satisfied. When  $\delta_y = 1$  as well,  $y$  is also mean stationary and (3.10) is fulfilled.

Under our hypothetical earnings determination scenario, one would expect that  $0 \leq \delta_x < 1$  since experience increases gradually over time and  $\eta_i > 0$ . On the other hand, suppose that the initial model in (3.5) represents a cost function with  $y_{it}$  denoting total cost,  $x$  denoting output, together with input prices, and  $\eta_i$  capturing the effect of cost inefficiency (so one would anticipate  $\eta_i \geq 0$ ). In this case one might expect that  $\delta_y \geq 1$ , i.e. firms' conditional expected cost in the beginning of the sample is at most equal to its long run mean, but not less. Under the hypothesis that firms adopt new work practices over time as an effort to cut expenditure, those firms which are economically more efficient are likely to be able to reduce total cost by a larger proportion. Hence,  $\delta_y$  would be strictly larger than one in this case and the series would approach its steady-state level from above.

One can provide an alternative interpretation of  $\delta_x$  and  $\delta_y$  when the initial conditions are perfectly correlated with the steady state levels. In particular, setting  $\text{var}(w_{i0}) = 0$  and  $\text{var}(e_{i0}) = 0$ , we have

$$\delta_x = \frac{\sqrt{\text{var}(x_{i0})}}{\tau \sigma_\eta / (1 - \rho)}, \quad (3.15)$$

and

$$\delta_y = \frac{\sqrt{\text{var}(y_{i0})}}{(1 - \rho + \beta\tau) \sigma_\eta / [(1 - \alpha)(1 - \rho)]}. \quad (3.16)$$

It can be seen that  $\delta_x$  and  $\delta_y$  equal the ratio of the standard deviation of the initial observations on  $x$  and  $y$ , respectively, over the standard deviation of the corresponding steady state levels. When there is more dispersion in the initial conditions than in the distribution of the steady state levels,  $\delta_x$  and  $\delta_y$  will be larger than one. In the economic growth literature, for example, this property is known as sigma convergence.

### 3.3.3 Consequences of Departures from Steady State Behavior

The magnitude of  $\delta_y$  and  $\delta_x$  turns out to be very important for the finite sample properties of various GMM estimators. For instance, for  $\delta_y = 1$  the correlation between  $\Delta y_{it-1}$  and  $y_{is}$ ,  $s < t - 1$  and  $t = 2, \dots, T$ , converges to zero when the variance of the  $\eta_i$  component of the error grows large. This is due to the fact that the total variation in  $y_{is}$  is dominated in this case by the variation in  $\eta_i$ , which, however,  $\Delta y_{it-1}$  is free from. This can have adverse consequences for GMM estimators that use lagged values of  $y$  in levels as instruments for first-differenced regressors when there is large unobserved heterogeneity present in the data. Hayakawa (2009), based on a pure AR(1) model, shows that the situation can be starkly different when  $\delta_y \neq 1$ . To see this, consider again for simplicity the case of an AR(1) model, i.e. impose  $\beta = 0$  in (2.1), and  $T = 2$ , and consider the DIF moment condition given in (2.13). Assuming time series homoskedasticity for idiosyncratic errors, the covariance between  $\Delta y_{i1}$  and  $y_{i0}$  is then:

$$\begin{aligned} \text{cov}(\Delta y_{i1}, y_{i0}) &= E[(\alpha y_{i0} + \eta_i + \varepsilon_{i1} - y_{i0}) y_{i0}] \\ &= -(1 - \alpha) E[y_{i0}^2] + E[\eta_i y_{i0}] \\ &= -(1 - \alpha) \left[ \delta_y^2 \frac{\sigma_\eta^2}{(1 - \alpha)^2} + \frac{\sigma_\varepsilon^2}{1 - \alpha^2} \right] + \delta_y \frac{\sigma_\eta^2}{1 - \alpha} \\ &= -\frac{\sigma_\varepsilon^2}{1 + \alpha} + \frac{\sigma_\eta^2}{1 - \alpha} \delta_y (1 - \delta_y). \end{aligned} \quad (3.17)$$

The first term is always negative while the sign of the second term depends on  $\delta_y$ .<sup>7</sup> For  $\delta_y > 1$  or  $\delta_y < 0$  the second term is always negative and thereby the correlation between  $\Delta y_{i1}$  and  $y_{i0}$  that is due to  $\eta_i$  adds up to the correlation that is due to the idiosyncratic component. Thus, the instruments should become stronger. This is not necessarily true when  $0 < \delta_y < 1$ , since in this case the second term is positive and the total effect on the correlation between  $\Delta y_{i1}$  and  $y_{i0}$  depends on the relative magnitude of  $\sigma_\eta^2$  and  $\sigma_\varepsilon^2$ , for a given value of  $\alpha$ . Thus, for  $\sigma_\eta^2/\sigma_\varepsilon^2 \rightarrow \infty$  any deviation from mean stationarity is likely to improve dramatically the performance of GMM estimators that do not require mean stationarity at first place, such as DIF and AS.

### 3.3.4 Testing for Constant-Correlated Effects

Moment conditions can be tested using the Sargan (1958)/Hansen (1982) overidentifying restrictions (OIR) statistic, which equals  $N$  times the value for the GMM objective function evaluated at the efficient two step GMM estimates. Asymptotically the OIR test statistic is chi-squared distributed with degrees of freedom equal to the number of overidentifying restrictions. It is clear that when there is no mean stationarity the linear moment conditions in (2.7) cannot be exploited. And an OIR test based on optimal system GMM should detect any deviations from assumption (2.6).

It has been shown, however, that the OIR test may be subject to low power due to many instruments (Bowsher 2002; Windmeijer 2005). Therefore, it is often suggested to use incremental or difference OIR tests. For example, assumption (2.6) implies extra moment conditions on top of those derived from (2.2). Hence, the difference between OIR SYS and DIF GMM statistics can be used, which is expected to have more discriminatory power compared with the SYS OIR test. Alternatively, the difference between the OIR SYS and AS GMM can be used, which is expected to have even better power properties. In the next section we will investigate by Monte Carlo simulation to what extent these and other predictions hold in finite samples.

## 3.4 SIMULATION RESULTS

In this section we first set out our Monte Carlo design, which is inspired by those of Blundell, Bond and Windmeijer (2001), Bun and Kiviet (2006) and Hayakawa and Nagata (2013). We allow for deviations from mean stationarity and pay special attention to some of the rules described by Kiviet (2007, 2012) for enhancing the scope of a simulation study and the interpretation of simulation results. For example, many existing Monte Carlo designs in the dynamic panel data literature do not obey any orthogonalization of the parameter space, which may hamper the interpretation of simulation results across experiments. Next, we discuss existing Monte Carlo studies simulating under deviations from mean stationarity. Finally, we report new simulation results investigating the impact of deviations from mean stationarity on various GMM coefficient estimators, corresponding Wald tests and Sargan statistics.

### 3.4.1 Monte Carlo Design

The data-generating process (dgp) is given by (3.5) and (3.7), which we replicate here for convenience:

$$y_{it} = \alpha y_{i,t-1} + \beta x_{it} + \eta_i + \varepsilon_{it}, |\alpha| < 1, \quad (4.1)$$

with

$$\begin{aligned} x_{it} &= \rho x_{i,t-1} + \tau \eta_i + v_{it}, |\rho| < 1, \\ v_{it} &= v_{it} + \phi_0 \varepsilon_{it} + \phi_1 \varepsilon_{i,t-1}, \end{aligned} \quad (4.2)$$

such that

$$\sigma_v^2 \equiv \text{var}(v_{it}) = \sigma_v^2 + (\phi_0^2 + \phi_1^2) \sigma_\varepsilon^2. \quad (4.3)$$

The long run coefficient of  $x$  on  $y$  equals  $\frac{\beta}{1-\alpha}$ . The initial condition for  $x$  in (3.13) is specified as

$$x_{i0} = \delta_x \xi \eta_i + w_{i0}, w_{i0} \sim i.i.d. (0, \sigma_w^2), E(w_{i0} | \eta_i) = 0, \quad (4.4)$$

where  $\xi = \frac{\tau}{1-\rho}$  and  $\sigma_w^2 = \frac{1}{1-\rho^2} \sigma_v^2$ .  $\xi \eta_i$  is the long run conditional mean, or steady state path, of  $x_{it}$  given  $\eta_i$ . Let  $r_x$  denote the correlation between the deviation of the initial condition of  $x$  from its long run steady state path and the level of the steady state path itself:

$$r_x = \text{corr}(x_{i0} - \xi \eta_i, \xi \eta_i) = \frac{\xi (\delta_x - 1) \sigma_\eta^2}{\sqrt{[\xi^2 (\delta_x - 1)^2 \sigma_\eta^2 + \sigma_w^2] \sigma_\eta^2}}. \quad (4.5)$$

Solving for  $\delta_x$  yields

$$\delta_x = \frac{r_x}{(1 - r_x^2)^{1/2}} \frac{\sigma_w}{\sigma_\eta} \frac{1}{\xi} + 1. \quad (4.6)$$

Thus, instead of setting an arbitrary value of  $\delta_x$  in order to investigate departures from steady state behavior, as it is common practice in the literature, we can set  $\delta_x$  according to  $r_x$ , which is more meaningful. For a fixed value of  $r_x$ , different values of  $\sigma_\eta^2$  change  $\delta_x$ . Similarly, larger values of  $\sigma_v^2$ , and hence of  $\sigma_w^2$ , increase the signal-to-noise ratio of the model and so  $\delta_x$  changes accordingly. When  $r_x = 0$ ,  $\delta_x = 1$  under the current design.<sup>8</sup>

We further specify the initial condition for  $y$  as

$$y_{i0} = \delta_y (\beta \xi + 1) \mu_i + e_{i0}, e_{i0} \sim i.i.d. (0, \sigma_e^2), E(e_{i0} | \eta_i) = 0, \quad (4.7)$$

where  $\mu_i = \eta_i / (1 - \alpha)$ ;  $(\beta \xi + 1) \mu_i$  is the long run conditional mean, or steady state path, of  $y_{it}$  given  $\eta_i$ . Thus, the process for  $y_{it}$  can be written as follows:

$$y_{it} = \varsigma_t \eta_i + \varpi_{it}, \quad (4.8)$$

where

$$\varsigma_t = \frac{1}{1-\alpha} [\alpha^t (\delta_y - 1) (\beta \xi + 1) + (\beta \xi + 1)] + \beta \xi (\delta_x - 1) \sum_{s=0}^{t-1} \alpha^s \rho^{t-s}, \quad (4.9)$$

and

$$\varpi_{it} = \beta \sum_{s=0}^{t-1} \alpha^s w_{i,t-s} + \alpha^t e_{i0} + \sum_{s=0}^{t-1} \alpha^s \varepsilon_{i,t-s}. \quad (4.10)$$

Let  $r_y$  denote the correlation coefficient between the deviation of the initial condition of the  $y$  process from its long run mean and the level of its long run mean:

$$\begin{aligned} r_y &= \text{corr}(y_{i0} - (\beta\xi + 1)\mu_i, (\beta\xi + 1)\mu_i) \\ &= \frac{(\delta_y - 1) \frac{1}{1-\alpha} (\beta\xi + 1) \sigma_\eta^2}{\sqrt{\left[ (\delta_y - 1)^2 (\beta\xi + 1)^2 \left( \frac{1}{1-\alpha} \right)^2 \sigma_\eta^2 + \sigma_\varpi^2 \right] \sigma_\eta^2}}, \end{aligned} \quad (4.11)$$

where  $\text{var}(\varpi_{it}) = \sigma_\varpi^2 = \sigma_v^2 c_v^2 + \sigma_\varepsilon^2 c_\varepsilon^2$ .<sup>9</sup> Solving for  $\delta_y$  yields

$$\delta_y = \frac{r_y}{\left(1 - r_y^2\right)^{1/2}} \frac{\sigma_\varpi}{\sigma_\eta} \frac{1 - \alpha}{\beta\xi + 1} + 1. \quad (4.12)$$

Thus, similarly to  $\delta_x$ ,  $\delta_y$  is set according to  $r_y$ . Clearly large values of  $\delta_y$  imply a high value for  $r_y$ , ceteris paribus.

Finally, as described in Kiviet (1995) and Bun and Kiviet (2006), the variances  $\sigma_v^2$  and  $\sigma_\eta^2$  are major determinants of the signal-to-noise ratio and the relative strength of the error components, respectively. The variance of  $y_{it}$  is equal to

$$\begin{aligned} \text{var}(y_{it}) &= \varsigma_t^2 \sigma_\eta^2 + \sigma_\varpi^2 \\ &= \varsigma_t^2 \sigma_\eta^2 + \sigma_v^2 c_v^2 + \sigma_\varepsilon^2 c_\varepsilon^2, \end{aligned} \quad (4.13)$$

A relationship between  $\sigma_\eta^2$  and  $\sigma_\varepsilon^2$  can be defined such that the cumulative impact on the average of  $\text{var}(y_{it})$  over time of the two error components  $\eta_i$  and  $\varepsilon_{it}$  is equal to the ‘variance ratio’ (VR):

$$\text{VR} = \frac{\overline{\varsigma^2} \sigma_\eta^2}{\sigma_\varepsilon^2 c_\varepsilon^2}, \quad (4.14)$$

where  $\overline{\varsigma^2} = T^{-1} \sum_{t=1}^T \varsigma_t^2$ . Solving for  $\sigma_\eta^2$  yields

$$\sigma_\eta^2 = \frac{\sigma_\varepsilon^2 c_\varepsilon^2 \text{VR}}{\overline{\varsigma^2}}. \quad (4.15)$$

Both  $c_\varepsilon^2$  and  $\overline{\varsigma^2}$  depend on the design parameters. Therefore, changes in these parameters will also affect the value of  $\sigma_\eta^2$  for a fixed variance ratio VR.

Consider the variance of the signal of the model at time  $t$ , conditionally on  $\eta_i$ , which can be written as

$$\sigma_{\text{signal},t}^2 = \text{var}(y_{it} | \eta_i) - \text{var}(\varepsilon_{it}) = \sigma_\varpi^2 - \sigma_\varepsilon^2. \quad (4.16)$$

The signal-to-noise ratio, defined conditionally on  $\eta_i$ , is now simply

$$\text{SNR} = \frac{\sigma_\varpi^2 - \sigma_\varepsilon^2}{\sigma_\varepsilon^2}. \quad (4.17)$$

SNR depends on the value of  $\sigma_{\omega}^2$ , which in turn is a function of  $\sigma_v^2$ . Hence, we may set  $\sigma_v^2$  such that SNR is controlled across experiments.

It should be noted that the proposed reparametrization of the parameter space to enhance the interpretability of Monte Carlo results is not unique. However, it follows closely the rules described in Kiviet (2007, 2012), notably the advice to reparametrize into an orthogonal autonomous design parameter space.

Setting  $\phi_1 = 0$  in (4.2), the dgp in (4.1) and (4.2) is equal to that of Blundell, Bond and, Windmeijer (2001). Setting  $\phi_0 = 0$  in (4.2), the dgp in (4.1) and (4.2) is equal to one of the schemes analyzed in Bun and Kiviet (2006). One can choose the vector of parameters  $(\delta_x, \delta_y, \sigma_{\eta}^2, \sigma_v^2)$  by choosing values for  $(r_x, r_y, VR, SNR)$  or vice versa. The advantage of fixing  $(r_x, r_y, VR, SNR)$  is that we control some important model characteristics across experiments. The remaining parameters in the design are  $(\alpha, \beta, \rho, \phi_0, \phi_1, \tau, \sigma_{\varepsilon}^2)$  and the dimensions are  $(T, N)$ .

### 3.4.2 Existing Results

Blundell and Bond (1998) already showed the vulnerability of system GMM to a deviation of the initial conditions from steady state behavior for the AR(1) model. The specification of their initial condition is such that the implied  $r_y \approx -0.65$  in their Table 6, which seems a quite strong deviation from cce. As a result the SYS GMM estimator of  $\alpha$  has a large upward bias, while DIF GMM is virtually unbiased and compared with the mean stationary case gets a much smaller Monte Carlo standard deviation. SYS Wald statistics are heavily size distorted, while rejection frequencies of DIF Wald statistics are close to nominal significance levels. Finally, SYS Sargan has power to detect the violation of the cce assumption.

Hayakawa (2009) and Hayakawa and Nagata (2013) also provide simulation results for the AR(1) model. In Hayakawa (2009) only coefficient bias for the DIF GMM estimator is investigated, while Hayakawa and Nagata (2013) analyze other estimators as well and investigate finite sample properties of Sargan tests too. They find favorable behavior of the DIF GMM estimator when  $\delta_y \neq 1$ . It should be noted, however, that these results are partly driven by the set-up of their Monte Carlo design. In particular, the variance of the individual effects,  $\sigma_{\eta}^2$ , instead of the variance ratio,  $VR$ , is fixed across experiments. This implies that when  $\alpha$  is close to one the correlation between the endogenous regressor and the instrument suddenly becomes very large when  $\delta_y \neq 1$ ; see result (3.17). In other words, minor deviations of initial conditions from steady state behavior have a huge impact on the relevance of the instruments.

Regarding the dynamic panel data model with additional regressors<sup>10</sup> Everaert (2013) provides simulation results (coefficient bias and t test) for a model with an additional strictly exogenous covariate. In other words,  $\phi_0 = \phi_1 = 0$  in (4.2). The only deviation from mean stationarity investigated is to set  $y_{i0} = 0$ . As a result SYS GMM becomes heavily biased, as expected.

Hayakawa and Nagata (2013) provide simulation results on coefficient bias for a model with an additional endogenous covariate. Also Sargan and incremental Sargan tests are investigated. They closely follow the design of Blundell, Bond, and Windmeijer (2001), that is,  $\phi_0 \neq 0$  and  $\phi_1 = 0$  in (4.2). DIF GMM shows favorable results when cce is violated, but the Monte Carlo design is again such that the individual effects dominate the idiosyncratic disturbances when persistence is high. In other words, important model characteristics like the variance ratio can achieve rather extreme values.

In Kiviet (1995) and Bun and Kiviet (2006) it has been shown that a proper comparison of simulation results over different parameter values requires control over basic model characteristics like goodness of fit and relative strength of error components. In the above design these are quantified by SNR and VR respectively, which in turn determine the values of the variances  $\sigma_v^2$  and  $\sigma_\eta^2$ . It is instructive to analyze what implied values for VR and SNR other studies have chosen. Blundell, Bond, and Windmeijer (2001) choose  $\beta = 1, \tau = 0.25, \phi_0 = -0.1, \phi_1 = 0, \sigma_\eta^2 = 1, \sigma_\varepsilon^2 = 1$  and  $\sigma_v^2 = 0.16$ . Furthermore, they consider four designs with  $\alpha$  and  $\rho$  either 0.5 or 0.95. Choosing  $\alpha = \rho = 0.5$  implies  $SNR = 0.48$  and  $VR = 9$ . Increasing  $\rho$  to 0.95 results in  $SNR = 6.36$  and  $VR = 119$ , a large proportional increase in both signal-to-noise ratio and variance ratio. Setting  $\alpha$  equal to 0.95 (but keeping  $\rho = 0.5$ ) results in  $SNR = 11.88$  and  $VR = 134$ , again a large increase in both signal and relative strength of individual effects. Finally, increasing both  $\alpha$  and  $\rho$  to 0.95 results in  $SNR = 337.17$  and  $VR = 1478$ , a huge increase in both variance ratio and signal-to-noise ratio. It is clear from these calculations that changing the autoregressive dynamics has substantial consequences for both explained variation and unobserved heterogeneity in the model. For proper comparison across experiments it is therefore necessary to control at least VR and SNR, but preferably other model characteristics as well. Similar calculations can be made for the Monte Carlo design of Hayakawa and Nagata (2013), which is basically that of Blundell, Bond, and Windmeijer (2001), but choosing  $\sigma_\eta^2, \delta_y$  and  $\delta_x$  different from 1 too.

### 3.4.3 New Simulation Results

We report results for the within group or Least Squares Dummy Variable (LSDV) estimator, DIF, AS and SYS estimators. We report coefficient bias (bias), standard deviation (sd) and root mean squared error (rmse) as well as rejection frequencies (rf) of nominal 5% Wald significance tests and overidentifying restrictions (OIR) tests. The GMM Wald tests use the variance correction of Windmeijer (2005), since it is well known (Arellano and Bond 1991) that two step GMM variance estimators are heavily downward biased. We also apply this finite sample variance correction to the nonlinear moment conditions of Ahn and Schmidt (1995). It can be expected that this leads to an improvement in the estimation of variances, although theoretically that is only the case for linear moment conditions (Windmeijer 2005). The incremental OIR tests are based on either the difference between SYS and AS, or between SYS and DIF.<sup>11</sup>

One problem with existing simulation results is that a comparison across experiments is hampered by the fact that typically more than one model characteristic is changed. Furthermore, it seems that the chosen design parameters often imply rather extreme values for  $VR$ ,  $SNR$ ,  $r_y$  and  $r_x$ . Therefore, we control for these four model characteristics across experiments.

Regarding the error components we specify  $\eta_i$  and  $\varepsilon_{it}$  *i.i.d.*  $N(0, \sigma_\eta^2)$  and  $N(0, \sigma_\varepsilon^2)$  respectively, with  $\sigma_\varepsilon^2 = 1$  and  $\sigma_\eta^2$  determined by (4.15). We consider  $VR = \{3, 100\}$ , and we set  $SNR = 3$ . We have also experimented with  $SNR = 9$  and generally the precision of all GMM estimators improves substantially for sufficiently high values of  $SNR$ . For  $VR = 3$ , we report simulation results for four parameter configurations: (1)  $\alpha = 0.2, \rho = 0.5$ ; (2)  $\alpha = 0.2, \rho = 0.95$ ; (3)  $\alpha = 0.8, \rho = 0.5$ ; (4)  $\alpha = 0.8, \rho = 0.95$ . For  $VR = 100$ , we report results only for (4)  $\alpha = 0.8, \rho = 0.95$ , in order to save space.<sup>12</sup>

We set  $\beta = 1 - \alpha$  across all experiments, so that the long run effect of  $x$  on  $y$  is one. Following Blundell, Bond, and Windmeijer (2001) we fix  $\tau = 0.25$ ,  $\phi_0 = -0.1$  and  $\phi_1 = 0$ . We have also experimented with other types of endogeneity: (1)  $\tau = 0$ , i.e. no correlation between  $\eta_i$  and  $x_{it}$ ; (2)  $\phi_0 = 0, \phi_1 = -0.1$ , i.e. weak exogeneity. The results are qualitatively similar for these cases. Finally, we only report results for  $T = 3$  and  $N = 500$ . For smaller  $N$  larger biases are seen for all GMM estimators. A larger value of  $T$  introduces instrument proliferation issues, as discussed in Section 3.2.

The pattern of the 9 columns within each table is: (1) baseline of cce, i.e.  $r_y = r_x = 0$ ; (2)  $r_y = 0.5$ ; (3)  $r_y = -0.5$ ; (4)  $r_x = 0.5$ ; (5)  $r_x = -0.5$ ; (6)  $r_y = 0.5, r_x = 0.5$ ; (7)  $r_y = 0.5, r_x = -0.5$ ; (8)  $r_y = -0.5, r_x = 0.5$ ; (9)  $r_y = -0.5, r_x = -0.5$ . Hence, columns (2)–(9) investigate all possible combinations of deviations from the cce assumption.

The following observations can be made regarding bias and precision of coefficient estimators:

1. LSDV coefficient bias is negative.
2. Unless it is negligible, DIF GMM coefficient bias is almost always negative.
3. SYS GMM coefficient bias can have either sign, but tends to be positive in most cases.
4. Under cce, coefficient biases for all GMM estimators are larger for  $VR = 100$  showing their lack of invariance to  $\sigma_\eta^2$ .
5. SYS GMM coefficient bias for  $\alpha$  is always larger under deviations from the cce assumption. In a few cases, however, it happens that coefficient bias is smaller for  $\beta$ , most notably in Table 3.5, where both  $y$  and  $x$  are highly persistent, unless  $r_y, r_x$  are both negative.
6. DIF and AS GMM coefficient biases are affected under deviations from the cce assumption, but there is no clear pattern in the simulation results. Hence, benefits for the location of the DIF GMM estimator may or may not occur depending on the particular parameter configuration.
7. AS GMM often performs equally well or better compared with SYS GMM. This somewhat remarkable result appears to hold even under cce. Actually the only

case in which AS GMM is noticeably outperformed by SYS GMM in terms of bias is Table 3.3, column 1.

8. When  $\alpha = 0.8$  (Tables 3.3–5) DIF GMM can have large coefficient bias or large standard deviation or both, indicating a weak instrument problem.
9. In Tables 3.1, 3.2 and 3.3 SYS GMM has often smaller standard deviation than AS GMM, which in turn has smaller dispersion than DIF GMM; this is unless there is moderate persistence, in which case DIF and AS GMM have similar or smaller standard deviation than SYS GMM.
10. Under cce, a weak instrument problem seems present for SYS GMM too when there is strong unobserved heterogeneity (i.e.,  $VR = 100$ .) Although coefficient bias seems limited, in relative terms (i.e., compared with the standard deviation of the estimator) it becomes large. This is consistent with the results in Bun and Windmeijer (2010), who show that also for moderate autoregressive dynamics LEV moment conditions may become less informative when  $VR$  gets large.

Regarding rejection frequencies of Wald statistics the following can be observed:

1. LSDV size distortions are large. For hypothesis testing on  $\alpha$  the actual rejection frequency is always 1.
2. DIF GMM rejection frequencies under the null hypothesis are close to the nominal significance level of 5%. In those cases where there appears to be a size distortion, it is probably caused by the weak instrument problem, as documented above.
3. The same weak instruments problem appears to hold for SYS GMM. Note that even under cce size distortions can be large.
4. The vulnerability of SYS GMM to deviations from cce is obvious. Rejection frequencies under the null hypothesis can become 1.
5. AS GMM rejection frequencies under the null hypothesis are often close to the nominal significance level of 5%, but sometimes size distortions appear.

Finally, regarding OIR test statistics the following observations can be made:

1. The performance of DIF and AS OIR test statistics under the null hypothesis is satisfactory. We didn't examine power, but it can be expected that power is low when persistence is high.
2. The performance of the SYS OIR test statistic under the null hypothesis is satisfactory. Also its power is high in case of moderate persistence (Table 3.1). However, as can be seen from Tables 3.4–5, power is low when there is a lot of persistence in both  $y$  and  $x$ . In between, it depends on the particular deviation from cce.
3. Similar conclusions hold for incremental SYS-DIF and SYS-AS statistics, but they outperform SYS OIR statistic in terms of power.

**Table 3.1 Simulation results for  $\alpha=0.2$ ,  $\rho=0.5$  and  $VR=3$**

	1	2	3	4	5	6	7	8	9
$\sigma_\eta$	0.922	0.879	0.961	0.770	1.067	0.723	1.028	0.814	1.102
$\sigma_\nu$	1.698	1.698	1.698	1.698	1.698	1.698	1.698	1.698	1.698
$r_y$	0.000	0.500	-0.500	0.000	0.000	0.500	0.500	-0.500	-0.500
$r_x$	0.000	0.000	0.000	0.500	-0.500	0.500	-0.500	0.500	-0.500
<b>bias <math>\alpha</math></b>	lsdv	-0.153	-0.132	-0.132	-0.150	-0.150	-0.141	-0.122	-0.121
	dif	-0.007	-0.003	-0.006	-0.004	-0.004	-0.004	-0.003	-0.001
	as	-0.000	-0.001	0.001	-0.000	-0.002	-0.001	-0.002	0.000
	sys	0.003	0.112	0.311	0.226	0.135	0.043	0.204	0.152
<b>bias <math>\beta</math></b>	lsdv	-0.046	-0.045	-0.045	-0.048	-0.048	-0.039	-0.052	-0.052
	dif	0.004	-0.001	0.000	-0.001	0.002	0.001	-0.001	-0.004
	as	-0.003	-0.002	-0.005	-0.004	0.000	-0.002	-0.001	-0.006
	sys	-0.005	-0.012	-0.200	-0.139	-0.014	-0.126	-0.036	0.154
<b>sd <math>\alpha</math></b>	lsdv	0.021	0.019	0.019	0.020	0.020	0.020	0.019	0.018
	dif	0.065	0.037	0.058	0.050	0.044	0.052	0.028	0.028
	as	0.050	0.035	0.038	0.041	0.041	0.046	0.027	0.027
	sys	0.047	0.053	0.062	0.049	0.058	0.054	0.040	0.036
<b>sd <math>\beta</math></b>	lsdv	0.021	0.021	0.021	0.021	0.021	0.021	0.021	0.021
	dif	0.103	0.087	0.092	0.074	0.075	0.095	0.067	0.066
	as	0.093	0.085	0.085	0.072	0.073	0.088	0.067	0.066
	sys	0.085	0.121	0.138	0.110	0.116	0.097	0.111	0.099
<b>rmse <math>\alpha</math></b>	lsdv	0.154	0.134	0.133	0.152	0.152	0.143	0.123	0.122
	dif	0.065	0.037	0.059	0.050	0.044	0.053	0.028	0.028
	as	0.050	0.035	0.038	0.041	0.041	0.046	0.027	0.027
	sys	0.047	0.124	0.317	0.231	0.147	0.069	0.208	0.157
<b>rmse <math>\beta</math></b>	lsdv	0.051	0.049	0.049	0.052	0.052	0.044	0.056	0.056
	dif	0.103	0.087	0.092	0.074	0.075	0.095	0.067	0.066
	as	0.093	0.085	0.086	0.072	0.073	0.088	0.067	0.066
	sys	0.085	0.121	0.243	0.177	0.117	0.160	0.116	0.183
<b>rf <math>\alpha</math></b>	lsdv	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	dif	0.044	0.047	0.045	0.045	0.054	0.044	0.057	0.047
	as	0.042	0.044	0.039	0.044	0.051	0.043	0.060	0.039
	sys	0.048	0.669	1.000	0.997	0.730	0.129	1.000	0.982
<b>rf <math>\beta</math></b>	lsdv	0.599	0.563	0.563	0.630	0.631	0.459	0.696	0.700
	dif	0.044	0.050	0.040	0.040	0.047	0.041	0.051	0.051
	as	0.048	0.061	0.045	0.045	0.053	0.045	0.062	0.055
	sys	0.048	0.118	0.462	0.339	0.120	0.280	0.160	0.443
<b>rf OIR</b>	dif	0.051	0.055	0.049	0.052	0.056	0.052	0.057	0.052
	as	0.045	0.052	0.048	0.055	0.051	0.050	0.052	0.050
	sys	0.047	1.000	1.000	1.000	1.000	0.955	1.000	1.000
	sys-as	0.058	1.000	1.000	1.000	1.000	0.990	1.000	1.000
	sys-dif	0.052	1.000	1.000	1.000	1.000	0.986	1.000	1.000

Note :  $\beta = 1 - \alpha$ ,  $\sigma_\varepsilon^2 = 1$ ,  $T = 3$ ,  $N = 500$ ,  $\phi_0 = -0.1$ ,  $\phi_1 = 0$ ,  $\tau = 0.25$  and  $SNR = 3$ .

**Table 3.2 Simulation results for  $\alpha = 0.2$ ,  $\rho = 0.95$  and  $VR = 3$**

	1	2	3	4	5	6	7	8	9
$\sigma_\eta$	0.268	0.256	0.279	0.130	0.406	0.119	0.394	0.141	0.417
$\sigma_v$	0.552	0.552	0.552	0.552	0.552	0.552	0.552	0.552	0.552
$r_y$	0.000	0.500	-0.500	0.000	0.000	0.500	0.500	-0.500	-0.500
$r_x$	0.000	0.000	0.000	0.500	-0.500	0.500	-0.500	0.500	-0.500
<b>bias <math>\alpha</math></b>	lsdv	-0.333	-0.246	-0.242	-0.269	-0.273	-0.329	-0.158	-0.154
	dif	-0.006	-0.004	-0.002	-0.003	-0.008	-0.003	-0.004	-0.002
	as	0.004	0.001	0.002	0.000	0.002	0.002	-0.000	0.007
	sys	-0.001	0.020	0.020	0.010	0.043	-0.002	0.038	0.005
<b>bias <math>\beta</math></b>	lsdv	-0.221	-0.231	-0.226	-0.251	-0.254	-0.212	-0.273	-0.268
	dif	-0.083	-0.064	-0.075	-0.053	-0.075	-0.044	-0.061	-0.057
	as	-0.019	-0.008	-0.048	-0.046	0.001	-0.026	-0.012	-0.047
	sys	0.007	0.166	0.154	0.063	0.270	-0.008	0.305	0.079
<b>sd <math>\alpha</math></b>	lsdv	0.028	0.026	0.025	0.026	0.027	0.028	0.021	0.020
	dif	0.052	0.037	0.036	0.039	0.046	0.047	0.029	0.026
	as	0.047	0.036	0.034	0.036	0.040	0.045	0.027	0.025
	sys	0.041	0.035	0.035	0.035	0.039	0.042	0.025	0.024
<b>sd <math>\beta</math></b>	lsdv	0.068	0.068	0.068	0.068	0.068	0.067	0.068	0.068
	dif	0.505	0.427	0.470	0.365	0.439	0.350	0.399	0.377
	as	0.287	0.280	0.283	0.262	0.270	0.266	0.267	0.284
	sys	0.105	0.061	0.052	0.048	0.076	0.146	0.071	0.026
<b>rmse <math>\alpha</math></b>	lsdv	0.334	0.247	0.243	0.270	0.274	0.330	0.159	0.156
	dif	0.052	0.037	0.036	0.039	0.047	0.048	0.030	0.026
	as	0.047	0.036	0.034	0.036	0.040	0.045	0.027	0.025
	sys	0.042	0.041	0.040	0.036	0.058	0.042	0.045	0.024
<b>rmse <math>\beta</math></b>	lsdv	0.231	0.241	0.236	0.260	0.263	0.223	0.282	0.277
	dif	0.511	0.432	0.476	0.369	0.445	0.353	0.404	0.381
	as	0.288	0.280	0.287	0.266	0.270	0.268	0.267	0.271
	sys	0.105	0.176	0.163	0.079	0.280	0.147	0.314	0.083
<b>rf <math>\alpha</math></b>	lsdv	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	dif	0.046	0.049	0.041	0.046	0.054	0.049	0.052	0.041
	as	0.047	0.039	0.039	0.043	0.037	0.049	0.046	0.047
	sys	0.056	0.087	0.083	0.057	0.211	0.051	0.359	0.047
<b>rf <math>\beta</math></b>	lsdv	0.914	0.927	0.918	0.961	0.963	0.894	0.979	0.973
	dif	0.041	0.047	0.040	0.051	0.043	0.043	0.046	0.041
	as	0.156	0.146	0.154	0.132	0.140	0.131	0.138	0.136
	sys	0.043	0.819	0.843	0.254	0.973	0.043	0.996	0.856
<b>rf OIR</b>	dif	0.050	0.048	0.052	0.056	0.050	0.058	0.050	0.056
	as	0.069	0.064	0.066	0.070	0.068	0.068	0.065	0.071
	sys	0.048	0.223	0.199	0.075	0.606	0.049	0.579	0.085
	sys-as	0.048	0.332	0.279	0.090	0.741	0.049	0.736	0.102
	sys-dif	0.057	0.300	0.265	0.086	0.718	0.049	0.705	0.100

Note : see Table 3.1.

**Table 3.3 Simulation results for  $\alpha=0.8$ ,  $\rho=0.5$  and  $VR=3$**

	1	2	3	4	5	6	7	8	9
$\sigma_\eta$	0.507	0.383	0.630	0.478	0.536	0.354	0.411	0.601	0.660
$\sigma_\nu$	2.015	2.015	2.015	2.015	2.015	2.015	2.015	2.015	2.015
$r_y$	0.000	0.500	-0.500	0.000	0.000	0.500	0.500	-0.500	-0.500
$r_x$	0.000	0.000	0.000	0.500	-0.500	0.500	-0.500	0.500	-0.500
<b>bias <math>\alpha</math></b>	lsdv	-0.564	-0.536	-0.536	-0.561	-0.561	-0.551	-0.517	-0.517
	dif	-0.049	-0.017	-0.286	-0.042	-0.038	-0.018	-0.017	-0.111
	as	-0.011	-0.012	-0.016	-0.012	-0.014	-0.012	-0.012	-0.011
	sys	-0.002	0.050	0.110	0.111	0.176	0.069	0.152	0.161
<b>bias <math>\beta</math></b>	lsdv	-0.062	-0.060	-0.060	-0.065	-0.064	-0.055	-0.070	-0.070
	dif	-0.010	-0.005	-0.059	-0.011	-0.008	-0.004	-0.007	-0.044
	as	-0.004	-0.005	-0.006	-0.005	-0.004	-0.004	-0.006	-0.008
	sys	-0.002	-0.027	0.014	-0.038	-0.042	-0.046	-0.028	-0.015
<b>sd <math>\alpha</math></b>	lsdv	0.031	0.031	0.031	0.031	0.031	0.031	0.030	0.031
	dif	0.153	0.087	0.383	0.149	0.129	0.097	0.082	0.240
	as	0.110	0.080	0.102	0.109	0.101	0.088	0.077	0.102
	sys	0.056	0.076	0.029	0.023	0.047	0.029	0.027	0.020
<b>sd <math>\beta</math></b>	lsdv	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.016
	dif	0.067	0.064	0.105	0.062	0.061	0.056	0.063	0.106
	as	0.061	0.062	0.060	0.056	0.057	0.055	0.060	0.062
	sys	0.045	0.045	0.046	0.046	0.052	0.047	0.043	0.046
<b>rmse <math>\alpha</math></b>	lsdv	0.565	0.537	0.537	0.562	0.562	0.551	0.518	0.518
	dif	0.160	0.089	0.478	0.155	0.134	0.098	0.084	0.265
	as	0.110	0.081	0.104	0.110	0.102	0.089	0.078	0.102
	sys	0.056	0.091	0.114	0.113	0.182	0.074	0.154	0.163
<b>rmse <math>\beta</math></b>	lsdv	0.064	0.062	0.062	0.067	0.066	0.057	0.072	0.072
	dif	0.068	0.064	0.121	0.063	0.062	0.056	0.063	0.115
	as	0.061	0.062	0.061	0.057	0.057	0.055	0.060	0.062
	sys	0.045	0.052	0.048	0.059	0.067	0.066	0.051	0.048
<b>rf <math>\alpha</math></b>	lsdv	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	dif	0.068	0.058	0.150	0.067	0.064	0.058	0.060	0.081
	as	0.101	0.077	0.128	0.117	0.099	0.082	0.076	0.139
	sys	0.061	0.153	0.940	0.986	0.909	0.661	0.999	1.000
<b>rf <math>\beta</math></b>	lsdv	0.975	0.965	0.965	0.981	0.979	0.932	0.989	0.988
	dif	0.052	0.055	0.088	0.050	0.055	0.049	0.055	0.082
	as	0.065	0.067	0.057	0.061	0.065	0.057	0.067	0.067
	sys	0.044	0.081	0.058	0.144	0.174	0.175	0.114	0.072
<b>rf OIR</b>	dif	0.059	0.052	0.065	0.055	0.056	0.055	0.052	0.059
	as	0.075	0.058	0.082	0.077	0.072	0.063	0.059	0.083
	sys	0.049	0.312	0.060	0.237	0.820	0.211	0.503	0.332
	sys-as	0.050	0.477	0.061	0.345	0.910	0.315	0.670	0.479
	sys-dif	0.050	0.416	0.074	0.305	0.896	0.279	0.622	0.431
									0.883

Note : see Table 3.1.

**Table 3.4 Simulation results for  $\alpha=0.8$ ,  $\rho=0.95$  and  $VR=3$**

	1	2	3	4	5	6	7	8	9
$\sigma_\eta$	0.268	0.200	0.336	0.237	0.299	0.169	0.229	0.304	0.367
$\sigma_\nu$	0.438	0.438	0.438	0.438	0.438	0.438	0.438	0.438	0.438
$r_y$	0.000	0.500	-0.500	0.000	0.000	0.500	0.500	-0.500	-0.500
$r_x$	0.000	0.000	0.000	0.500	-0.500	0.500	-0.500	0.500	-0.500
<b>bias <math>\alpha</math></b>	lsdv	-0.657	-0.620	-0.615	-0.641	-0.645	-0.651	-0.573	-0.565
	dif	-0.056	-0.021	-0.099	-0.033	-0.069	-0.019	-0.026	-0.039
	as	-0.015	-0.010	-0.015	-0.015	-0.016	-0.011	-0.012	-0.016
	sys	-0.005	0.037	0.040	0.031	-0.006	0.000	0.081	0.036
<b>bias <math>\beta</math></b>	lsdv	-0.458	-0.457	-0.448	-0.476	-0.477	-0.442	-0.504	-0.494
	dif	-0.302	-0.131	-0.696	-0.170	-0.310	-0.083	-0.154	-0.290
	as	-0.063	-0.048	-0.086	-0.058	-0.062	-0.044	-0.063	-0.069
	sys	0.005	0.087	0.081	0.025	0.088	-0.013	0.087	0.076
<b>sd <math>\alpha</math></b>	lsdv	0.032	0.032	0.031	0.032	0.032	0.032	0.031	0.032
	dif	0.148	0.090	0.201	0.117	0.158	0.097	0.090	0.113
	as	0.095	0.079	0.088	0.090	0.094	0.087	0.073	0.077
	sys	0.054	0.063	0.039	0.041	0.074	0.056	0.042	0.043
<b>sd <math>\beta</math></b>	lsdv	0.076	0.077	0.076	0.076	0.076	0.076	0.078	0.077
	dif	0.833	0.578	1.424	0.603	0.771	0.457	0.583	0.824
	as	0.344	0.345	0.340	0.329	0.330	0.322	0.339	0.344
	sys	0.131	0.124	0.080	0.110	0.109	0.222	0.116	0.066
<b>rmse <math>\alpha</math></b>	lsdv	0.657	0.621	0.616	0.642	0.646	0.652	0.574	0.566
	dif	0.158	0.092	0.224	0.122	0.173	0.098	0.094	0.119
	as	0.096	0.079	0.090	0.091	0.095	0.087	0.074	0.079
	sys	0.054	0.073	0.056	0.051	0.074	0.056	0.091	0.056
<b>rmse <math>\beta</math></b>	lsdv	0.464	0.464	0.454	0.482	0.483	0.449	0.509	0.500
	dif	0.886	0.593	1.585	0.627	0.832	0.465	0.603	0.874
	as	0.350	0.349	0.351	0.334	0.335	0.324	0.345	0.351
	sys	0.132	0.152	0.114	0.113	0.140	0.223	0.145	0.101
<b>rf <math>\alpha</math></b>	lsdv	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	dif	0.065	0.055	0.062	0.061	0.088	0.057	0.068	0.055
	as	0.067	0.047	0.074	0.079	0.066	0.051	0.069	0.100
	sys	0.048	0.089	0.196	0.127	0.058	0.047	0.534	0.144
<b>rf <math>\beta</math></b>	lsdv	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	dif	0.063	0.055	0.057	0.059	0.079	0.052	0.061	0.048
	as	0.147	0.147	0.162	0.100	0.164	0.108	0.164	0.110
	sys	0.043	0.088	0.162	0.057	0.115	0.059	0.075	0.227
<b>rf OIR</b>	dif	0.052	0.057	0.042	0.053	0.059	0.053	0.053	0.049
	as	0.102	0.083	0.107	0.103	0.108	0.083	0.082	0.098
	sys	0.048	0.089	0.055	0.050	0.122	0.050	0.133	0.058
	sys-as	0.041	0.105	0.050	0.053	0.143	0.042	0.191	0.058
	sys-dif	0.062	0.108	0.090	0.063	0.159	0.051	0.174	0.075
									0.099

Note : see Table 3.1.

**Table 3.5 Simulation results for  $\alpha=0.8$ ,  $\rho=0.95$  and  $VR=100$**

	1	2	3	4	5	6	7	8	9
$\sigma_\eta$	1.549	1.480	1.617	1.518	1.579	1.450	1.511	1.586	1.647
$\sigma_\nu$	0.438	0.438	0.438	0.438	0.438	0.438	0.438	0.438	0.438
$r_y$	0.000	0.500	-0.500	0.000	0.000	0.500	0.500	-0.500	-0.500
$r_x$	0.000	0.000	0.000	0.500	-0.500	0.500	-0.500	0.500	-0.500
<b>bias <math>\alpha</math></b>	lsdv	-0.655	-0.633	-0.600	-0.631	-0.652	-0.655	-0.591	-0.545
	dif	-0.120	-0.145	-0.018	-0.161	-0.126	-0.093	-0.179	-0.051
	as	0.006	-0.006	-0.002	-0.014	0.014	-0.001	-0.018	-0.013
	sys	0.015	0.067	0.043	0.051	0.053	0.035	0.067	0.047
<b>bias <math>\beta</math></b>	lsdv	-0.450	-0.474	-0.418	-0.464	-0.474	-0.455	-0.523	-0.462
	dif	-0.591	-0.812	-0.142	-0.997	-0.447	-0.501	-0.890	-0.517
	as	0.057	-0.003	0.031	0.013	0.060	0.048	-0.077	-0.030
	sys	0.132	0.093	0.116	0.079	0.121	0.139	0.092	0.103
<b>sd <math>\alpha</math></b>	lsdv	0.032	0.032	0.031	0.032	0.032	0.032	0.031	0.031
	dif	0.235	0.235	0.102	0.245	0.246	0.192	0.252	0.142
	as	0.102	0.143	0.083	0.109	0.117	0.114	0.169	0.080
	sys	0.049	0.047	0.036	0.039	0.061	0.051	0.040	0.037
<b>sd <math>\beta</math></b>	lsdv	0.076	0.076	0.076	0.076	0.075	0.076	0.077	0.078
	dif	1.190	1.357	0.735	1.491	0.901	1.034	1.289	1.285
	as	0.330	0.614	0.309	0.377	0.342	0.434	0.702	0.375
	sys	0.081	0.085	0.072	0.082	0.106	0.145	0.085	0.069
<b>rmse <math>\alpha</math></b>	lsdv	0.655	0.634	0.601	0.632	0.653	0.656	0.592	0.546
	dif	0.264	0.276	0.103	0.294	0.277	0.214	0.309	0.151
	as	0.102	0.143	0.083	0.109	0.117	0.114	0.170	0.081
	sys	0.051	0.082	0.056	0.064	0.081	0.062	0.078	0.060
<b>rmse <math>\beta</math></b>	lsdv	0.457	0.480	0.425	0.471	0.480	0.461	0.528	0.469
	dif	1.329	1.581	0.749	1.794	1.006	1.149	1.566	1.385
	as	0.335	0.614	0.310	0.377	0.347	0.436	0.707	0.376
	sys	0.155	0.126	0.136	0.114	0.160	0.201	0.126	0.124
<b>rf <math>\alpha</math></b>	lsdv	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	dif	0.083	0.084	0.044	0.081	0.105	0.076	0.111	0.039
	as	0.047	0.054	0.053	0.069	0.044	0.034	0.093	0.116
	sys	0.087	0.305	0.256	0.282	0.145	0.128	0.418	0.283
<b>rf OIR</b>	dif	0.047	0.044	0.052	0.037	0.063	0.051	0.039	0.044
	as	0.073	0.068	0.073	0.071	0.073	0.068	0.071	0.060
	sys	0.053	0.076	0.056	0.053	0.109	0.060	0.085	0.062
	sys-as	0.045	0.085	0.059	0.062	0.137	0.064	0.105	0.079
	sys-dif	0.073	0.111	0.072	0.093	0.143	0.077	0.124	0.091

Note : see Table 3.1.

- 
4. No clear ranking exists between SYS-AS and SYS-DIF statistics although the former has often a slightly higher rejection frequency under the alternative hypothesis.
  5. Perhaps surprisingly, sometimes OIR SYS, SYS-AS and SYS-DIF tests have complete lack of power against deviations from cce. Sometimes this is even the case when coefficient bias in the SYS GMM estimator is relatively large, e.g. Table 3.5, column 9, or Table 3.2, column 3.

### 3.5 CONCLUDING REMARKS

---

In this chapter we have reviewed the literature on dynamic panel data models estimated by GMM. We have focused on the analysis of GMM estimators in dynamic models with additional endogenous regressors. We have discussed in detail the assumptions underlying the validity of, especially, the system GMM estimator. Furthermore, we have embarked on the consequences of violation of mean stationarity for several GMM estimators. In cases where the constant correlated effects assumption is violated, individual-specific unobserved heterogeneity is only partially removed by taking first differences. Obviously, lagged differenced instruments for the model in levels are then not exogenous anymore, therefore invalidating the system GMM estimator. Additionally, the relevance of the lagged level instruments for the first-differenced model changes in a nontrivial manner. Apart from mean stationarity we have discussed briefly a number of other practical issues when applying GMM inference methods, for example how to determine the optimal number of moment conditions.

Our simulation results indicate that no universal ranking exists among first-differenced (DIF), nonlinear (AS), and system (SYS) GMM estimators. Some general observations can be made. First, DIF GMM has low precision and coefficient bias, especially when the series are persistent. Second, SYS GMM is vulnerable to nuisance parameters, and its performance deteriorates rapidly under deviations from cce. Even when absolute coefficient bias seems small, large size distortion can still occur. Third, the AS GMM estimator performs quite satisfactory in most experiments. It has higher precision than DIF GMM and only moderate coefficient bias and size distortion. Compared with SYS GMM, however, its root mean squared error is relatively large when the series are persistent. Fourth, in testing for cce all OIR tests appear to lack power in case of high persistence.

Summarizing, GMM estimators for dynamic panel data models can be vulnerable to important nuisance parameters and weak identification issues. Until recently, system GMM has been considered to be the solution to the first-differenced GMM estimator in case of persistent panel data. However, its additional restriction on the initial conditions has been criticized for being unrealistic precisely in case of persistent panel data. Additionally, tests for cce lack power when having persistent panel data and/or an

abundance of moment conditions. This may lead to acceptance of the levels moment conditions when this is not appropriate. But even in case of mean stationarity inference based on system GMM may be inaccurate. A straightforward advise for practitioners regarding which method to prefer in small samples does not emerge, but the non-linear AS GMM estimator seems a relatively safe choice. It is robust to deviations from cce, and more efficient than first-differenced linear GMM.

## ACKNOWLEDGMENTS

---

We would like to thank Artūras Juodis, Andrew Pua, two anonymous referees and the Editor for helpful comments and suggestions. The first author wants to thank Monash University for hospitality while working on this chapter. The research of the first author has been funded by the NWO Vernieuwingsimpuls research grant ‘Causal Inference with Panel Data’.

## NOTES

---

1. Time-specific effects can also be included explicitly or controlled for by cross-sectional demeaning of the data prior to estimation. We will discuss an example of a process with time-specific effects later.
2. This is not very restrictive, because in autoregressive models any nonzero correlation between individual effects and idiosyncratic errors tends to vanish over time (Arellano 2003a, p. 82).
3. Assumption (2.6) is often labeled the ‘mean stationarity’ assumption. Some authors (e.g., Kiviet, 2007) prefer to label it as ‘effect stationarity’, because it is an expectation conditional on the individual specific effect  $\eta_i$ . In the next section we use the term ‘constant correlated effects’ to describe this assumption. We believe this is more precise because the additional moment conditions do not require mean stationarity, as it will become clear.
4. A similar result holds for the DIF moment condition.
5. See also Hsiao (2003, Ch. 4) for an excellent discussion.
6. A deviation from mean stationarity may occur also as the result of a finite number of start-up periods.
7. Notice that when  $\delta_y = 1$ , the expression above depends only on  $\sigma_\varepsilon^2$ , for given  $\alpha$ , which confirms that in this case  $\Delta y_{i1}$  is free from  $\eta_i$ . In this case we have the earlier result (2.19).
8. Notice that setting  $r_x$  equal to a fixed value, say  $r_x = c$ , also captures the case where the  $x$  process is mean-stationary ( $\delta_x = 1$ ) for a proportion of individuals only. For example, if the proportion of individuals that satisfy  $\delta_x = 1$  is .5,  $r_x = 0$  for those individuals and  $r_x = 2c$  for the remaining ones, provided that  $|c| \leq 0.5$ .
9. We have

$$c_v^2 = \frac{(1 + \alpha\rho)\beta^2}{(1 - \rho^2)(1 - \alpha^2)(1 - \alpha\rho)},$$

and

$$\sigma^2_{\varepsilon} = \frac{(1+\alpha\rho)}{(1-\rho^2)(1-\alpha^2)(1-\alpha\rho)} (1+\beta\phi_0)^2 + (\beta\phi_1 - \rho)^2 + 2\frac{(\alpha+\rho)}{1+\alpha\rho} (\beta\phi_1 - \rho) (1+\beta\phi_0).$$

Thus, similarly to the process for  $x$ , we have assumed that the idiosyncratic component in  $y$ ,  $\varpi_{it}$ , is covariance-stationary.

10. Juodis (2013) provides simulation results under mean nonstationarity for panel VAR models.
11. For LSDV, DIF, and SYS estimation, we use the DPD for Ox package (Doornik et al. 2006). AS estimation is based on our own Ox code.
12. The results for the remaining configurations are available on the author's website.

## REFERENCES

---

- Ahn, S.C. and P. Schmidt (1995). Efficient estimation of models for dynamic panel data. *Journal of Econometrics* 68, 5–27.
- Ahn, S.C., Y.H. Lee, and P. Schmidt (2013). Panel data models with multiple time-varying individual effects. *Journal of Econometrics* 174, 1–14.
- Alvarez, J. and M. Arellano (2003). The time series and cross-sectional asymptotics of dynamic panel data estimators. *Econometrica* 71, 1121–1159.
- 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. (2003a). Panel Data Econometrics. Oxford: *Oxford University Press*.
- Arellano, M. (2003b). Modeling optimal instrumental variables for dynamic panel data models. Working Paper 0310, Centro de Estudios Monetarios y Financieros, Madrid.
- Arellano, M. and S. Bond (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* 58, 277–298.
- Arellano, M. and O. Bover (1995). Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics* 68, 29–51.
- Arellano, M. and B. Honoré (2001). Panel data models: some recent developments. In J. Heckman and E. Leamer (eds.), *Handbook of Econometrics*, Volume 5, Chapter 3, pages 3229–3296. Elsevier: North-Holland.
- Baltagi, B.H. (2013). Econometric analysis of panel data (5th edition). John Wiley, Chichester.
- Bekker, P.A. (1994). Alternative approximations to the distributions of Instrumental Variable estimators. *Econometrica* 62, 657–681.
- Besley, T. and A. Case (2000). Unnatural experiments? Estimating the incidence of endogenous policies. *The Economic Journal* 110, F672–F694.
- Bhargava, A. and J.D. Sargan (1983). Estimating dynamic random effects models from panel data covering short time periods. *Econometrica* 51, 1635–1659.
- Binder, M., C. Hsiao, and M. H. Pesaran (2005). Estimation and inference in short panel vector autoregressions with unit roots and cointegration. *Econometric Theory* 21, 795–837.
- Blundell, R. and S. Bond (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* 87, 115–143.

- Blundell, R. and S. Bond (2000). GMM Estimation with persistent panel data: an application to production functions. *Econometric Reviews* 19, 321–340.
- Blundell, R., S. Bond, and F. Windmeijer (2001). Estimation in dynamic panel data models: Improving on the performance of the standard GMM estimator. In, B.H. Baltagi, T.B. Fomby, and R. Carter Hill (eds.), *Nonstationary Panels, Panel Cointegration, and Dynamic Panels*. Advances in Econometrics, Volume 15, Emerald Group Publishing Limited, 53–91.
- Bond, S. and F. Windmeijer (2005). Reliable inference for GMM estimators? Finite sample properties of alternative test procedures in linear panel data models. *Econometric Reviews* 24, 1–37.
- Bond, S., A. Hoefller, and J. Temple (2001). GMM estimation of empirical growth models. Economics group working paper 2001-W21, University of Oxford.
- Bond, S., C. Naujas and F. Windmeijer (2005). Unit roots: identification and testing in micro panels. Cemmap Working Paper 07/05, London.
- 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, 443–450.
- Bowsher, C. G. (2002). On testing overidentifying restrictions in dynamic panel data models. *Economics Letters* 77, 211–220.
- Bun, M.J.G. and M.A. Carree (2005). Bias-corrected estimation in dynamic panel data models. *Journal of Business & Economic Statistics* 23, 200–210.
- Bun, M.J.G. and J.F. Kiviet (2006). The effects of dynamic feedbacks on LS and MM estimator accuracy in panel data models. *Journal of Econometrics* 132, 409–444.
- Bun, M.J.G. and F. Kleibergen (2013). Identification and inference in moments based analysis of linear dynamic panel data models. UvA-Econometrics Discussion Paper 2013/07, University of Amsterdam.
- Bun, M.J.G. and F. Windmeijer (2010). The weak instrument problem of the system GMM estimator in dynamic panel data models. *Econometrics Journal* 13, 95–126.
- Dhaene, G. and K. Jochmans (2012). An adjusted profile likelihood for non-stationary panel data models with fixed effects. Working Paper, KU Leuven.
- Doornik, J.A., M. Arellano, and S. Bond (2006). Panel data estimation using DPD for Ox. Mimeo, University of Oxford.
- Everaert, G. (2013). Orthogonal to backward mean transformation for dynamic panel data models. *Econometrics Journal* 16, 179–221.
- Hahn, J., J. Hausman, and G. Kuersteiner (2007). Long difference instrumental variables estimation for dynamic panel models with fixed effects. *Journal of Econometrics* 140, 574–617.
- Han, C. and P.C.B. Phillips (2006). GMM with many moment conditions. *Econometrica* 74, 147–192.
- Han, C. and P.C.B. Phillips (2010). GMM estimation for dynamic panels with fixed effects and strong instruments at unity. *Econometric Theory* 26, 119–151.
- Han, C., P.C.B. Phillips, and D. Sul (2014). X-Differencing and dynamic panel model estimation. *Econometric Theory* 30, 201–251.
- Hansen, L.P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* 50, 1029–1054.
- Hause, J.C. (1980). The fine structure of earnings and the on-the-job training hypothesis. *Econometrica* 48, 1013–1029.

- Hayakawa, K. (2009). On the effect of mean-nonstationarity in dynamic panel data models. *Journal of Econometrics* 153, 133–135.
- Hayakawa, K. and S. Nagata (2013). On the Behavior of the GMM Estimator in Persistent Dynamic Panel Data Models with Unrestricted Initial Conditions. Working Paper, Hiroshima University.
- Hayakawa, K. and M.H. Pesaran (2014). Robust standard errors in transformed likelihood estimation of dynamic panel data models with cross-sectional heteroskedasticity. Working Paper, Hiroshima University.
- Holtz-Eakin, D., W. Newey, and H.S. Rosen (1988). Estimating vector autoregressions with panel data. *Econometrica*, 56, 1371–1395.
- Hsiao, C. (2003). Analysis of panel data. Cambridge: Cambridge University Press.
- Hsiao, C., M.H. Pesaran, and A.K. Tahmisioglu (2002). Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods. *Journal of Econometrics* 109, 107–150.
- Juodis, A. (2013). First difference transformation in panel VAR models: robustness, estimation and inference. UvA-Econometrics Discussion Paper 2013/06, University of Amsterdam.
- Kiviet, J.F. (1995). On bias, inconsistency and efficiency of various estimators in dynamic panel data models. *Journal of Econometrics* 68, 53–78.
- Kiviet, J.F. (2007). Judging contending estimators by simulation: tournaments in dynamic panel data models. In *The Refinement of Econometric Estimation and Test Procedures* (eds.: G.D.A. Phillips and E. Tzavalis), 282–318. Cambridge: Cambridge University Press.
- Kiviet, J.F. (2012). Monte Carlo simulation for Econometricians. Foundations and Trends® in Econometrics Vol 5, Nos. 1–2, 1–181. Now Publishers, Boston-Delft.
- Kiviet, J.F., M. Pleus, and R. Poldermans (2013). Accuracy and efficiency of various GMM inference techniques in dynamic micro panel data models. Mimeo, University of Amsterdam.
- Kleibergen, F. (2005). Testing parameters in GMM without assuming that they are identified. *Econometrica* 73, 1103–1124.
- Koenker, R. and J.A.F. Machado (1999). GMM inference when the number of moment conditions is large. *Journal of Econometrics* 93, 327–344.
- Kruiniger, H. (2009). GMM estimation and inference in dynamic panel data models with persistent data. *Econometric Theory* 25, 1348–1391.
- Lancaster, T. (2002). Orthogonal parameters and panel data. *Review of Economic Studies* 69, 647–666.
- Mátyás, L. and P. Sevestre (2008). The Econometrics of panel data. Springer: Berlin Heidelberg.
- Newey, W.K. and K.D. West (1987). Hypothesis testing with efficient method of moments estimation. *International Economic Review* 28, 777–787.
- Neyman, J. and E.L. Scott (1948). Consistent estimates based on partially consistent observations. *Econometrica* 16, 1–32.
- Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica*, 49, 1417–1426.
- Robertson, D. and V. Sarafidis (2013). IV Estimation of Panels with Factor Residuals. Cambridge Working Papers in Economics No. 1321, University of Cambridge.
- Roodman, D. (2009). A note on the theme of too many instruments. *Oxford Bulletin of Economics and Statistics* 71, 135–158.

- Sarafidis V., and T. Wansbeek (2012) Cross-section dependence in panel data analysis. *Econometric Reviews* 31, 483–531.
- Sargan, J.D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica* 26, 393–415.
- Staiger, D. and J.H. Stock (1997). Instrumental variables regression with weak instruments. *Econometrica* 65, 557–586.
- Stock, J.H. and J.H. Wright (2000). GMM with weak identification. *Econometrica* 68, 1055–1096.
- Stock, J.H., J.H. Wright, and M. Yogo (2002). A survey of weak instruments and weak identification in Generalized Method of Moments, *Journal of Business & Economic Statistics*, 20, 518–529.
- Wansbeek, T.J. and T. Knaap (1999). Estimating a dynamic panel data model with heterogeneous trends. *Annales d'Economie et de Statistique* 55–56, 331–350.
- Windmeijer, F. (2005). A finite sample correction for the variance of linear efficient two step GMM estimators. *Journal of Econometrics* 126, 25–51.
- Ziliak, J.P. (1997). Efficient estimation with panel data when instruments are predetermined: An empirical comparison of moment-condition estimators. *Journal of Business & Economic Statistics* 15, 419–431.

## CHAPTER 4

---

# INCIDENTAL PARAMETERS AND DYNAMIC PANEL MODELING

---

HYUNGSIK ROGER MOON, BENOIT PERRON, AND  
PETER C.B. PHILLIPS

### 4.1 INTRODUCTION

---

PANEL data offers great opportunities for empirical research throughout the social and business sciences, as such data has done for many decades in longitudinal medical studies. In economics and business, just as in medicine, there is often intense interest in the effects of policy measures or treatments on individual consumer and firm behavior over time. The pooling of data records across wide panels of individuals has the potential to deliver substantial econometric power in estimation through cross-section averaging to sharpen estimates of common response patterns.

With these great opportunities for studying individual behavior come many challenges. The chapter focuses on one of these challenges—the role and effects of incidental parameters that capture the idiosyncratic features of individual entities within a panel. Adding a new individual to a panel brings new idiosyncratic elements to be explained in the data just as it also brings observations that enhance the power of cross-section averaging for the common elements of behavior. Exploring the effects of such additions is the subject of this chapter.

The problem of incidental parameters in statistical inference was first pointed out in a classic article by Neyman and Scott (1948).<sup>1</sup> According to their characterization, an incidental parameter only figures in a finite dimensional probability law—thereby involving only a finite number of observations and, in consequence, rendering the corresponding maximum likelihood estimator inconsistent. Interestingly, in its attempt to deliver the best possible estimates of all of the parameters in a model, including

incidental parameters that are specific to a single cross-section observation, maximum likelihood estimates of other parameters in the model are inevitably affected and may be inconsistent.

In the context of panel data, the incidental parameter problem typically arises from the presence of individual-specific parameters. These may relate to individual consumer, firm, or country fixed intercept (or mean) effects in a panel. They may also involve incidental trends that are specific to each individual in the sample. The challenges presented by incidental parameters are particularly acute in dynamic panels where behavioral effects over time are being measured in conjunction with individual effects. Nickell (1981) discovered that maximum likelihood estimation of such models leads to inconsistent estimates of the parameters that govern the dynamics. The problems are even more acute in nonstationary panels and panels with incidental trends. Much methodological research on dynamic panel modeling has been devoted to understanding these problems and to developing econometric techniques that address them.

A secondary but no less important issue arises in matters of dynamic model specification. Curiously, some standard model selection procedures such as the BIC information criterion are inconsistent in the presence of incidental parameters. The problem was first identified in a simple example by Stone (1979) and later shown to apply in all dynamic panel models with fixed effects. Model choice is a fundamental aspect of good empirical model building and much applied econometric work relies heavily on standard selection methods such as BIC. The failure of these methods in the context of dynamic panels therefore is a major obstacle to good empirical practice. Approaches to overcome the inconsistency of BIC and related Bayesian procedures have been developed only very recently. Typically one needs to use a special prior in the conventional approach (see Berger, Ghosh, and Mukhopadhyay 2003). Or in a Kullback Leibler approach, one can increase the penalty. Lee (2012) has some discussion of these possibilities in econometrics. A wider literature is now emerging—see Casella and Moreno (2006, 2009) and Moreno, Girón, and Casella (2010). Recent work by Han, Phillips, and Sul (2015) has shown that BIC is also inconsistent as an order estimator in dynamic panels even when no fixed effects are present. So the issues presented by model choice in dynamic panels are wider than those involving incidental parameters alone.

The chapter will discuss these issues in the context of a prototype model under different assumptions on the size of the panel and the properties of the variables in the panel. The prototype of central focus is a simple linear dynamic panel regression with various forms of unobserved characteristics. The model includes only a lagged dependent variable as a regressor but it is the kernel of all dynamic panel models. As it stands, the prototype model is too simplistic for empirical applications where it is often necessary to control for the effects of additional regressors. But this simple case helps us to understand and illustrate the nature of incidental parameter problems in dynamic panel regressions under various fixed effect assumptions. Some of the original papers that are cited in each section analyze an extended

framework with extra regressors, and readers are encouraged to refer to these as needed.

The prototype model is as follows. Suppose a double indexed random variable  $y_{it}$  is observed over  $N$  cross-section individuals  $i = 1, \dots, N$  and  $T$  time periods  $t = 1, \dots, T$ . We consider two forms of dynamic specification. The first is the component model where the observed data comprise an individual and time effect plus a disturbance that follows an autoregressive model:

$$y_{it} = \delta(\lambda_i, f_t) + \varepsilon_{it} \quad (4.1)$$

$$\varepsilon_{it} = \rho \varepsilon_{it-1} + u_{it}. \quad (4.2)$$

The second is an augmented regression model form:

$$y_{it} = \rho y_{it-1} + \alpha(\lambda_i, f_t) + u_{it}, \quad (4.3)$$

These two formulations are related when  $\alpha(\lambda_i, f_t) = \delta(\lambda_i, f_t) - \rho \delta(\lambda_i, f_{t-1})$  and the error dynamics in (4.2) are absorbed into (4.1) giving (4.3). The distinction is particularly relevant when the data are nonstationary.

In models (4.1) and (4.3),  $\lambda_i$  and  $f_t$  represent individual specific effects and time specific effects, respectively. In this chapter, we treat  $\lambda_i$  and  $f_t$  as fixed, so that they can be arbitrarily correlated with the initial condition and are unknown parameters that need to be estimated. While these individual and time specific parameters are relevant and important, it is usually the parameter  $\rho$  that governs the behavioral dynamics that is of primary interest in practical work. This chapter therefore focuses on issues of statistical estimation and inference concerning the common behavioral parameter  $\rho$ .

## 4.2 FORMULATIONS AND PROBLEMS OF INTEREST

---

Several forms of specification for fixed effects have been used in practical work. In the augmented regression model (4.3), the most common form is a simple additive individual effect with  $\alpha(\lambda_i, f_t) = \lambda_i$ . This specification is typically used to model unobservable differences in characteristics that do not vary over time but that would invalidate results if they were not taken into account in the regression. For example, intrinsic ability in a wage equation is an unobserved individual characteristic that materially affects the wage of an individual. Inclusion of such individual effects in regression helps to model situations where individuals with the same observable characteristics have different outcomes (here, wages) for reasons that the econometrician cannot observe. In such cases, the

individual effect serves the role of an individual specific dummy variable in the regression.

Other common specifications of fixed effects in the augmented regression are combined individual-specific and time-specific effects, as in the additive form  $\alpha(\lambda_i, f_t) = \lambda_i + f_t$ , and interactive fixed effects, as in the multiplicative form  $\alpha(\lambda_i, f_t) = \lambda_i f_t$  (or  $\lambda'_i f_t$  if  $\lambda_i$  and  $f_t$  are vectors). The time-specific variable  $f_t$  may represent nonstationarity (or time evolution) in the time series of the panel  $y_{it}$ , or it may represent a common shock that provides a source of cross-sectional dependence in the panel  $y_{it}$ . For example, all individual consumers in a panel may face a common interest rate shock or individual firms may all be subjected to a common exchange rate shock.

It is also natural to allow for the possibility of time trends in the variables of a panel that might be captured through a deterministic trend. Such effects can be incorporated in a component model of the form  $\delta(\lambda_i, f_t) = \lambda_{i0} + \lambda_{i1} t$ , so that each unit in the panel embodies a linear trend with its own unit-specific slope. This formulation is called an incidental trend effect. Such effects commonly appear in panel models of cross country economic growth.

The impact of incidental parameters on dynamic panel estimation was studied by Nickell (1981) who considered a panel autoregression of order one with common coefficient  $\rho$  in the presence of individual-specific intercepts  $\alpha(\lambda_i, f_t) = \lambda_i$ . Analytical expressions showed that the pooled least squares (OLS) estimator or Gaussian MLE is asymptotically (as the cross-section dimension  $N \rightarrow \infty$ ) biased downward (if  $\rho > 0$ ), and that the bias decreases as the time series sample size  $T$  increases. Thus, using OLS or MLE results in inconsistent estimates of  $\rho$  when  $N \rightarrow \infty$  whenever there is a finite time series sample for each cross-sectional unit. The phenomenon arises as a consequence of having only a finite number of observations from which to estimate the individual-specific parameter  $\lambda_i$ , which contaminates the estimation of  $\rho$ . In effect, as intimated in the Introduction, maximum likelihood in attempting to get the best estimates of all the parameters in the model (including the individual effects) fails to achieve consistent estimation even for the common parameter  $\rho$ .

The following sections consider several instances of incidental parameter problems in dynamic panels, as well as various solutions that have been proposed to circumvent them. The remainder of the chapter is divided into four parts. The next section deals with the estimation of the panel autoregressive coefficient  $\rho$ , while Section 4.4 is concerned with inference about  $\rho$  and tests of specific null hypotheses such as  $H_0 : \rho = \rho_0$ , i.e., that  $\rho$  equals a certain constant value of interest such as zero or unity. Much attention has been given to the unit root case where the hypothesis of interest is that members of the panel follow dynamic paths described by random walks in which  $\rho = 1$ . We discuss recent work on nonlinear panels and on model selection in dynamic panels in Sections 4.5 and 4.6. Throughout the chapter we consider large  $N$  asymptotics and cases where  $T$  is finite and  $T \rightarrow \infty$ . Only Classical methods are discussed, although Bayesian principles underlie some of the model selection criteria.<sup>2</sup>

## 4.3 ESTIMATING $\rho$ WITH INCIDENTAL PARAMETERS

---

### 4.3.1 Individual Fixed Effects: $\alpha(\lambda_i, f_t) = \lambda_i$

We consider the dynamic panel regression model (4.3) with time invariant additively separable fixed effects,  $\alpha(\lambda_i, f_t) = \lambda_i$  and  $|\rho| < 1$ :

$$y_{it} = \rho y_{it-1} + \lambda_i + u_{it}, \quad (4.4)$$

where  $y_{i0}$  is the initial condition and the error term  $u_{it}$  is uncorrelated with  $\{y_{it-s} : s \geq 1\}$ . The individual effect  $\lambda_i$  in model (4.4) is fixed in the sense that it can be arbitrarily correlated with the initial condition  $y_{i0}$  and we treat these individual effects as parameters to be estimated. The parameter of interest is the common dynamic coefficient  $\rho$ .

The Gaussian quasi-maximum likelihood estimator (QMLE) of  $\rho$  is the maximum likelihood estimator when  $u_{it} \sim iidN(0, \sigma^2)$  and the  $\lambda_i$  are nuisance parameters. Nickell (1981) showed that the QMLE of  $\rho$  is inconsistent as  $N \rightarrow \infty$  with  $T$  fixed and finite, due to the presence of the incidental parameters  $\lambda_i$  in the regression (4.4).

#### 4.3.1.1 Explaining the Nickell Bias

To explain the source of the asymptotic bias, it is useful to simplify the algebra before developing the large  $N$  asymptotics in detail. The Gaussian QMLE of  $\rho$  is equivalent to pooled least squares on the panel and has the form

$$\hat{\rho} = \left( \sum_{i=1}^N \sum_{t=1}^T (y_{it-1} - \bar{y}_{i,-1})^2 \right)^{-1} \sum_{i=1}^N \sum_{t=1}^T (y_{it-1} - \bar{y}_{i,-1})(y_{it} - \bar{y}_i), \quad (4.5)$$

where  $\bar{y}_i = T^{-1} \sum_{t=1}^T y_{it}$  and  $\bar{y}_{i,-1} = T^{-1} \sum_{t=1}^T y_{it-1}$ . In view of the demeaning transformation within the panel for each individual  $i$ , the estimator is also called the within estimator (or fixed effect or least squares dummy variable estimator) in the literature. In taking the time series mean from the data, the within transformation eliminates the fixed effects  $\lambda_i$  because

$$y_{it} - \bar{y}_i = \rho(y_{it-1} - \bar{y}_{i,-1}) + u_{it} - \bar{u}_i. \quad (4.6)$$

When  $|\rho| < 1$ , the solution of  $y_{it}$  in (4.4) is

$$y_{it} = \frac{\lambda_i}{1-\rho} + y_{it}^0, \quad \text{where } y_{it}^0 = \rho y_{it-1}^0 + u_{it} = \sum_{j=0}^{\infty} \rho^j u_{it-j}.$$

It follows that  $y_{it} - \bar{y}_i = y_{it}^0 - \bar{y}_i^0$  and  $y_{it-1} - \bar{y}_{i,-1} = y_{it-1}^0 - \bar{y}_{i,-1}^0$ . The estimation error in  $\hat{\rho}$  therefore has the following form

$$\begin{aligned}\hat{\rho} - \rho &= \left( \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T (y_{it-1}^0 - \bar{y}_{i,-1}^0)^2 \right)^{-1} \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T (y_{it-1}^0 - \bar{y}_{i,-1}^0) (u_{it} - \bar{u}_i) \\ &= \left( \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T (y_{it-1}^0 - \bar{y}_{i,-1}^0)^2 \right)^{-1} \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T (y_{it-1}^0 - \bar{y}_{i,-1}^0) u_{it}. \quad (4.7)\end{aligned}$$

Since  $\sum_{t=1}^T (y_{it-1}^0 - \bar{y}_{i,-1}^0)^2$  is *iid* over  $i$ , the ergodic theorem implies that

$$\begin{aligned}\frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T (y_{it-1}^0 - \bar{y}_{i,-1}^0)^2 &\xrightarrow{\text{a.s.}} \mathbb{E} \left\{ \sum_{t=1}^T (y_{it-1}^0 - \bar{y}_{i,-1}^0)^2 \right\} = T \mathbb{E} (y_{it-1}^0 - \bar{y}_{i,-1}^0)^2 \\ &= T \mathbb{E} (y_{it-1}^0)^2 - T \mathbb{E} (\bar{y}_{i,-1}^0)^2 \\ &= T \frac{\sigma^2}{1 - \rho^2} - \frac{1}{T} \sum_{t,s=1}^T \mathbb{E} (y_{it} y_{is}) \\ &= T \frac{\sigma^2}{1 - \rho^2} - \frac{1}{T} \sum_{t,s=1}^T \rho^{|t-s|} \frac{\sigma^2}{1 - \rho^2} \\ &= T \frac{\sigma^2}{1 - \rho^2} - \frac{1}{T} \sum_{j=-T+1}^{T-1} \rho^{|j|} \left[ \sum_{t,s=1}^T \mathbf{1}\{t-s=j\} \right] \frac{\sigma^2}{1 - \rho^2} \\ &= T \frac{\sigma^2}{1 - \rho^2} - \frac{1}{T} \sum_{j=-T+1}^{T-1} \rho^{|j|} [T - |j|] \frac{\sigma^2}{1 - \rho^2} \\ &= T \frac{\sigma^2}{1 - \rho^2} + O(1), \text{ because } |\rho| < 1 \quad (4.8)\end{aligned}$$

and

$$\begin{aligned}\frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T (y_{it-1}^0 - \bar{y}_{i,-1}^0) u_{it} &\xrightarrow{\text{a.s.}} \mathbb{E} \left\{ \sum_{t=1}^T (y_{it-1}^0 - \bar{y}_{i,-1}^0) u_{it} \right\} \\ &= \sum_{t=1}^T \mathbb{E} (y_{it-1}^0 u_{it} - \bar{y}_{i,-1}^0 u_{it}) \\ &= - \sum_{t=1}^T \mathbb{E} (\bar{y}_{i,-1}^0 u_{it}), \text{ because } \mathbb{E} y_{it-1}^0 u_{it} = 0\end{aligned}$$

$$\begin{aligned}
&= -\frac{1}{T} \sum_{t=1}^T \mathbb{E} \left( \sum_{s=1}^T y_{is-1}^0 u_{it} \right) \\
&= -\frac{1}{T} \sum_{t=1}^T \mathbb{E} \left( \sum_{s=t+1}^T y_{is-1}^0 u_{it} \right) \\
&= -\frac{1}{T} \sum_{t=1}^T \mathbb{E} \left\{ \sum_{s=t+1}^T \sum_{j=0}^{\infty} \rho^j u_{is-1-j} u_{it} \right\} \\
&= -\sigma^2 \frac{1}{T} \sum_{t=1}^T \sum_{j=0}^{T-t-1} \rho^j \\
&\quad \text{since } \mathbb{E} \{ u_{is-1-j} u_{it} \} \\
&= \sigma^2 \mathbf{1} \{ s = t + 1 + j \} \\
&= -\frac{\sigma^2}{1-\rho} + \frac{\sigma^2}{1-\rho} \frac{1}{T} \sum_{t=1}^T \rho^{T-t} \\
&= -\frac{\sigma^2}{1-\rho} + \frac{\sigma^2}{1-\rho} \frac{1}{T} \sum_{j=0}^{T-1} \rho^j \\
&= -\frac{\sigma^2}{1-\rho} + O\left(\frac{1}{T}\right). \tag{4.9}
\end{aligned}$$

It follows directly from (4.8) and (4.9) that as  $N \rightarrow \infty$

$$\hat{\rho} - \rho \xrightarrow{a.s} \frac{-\frac{\sigma^2}{1-\rho} + O\left(\frac{1}{T}\right)}{T \frac{\sigma^2}{1-\rho} + O(1)} = -\frac{1+\rho}{T} + O\left(\frac{1}{T^2}\right), \tag{4.10}$$

and  $\hat{\rho}$  is inconsistent. The inconsistency is non-trivial for small  $T$ . For example when  $\rho = 0.5$  and  $T = 5$ , the asymptotic bias is  $-\frac{1+\rho}{T} = -0.3$ , so in this case  $\hat{\rho} \xrightarrow{a.s} 0.2$ . When  $N$  is large, because of the small variance in  $\hat{\rho}$ , almost all of the distribution of  $\hat{\rho}$  then lies to the left of the true value  $\rho = 0.5$  and confidence intervals have a coverage probability close to zero. In short wide panels (with  $T$  small and  $N$  large), the inconsistency can therefore have dramatic effects on inference.

The bias expression (4.10) holds only for the stationary panel case where  $|\rho| < 1$ . The unit root case  $\rho = 1$  is handled in a similar way and we obtain (see Phillips and Sul, 2007)

$$\hat{\rho} - \rho \xrightarrow{a.s} -\frac{3}{T} + O\left(\frac{1}{T^2}\right), \tag{4.11}$$

so the bias effects are magnified in the unit root case and there is no continuity in the asymptotic bias expression as  $\rho$  moves to unity. The bias in the unit case can clearly be very large when  $T$  is small. For example, when  $T = 3$  and 4 the respective bias is  $-1$  and  $-0.75$ , almost sufficient to change the sign of  $\hat{\rho}$ .

As is apparent from the calculation leading to (4.9),  $u_{it}$  and  $\bar{y}_{i,-1}^0$  are correlated and it is this correlation that produces the inconsistency in  $\hat{\rho}$  as  $N \rightarrow \infty$  for fixed  $T$ . Thus, it is the removal of the fixed effects by the within regression transformation of demeaning (manifest in the presence of  $\bar{y}_{i,-1}^0$  in (4.7)) that induces the correlation. Notice also that the QMLE of  $\lambda_i$  is  $\hat{\lambda}_i = \bar{y}_i - \hat{\rho}\bar{y}_{i,-1}$ , so that there is a further induced bias in the estimation of the individual effects. The numerator of  $\hat{\rho} - \rho$ ,  $\sum_{i=1}^N \sum_{t=1}^T (y_{it-1} - \bar{y}_{i,-1}) u_{it}$ , is the score function of the concentrated Gaussian quasi-likelihood function and it too is biased with consequential effects on testing. Thus, the incidental parameter problem arising from the presence of fixed effects in (4.4) has a manifold deleterious impact on estimation and inference in dynamic panel regression.

#### 4.3.1.2 Panels with Fixed $T$ : First Differencing and Instrumental Variables

When the number of time series observations  $T$  is fixed, methods to overcome the incidental parameter problem described in the previous analysis are well studied and many solutions have been suggested since the early 1980's. Excellent overviews of the subject are available, including the chapter by Arellano and Honore (2001), and the books by Arellano (2003), Hsiao (2003), and Baltagi (2008).

The most common approach is to rely on first differencing to eliminate the fixed effect instead of the within transformation. The resulting equation is then:

$$\Delta y_{it} = \rho \Delta y_{it-1} + \Delta u_{it}. \quad (4.12)$$

However, while this transformation removes the individual effect, it introduces a moving average component of order 1 in the error term which brings about correlation with the regressor in (4.12). However, as noted by Anderson and Hsiao (1982) past values of the dependent variable satisfy the necessary moment conditions for a valid instrument, i.e.

$$E(\Delta u_{it} y_{it-s}) = E((u_{it} - u_{it-1}) y_{it-s}) = 0, \quad \text{for any } s > 1. \quad (4.13)$$

They proposed instrumental variable estimation with  $y_{it-2}$  as instrument, leading to

$$\hat{\rho}_{IV} = \left( \sum_{i=1}^N \sum_{t=1}^T y_{it-2} \Delta y_{it-1} \right)^{-1} \sum_{i=1}^N \sum_{t=1}^T y_{it-2} \Delta y_{it}$$

which is consistent and asymptotically normal as  $N \rightarrow \infty$  for  $T$  fixed.

Of course, because all lagged values of  $y_{it-1}$  are valid instruments, one could use many more instruments among the overall  $T(T-1)/2$  moment conditions in the (4.13) class. These may be mobilized in a GMM framework as suggested by Arellano and Bond (1991). Ahn and Schmidt (1995) further added the  $T-2$  moments  $E(\Delta u_{it} u_{iT}) = 0$  since  $u_{it}$  is assumed to be serially uncorrelated. Ahn and Schmidt also show that the estimator that uses these  $T(T-1)/2 + (T-2)$  moments uses

all the moments implied by the basic assumptions, and that the resulting estimator correspondingly reaches the semi-parametric efficiency bound of Chamberlain (1982, 1984).

Han, Phillips, and Sul (2014)<sup>3</sup> generalize this idea and introduce the new concept of X-differencing to generate moment conditions. The procedure eliminates the fixed effects like conventional first differencing while making the regressor and error uncorrelated after the transformation. Hence, there is no need for instrumental variables, and the method does not suffer from the weak identification problem that arises as the autoregressive parameter approaches unity (a problem originally noted by Blundell and Bond 1998). The method combines the basic equation (4.4) with the forward-looking regression

$$y_{is} = \lambda_i + \rho y_{is+1} + \varepsilon_{is}^* \quad (4.14)$$

where  $\varepsilon_{is}^* = \varepsilon_{is} - \rho(y_{is+1} - y_{is-1})$ , which is uncorrelated with  $y_{is+1}$  if  $\varepsilon_{it}$  is serially uncorrelated and uncorrelated with the individual effect  $E(\lambda_i \varepsilon_{is}) = 0$ , though this condition is not needed for the properties of the estimator as the individual effects are eventually eliminated. The same orthogonality condition applies when replacing  $s+1$  by  $t > s$ .

Subtracting (4.14) from (4.4), leads to the simple regression equation

$$y_{it} - y_{is} = \rho(y_{it-1} - y_{is+1}) + (\varepsilon_{it} - \varepsilon_{is}^*) \quad (4.15)$$

where the regressor and error are uncorrelated for any  $s < t-1$  and any  $-1 < \rho \leq 1$  so that the approach accommodates the unit root case  $\rho = 1$  within the same framework. The X-differencing terminology is suggested by virtue of the fact that the regressand is differenced by  $X = t-s$  periods while the regressor is differenced by  $X-2$  periods. All admissible values of  $s = 1, \dots, t-3$  or  $X = 3, \dots, t-1$  can be considered.

Based on these X-differences, Han, Phillips, and Sul construct a panel fully aggregated estimator (PFAE) as the pooled regression estimator in (4.15) for all  $i$ ,  $t$ , and  $s$

$$\hat{\rho}_{PFAE} = \left( \sum_{i=1}^N \sum_{t=1}^T \sum_{s=1}^{t-3} (y_{it-1} - y_{is+1})^2 \right)^{-1} \sum_{i=1}^N \sum_{t=1}^T \sum_{s=1}^{t-3} (y_{it-1} - y_{is+1})(y_{it} - y_{is}).$$

This estimator is consistent and asymptotically normal as long as the number of observations  $NT$  goes to infinity. Thus, it is consistent for fixed  $T$  as  $N \rightarrow \infty$ . More importantly, given the orthogonality between the error and regressor after the transformation, it is essentially unbiased in finite samples as confirmed in simulations. The FAE estimator has other appealing properties, including improved efficiency when  $\rho$  is in the vicinity of unity.

Other transformations have been considered to eliminate the fixed effects and allow for consistent estimation of  $\rho$  even with a finite number of time series observations.

For example, Arellano and Bover (1995) proposed forward orthogonal differences:

$$y_{it}^* = \sqrt{\frac{T-t}{T-t+1}} \left[ y_{it} - \frac{1}{T-t} (y_{it+1} + \dots + y_{iT}) \right]. \quad (4.16)$$

For detailed results of the analysis of the various IV and GMM estimators of  $\rho$  when  $T$  is fixed, readers may refer to several standard textbook treatments such as Arellano (2003), Baltagi (2008), and Hsiao (2003). Another approach using indirect inference methods has been suggested recently by Gourieroux, Phillips, and Yu (2010).

#### 4.3.1.3 When $T$ is Large

When the time series dimension of the panel is large, Hahn and Kuersteiner (2002) proposed to use joint  $(N, T) \rightarrow \infty$  asymptotics to characterize the bias of the fixed effect estimator that arises from the incidental parameters  $\lambda_i$ . More specifically, they derived the limit distribution of  $\hat{\rho}$  by allowing  $N, T \rightarrow \infty$  jointly under the rate condition  $\frac{N}{T} \rightarrow \kappa^2$ , where  $0 < \kappa < \infty$ , so that  $N$  and  $T$  pass to infinity at the same rate. For expositional simplicity here,<sup>4</sup> we assume that  $u_{it} \sim iid(0, \sigma_u^2)$  across  $i$  and over  $t$  with finite fourth moments. We further assume that  $\frac{1}{N} \sum_{i=1}^N y_{i0}^2 = O_p(1)$  and the fixed effects  $\lambda_i$  satisfy  $\frac{1}{N} \sum_{i=1}^N \lambda_i^2 = O_p(1)$ .

The centered and normalized within estimator (4.5) has the form

$$\sqrt{NT}(\hat{\rho} - \rho) = \left( \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T (y_{it-1} - \bar{y}_{i,-1})^2 \right)^{-1} \frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T (y_{it-1} - \bar{y}_{i,-1}) u_{it}. \quad (4.17)$$

As  $N, T \rightarrow \infty$ , the denominator of  $\sqrt{NT}(\hat{\rho} - \rho)$  has the following limit

$$\frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T (y_{it-1} - \bar{y}_{i,-1})^2 \rightarrow_p \frac{\sigma_u^2}{1 - \rho^2}, \quad (4.18)$$

mirroring our earlier result (4.8) for fixed  $T$ . Defining  $u_{it}(\rho) = \sum_{s=0}^{t-1} \rho^s u_{it-s}$ , the numerator of (4.17) decomposes as follows

$$\begin{aligned} & \frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T (y_{it-1} - \bar{y}_{i,-1}) u_{it} \\ &= \frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T u_{it-1}(\rho) u_{it} - \sqrt{\frac{N}{T}} \left( \frac{1}{N} \sum_{i=1}^N \frac{1}{T} \sum_{t=1}^T \sum_{s=1}^T u_{it-1}(\rho) u_{is} \right) + o_p(1). \end{aligned}$$

Since  $u_{it-1}(\rho) u_{it}$  is a martingale difference, the first term satisfies an extended version of the martingale central limit theorem (c.f. Phillips and Moon 1999)

$$\frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T u_{it-1}(\rho) u_{it} \Rightarrow \mathcal{N}\left(0, \frac{\sigma_u^4}{1 - \rho^2}\right),$$

whereas the second term converges in probability to a constant

$$\sqrt{\frac{N}{T}} \left( \frac{1}{N} \sum_{i=1}^N \frac{1}{T} \sum_{t=1}^T \sum_{s=1}^T u_{it-1}(\rho) u_{is} \right) \xrightarrow{p} \kappa \frac{\sigma_u^2}{1-\rho},$$

which mirrors our earlier result (4.9) for the numerator as  $N \rightarrow \infty$  for fixed  $T$ . Therefore, as  $N, T \rightarrow \infty$  with  $\frac{N}{T} \rightarrow \kappa^2$ , the weak limit of the numerator of  $\sqrt{NT}(\hat{\rho} - \rho)$  is

$$\frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T (y_{it-1} - \bar{y}_{i,-1}) (u_{it} - \bar{u}_i) \Rightarrow \mathcal{N}\left(-\kappa \frac{\sigma_u^2}{1-\rho}, \frac{\sigma_u^4}{1-\rho^2}\right). \quad (4.19)$$

Combining (4.18) and (4.19) as  $N, T \rightarrow \infty$  with  $\frac{N}{T} \rightarrow \kappa^2$  and where  $0 < \kappa < \infty$ , we have

$$\sqrt{NT}(\hat{\rho} - \rho) \Rightarrow \mathcal{N}(-\kappa(1+\rho), 1-\rho^2). \quad (4.20)$$

An important aspect of the limit distribution of the fixed effects estimator is the bias involved in the miscentred normal limit of (4.20). This bias, like that for the  $N \rightarrow \infty$  case, is due to the presence of the incidental parameters  $\lambda_i$  in (4.4) and arises from the limiting correlation apparent in (4.19) that is induced by the (within) transformation to eliminate the nuisance parameters. A finite sample implication of (4.20) is that the bias of  $\hat{\rho}$  can approximated by  $-\frac{1+\rho}{T}$ , just as suggested in the earlier limit (4.10) for fixed  $T$ .

Accordingly Hahn and Kuersteiner (2002) suggested the following bias-corrected estimator

$$\check{\rho} = \hat{\rho} + \frac{1}{T}(1+\hat{\rho}) = \frac{T+1}{T}\hat{\rho} + \frac{1}{T},$$

which, as  $\frac{N}{T} + \frac{1}{N} \rightarrow \kappa^2$  with  $0 < \kappa < \infty$ , has the following correctly centred limit distribution

$$\sqrt{NT}(\check{\rho} - \rho) \Rightarrow \mathcal{N}(0, 1-\rho^2). \quad (4.21)$$

This bias corrected estimator is shown in Hahn and Kuersteiner to be asymptotically efficient under Gaussian errors provided  $|\rho| < 1$ . The unit case  $\rho = 1$  is more complex and the limit theory (4.21) no longer holds. While the limit distribution is still normal, the rate of approach to the limit theory is no longer  $O(\sqrt{NT})$  but is instead  $O(\sqrt{NT^2})$ , reflecting the stronger time series signal in the regressors in the unit root case. Hahn and Kuersteiner proved that  $\hat{\rho}$  has the following limit distribution in this case when  $\rho = 1$  and  $(N, T) \rightarrow \infty$

$$\sqrt{NT^2} \left( \hat{\rho} - \rho + \frac{3}{T} \right) \Rightarrow \mathcal{N}\left(0, \frac{51}{5}\right).$$

Alvarez and Arellano (2003) studied the GMM estimator  $\hat{\rho}_{GMM}$  based on the transformed data (4.16) and using the lagged dependent variables  $z_{it} = (y_{it-1}, \dots, y_{i0})$  as IVs expressed as

$$\begin{aligned}\hat{\rho}_{GMM} &= \left( \sum_{t=1}^T \left( \sum_{i=1}^N x_{it}^* z_{it}' \right) \left( \sum_{i=1}^N z_{it} z_{it}' \right)^{-1} \left( \sum_{i=1}^N z_{it} x_{it}^* \right) \right)^{-1} \\ &\quad \times \left( \sum_{t=1}^T \left( \sum_{i=1}^N x_{it}^* z_{it}' \right) \left( \sum_{i=1}^N z_{it} z_{it}' \right)^{-1} \left( \sum_{i=1}^N z_{it} y_{it}^* \right) \right),\end{aligned}$$

where  $x_{it}^* = \sqrt{\frac{T-t}{T-t+1}} \left[ y_{it-1} - \frac{1}{T-t} (y_{it} + \dots + y_{iT-1}) \right]$ . They showed that as  $T \rightarrow \infty$  the bias of  $\hat{\rho}_{GMM}$  may be approximated by

$$-\frac{1}{N} (1 + \rho)$$

and its limiting distribution is

$$\sqrt{NT} \left( \hat{\rho}_{GMM} - \rho + \frac{1}{N} (1 + \rho) \right) \Rightarrow \mathcal{N}(0, 1 - \rho^2)$$

which can be written as

$$\sqrt{NT} (\hat{\rho}_{GMM} - \rho) \Rightarrow \mathcal{N}(-\kappa^{-1} (1 + \rho), 1 - \rho^2).$$

**General Motivation of the Alternative Asymptotics:** The fixed effect estimator  $\hat{\rho}$  corresponds to the maximum likelihood estimator based on the conditional Gaussian likelihood. To see this, suppose that conditional on  $y_{i0}$  and  $\lambda_i$ ,  $u_{it} \sim iid N(0, \sigma_u^2)$  with known  $\sigma_u^2$ . Then, the conditional log-likelihood of  $(y_{iT}, \dots, y_{i1})$  on  $(y_{i0}, \lambda_i)$  is

$$l_{NT}(\rho, \lambda^N) = \sum_{i=1}^N \ln f(y_{iT}, \dots, y_{i1} | y_{i0}, \rho, \lambda_i) = \sum_{i=1}^N \sum_{t=1}^T l_{it}(\rho, \lambda_i),$$

where

$$l_{it}(\rho, \lambda_i) = -\frac{1}{2\sigma_u^2} (y_{it} - \rho y_{it-1} - \lambda_i)^2.$$

Without loss of generality, assume that  $\sigma_u^2 = 1$  for simplicity. Then, for given  $\rho$  the MLE of  $\lambda_i$  is

$$\hat{\lambda}_i(\rho) = \arg \max_{\lambda} \sum_{t=1}^T l_{it}(\rho, \lambda) = \bar{y}_i - \rho \bar{y}_{i,-1}.$$

Plugging  $\hat{\lambda}_i(\rho)$  into the likelihood, we have the (concentrated) profile likelihood of  $\rho$ :

$$\begin{aligned} l_{NT}(\rho) &= l_{NT}\left(\rho, \hat{\lambda}_i(\rho)\right) = \sum_{i=1}^N \sum_{t=1}^T l_{it}\left(\rho, \hat{\lambda}_i(\rho)\right) \\ &= -\frac{1}{2} \sum_{i=1}^N \sum_{t=1}^T (y_{it} - \bar{y}_i - \rho(y_{it-1} - \bar{y}_{i,-1}))^2. \end{aligned}$$

The fixed effect estimator  $\hat{\rho}$  is simply the MLE in this case since

$$\hat{\rho} = \arg \max_{\rho} l_{NT}(\rho).$$

Define

$$L_T(\rho) = \lim_N \frac{1}{N} E(l_{NT}(\rho)) = \lim_N \frac{1}{N} \sum_{i=1}^N E\left[\sum_{t=1}^T l_{it}\left(\rho, \hat{\lambda}_i(\rho)\right)\right].$$

As follows by standard asymptotic theory for extremum estimators, when  $N \rightarrow \infty$  but  $T$  is fixed, the MLE  $\hat{\rho}$  converges in probability to

$$\hat{\rho} \xrightarrow{p} \rho_T = \arg \max_{\rho} L_T(\rho).$$

Denote by  $(\rho_0, \lambda_i^0)$  the true parameters of  $(\rho, \lambda_i)$ . Then, since

$$\sum_{t=1}^T l_{it}\left(\rho, \hat{\lambda}_i(\rho)\right) = -\frac{1}{2} \sum_{t=1}^T \{(u_{it} - \bar{u}_i) - (\rho - \rho_0)(y_{it-1} - \bar{y}_{i,-1})\}^2$$

we have

$$\begin{aligned} L_T(\rho) &= -\frac{1}{2} (\rho - \rho_0)^2 \lim_N \frac{1}{N} \sum_{i=1}^N E\left(\sum_{t=1}^T (y_{it-1} - \bar{y}_{i,-1})\right)^2 \\ &\quad + (\rho - \rho_0) \lim_N \frac{1}{N} \sum_{i=1}^N E\left(\sum_{t=1}^T (y_{it-1} - \bar{y}_{i,-1})(u_{it} - \bar{u}_i)\right) \\ &\quad - \frac{1}{2} \lim_N \frac{1}{N} \sum_{i=1}^N E\left(\sum_{t=1}^T (u_{it} - \bar{u}_i)^2\right). \end{aligned} \tag{4.22}$$

For fixed  $T$ , the expected score of the profile likelihood  $E(\sum_{t=1}^T (y_{it-1} - \bar{y}_{i,-1})(u_{it} - \bar{u}_i)) \neq 0$ , and we have

$$\rho_T \neq \rho_0.$$

However, when  $T \rightarrow \infty$  as well as  $N \rightarrow \infty$ , we have

$$\rho_T \rightarrow \rho_0.$$

Arellano and Hahn (2006) observed that when the profile likelihood function  $l_{NT}(\rho)$  is smooth, we usually have the expansion

$$\rho_T = \rho_0 + \frac{B}{T} + O\left(\frac{1}{T^2}\right), \quad (4.23)$$

and the re-centred limit theory

$$\sqrt{NT}(\hat{\rho} - \rho_T) \Rightarrow \mathcal{N}(0, \Omega)$$

for some  $\Omega > 0$ . Then, as  $N, T \rightarrow \infty$  with  $\frac{N}{T} \rightarrow \kappa^2$ , where  $0 < \kappa < \infty$ , we have

$$\begin{aligned} \sqrt{NT}(\hat{\rho} - \rho_0) &= \sqrt{NT}(\hat{\rho} - \rho_T) + \sqrt{NT}\frac{B}{T} + O\left(\sqrt{\frac{N}{T^3}}\right) \\ &\Rightarrow \mathcal{N}(\kappa B, \Omega). \end{aligned}$$

Here the bias  $B$  of the fixed effect estimator is characterized as the bias of the asymptotic distribution under these joint asymptotics.

#### 4.3.1.4 Alternative Bias Correction Method

Dhaene and Jochmans (2012) proposed a jackknife method to reduce the order of the bias of the fixed effect estimator in nonlinear dynamic panel regression models. To apply their ideas to the dynamic linear set up in (4.4), recall that the pseudo true value  $\rho_T$  that maximizes the limit of the profile likelihood  $L_T(\rho)$  in (4.22) is  $\rho_T = \rho_0 + \frac{B}{T} + O(\frac{1}{T^2})$ . We now split the  $(N \times T)$  panel into two  $(N \times \frac{T}{2})$  dimensional pieces and denote by  $\hat{\rho}_1$  and  $\hat{\rho}_2$  the fixed effect estimators in these respective subpanels. Define

$$\bar{\rho}_{1/2} = \frac{1}{2}(\hat{\rho}_1 + \hat{\rho}_2).$$

Dhaene and Jochmans's bias corrected estimator is then based on the usual jackknife formula

$$\hat{\rho}_{1/2} = 2\hat{\rho} - \bar{\rho}_{1/2}.$$

We can expect this estimator to correct the bias  $B$  because by using the expansion (4.23) we have

$$\begin{aligned} \hat{\rho}_{1/2} &= 2\rho_0 + \frac{2B}{T} + O\left(\frac{1}{T^2}\right) - \frac{1}{2}\left(2\rho_0 + \frac{4B}{T} + O\left(\frac{1}{T^2}\right)\right) \\ &= \rho_0 + O\left(\frac{1}{T^2}\right). \end{aligned}$$

Dhaene and Jochmans (2012) showed that this idea can be applied to a more general nonlinear dynamic panel regression models with fixed effects and show how to reduce the bias to a higher order.

### 4.3.2 Additive Individual Effects and Time Effects: $\alpha(\lambda_i, f_t) = \lambda_i + f_t$

Hahn and Moon (2006) extended Hahn and Kuersteiner's (2002) results by considering time effects as additional incidental parameters in the model

$$y_{it} = \rho y_{it-1} + \lambda_i + f_t + u_{it}, \quad (4.24)$$

where the model satisfies the conditions in the previous section except that it now includes time specific fixed effects  $f_t$ . In many empirical applications, the time effect  $f_t$  is included to model a simple form of nonstationarity in the time series  $y_{it}$  or to represent an aggregate shock (e.g., a common macro shock) that is common to all the cross-section units. In the latter case, when the common shock  $f_t$  is random, the cross-sectional observations  $y_{it}$  have cross-sectional dependence. In model (4.24) incidental parameters exist in both the cross-sectional direction ( $\lambda_i$ ) and the time series direction ( $f_t$ ).

Define  $\bar{y}_{\cdot,t} = \frac{1}{N} \sum_{i=1}^N y_{it}$ ,  $\bar{y}_{i,\cdot} = \frac{1}{T} \sum_{t=1}^T y_{it}$ ,  $\bar{y} = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T y_{it}$ ,  $\bar{y}_{i,-1} = \frac{1}{T} \sum_{t=1}^T y_{it-1}$ ,  $\bar{y}_{-1} = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T y_{it-1}$ . Similar definitions are used for  $\lambda$ ,  $\bar{f}$ ,  $\bar{u}_{\cdot,t}$ ,  $\bar{u}_{i,\cdot}$ , and  $\bar{u}$ . In this model, the fixed effect estimator is

$$\hat{\rho} = \left( \sum_{i=1}^N \sum_{t=1}^T (y_{it-1} - \bar{y}_{\cdot,t-1} - \bar{y}_{i,-1} + \bar{y}_{-1})^2 \right)^{-1} \times \left( \sum_{i=1}^N \sum_{t=1}^T (y_{it-1} - \bar{y}_{\cdot,t-1} - \bar{y}_{i,-1} + \bar{y}_{-1}) (y_{it} - \bar{y}_{\cdot,t} - \bar{y}_{i,\cdot} + \bar{y}) \right).$$

Hahn and Moon (2006) showed that under joint asymptotics as  $N, T \rightarrow \infty$  with  $\frac{N}{T} \rightarrow \kappa^2$ , where  $0 < \kappa < \infty$ ,  $\sqrt{NT}(\hat{\rho} - \rho)$  has the limit distribution

$$\sqrt{NT}(\hat{\rho} - \rho) \Rightarrow \mathcal{N}(-\kappa(1 + \rho), 1 - \rho^2),$$

identical to the limit given earlier in (4.20). Hence, the asymptotic bias of the fixed effect estimator with  $\lambda_i + f_t$  is the same as the asymptotic bias of the fixed effect estimator only with  $\lambda_i$ .

In this case the fixed effect estimator  $\hat{\rho}$  eliminates  $f_t$  by taking out the cross-sectional mean of the panel data  $y_{it}$ . The regressor after this demeaning transformation becomes  $y_{it-1} - \bar{y}_{\cdot,t-1}$  and the error becomes  $u_{it} - \bar{u}_{\cdot,t}$ . From a time series perspective the regressor  $y_{it-1} - \bar{y}_{\cdot,t-1}$  is still predetermined and uncorrelated with the error  $u_{it} - \bar{u}_{\cdot,t}$ . From a cross-sectional perspective, the influence of the time specific effect  $f_t$  in the regressor  $y_{it-1} - \bar{y}_{\cdot,t-1}$  and the error  $u_{it} - \bar{u}_{\cdot,t}$  is removed. As a consequence, the limit of the estimator  $\hat{\rho}$  is not affected by the presence of  $f_t$ .

This finding confirms that the asymptotic bias of the fixed effect estimator in the linear dynamic panel regression model is sourced in the presence of the individual

effect  $\lambda_i$  nor the presence of the time effect  $f_t$  when these components enter the model in an additive and separable form. However, when the panel model is nonlinear and/or  $\lambda_i$  and  $f_t$  enter the model in a more general functional form, it may be the case that both  $\lambda_i$  and  $f_t$  contribute to the asymptotic bias of the fixed effect estimator as we will discuss in the following sections.

### 4.3.3 Interactive Fixed Effects: $\alpha(\lambda_i, f_t) = \lambda'_i f_t$

In this section, we discuss the case where the fixed effects take a multiplicative form involving  $\lambda_i$  and  $f_t$ , viz.,

$$y_{it} = \rho y_{it-1} + \lambda'_i f_t + u_{it}, \quad (4.25)$$

where  $\lambda_i$  and  $f_t$  are unknown fixed effects for  $i$  and  $t$ , respectively and  $u_{it}$  are idiosyncratic shocks. We denote by  $\lambda_i^0, f_t^0, \rho_0$  the true values of  $\lambda_i, f_t$ , and  $\rho$ , respectively. In this section we assume the dimension of  $f_t$  and  $\lambda_i$  are known, say  $R^0$ . Also assume that  $u_{it}$  are independent across  $i$  and  $t$  with mean zero and higher moments finite. The multiplicative form of the fixed effects appearing in  $\lambda'_i f_t$  are often called interactive fixed effects or common factors in the literature.

The linear panel regression with interactive fixed effects was studied by Kiefer (1980), Lee (1991), Ahn, Lee, and Schmidt (2001), and Bai (2009) when the regressors are strictly exogenous with respect to  $u_{it}$ , and by Holtz-Eakin, Newey, and Rosen (1988), Phillips and Sul (2003), and Moon and Weidner (2010, 2014) when the regressors are lagged dependent variables.<sup>5</sup>

#### 4.3.3.1 Quasi-Differencing Approach

Holtz-Eakin, Newey, and Rosen (1988) suggest that the interactive fixed effects be eliminated by taking a quasi-difference of the data. Suppose that  $R_0 = 1$  for expositional convenience and that we normalize  $f_1 = 1$ . Since

$$\frac{y_{it}}{f_t} = \rho \frac{y_{it-1}}{f_t} + \lambda_i + \frac{u_{it}}{f_t},$$

we have

$$\frac{y_{it}}{f_t} - \frac{y_{it-1}}{f_{t-1}} = \rho \left( \frac{y_{it-1}}{f_t} - \frac{y_{it-2}}{f_{t-1}} \right) + \frac{u_{it}}{f_t} - \frac{u_{it-1}}{f_{t-1}}.$$

They suggested using lagged variables  $\{y_{it-s}\}_{s \geq 2}$  as instruments to estimate  $\rho$  and  $(f_2, \dots, f_T)$  using the GMM method. This approach may work well when  $T$  is small. However, if  $T$  is large, this approach becomes problematic because  $\{f_2, \dots, f_T\}$  become another set of incidental parameters to be accommodated. One must also address the issue that the number of the instruments then increases at the order  $O(T^2)$ .<sup>6</sup>

### 4.3.3.2 Principal Component Approach with Joint Asymptotics

Another approach is to estimate the model (4.25) together with  $\{\lambda'_i f_t\}_{i,t}$  by least squares and apply joint asymptotics as  $N, T \rightarrow \infty$  with  $\frac{N}{T} \rightarrow \kappa^2 > 0$  to characterize the form of the bias. The estimate may then be bias corrected using a plug in estimate of the resulting bias.

To fix ideas, let  $Y, Y_{-1}, U$  be  $N \times T$  matrices whose  $(i, t)^{th}$  elements are  $y_{it}$ ,  $y_{it-1}$ , and  $u_{it}$ , respectively. Let  $\lambda$  and  $f$  be the  $(N \times R)$  and the  $(T \times R)$  matrices that stack the  $R$ -row vectors  $\lambda'_i$  and  $f'_t$ , respectively. We may then write the model (4.25) in matrix notation as

$$Y = \rho Y_{-1} + \lambda f' + U.$$

Notice that  $\lambda$  and  $f$  are not separately identified. One can always transform or rotate with an invertible symmetric matrix giving  $\tilde{\lambda} = \lambda S$  and  $\tilde{f} = f S^{-1}$  such that  $\lambda f' = \tilde{\lambda} \tilde{f}'$ . However, to identify the parameter of interest  $\rho$ , we do not need to separately identify  $\lambda$  and  $f$ . Instead, identification of  $\lambda f'$  is enough. Details of the identification of  $\rho$  are given in Moon and Weidner (2014).

The (negative) Gaussian quasi-loglikelihood function conditioned on the initial conditions  $\{y_{io}\}$  and the interactive fixed effects  $\{\lambda'_i f_t\}$  is (up to a constant)

$$l_{NT}(\rho, \lambda, f) = \text{tr} \left[ (Y - \rho Y_{-1} - \lambda f')' (Y - \rho Y_{-1} - \lambda f') \right].$$

The profile quasi-loglikelihood function is

$$\begin{aligned} l_{NT}(\rho) &= \min_{\lambda, f} l_{NT}(\rho, \lambda, f) \\ &= \min_f \text{tr} \left[ (Y - \rho Y_{-1}) M_f (Y - \rho Y_{-1})' \right] \\ &= \sum_{t=R+1}^T \mu_t \left[ (Y - \rho Y_{-1})' (Y - \rho Y_{-1}) \right], \end{aligned}$$

where  $M_f = I_T - f(f'f)^{-1}f'$  and  $\mu_k[A]$  is the  $k^{th}$  smallest eigenvalue of matrix  $A$ . The QMLE or the fixed effect estimator  $\hat{\rho}$  minimizes  $l_{NT}(\rho)$ :

$$\hat{\rho} = \arg \min l_{NT}(\rho).$$

Moon and Weidner (2010, 2013) analyzed the properties of this estimator. As before, suppose that  $u_{it} \sim$  independent across  $i$  and  $t$  with a finite uniform  $8^{th}$  moment, that is,  $\sup_{i,t} \mathbb{E}(u_{it}^8) < \infty$ . Also, assume that  $\lambda_i^0$  and  $f_t^0$  are  $R$ -vector strong factors in the sense that

$$\frac{1}{N} \lambda^{0'} \lambda^0 \rightarrow_p \Sigma_\lambda > 0 \text{ and } \frac{1}{T} f^{0'} f^0 \rightarrow_p \Sigma_f > 0.$$

Under these conditions, they showed that  $\hat{\rho}$  is consistent, that is,

$$\hat{\rho} \rightarrow_p \rho_0.$$

Their consistency proof is different from the conventional consistency proof of an extremum estimator that uses a uniform law of large numbers and an identification condition. Moon and Weidner (2013)'s proof is to bound  $l_{NT}(\hat{\rho})$  by a lower and an upper bound as follows

$$\begin{aligned} & c(\hat{\rho} - \rho_0)^2 + O_p\left(\frac{|\hat{\rho} - \rho_0|}{\sqrt{\min\{N, T\}}}\right) + \frac{1}{NT} \text{tr}(U'U) + O_p\left(\frac{1}{\min\{N, T\}}\right) \\ & \leq \frac{l_{NT}(\hat{\rho})}{NT} \leq \frac{l_{NT}(\rho_0)}{NT} \\ & = \frac{1}{NT} \sum_{t=R+1}^T \mu_t \left[ (\lambda^0 f^{0'} + U)' (\lambda^0 f^{0'} + U) \right] \leq \frac{1}{NT} \text{tr}(U'U) \leq 0, \end{aligned}$$

from which expression and the given rates we can deduce that

$$\hat{\rho} - \rho_0 = O_p\left(\frac{1}{\sqrt{\min\{N, T\}}}\right) = o_p(1).$$

A remaining challenge is to derive the limit distribution of  $\hat{\rho} - \rho_0$ . The problem is challenging because the conventional approach to deriving the limiting distribution of an extremum estimator typically uses a quadratic approximation of the objective function obtained, for example, via a Taylor approximation. Instead, Moon and Weidner (2010, 2013) use the perturbation theory of a linear operator to approximate  $l_{NT}(\rho)$  with a quadratic function of  $\rho$  as

$$l_{NT}(\rho) - l_{NT}(\rho_0) = l_{q,NT}(\rho) + \mathcal{R}_{NT}(\beta),$$

where

$$\begin{aligned} l_{q,NT}(\rho) &= -2\sqrt{NT}(\rho - \rho_0) C_{NT} + \left(\sqrt{NT}(\rho - \rho_0)\right)^2 W_{NT} \\ C_{NT} &= C^{(1)}(\lambda^0, f^0, Y_{-1}, U) + C^{(2)}(\lambda^0, f^0, Y_{-1}, U) \\ C^{(1)}(\lambda^0, f^0, Y_{-1}, U) &= \frac{1}{\sqrt{NT}} \text{tr}\left(M_{f^0} U' M_{\lambda^0} Y_{-1}\right) \\ C^{(2)}(\lambda^0, f^0, Y_{-1}, U) &= -\frac{1}{\sqrt{NT}} \left[ \begin{array}{l} \text{tr}\left(U M_{f^0} U' M_{\lambda^0} Y_{-1} f^0 (f^0 f^0)^{-1} (\lambda^0 \lambda^0)^{-1} \lambda^0\right) \\ + \text{tr}\left(U' M_{\lambda^0} U M_{f^0} Y_{-1} \lambda^0 (\lambda^0 \lambda^0)^{-1} (f^0 f^0)^{-1} f^0\right) \\ + \text{tr}\left(U' M_{\lambda^0} Y_{-1} M_{f^0} U' \lambda^0 (\lambda^0 \lambda^0)^{-1} (f^0 f^0)^{-1} f^0\right) \end{array} \right] \\ W_{NT} &= \frac{1}{NT} \text{tr}\left(M_{f^0} Y_{-1}' M_{\lambda^0} Y_{-1}\right), \end{aligned}$$

and

$$\sup_{|\rho - \rho_0| \leq \eta_{NT}} \frac{\mathcal{R}_{NT}(\beta)}{\left(1 + \sqrt{NT}(\rho - \rho_0)^2\right)} = o_p(1),$$

for any sequence  $\eta_{NT} \rightarrow 0$ . An immediate consequence of the quadratic approximation is that if  $C_{NT} = O_p(1)$ , then

$$\sqrt{NT}(\hat{\rho} - \rho_0) = W_{NT}^{-1}C_{NT} + o_p(1).$$

Let  $v_{it} = \sum_{\tau=0}^{\infty} \rho_0^\tau u_{it-\tau}$  and define the  $(N \times T)$  matrix  $V$  with  $v_{it}$ . Also, let  $F_t^0 = \sum_{\tau=0}^{\infty} \rho_0^\tau f_{t-\tau}^0$  and define the  $(T \times R)$  matrix  $F^0$ . Let<sup>7</sup>

$$\begin{aligned} W &= \lim_{N,T} \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \mathbb{E}(v_{it-1}^2), \\ \Omega &= \lim_{N,T} \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \mathbb{E}(u_{it}^2) \mathbb{E}(v_{it-1}^2), \\ B_1 &= p \lim_{N,T} \frac{1}{N} \text{tr} \left[ P_{f^0} \mathbb{E}(U'V) \right], \\ B_3 &= p \lim_{N,T} \frac{1}{N} \text{tr} \left[ \mathbb{E}(U'U) M_{f^0} F^0 (f^{0'} f^0)^{-1} f^{0'} \right]. \end{aligned}$$

Suppose that  $|\lambda_i^0|$  and  $|f_t^0|$  are uniformly bounded across  $i, t$ . Then, according to Moon and Weidner (2013), as  $N, T \rightarrow \infty$  with  $\frac{N}{T} \rightarrow \kappa^2 > 0$ ,

$$\sqrt{NT}(\hat{\rho} - \rho_0) \Rightarrow N(-\kappa W^{-1}(B_1 + B_3), W^{-1}\Omega W^{-1}).$$

Here the first bias component  $B_1$  arises because the regressor is a lagged dependent variable and so that the regressor is sequentially exogenous, not strictly exogenous (Bai (2009) does not have this bias since that work assumed only strictly exogenous regressors). The second bias component  $B_3$  arises when the error term  $u_{it}$  is not homoskedastic. (see Bai 2009 and Moon and Weidner 2010).

Nonparametric estimators of  $W, \Omega, B_1$ , and  $B_3$  are proposed by Moon and Weidner (2010) to achieve bias correction. The resulting bias corrected estimator is

$$\check{\rho}^+ = \hat{\rho} + \hat{W}^{-1} \left( \frac{\hat{B}_1}{T} + \frac{\hat{B}_3}{T} \right),$$

and it is shown that this estimator has the centred limit theory

$$\sqrt{NT}(\check{\rho}^+ - \rho_0) \Rightarrow N(0, W^{-1}\Omega W^{-1}).$$

When the error  $u_{it}$  is homoskedastic with  $\mathbb{E}(u_{it}^2) = \sigma^2$ , we have

$$\sqrt{NT}(\check{\rho}^+ - \rho_0) \Rightarrow N(0, 1 - \rho^2)$$

which is the same distribution as the bias-corrected estimator of Hahn and Kuersteiner (2002) with incidental individual effects.

### 4.3.4 Incidental Trends: $\delta(\lambda_i, f_t)$ or $\alpha(\lambda_i, f_t) = \lambda_{i0} + \lambda_{i1}t$

In this section, we turn our attention to panel models with trends. To do so it is convenient to consider a components model specification in which the observed panel data  $y_{it}$  consist of a cross-sectionally heterogenous linear trend superposed with serially correlated errors. Our discussion of this model is in two parts, depending on whether the serial correlation in the errors is weak or strong.

#### 4.3.4.1 Weakly Serially Correlated Case

Suppose that the observed panel  $y_{it}$  is generated by the system

$$\begin{aligned} y_{it} &= \lambda_{i0} + \lambda_{i1}t + \varepsilon_{it} \\ \varepsilon_{it} &= \rho \varepsilon_{it-1} + u_{it}, \end{aligned} \tag{4.26}$$

where  $|\rho| < 1$  and  $u_{it} \sim iid(0, \sigma^2)$  with  $\mathbb{E}|u_{it}|^{4+\varsigma} < \infty$  for some  $\varsigma > 0$ . In this model the individual time series  $y_{it}$  are stationary with deterministic trends. The component panel model (4.26) can be written in an augmented regression form. Subtracting  $\rho y_{it-1}$  from  $y_{it}$ , we have

$$\begin{aligned} y_{it} &= \lambda_{i0}(1 - \rho) + \lambda_{i1}(t - \rho(t - 1)) + \rho y_{it-1} + u_{it} \\ &= \tilde{\lambda}_{i0} + \tilde{\lambda}_{i1}t + \rho y_{it-1} + u_{it}. \end{aligned} \tag{4.27}$$

The main difference between (4.4) and (4.27) is that the individual effects in (4.27) are now time varying.

Again, suppose that the object of interest with model (4.27) is to estimate  $\rho$ . Phillips and Sul (2007)<sup>8</sup> showed that the QMLE estimator of  $\rho$  is asymptotically biased for finite  $T$  due to the presence of the incidental trends. The bias in the stationary case is

$$\hat{\rho} - \rho \rightarrow_p -2 \frac{1 + \rho}{T - 1} \left( 1 + O\left(\frac{1}{T}\right) \right).$$

One way to address the incidental trend problem is to eliminate the incidental trends  $\lambda_{i0} + \lambda_{i1}t$  by a double difference transform, instead of first differencing, as

$$\Delta^2 y_{it} = \rho \Delta^2 y_{it-1} + \Delta^2 u_{it}, \tag{4.28}$$

where  $\Delta^2 y_{it} = \Delta(\Delta y_{it}) = y_{it} - 2y_{it-1} + y_{it-2}$ , as suggested in Wansbeek and Knapp (1999). We may then use  $\{y_{it-s}\}_{s \geq 3}$  as instruments to estimate  $\rho$  in (4.28).

Another approach is to estimate  $\rho$  in (4.27) by least squares and then apply joint asymptotics as  $N, T \rightarrow \infty$  with  $\frac{N}{T} \rightarrow \kappa^2 > 0$  to derive a limit theory. In this case we can show that

$$\sqrt{NT}(\hat{\rho} - \rho) \Rightarrow N(-2\kappa(1 + \rho), 1 - \rho^2).$$

The bias corrected estimator is then simply

$$\check{\rho}^{++} = \hat{\rho} + \frac{2}{T} (1 + \hat{\rho}) = \frac{T+2}{T} \hat{\rho} + \frac{2}{T}.$$

It follows that as  $N, T \rightarrow \infty$  with  $\frac{N}{T} \rightarrow \kappa^2$ , where  $0 < \kappa < \infty$ ,

$$\sqrt{NT} (\check{\rho}^{++} - \rho) \Rightarrow \mathcal{N}(0, 1 - \rho^2)$$

which is again the same distribution as in Hahn and Kuersteiner (2002).

#### 4.3.4.2 Strongly Serially Correlated Case

For a panel whose time series has both deterministic trends and stochastic trends, Moon and Phillips (1999, 2000, 2004) and Phillips and Sul (2007) considered the following model:

$$y_{it} = \lambda_{i0} + \lambda_{i1} t + \varepsilon_{it} \quad (4.29)$$

$$\varepsilon_{it} = \rho \varepsilon_{it-1} + u_{it}, \quad (4.30)$$

where

$$\rho = \left(1 - \frac{\theta}{T}\right).$$

In (4.29), the time series of panel  $y_{it}$  consists of cross-sectionally heterogeneous deterministic trends and highly persistent errors (or stochastic trends). The modeling of autoregressive roots as local to unity is common in the analysis of time series (e.g., Phillips, 1987). While it is known that the parameter  $\theta$  cannot be consistently estimated from a single time series, Moon and Phillips (1999, 2000, 2004) consider estimation possibilities using panel data. For this, they assume that both the cross-sectional dimension  $N$  and the time dimension  $T$  are large such that  $\frac{N}{T} \rightarrow 0$ . We also assume that  $\{u_{it}\}_{i,t}$  are randomly drawn.

First, suppose that  $\varepsilon_{it}$  is observed (equivalently, that  $\lambda_{i0} + \lambda_{i1} t$  is known). Then, we can estimate  $\theta$  consistently. To see this, consider

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

where  $\hat{\rho}$  is the least squares estimator of (4.30):

$$\hat{\rho} = \frac{\sum_{i=1}^N \sum_{t=1}^T \varepsilon_{it-1} \varepsilon_{it}}{\sum_{i=1}^N \sum_{t=1}^T \varepsilon_{it-1}^2}.$$

As  $N, T \rightarrow \infty$ , since

$$T(\hat{\rho} - \rho) = \frac{\frac{1}{N} \sum_{i=1}^N \frac{1}{T} \sum_{t=1}^T \varepsilon_{it-1} u_{it}}{\frac{1}{N} \sum_{i=1}^N \frac{1}{T^2} \sum_{t=1}^T \varepsilon_{it-1}^2} \xrightarrow{p} 0,$$

it follows that

$$\hat{\theta} \rightarrow_p \theta$$

contrary to the case where only time series observations are available. An implication is that when panel data are available, using the cross-sectional variation, we can estimate strong serial dependence in the data, measured by  $\rho$  in the vicinity of unity, much more accurately than when only a single time series is available.

Now suppose that the true trends are heterogeneous and unknown, so that  $\lambda_{i0} + \lambda_{i1}t$  become incidental trends. Let  $\Delta_c$  be the quasi-difference operator for some local-to-unity parameter  $c$ , so that  $\Delta_c y_{it} = y_{it} - (1 - \frac{c}{T})y_{it-1} = \Delta y_{it} + \frac{c}{T}y_{it-1}$ . In this section, we shall denote  $\theta$  as the true localizing coefficient parameter and  $c$  as the parameter used in estimation. Then, the Gaussian quasi log-likelihood function conditional on the initial condition  $y_{i0}$  is

$$l_{NT}(c, \lambda_1, \dots, \lambda_N) = -\frac{1}{2} \sum_{i=1}^N \sum_{t=1}^T \left( \Delta_c y_{it} - \lambda_{i0} \frac{c}{T} - \lambda_{i1} \left( 1 + c \frac{t-1}{T} \right) \right)^2.$$

Given  $c$ , the MLE for  $\hat{\lambda}_i(c)$  is the OLS estimator of  $\Delta_c y_{it}$  on  $(\frac{c}{T}, 1 + c \frac{t-1}{T})$ . Plugging this into  $l_{NT}(c, \lambda_1, \dots, \lambda_N)$ , we have the concentrated log-likelihood function

$$l_{NT}\left(c, \hat{\lambda}_1(c), \dots, \hat{\lambda}_N(c)\right) = -\frac{1}{2} \sum_{i=1}^N \sum_{t=1}^T \left( \Delta_c y_{it} - \hat{\lambda}_{i0}(c) \frac{c}{T} - \hat{\lambda}_{i1}(c) \left( 1 + c \frac{t-1}{T} \right) \right)^2.$$

The QMLE of  $\theta$  is

$$\hat{\theta} = \arg \max_c l_{NT}\left(c, \hat{\lambda}_1(c), \dots, \hat{\lambda}_N(c)\right).$$

Moon and Phillips (1999) showed that

$$\frac{1}{NT} l_{NT}\left(c, \hat{\lambda}_1(c), \dots, \hat{\lambda}_N(c)\right) \rightarrow_p l(c; \theta)$$

for some function  $l(c; \theta)$  uniformly in  $c$ , and  $\theta$  does not maximize the limit function  $l(c; \theta)$ . This implies that the probability limit of  $\hat{\theta}$  is not  $\theta$ . The inconsistency of  $\hat{\theta}$  is due to presence of the unknown incidental trends  $\lambda_{i0} + \lambda_{i1}t$  in the panel data. Moon and Phillips (1999) called this inconsistency the “incidental trend” problem.

Moon and Phillips (2000, 2004) investigated how to correct for the bias that arises in the presence of the incidental trends in estimating  $\theta$  in (4.29). Moon and Phillips (2000) proposed several estimators based on the OLS detrended data. They showed that the estimators are consistent and asymptotically normal when the true parameter  $\theta < 0$  but not when  $\theta = 0$  (the unit root case). For example, consider the pooled OLS panel estimator  $\hat{\theta}^+$  that corrects for the bias due to the time series serial correlation in  $\varepsilon_{it}$ . Moon and Phillips (2000) showed that

$$\hat{\theta}^+ \rightarrow_p F(\theta),$$

where  $F(\theta) \neq \theta$ . The inconsistency arises because in the regression

$$\hat{\varepsilon}_{it} = \left(1 - \frac{\hat{\theta}^+}{T}\right) \hat{\varepsilon}_{it-1},$$

where

$$\begin{aligned} \hat{\varepsilon}_{it} &= y_{it} - \hat{\lambda}_{i0} - \hat{\lambda}_{i1} t \\ \begin{pmatrix} \hat{\lambda}_{i0} \\ \hat{\lambda}_{i1} \end{pmatrix} &= \begin{pmatrix} T & \sum_{t=1}^T t \\ \sum_{t=1}^T t & \sum_{t=1}^T t^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{t=1}^T y_{it} \\ \sum_{t=1}^T t y_{it} \end{pmatrix}, \end{aligned}$$

(here  $\hat{\lambda}_{i0} + \hat{\lambda}_{i1} t$  is the OLS estimator of the incidental trend) the detrended regressor  $\hat{\varepsilon}_{it-1}$  ends up being correlated with the error term  $u_{it}$  even after correcting for the bias due to the serial correlation in  $u_{it}$ .

The first estimator they proposed to resolve this difficulty is to invert the bias function  $F(\theta)$  as

$$\tilde{\theta} = F^{-1}(\hat{\theta}^+).$$

Through numerical analysis, Moon and Phillips (2000) showed that  $F^{-1}(\bullet)$  is well defined unless  $\theta = 0$ . A second estimation method they proposed is to correct for the asymptotic bias of  $\hat{\theta}^+$  as an approximately linear function of the parameter  $\theta$ . They showed that both estimators are consistent and asymptotically normal when  $\theta < 0$ , but these estimators become invalid when  $\theta = 0$ .

To overcome the problem at  $\theta = 0$ , Moon and Phillips (2004) considered an estimation method based on two asymptotic moment conditions. The first moment condition was considered in Moon and Phillips (2000). Let  $\omega_T(c)$  and  $\lambda_T(c)$  be the biases of the score functions of the concentrated likelihoods of the OLS detrended panel and GLS detrended panel, respectively:

$$\begin{aligned} \omega_T(c) &= \mathbb{E} \left[ \frac{1}{T} \sum_{t=1}^T \left( \hat{\varepsilon}_{it} - \left(1 - \frac{c}{T}\right) \hat{\varepsilon}_{it-1} \right) \hat{\varepsilon}_{it-1} \right] \\ \lambda_T(c) &= \mathbb{E} \left[ \frac{1}{T} \sum_{t=1}^T u_{it} \left( \theta, \hat{\lambda}_i(c) \right) \varepsilon_{it-1} \left( \hat{\lambda}_i(c) \right) \right], \end{aligned}$$

where

$$\begin{aligned} u_{it}(c, \lambda_i) &= \Delta_c y_{it} - \Delta_c \lambda_{i0} - \lambda_{i1} \Delta_c t \\ \varepsilon_{it-1}(\lambda_i) &= y_{it-1} - \lambda_{i0} - \lambda_{i1}(t-1). \end{aligned}$$

The two moment conditions that Moon and Phillips (2004) considered are

$$\begin{aligned} m_{1,iT}(c) &= \frac{1}{T} \sum_{t=1}^T \left( \hat{\varepsilon}_{it} - \left(1 - \frac{c}{T}\right) \hat{\varepsilon}_{it-1} \right) \hat{\varepsilon}_{it-1} - \omega_T(c) \\ m_{2,iT}(c) &= \frac{1}{T} \sum_{t=1}^T u_{it} \left( c, \hat{\lambda}_i(c) \right) \varepsilon_{it-1} \left( \hat{\lambda}_i(c) \right) - \lambda_T(c), \end{aligned}$$

and the corresponding GMM estimator is

$$\hat{\theta}_{GMM} = \arg \min_{c \leq 0} \left( \frac{1}{N} \sum_{i=1}^N m_{iT}(c) \right)' \hat{W} \left( \frac{1}{N} \sum_{i=1}^N m_{iT}(c) \right),$$

where  $m_{iT}(c) = (m_{1,iT}(c), m_{2,iT}(c))'$  and  $\hat{W} \rightarrow_p W > 0$ . Moon and Phillips (2004) showed that when  $\theta < 0$ , the GMM estimator  $\hat{\theta}_{GMM}$  is  $\sqrt{N}$ -consistent and asymptotically normal as  $N, T \rightarrow \infty$  with  $\frac{N}{T} \rightarrow 0$ . When  $\theta = 0$ ,  $\hat{\theta}_{GMM}$  is  $N^{1/6}$ -consistent and has a nonstandard limiting distribution. So,  $\theta$  is consistently estimable at  $\theta = 0$ . However, an important implication of the limit theory is that the presence of incidental trends complicates the identification of a unit root in panels - making it difficult to discriminate locally in the vicinity of unity. This difficulty motivates the investigation of the power of panel unit root tests, as in Moon, Perron, and Phillips (2007).

#### 4.4 TESTING FOR UNIT ROOTS WITH INCIDENTAL PARAMETERS

---

A large literature has developed in testing for unit roots in dynamic panels over the past two decades. Information from cross-section observations should be useful in helping to improve inference regarding the long-run properties of data relative to the standard time series tests. However, as the work on local asymptotics described in the previous section makes clear, there are still substantial difficulties in getting good discriminatory power in the immediate vicinity of unity. Our discussion in what follows will focus on the consequences of incidental parameters on the testing problem. Readers should refer to Breitung and Pesaran (2008) for a thorough survey of the area.

We consider a component model where the autoregressive parameter is allowed to be heterogeneous:

$$y_{it} = \lambda_{0i} + \lambda_{1i}t + u_{it} \quad (4.31)$$

$$u_{it} = \rho_i u_{it-1} + \varepsilon_{it},$$

where the initial conditions are  $y_{i,0} = 0$  for all  $i$ . For expositional simplicity, we start by assuming that  $\varepsilon_{it}$  is potentially heteroskedastic with mean 0 and variance  $\sigma_i^2$ , but that

$\varepsilon_{it}$  is independent across  $i$  and over  $t$  with finite fourth moments. The case with serial correlation will be considered later.

The focus of interest is the problem of testing for the presence of a common unit root in the panel when both  $N$  and  $T$  are large, which we express as the null hypothesis

$$\mathbb{H}_0 : \rho_i = 1 \text{ for all } i, \quad (4.32)$$

against the alternative

$$\mathbb{H}_1 : \rho_i \neq 1 \text{ for some } i's. \quad (4.33)$$

It turns out to be convenient to consider a local alternative specification. We assume that

$$\rho(\theta_i) = 1 - \frac{\theta_i}{N^\kappa T} \quad \text{for some constant } \kappa > 0, \quad (4.34)$$

where  $\theta_i$  is a sequence of iid random variables. As we see below, the constant  $\kappa$  will depend on the nature of the incidental parameters.

Moon, Perron, and Phillips (2007) considered efficient tests for the null hypothesis of a unit root for all individuals in the panel. In terms of the local specification, the null and alternative hypotheses can be formulated as:

$$\mathbb{H}_0 : \theta_i = 0 \quad \text{for all } i, \quad (4.35)$$

against the alternative

$$\mathbb{H}_1 : \theta_i \neq 0 \quad \text{for some } i's. \quad (4.36)$$

If the alternative was a singleton, i.e. if the alternative were some specific value of the local-to-unity parameters which we may call  $c = (c_1, c_2, \dots, c_N)'$ , then by the Neyman-Pearson lemma, the most powerful test is given by the likelihood ratio. As in Elliott, Rothenberg, and Stock (1996) for the time series case, changing the values of  $c$  enables the computation of a power envelope from which we may trace out the maximum power of a test for any point alternative. Under Gaussianity, the log-likelihood is given (up to a constant) by

$$\begin{aligned} L_{NT}(c) = & -\frac{1}{2} \sum_{i=1}^N \sum_{t=1}^T \frac{1}{\sigma_i^2} \\ & \times \left\{ \Delta_{ci} y_{it} - \lambda_{i0} (1 - \rho(c_i)) - \lambda_{i1} [t - \rho(c_i)(t-1)] \right\}^2, \end{aligned} \quad (4.37)$$

and the test statistic is

$$V_{fe,NT} = 2 [L_{NT}(c) - L_{NT}(0)] - \mu(c),$$

where  $\mu(c)$  is a centering term that ensures that the statistic has mean 0. When deterministic components are included, we appeal to invariance and use the maximum of the likelihood with respect to these parameters.

One interesting result is that the value of  $\kappa$  that defines local neighborhoods is different according to the specification of the deterministic components. Hence, if  $\lambda_{i1} = 0$  and only individual intercepts are present, we find that  $\kappa = 1/2$  so that the likelihood ratio test can detect alternatives that converge to the null hypothesis of a panel unit root at the rate  $\frac{1}{\sqrt{NT}}$ . However, if individual trends are present with  $\lambda_{i1} \neq \lambda$  for all  $i$ , we find that  $\kappa = 1/4$  which means that the maximal power that can be achieved by any test is much lower for a given alternative. Thus, the presence of incidental trends reduces the potential power in discriminating between panels where all individual series have unit roots and panels where some of the individual series have highly persistent but stationary dynamics. Another interesting result is that the presence of incidental intercepts does not change the asymptotic distribution of the test statistic, i.e. the power envelope and distribution of the test statistic is the same whether  $\lambda_{i0} = 0$  or  $\lambda_{i0} \neq 0$ .

As a result, in the incidental intercepts case, Moon, Perron, and Phillips (2007) parametrize the autoregressive parameters as:

$$\rho(\theta_i) = 1 - \frac{\theta_i}{\sqrt{NT}},$$

and the null hypothesis can be written as:

$$\mathbb{H}_0 : \mathbb{E}(\theta_i) = 0$$

while the alternative hypothesis is:

$$\mathbb{H}_1 : \mathbb{E}(\theta_i) > 0$$

where the  $\theta'_i$ 's are assumed to be independent across  $i$  and lie in the bounded interval  $[0, M_\theta]$  for some  $M_\theta \geq 0$ . The centering term in the likelihood ratio statistic (4.37) is  $\mu(c) = \frac{1}{2N} \sum_{i=1}^N c_i^2$ , and its asymptotic distribution is  $N(-\mathbb{E}(c_i \theta_i), 2\mu(c))$  which reflects the impact of both the values of the local-to-unity parameters in the population ( $\theta_i$ ) and those used to set up the test and compute the likelihood ratio statistic ( $c_i$ ).

Moon, Perron, and Phillips (2007) make the above test operational by proposing an estimator for  $\sigma_i^2$  and suggesting a common-point-optimal test in which one chooses all  $c_i$  to be the same. In that case, the asymptotic power of a test at level  $\alpha$  is  $\Phi\left(\frac{\mathbb{E}(\theta_i)}{\sqrt{2}} - \bar{z}_\alpha\right)$  where  $\bar{z}_\alpha$  is the  $(1 - \alpha)$  quantile from the standard normal distribution. A remarkable feature of this result is that power is independent of the value of the common  $c_i$  chosen, in contrast to the time series case where power depends on the choice of the local to unity parameter used to construct the test.

In the incidental trends case, that is when  $\lambda_{i1} \neq \lambda$  for all  $i$ , as already mentioned, the local neighborhoods must shrink to 1 at a slower rate and the autoregressive parameters are parametrized as (see Ploberger and Phillips, 2002):

$$\rho(\theta_i) = 1 - \frac{\theta_i}{N^{1/4} T}.$$

In this instance, it is possible to allow for some explosive behavior and the assumption made on  $\theta_i$  is that it is contained in a bounded interval  $[-M_{l\theta}, M_{u\theta}]$  where  $M_{l\theta}$  and  $M_{u\theta}$  are non-negative constants. Under this assumption, the null hypothesis of a panel unit root can be expressed as:

$$\mathbb{H}_0 : \mathbb{E}(\theta_i^2) = 0$$

while the alternative hypothesis is:

$$\mathbb{H}_1 : \mathbb{E}(\theta_i^2) > 0.$$

The use of the second moment is necessary since the requirement that  $\mathbb{E}(\theta_i) = 0$  does not imply that all  $\theta_i$ 's are 0.

Under this scenario, the centering term in the likelihood ratio statistic (4.37) is

$$\mu(c) = -\frac{1}{N^{1/4}} \sum_{i=1}^N c_i - \frac{1}{N^{1/2}} \sum_{i=1}^N c_i^2 \omega_{p2T} - \frac{1}{N} \sum_{i=1}^N c_i^4 \omega_{p4T},$$

where

$$\begin{aligned} \omega_{p2T} &= -\frac{1}{T} \sum_{t=1}^T \frac{t-1}{T} + \frac{2}{T} \sum_{t=1}^T \frac{t}{T} \left( \frac{t-1}{T} \right) - \frac{1}{3}, \\ \omega_{p4T} &= \frac{1}{T^2} \sum_{t=1}^T \sum_{s=1}^T \frac{t-1}{T} \frac{s-1}{T} \min\left(\frac{t-1}{T}, \frac{s-1}{T}\right) - \frac{2}{3} \frac{1}{T} \sum_{t=1}^T \left( \frac{t-1}{T} \right)^2 + \frac{1}{9}. \end{aligned}$$

The asymptotic distribution is now  $N(-\frac{1}{90}\mathbb{E}(c_i^2\theta_i^2), \frac{1}{45}\mathbb{E}(c_i^4))$ . A common point optimal test that is constructed by imposing the same value of  $c_i$  for all units would have power  $\Phi(\frac{1}{6\sqrt{5}}\mathbb{E}(\theta_i^2) - \bar{z}_\alpha)$ , again independent of the particular choice of common local-to-unity parameter used to set up the test. However, this test has power below the power envelope unless the alternative hypothesis is homogeneous, i.e. all  $\theta$ 's are the same. Based on simulation results, Moon, Perron, and Phillips (2007) recommend the use of  $c_i = 1$  to construct the test.

Moon, Perron, and Phillips (2014) extend these results to the case where  $\varepsilon_{it}$  is serially correlated. They show that optimal tests under serial correlation involve two adjustments to the above statistics. First, the error variance in the denominator must be replaced by the corresponding long-run variance denoted  $\omega_i^2$ , and, second, the centering of the statistic must be adjusted to account for the correlation between the error and the lagged dependent variable as in the standard Phillips-Perron statistic in the time series case. This adjustment depends on the one-sided long-run variance  $\Lambda_i = \sum_{j=1}^{\infty} \mathbb{E}(\varepsilon_{it}\varepsilon_{t-j})$ .

In the case of incidental intercepts, the new centering is  $\mu(c) = \frac{1}{2N} \sum_{i=1}^N c_i^2 + \frac{2}{\sqrt{N}} \sum_{i=1}^N c_i \frac{\Lambda_i}{\omega_i^2}$ , whereas with incidental trends, the centering is

$$\begin{aligned}\mu(c) = & -\frac{1}{N^{1/4}} \sum_{i=1}^N c_i \frac{\sigma_i^2}{\omega_i^2} - \frac{1}{N^{1/2}} \sum_{i=1}^N c_i^2 \omega_{p2T} \\ & - \frac{1}{N} \sum_{i=1}^N c_i^4 \omega_{p4T}.\end{aligned}$$

In both cases, the new centering reduces to the centering term in Moon et al. (2007) as  $\Lambda_i = 0$  and  $\sigma_i^2 = \omega_i^2$  if no serial correlation is present.

## 4.5 NONLINEAR DYNAMIC PANELS

The incidental parameter problem in nonlinear panel regressions where the incidental parameters are not additively separable is a further challenging problem, in particular, when  $T$  is fixed. Surveys of the early literature are available in Arellano and Honore (2001) and Hsiao (2003). For more recent research results, readers may refer to Arellano and Hahn (2006) and Arellano and Bonhomme (2011). This section provides a selective survey on more recent developments of nonlinear dynamic panel regression research. In particular, we focus on estimating the common parameter<sup>9</sup> of the nonlinear dynamic panel regression model and on bias reduction methods.

### 4.5.1 Concentrated Likelihood Approach

Hahn and Newey (2004), Arellano and Hahn (2006), and Hahn and Kuersteiner (2011) extended the idea of Hahn and Kuersteiner (2002) to nonlinear panel regression models and characterize the bias due to the incidental parameters using joint asymptotics as  $N, T \rightarrow \infty$  with  $\frac{N}{T} \rightarrow \kappa^2$ . In particular, Hahn and Kuersteiner (2011) allowed the panel regression to be nonlinear and dynamic in which the individual fixed effect parameters and the lagged dependent variables enter the regression model nonlinearly.

Suppose that the observed panel data is  $x_{it}$ . The typical composition of  $x_{it}$  is  $x_{it} = (y_{it}, y_{it-1}, z_{it})$ , where  $y_{it}$  is the dependent variable in the panel regression and  $(y_{it-1}, z_{it})$  are regressors. Let  $\theta$  be the common parameters of interest, including the coefficient of the lagged dependent regressor  $y_{it-1}$ , and  $\lambda_i$  be the individual fixed effects. Let  $(\theta_0, \lambda_1^0, \dots, \lambda_N^0)$  be the true parameters. Consider the fixed effect estimator that maximizes some objective function

$$\left(\hat{\theta}, \hat{\lambda}_1, \dots, \hat{\lambda}_N\right) = \arg \max \sum_{i=1}^N \sum_{t=1}^T \psi(x_{it}; \theta, \lambda_i).$$

Denote

$$l_{it}(\theta, \lambda_i) = \psi(x_{it}; \theta, \lambda_i).$$

An example of  $l_{it}(\theta, \lambda_i)$  is the conditional log likelihood function. Define

$$\begin{aligned} w_{it}(\theta, \lambda_i) &= \frac{\partial l_{it}(\theta, \lambda_i)}{\partial \theta}, \quad w_{it} = w_{it}(\theta_0, \lambda_i^0), \quad v_{it}(\theta, \lambda_i) = \frac{\partial l_{it}(\theta, \lambda_i)}{\partial \lambda_i}, \quad v_{it} = v_{it}(\theta_0, \lambda_i^0), \\ w_{it}^\lambda &= \frac{\partial w_{it}(\theta_0, \lambda_i^0)}{\partial \lambda_i}, \quad v_{it}^\lambda = \frac{\partial v_{it}(\theta_0, \lambda_i^0)}{\partial \lambda_i} \\ V_{2it}(\theta, \lambda_i) &= v_{it}^2(\theta, \lambda_i) + \frac{\partial v_{it}(\theta, \lambda_i)}{\partial \lambda_i}, \quad W_{it}(\theta, \lambda_i) = w_{it}(\theta, \lambda_i) \\ &\quad - v_{it}(\theta, \lambda_i) \mathbb{E}[v_{it}^\lambda]^{-1} \mathbb{E}[w_{it}^\lambda] \\ W_{it} &= W_{it}(\theta_0, \lambda_i^0), \quad W_{it}^\lambda = \frac{\partial W_{it}(\theta_0, \lambda_i^0)}{\partial \lambda_i}, \quad W_{it}^{\lambda\lambda} = \frac{\partial^2 W_{it}(\theta_0, \lambda_i^0)}{\partial \lambda_i^2}. \\ \mathcal{I}_i &= -\mathbb{E}\left[\frac{\partial W_{it}(\theta_0, \lambda_i^0)}{\partial \theta}\right]. \end{aligned}$$

Hahn and Kuersteiner (2011) showed that in a general nonlinear dynamic panel regression model, as  $N, T \rightarrow \infty$  with  $\frac{N}{T} \rightarrow \kappa^2$ ,

$$\sqrt{NT}(\hat{\theta} - \theta_0) \Rightarrow N(\kappa B, \mathcal{I}^{-1} \Omega \mathcal{I}^{-1}), \quad (4.38)$$

where

$$\begin{aligned} \mathcal{I} &= \lim_N \frac{1}{N} \sum_{i=1}^N \mathcal{I}_i, \quad \Omega = \lim_N \frac{1}{N} \sum_{i=1}^N \text{Var}\left(\frac{1}{\sqrt{T}} \sum_{t=1}^T W_{it}\right) \\ B &= -\mathcal{I}^{-1} \left( \lim_N \frac{1}{N} \sum_{i=1}^N \frac{\sum_{l=-\infty}^{\infty} \text{Cov}(v_{it}, W_{it-l}^\lambda)}{\mathbb{E}[v_{it}^\lambda]} \right. \\ &\quad \left. - \frac{1}{2} \lim_N \frac{1}{N} \sum_{i=1}^N \frac{\mathbb{E}[W_{it}^{\lambda\lambda}] \sum_{l=-\infty}^{\infty} \text{Cov}(v_{it}, v_{it-l})}{\mathbb{E}[v_{it}^\lambda]^2} \right). \end{aligned}$$

Notice that in the dynamic linear regression model (4.4), we have

$$B = \frac{-(1-\rho_0)}{1-\rho_0^2} = -\frac{1}{1+\rho_0},$$

as shown in (4.20). Hahn and Kuersteiner (2011) also provide a consistent estimator  $\hat{B}$  of the bias  $B$ , and propose a bias corrected estimator

$$\hat{\theta}^+ = \hat{\theta} - \frac{1}{T} \hat{B}.$$

The intuition underlying (4.38) is as follows. Consider an infeasible estimator  $\tilde{\theta}$  based on  $\hat{\lambda}_i(\theta_0)$  rather than  $\hat{\lambda}_i(\hat{\theta})$ , where

$$0 = \sum_{i=1}^N \sum_{t=1}^T W_{it} (\tilde{\theta}, \hat{\lambda}_i(\theta_0)).$$

Then, conventional first order Taylor approximation yields

$$\tilde{\theta} - \theta_0 \simeq \left( \frac{1}{N} \sum_{i=1}^N \mathcal{I}_i \right)^{-1} \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T W_{it} (\theta_0, \hat{\lambda}_i(\theta_0)).$$

Applying a second order Taylor series approximation to  $W_{it}(\theta_0, \hat{\lambda}_i(\theta_0))$ , we have

$$\begin{aligned} \frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T W_{it} (\theta_0, \hat{\lambda}_i(\theta_0)) &\simeq \frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T W_{it} \\ &+ \frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T W_{it}^\lambda (\hat{\lambda}_i(\theta_0) - \lambda_i^0) \\ &+ \frac{1}{2\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T W_{it}^{\lambda\lambda} (\hat{\lambda}_i(\theta_0) - \lambda_i^0)^2. \end{aligned}$$

Since  $\hat{\lambda}_i(\theta_0) - \lambda_i^0 \simeq -\frac{1}{T} \sum_{t=1}^T v_{it} (\mathbb{E}(v_{it}^\lambda))^{-1}$ , we have

$$\begin{aligned} \sqrt{NT} (\tilde{\theta} - \theta_0) &\simeq \left( \frac{1}{N} \sum_{i=1}^N \mathcal{I}_i \right)^{-1} \left( \frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T W_{it} \right) \\ &- \sqrt{\frac{N}{T}} \left( \frac{1}{N} \sum_{i=1}^N \mathcal{I}_i \right)^{-1} \frac{1}{N} \sum_{i=1}^N \left[ \frac{\sum_{t=1}^T v_{it}}{\sqrt{T} \mathbb{E}(v_{it}^\lambda)} \right] \\ &\times \left[ \frac{1}{\sqrt{T}} \sum_{t=1}^T \left( W_{it}^\lambda - \frac{\mathbb{E}[W_{it}^{\lambda\lambda}]}{2\mathbb{E}(v_{it}^\lambda)} v_{it} \right) \right], \end{aligned}$$

and it follows that

$$\left( \frac{1}{N} \sum_{i=1}^N \mathcal{I}_i \right)^{-1} \left( \frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T W_{it} \right) \Rightarrow N(0, \mathcal{I}^{-1} \Omega \mathcal{I}^{-1})$$

and

$$-\sqrt{\frac{N}{T}} \left( \frac{1}{N} \sum_{i=1}^N \mathcal{I}_i \right)^{-1} \frac{1}{N} \sum_{i=1}^N \left[ \frac{\sum_{t=1}^T v_{it}}{\sqrt{T} \mathbb{E}(v_{it}^\lambda)} \right] \left[ \frac{1}{\sqrt{T}} \sum_{t=1}^T \left( W_{it}^\lambda - \frac{\mathbb{E}[W_{it}^{\lambda\lambda}]}{2\mathbb{E}(v_{it}^\lambda)} v_{it} \right) \right] \rightarrow_p \kappa B.$$

### 4.5.2 Integrated Likelihood Approach

The fixed effect estimator concentrates out the incidental parameters in the objective function. Another way to deal with the incidental parameters is to integrate out with a certain weight function or a prior for the parameters. This approach was studied by Arellano and Bonhomme (2009).

Using the notation in this section, suppose that  $l_{it}(\theta, \lambda)$  is the conditional log likelihood function. Let  $l_i(\theta, \lambda) = \frac{1}{T} \sum_{t=1}^T \ln l_{it}(\theta, \lambda)$ . Let  $\pi_i(\lambda_i | \theta)$  be a conditional prior distribution on the individual fixed effect given  $\theta$ . Here the dependence of the  $\pi_i$  on the index  $i$  allows for possible conditioning on strictly exogenous regressors and initial conditions. Assume that the support of  $\pi_i(\lambda_i | \theta)$  contains an open neighborhood of the true parameter  $(\theta_0, \lambda_i^0)$  and  $\sup_i \ln \pi_i(\lambda_i | \theta) = O(1)$  for all  $\theta$  and  $\lambda_i$  as  $T \rightarrow \infty$ .

The fixed effect estimator  $\hat{\theta}$  maximizes the concentrated likelihood function

$$\hat{\theta} = \arg \max \sum_{i=1}^N l_i^c(\theta),$$

where

$$l_i^c(\theta) = l_i\left(\theta, \hat{\lambda}_i(\theta)\right).$$

An alternative objective function to the concentrated likelihood is the individual log integrated likelihood given by

$$l_i^I(\theta) = \frac{1}{T} \ln \int \exp[Tl_i(\theta, \lambda_i)] \pi_i(\lambda_i | \theta) d\lambda_i.$$

This likelihood could be considered subjective Bayesian with a joint prior that is separable in the individual effects. The target likelihood is defined as

$$\bar{l}_i(\theta) = l_i\left(\theta, \bar{\lambda}_i(\theta)\right),$$

where

$$\bar{\lambda}_i(\theta) = \arg \max_{\lambda_i} p \lim_T l_i(\theta, \lambda_i).$$

Notice that the concentrated and target likelihood functions can be regarded as integrated likelihood functions with respect to the priors

$$\bar{\pi}_i(\lambda_i | \theta) = \delta(\lambda_i - \bar{\lambda}_i(\theta)) \text{ and } \pi_i^c(\lambda_i | \theta) = \delta(\lambda_i - \hat{\lambda}_i(\theta)),$$

where  $\delta(\cdot)$  is the Dirac delta function. Here  $\pi_i^c(\lambda_i|\theta)$  is a sample counterpart of  $\bar{\pi}_i(\lambda_i|\theta)$ . Define  $v_i(\theta, \lambda_i) = \frac{\partial l_i(\theta, \lambda_i)}{\partial \lambda_i}$  and  $v_i^{\lambda_i}(\theta, \lambda_i) = \frac{\partial v_i(\theta, \lambda_i)}{\partial \lambda_i}$ ,  $v_i^\theta(\theta, \lambda_i) = \frac{\partial v_i(\theta, \lambda_i)}{\partial \theta}$ , and  $v_i^{\lambda_i \lambda_i}(\theta, \lambda_i) = \frac{\partial^2 v_i(\theta, \lambda_i)}{\partial \lambda_i^2}$ .

Arellano and Bonhomme (2009) showed that the bias of the integrated likelihood is

$$\mathbb{E}\left[l_i^I(\theta) - \bar{l}_i(\theta)\right] = \text{const} + \frac{\beta_i(\theta)}{T} + O\left(\frac{1}{T^2}\right),$$

where

$$\begin{aligned}\beta_i(\theta) &= \frac{1}{2} \left\{ \mathbb{E}\left[-v_i^{\lambda_i}(\theta, \bar{\lambda}_i(\theta))\right] \right\}^{-1} \mathbb{E}\left[Tv_i^2(\theta, \bar{\lambda}_i(\theta))\right] \\ &\quad - \frac{1}{2} \ln \mathbb{E}\left[-v_i^{\lambda_i}(\theta, \bar{\lambda}_i(\theta))\right] + \ln \pi_i(\bar{\alpha}_i(\theta)|\theta).\end{aligned}$$

Let  $b_i(\theta_0) = \frac{\partial}{\partial \theta}|_{\theta_0} \beta_i(\theta)$  be the first order bias of the integrated score evaluated at the true value. Arellano and Bonhomme defined a prior family as bias reducing or robust, if and only if

$$b_\infty(\theta_0) = p \lim_N \frac{1}{N} \sum_{i=1}^N b_i(\theta_0) = o(1).$$

Since bias reduction of the moment equation (score function) implies bias reduction of the estimator, for a robust prior family, the mode of the integrated likelihood

$$\hat{\theta}_I = \arg \max_{\theta} \sum_{i=1}^N l_i^I(\theta)$$

has zero first-order bias, that is,

$$p \lim_{N \rightarrow \infty} \hat{\theta}_I = \theta_0 + o\left(\frac{1}{T}\right).$$

Arellano and Bonhomme showed that a prior  $\pi_i$  is bias-reducing if

$$\begin{aligned}\frac{\partial}{\partial \theta} \Big|_{\theta_0} \ln \pi_i(\bar{\alpha}_i(\theta)|\theta) &= \frac{\partial}{\partial \theta} \Big|_{\theta_0} \ln \left( \mathbb{E}\left[-v_i^{\lambda_i}(\theta, \bar{\lambda}_i(\theta))\right] \{\mathbb{E}[T v_i^2(\theta, \bar{\lambda}_i(\theta))]\}^{-1/2} \right) \\ &\quad + O\left(\frac{1}{T}\right).\end{aligned}$$

They suggested the following data-dependent prior:

$$\pi_i^R(\lambda_i|\theta) \propto \mathbb{E}\left[\widehat{-v_i^{\lambda_i}}(\theta, \lambda_i)\right] \left\{\mathbb{E}\left[\widehat{v_i^2}(\theta, \lambda_i)\right]\right\}^{-1/2},$$

where the hat denotes consistent estimators as  $T \rightarrow \infty$ . In the pseudo likelihood setting, they suggested

$$\pi_i^R(\lambda_i|\theta) \propto \mathbb{E}\left[-\widehat{v_i^{\lambda_i}}(\theta, \lambda_i)\right]^{1/2} \exp\left(-\frac{T}{2}\left\{\mathbb{E}\left[-\widehat{v_i^{\lambda_i}}(\theta, \lambda_i)\right]\right\}^{-1}\mathbb{E}\left[\widehat{v_i^2}(\theta, \lambda_i)\right]\right).$$

Under these robust priors, as  $N, T \rightarrow \infty$  with  $\frac{N}{T} \rightarrow \kappa^2$ ,

$$\sqrt{NT}\left(\hat{\theta}_I - \bar{\theta}\right) = o_p(1),$$

where  $\bar{\theta} = \arg \max_{\theta} \sum_{i=1}^N \bar{l}_i(\theta)$ . Therefore,

$$\sqrt{NT}\left(\hat{\theta}_I - \theta_0\right) = \sqrt{NT}\left(\bar{\theta} - \theta_0\right) + o_p(1).$$

## 4.6 ORDER SELECTION IN DYNAMIC PANELS

---

As indicated in the Introduction, the presence of incidental parameters further complicates model selection. Standard procedures such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), or Hannan-Quinn (HQ) suppose that the number of parameters is finite or grows slowly as sample size increases. The presence of incidental parameters violates this assumption, and Stone (1979) showed that BIC is inconsistent with incidental parameters. The same problem occurs in dynamic panels with fixed effects. The reason why BIC is inconsistent in this context is that the usual Laplace approximation does not hold for an infinity of parameters, thereby requiring an infinite dimensional integration.

It is particularly remarkable that BIC is also inconsistent as an order estimator in dynamic panels without fixed effects as shown recently by Han, Phillips, and Sul (2015). The reason for this curious finding is that in a panel as  $N \rightarrow \infty$  there are an infinity of observations relevant to models with a lower lag order than those with higher lag order. It is this discrepancy in the use of the data that can lead to inconsistency in conventional order selectors like BIC. Han, Phillips, and Sul show that the overestimation probability in the BIC order selector is 50%. The remedy is to raise the penalty in the BIC criterion to take account of this difference.

How to perform model selection with incidental parameters remains an ongoing issue in the panel literature although some approaches have been developed recently. Berger, Ghosh, and Mukhopadhyay (2003) have shown that the choice of priors is important with incidental parameters, and that a suitable choice can make BIC consistent in this context. They also propose a different approximation that leads to a criterion that is consistent in model selection.

Alternatively, in a Kullback-Leibler approach, one must increase the penalty used because standard methods impose penalties that are too small if the problem of incidental parameters is not taken into account. For example, Lee (2014) considers the problem of selecting a model among a set of candidate models that may not contain the true one. Suppose that there are two sets of parameters, those of interest denoted by  $\psi_k$  with dimension  $r_k$  for model  $k$  and the incidental parameters  $\lambda_i$ . He defines a set of information criteria for model  $k$  of the form:

$$LIC^h(\mathcal{M}^k) = -\frac{2}{NT} \sum_{i=1}^N \sum_{t=1}^T \log f_{it}(z_{it}; \hat{\psi}_M^k, \hat{\lambda}_i^k) + r_k \frac{h(N, T)}{NT} + \frac{2}{NT} \sum_{i=1}^N M_i(\hat{\psi}_M^k)$$

where  $h(N, T)$  is a penalty function that differentiates the criteria. The estimator  $\hat{\psi}_M$  is based on the modified profile likelihood that corrects for the fact that the score of the profile likelihood is not 0 due to the incidental parameters.

The penalty in these criteria has two components. The first one is the same as in the standard AIC and BIC and is proportional to the number of parameters. For the AIC, one would set  $h(N, T) = 2$  while for BIC, one would set  $h(N, T) = \log(NT)$ . The second term is the contribution associated with the presence of incidental parameters. This term is always positive so that a more severe penalty is imposed relative to the standard criteria.

## 4.7 CONCLUSIONS

---

Practical empirical work with dynamic panel models offers many opportunities for learning about individual behavior over time and the common elements that figure in that behavior. This work also faces many challenges, ranging from the impact of individual effects and incidental trends on estimation bias in short wide panels, through to the difficulties of treating nonlinear dynamic models with nonseparable fixed effects, and the problems of inconsistency in dynamic model specification. This chapter has overviewed some of the established methodology in the field, the ground that has been won in developing a theory of inference for dynamic panel modeling, as well as some exciting ongoing research that seeks to address some of the many challenges that remain.

## ACKNOWLEDGMENTS

---

The authors thank the referee and Editor for helpful comments on the original version of this chapter. Moon acknowledges that this work was supported by the National Research Foundation of Korea Grant funded by the Korean Government

(NRF2014S1A5A8012177). Perron acknowledges financial support from SSHRC and FQRSC-ANR. Phillips acknowledges NSF support under Grant Nos. SES 09-56687 and 12-58258.

## NOTES

---

1. See Lancaster (2000) for an historical overview.
2. For Bayesian analysis of dynamic panels, one could refer to Lancaster (2002) for example.
3. See also Han and Phillips (2010) and Han, Phillips, and Sul (2011).
4. Hahn and Kuersteiner (2002) also considered the more general case where the regression errors may be conditionally heteroskedastic.
5. Pesaran (2006) studied a common correlated random effects model that allows for heterogenous regression coefficients.
6. See, for example, Han and Phillips (2006) and Newey and Windmeijer (2009) for a treatment of the problem of many IVs.
7. The bias term  $B_2$  in Moon and Weidner (2010) is zero when the regressor is a lagged dependent variable.
8. Phillips and Sul (2007) also proposed a median unbiased estimator.
9. Another parameter of interest that has been widely studied in nonlinear panel regression models is the average marginal effect (or average treatment effect). See, for example, Fernandez-Val (2009), Chernazhukov, Fernandez-Val, Hahn, and Newey (2013).

## REFERENCES

---

- Ahn, S.C., Lee, Y.H., and P. Schmidt (2001). GMM Estimation of Linear Panel Data Models with Time-varying Individual Effects, *Journal of Econometrics*, 101, 219–255.
- Ahn, S., and P. Schmidt (1995). Efficient Estimation of Models for Dynamic Panel Data, *Journal of Econometrics*, 68, 5–28.
- Alvarez, J., and Arellano, M. (2003). The Time Series and Cross-Section Asymptotics of Dynamic Panel Data Estimators. *Econometrica*, 71, 1121–1159.
- Anderson, T.W. and C. Hsiao (1982). Formulation and Estimation of Dynamic Models Using Panel Data, *Journal of Econometrics*, 18, 47–82.
- Arellano, M. (2003). Panel data econometrics. Oxford: Oxford University Press.
- Arellano, M., and S. Bond (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations, *Review of Economic Studies*, 58, 277–297.
- Arellano, M., and S. Bonhomme (2009). Robust Priors in Nonlinear Panel Data Models, *Econometrica*, 77, 489–539.
- Arellano, M., and S. Bonhomme (2011). Nonlinear Panel Data Analysis, *Annual Review of Economics*, 3, 395–424.
- Arellano, M., and O. Bover (1995). Another Look at the Instrumental Variable Estimation of Error-Component Models, *Journal of Econometrics*, 68, 29–51.

- Arellano, M., and J. Hahn (2006). Understanding Bias in Nonlinear Panel Models: Some Recent Developments, in R. Blundell, W.K. Newey, and T. Persson, eds. *Advances in Economics and Econometrics*, Cambridge: Cambridge University Press, 381–409.
- Arellano, M., and B. Honore (2001). Panel Data Models: Some Recent Development in J.J. Heckman and E.E. Leamer, eds., *Handbook of Econometrics*, vol. 5, Elsevier, Amsterdam, 3229–3296.
- Bai, J. (2009). Panel Data Models with Interactive Fixed Effects. *Econometrica*, 67, 1341–1384.
- Baltagi, B.H (2008). *Econometric Analysis of Panel Data (Fourth Edition)*, John Wiley and Sons, Chichester.
- Berger, J.O., J.K. Ghosh, and N. Mukhopadhyay, (2003). Approximations and Consistency of Bayes Factors as the Model Dimension Grows. *Journal of Statistical Planning and Inference*, 112, 241–258.
- Blundell, R., and S. Bond (1998). Initial Conditions and Moment Restrictions in Dynamic Panel Data Models, *Journal of Econometrics*, 87, 115–143.
- Breitung, J., and M.H. Pesaran (2008). Unit Roots and Cointegration in Panels, in L. Matyas and P. Sevestre, *The Econometrics of Panel Data (Third Edition)*, Springer-Verlag, Berlin.
- Casella, G., F.J. Girón, M.L. Martínez, and E. Moreno (2009). Consistency of Bayesian Procedures for Variable Selection. *Annals of Statistics*, 37, 1207–1228.
- Casella, G., and E. Moreno (2006). Objective Bayesian variable selection. *Journal of the American Statistical Association*, 101, 157–167.
- Casella G., and E. Moreno (2009). Assessing robustness of intrinsic test of independence in two-way contingency tables. *Journal of the American Statistical Association*, 104, 1261–1271.
- Chamberlain, G. (1982). Multivariate Regression Models for Panel Data, *Journal of Econometrics*, 18, 5–46.
- Chamberlain, G. (1984). Panel Data, in Z. Griliches and M. Intriligator, eds., *Handbook of Econometrics, Volume 2*, Amsterdam: North-Holland, 1247–1318.
- Chernozhukov, V., Fernandez-Val, I., Hahn, J., and Newey, W. (2013). Average and quantile effects in nonseparable panel models. *Econometrica*, 81, 535–580.
- Dhaene, G., and K. Jochmans (2012). Split-Panel Jackknife Estimation of Fixed-Effect Models, Working Paper.
- Elliott, G., T. Rothenberg, and J. Stock (1996). Efficient Tests for an Autoregressive Unit Root, *Econometrica*, 64, 813–836.
- Fernandez-Val, I. (2009). Fixed effects estimation of structural parameters and marginal effects in panel probit models. *Journal of Econometrics*, 150, 71–85.
- Gourieroux, C., P.C.B. Phillips, and J. Yu (2010). Indirect Inference for Dynamic Panel Models, *Journal of Econometrics*, 157, 68–77.
- Hahn, J., and G.M. Kuersteiner (2002). Asymptotically Unbiased Inference for a Dynamic Panel Model with Fixed Effects when Both  $n$  and  $T$  are Large, *Econometrica*, 70, 1639–1657.
- Hahn, J. and G.M. Kuersteiner (2011). Bias Reduction for Dynamic Nonlinear Panel Models with Fixed Effects, *Econometric Theory*, 27, 1152–1191.
- Hahn, J., and H.R. Moon (2006). Reducing Bias of MLE in a Dynamic Panel Model, *Econometric Theory*, 22, 499–512.

- Hahn, J., and W. Newey (2004). Jackknife and Analytical Bias Reduction for Nonlinear Panel Models, *Econometrica*, 72, 1295–1319.
- Han, C., and P.C.B. Phillips (2006). GMM with Many Moment Conditions, *Econometrica*, 74, 147–192.
- Han, C., and P.C.B. Phillips (2010). GMM Estimation for Dynamic Panels with Fixed Effects and Strong Instruments at Unity, *Econometric Theory*, 26, 119–151.
- Han, C., P.C.B. Phillips, and D. Sul. (2011). Uniform Asymptotic Normality in Stationary and Unit Root Autoregression. *Econometric Theory* 27, 1117–1151.
- Han, C., P.C.B. Phillips, and D. Sul (2014). X-Differencing and Dynamic Panel Model Estimation. *Econometric Theory*, 201–251.
- Han, C., P.C.B. Phillips, and D. Sul (2015). Lag Length Selection in Panel Autoregression. Working paper, Yale University.
- Holtz-Eakin, D., W. Newey, and H. Rosen (1988). Estimating Vector Autoregressions with Panel Data, *Econometrica*, 56, 1371–1395.
- Hsiao, C. (2003). *Analysis of Panel Data*, Cambridge: Cambridge University Press.
- Kiefer, N. (1980). A Time Series-Cross Section Model with Fixed Effects with an Intertemporal Factor Structure, Mimeo, Department of Economics, Cornell University.
- Lancaster, T. (2000). The Incidental Parameter Problem Since 1948, *Journal of Econometrics*, 95, 391–413.
- Lee, Y.H. (1991). Panel Data Models with Multiplicative Individual and Time Effects: Application to Compensation and Frontier Production Functions, Ph.D. Dissertation, Michigan State University.
- Lee, Y. (2014). Model selection in the presence of incidental parameters. Unpublished manuscript, University of Michigan.
- Moon, H.R., B. Perron, and P.C.B. Phillips (2007). Incidental Trends and the Power of Panel Unit Root Tests, *Journal of Econometrics*, 141, 416–459.
- Moon, H.R., B. Perron, and P.C.B. Phillips (2014). Point Optimal Econometrics Journal (forthcoming). Panel Unit Root Tests with Serially Correlated Errors, Mimeo.
- Moon, H.R., and P.C.B. Phillips (1999). Maximum Likelihood Estimation in Panels with Incidental Trends, *Oxford Bulletin of Economics and Statistics*, 61, 711–747.
- Moon, H.R., and P.C.B. Phillips (2000). Estimation of Autoregressive Roots near Unity using Panel Data, *Econometric Theory*, 16, 927–997.
- Moon, H.R., and P.C.B. Phillips (2004). GMM Estimation of Autoregressive Roots Near Unity with Panel Data, *Econometrica*, 72, 467–522.
- Moon, H.R., and M. Weidner (2013). Dynamic Linear Panel Regression Models with Interactive Fixed Effects, Working Paper.
- Moon, H.R., and M. Weidner (2014). Linear Regression for Panel with Unknown Number of Factors as Interactive Fixed Effects, Working Paper.
- Moreno, E.F.J. Girón, and G. Casella (2010). Consistency of Objective Bayes Factors as the Model Dimension Grows, *Annals of Statistics*, 38, 4, 1937–1952.
- Newey, W. K., and Windmeijer, F. (2009). Generalized method of moments with many weak moment conditions. *Econometrica*, 77, 687–719.
- Neyman J., and E. Scott (1948). Consistent Estimates Based on Partially Consistent Observations, *Econometrica*, 16, 1–31.
- Nickell, S. (1981). Biases in Dynamic Models with Fixed Effects, *Econometrica*, 49, 1417–1426.

- Lancaster, T. (2002). Orthogonal parameters and panel data. *Review of Economic Studies*, 69, 647–666.
- Pesaran, M.H. (2006). Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure, *Econometrica*, 74, 967–1012.
- Phillips, P.C.B. (1987). Towards a Unified Asymptotic Theory for Autoregression, *Biometrika*, 74, 535–547.
- Phillips, P.C.B., and H.R. Moon (1999). Linear Regression Limit Theory for Nonstationary Panel Data, *Econometrica*, 67, 1057–1111.
- Phillips, P.C.B., and D. Sul (2003). Dynamic Panel Estimation and Homogeneity Testing under Cross Section Dependence, *Econometrics Journal*, 6, 217–259.
- Phillips, P.C.B., and D. Sul (2007). Bias in Dynamic Panel Estimation with Fixed Effects, Incidental Trends and Cross Section Dependence, *Journal of Econometrics*, 137, 162–188.
- Ploberger, W., and P.C.B. Phillips (2002). Optimal Testing for Unit Roots in Panel Data. Unpublished working paper.
- Stone, M. (1979). Comments on Model Selection Criteria of Akaike and Schwarz. *Journal of the Royal Statistical Society, Series B*, 41, 276–278.
- Wansbeek, T., and T. Knapp (1999). Estimating a Dynamic Panel Data Model with Heterogeneous Trends, *Annales d'Économie et de Statistique*, 55/56, 331–349.

## CHAPTER 5

---

# UNBALANCED PANEL DATA MODELS WITH INTERACTIVE EFFECTS

---

JUSHAN BAI, YUAN LIAO, AND JISHENG YANG

## 5.1 INTRODUCTION

---

PANEL data models are widely used by researchers. In practice, it is frequently the case that researchers may encounter missing observations in the collected data. One common source of unbalanced data is attrition. For example, individuals may disappear from a panel after a few waves because they leave the household that is participating in the panel. Moreover, in some sample designs, in the hope of mitigating the effects of attrition, a portion of a panel is replaced by new units in each wave. In labor economics, the empirical data on wages often have missing observations because individuals may drop out of the labor force due to either retirement or lack of job prospects. In treatment effect studies, some subjects may not comply with assigned treatments which prevents experimenters from collecting complete outcome data. In macroeconomics, observations of time series data may be missing due to the aggregation to a lower frequency. Some variables are observed quarterly while others are collected and recorded monthly.

While much of the literature assumes the individual effects and the time effects enter the model additively, a panel data model with interactive effects is also useful in many applications, because it allows the individual effects to interact with time effects. In this chapter, we investigate the following unbalanced panel data model with interactive effects:

$$\begin{aligned} y_{it} &= X'_{it}\beta + \varepsilon_{it}, \\ \varepsilon_{it} &= \alpha_i + \lambda'_i F_t + \xi_{it}, \\ i &= 1, \dots, N, \quad t = t_i = t_i(1), \dots, t_i(T_i), \end{aligned} \tag{5.1}$$

where  $y_{it}$  is the dependent variable;  $X_{it}$  is a  $K \times 1$  vector of observable regressors;  $\alpha_i$  represents the individual effect;  $F_t$  ( $r \times 1$ ) is a vector of common factors with loadings  $\lambda'_i$  ( $1 \times r$ ), and  $\xi_{it}$  is the idiosyncratic error. Note that the model might be unbalanced because for each individual  $i$ , there are  $T_i$  observations available at times  $(t_i(1), \dots, t_i(T_i))$ , and  $T_i$  can be different across  $i$ . Hence observations are subject to missing. Here  $\lambda'_i F_t$  represents the interactive effect. We assume that  $\xi_{it}$  is uncorrelated with  $\lambda_i$  and  $F_s$  for all  $i, t$ , and  $s$ ; but we allow arbitrary correlation between  $X_{it}$  and  $\lambda_i$  and  $F_t$ . In the model, only  $y_{it}$  and  $X_{it}$  are observable. Unlike in the model with additive effects only, the problem of missing data becomes challenging when interactive effects are present. In addition, due to the correlation between  $X_{it}$  and  $(F_t, \lambda_i)$ , the regressor in the first equation can be endogenous. As a result, directly regressing  $y_{it}$  on  $X_{it}$  does not lead to a consistent estimate of  $\beta$ .

We propose new algorithms to estimate the model when missing data are present, allowing various types of missing patterns such as block missing, regular missing, and random (irregular) missing. We adapt the EM algorithm. In particular, when the common factors are deterministic (smooth in  $t$ ), the functional principal components method in Peng and Paul (2009) can be applied. Our proposed algorithms also work for stochastic (nonsmooth) common factors, and therefore are applicable to a broad class of panel data models.

We also consider the dynamic model with:

$$y_{it} = \sum_{j=1}^p \rho_j y_{i,t-j} + X'_{it} \beta + \varepsilon_{it}, \quad \varepsilon_{it} = \alpha_i + \lambda'_i F_t + \xi_{it}.$$

In this case, the instrumental variables (IV) method is applied, with  $\{X_{i,t_i}\}_{t_i \leq t-1}$  as the valid IVs. Extensive simulation studies are carried out to demonstrate the finite-sample properties of the proposed algorithms.

There is an extensive literature on studying the missing data problem in statistics. For example, Biorn (1981), Trawinski and Bargmann (1964), Afifi and Elashoff (1966), Hocking and Smith (1968), and Beale and Little (1975) tackled this problem under multivariate analysis. The problem was also addressed under *seemingly unrelated regressions* by Schmidt (1977), Conniffe (1985), Baltagi, Garvin, and Kerman (1989), Hwang (1990), Hwang and Schulman (1996), etc.

In pure factor models, the functional principal component analysis (FPCA) is often used to deal with missing data. FPCA is a useful tool for factor analysis when the factors are smooth functions of the time. The method is usually based on either kernel smoothing (Boente and Fraiman 2000 and Yao, Müller, and Wang 2005) or sieve representation (Cardot 2000, James, Hastie, and Sugar 2000, and Rice and Wu 2001). Peng and Paul (2009) provide a restricted maximum likelihood method for factor models with smooth factors, but they do not study the panel data with explanatory variables. Stock and Watson (1998) estimate the unobserved factors in unbalanced panel data using the EM algorithm based on principal component analysis (PCA).

For panel data models, Hsiao (2003), Arellano (2003), and Baltagi (2008) provide excellent textbook treatments for general methodologies and applications. Baltagi and Chang (1994) compare various estimation methods for unbalanced panel data models with the one-way error component (individual effect); Wansbeek and Kapteyn (1989) study the two-way error component. Moreover, Davis (2001) allows arbitrary number of error components. Also see Baltagi, Song, and Jung (2001) and Antweiler (2001). For a survey on panel data with missing observations, see Baltagi and Song (2006). Note that all the aforementioned works in the literature focus on models without interactive effects. In contrast, Ahn, Lee, and Schmidt (2013), Pesaran (2006), Bai (2009, 2010), Bai and Li (2012), and Moon and Weidner (2010) consider panel data models with interactive effects. More recently, Kneip, Sickles, Song (2012) employ the PCA based on sieve representations. However, they do not consider the missing data problem.

The remainder of the chapter is organized as follows. Section 5.2 describes the static panel data model with various patterns of missing data. Sections 5.3 and 5.4 propose two types of estimators, FPCA-based and EM-based, to estimate the model, dealing with smooth and non-smooth factors respectively. Section 5.5 discusses the extension to the dynamic model with the instrumental variable method. Simulation results are presented in Section 5.6. Finally, Section 5.7 concludes.

## 5.2 MODEL AND IDENTIFICATION

### 5.2.1 Panel Data Model with Interactive Effects

Consider the static unbalanced panel data model (5.1), which can be written as

$$y_{it} = X'_{it}\beta + \alpha_i + \lambda'_i F_t + \xi_{it}, \quad (5.2)$$

where  $\lambda'_i F_t$  represents the individual-time interactive effect. In particular, for  $r = 1$  and  $\lambda_i = 1$ , model (5.2) becomes

$$y_{it} = X'_{it}\beta + \alpha_i + F_t + \xi_{it}.$$

Thus additive effects are special cases of interactive effects. The interactive effect brings new challenges to the estimation problem, because  $X_{it}$  is generally endogenous in  $y_{it} = X'_{it}\beta + \varepsilon_{it}$  due to the correlations between  $X_{it}$  and the common factor  $F_t$  and loading  $\lambda_i$  in the error  $\varepsilon_{it}$ . In this case, traditional techniques such as differencing and within-group transformation can no longer remove the individual and time effects, which cause the endogeneity. As a result, methods based on either OLS or the within-group transformations are no longer consistent. The problem becomes more difficult when missing observations are present.

We allow one or all regressors affected by both unobserved factors and loadings, for example,

$$x_{1,it} = \theta x_{1,i,t-1} + \Gamma_i' F_t + S_i' \lambda_i + \kappa \lambda_i' F_t + u_{it},$$

in which  $u_{it}$  are idiosyncratic errors such that  $E(\xi_{it} u_{js}) = 0$  for any  $i, j, t, s$ . When the panel data are balanced across  $i$  (that is,  $T_i = T$  and  $t_i = 1, 2, \dots, T$  for each  $i$ ), and no observations are missing, Bai (2009) and Moon and Weidner (2010) estimated the model by minimizing

$$\min_{\beta, \alpha_i, \lambda_i, F_t} \sum_{i=1}^N \sum_{t=1}^T (y_{it} - X_{it}' \beta - \alpha_i - \lambda_i' F_t)^2 \quad (5.3)$$

subject to normalization constraints on  $\lambda_i$  and  $F_t$ .

In this chapter, we propose new algorithms for a more general model where not only the number of observations ( $T_i, 1 \leq i \leq N$ ) but also the observed time locations vary from individual to individual. In other words, the missing observations may exist arbitrarily for the individuals.

## 5.2.2 Missing Observations in Panel Data

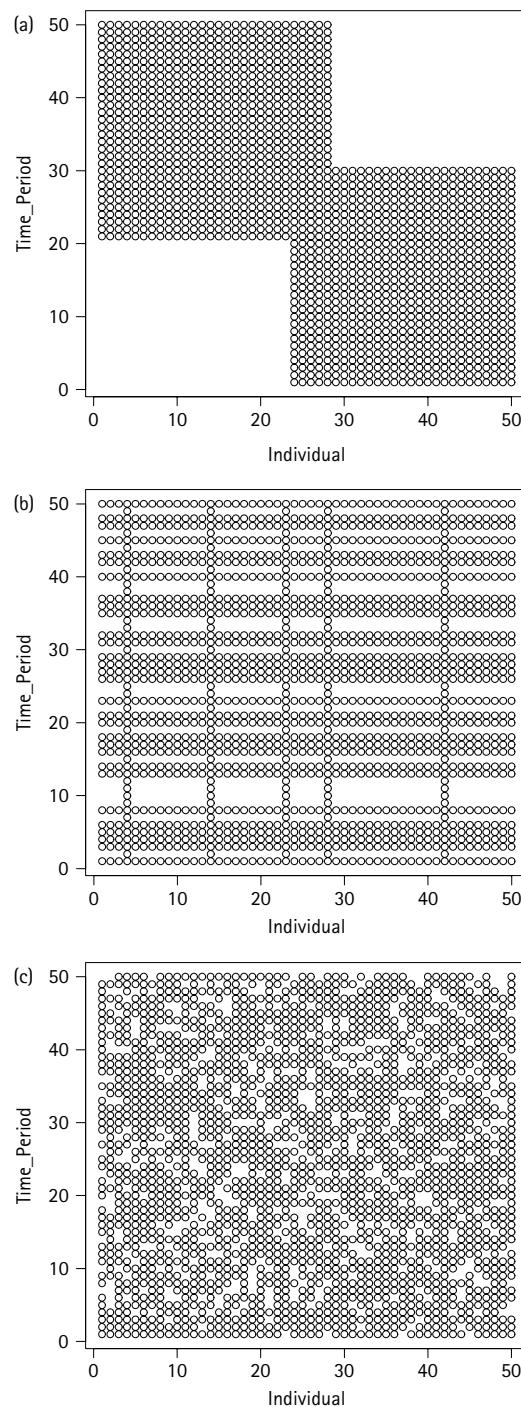
Missing data are particularly common when data are based on observational experiments. For example, in the security market, holidays are often a source of missing data. Firms drop out of the sample due to mergers and acquisitions; new companies emerge over time. Patterns of missing observations differ, which can be block missing, regular missing, or even random missing. This chapter considers three types of missing patterns, as illustrated in Figure 5.1.

The first type of missing is block missing, see Figure 5.1(a). Figure 5.1(a) depicts that the first 23 individuals join the sample at  $t = 21$  whereas another 22 individuals drop out at  $t = 31$ . More specifically, the data of individuals  $i = 1, \dots, 23$  for  $t = 1, \dots, 20$  and those of individuals  $i = 29, \dots, 50$  for  $t = 31, \dots, 50$  are missing, so there are 36% missing data in total. Such a synchronism often arises from rule changes in policy studies.

The second type of missing is regular where the missing event, if any, occurs on the same time frequency for all the individuals. As in Figure 5.1(b), about 90% of the individuals have missing observations, and each misses forty percent of data from its original time series. Such a missing pattern is more common in economics.

The third pattern of missing is random (irregular), see Figure 5.1(c), in which twenty percent of the data are missing randomly without an obvious pattern. This is a common missing mechanism for households data. Note that this type of missing is particularly troublesome in models with interactive effects, compared to the regular model with additive effects only.

We emphasize that our proposed algorithms do not need to know in advance which missing pattern the data possess. In the computations of this chapter, the missing data



**FIGURE 5.1** Missing patterns

patterns are only used for simulating the data generating process, and the algorithms can handle all the cases regardless of the underlying missing patterns.

### 5.2.3 Identification and Iterative Estimation

In the interactive effect, the factors and loadings are not separately identified. To see this, let  $T$  denote the length of the complete time period, which is  $\max\{t_i(T_i)\}_{i \leq N} - \min\{t_i(T_i)\}_{i \leq N} + 1$ . Define  $F = (F_1, F_2, \dots, F_T)'$  and  $\Lambda = (\lambda_1, \dots, \lambda_N)'$ . Given an arbitrary  $r \times r$  invertible matrix  $A$ , we have  $FA' = FAA^{-1}\Lambda'$ . Hence  $(F, \Lambda')$  and  $(FA, A^{-1}\Lambda')$  produce observationally equivalent models. For the identification purpose, we orthonormalize the factors and loadings as

$$\frac{F'F}{T} = I_r, \quad \frac{\Lambda'\Lambda}{N} \text{ is diagonal.} \quad (5.4)$$

Let  $M_F = I_T - FF'/T$ . Under condition (5.4), left multiplying  $M_F$  to (5.2) implies

$$M_F y_{it} = M_F \alpha_i + M_F X'_{it} \beta + M_F \xi_{it}.$$

Removing individual means further gives

$$M_F(y_{it} - \bar{y}_i) = M_F(X_{it} - \bar{X}_i)' \beta + M_F(\xi_{it} - \bar{\xi}_i), \quad (5.5)$$

where  $\bar{y}_i, \bar{X}_i, \bar{\xi}_i$  denote the averages on the  $i$ th individual. Suppose  $F$  is known, then  $X_{it}$  becomes exogenous given the condition that  $E(X_{it}\xi_{it}) = 0$ . Hence one can consistently estimate  $\beta$  using OLS. Equivalently, given  $F_t$  and  $\lambda_i$ , the model reduces to

$$y_{it} - \lambda'_i F_t \equiv y_{it}^F = \alpha_i + X'_{it} \beta + \xi_{it},$$

which can be consistently estimated by the least squares dummy variable (LSDV).

On the other hand, if  $\beta$  and  $\alpha_i$  are known, one has the typical factor model

$$y_{it} - X'_{it} \beta - \alpha_i \equiv y_{it}^\beta = \lambda'_i F_t + \xi_{it}, \quad (5.6)$$

in which  $F_t$  and  $\lambda_i$  can be estimated by methods considered by Stock and Watson (1998), Peng and Paul (2009) and Bańbura and Modugno (2010), etc., under different settings of missing data.

Therefore, model (5.1) can be estimated by iterating (5.5) and (5.6). In Sections 5.3 and 5.4, we propose two new iterative algorithms: one based on smoothing functional PCA and the other based on the EM-algorithm. While the smoothing functional PCA method (to be introduced in Section 5.3) works well only when the factors are deterministic and  $F_t$  is a smooth function of  $t$ , the EM algorithm (Section 5.4) can deal with both smooth (deterministic) and non-smooth (stochastic) factors.

## 5.3 LS-FPCA ESTIMATION

---

Throughout Section 5.3, we assume the factor  $F_t$  to be deterministic, and treat it as an unknown smooth function of  $t$ , which is particularly useful when missing data are present in the time domain. We first review the *Functional PCA* (FPCA) method for the missing data problem in factor models, proposed by Peng and Paul (2009). In Section 5.3.2, we propose an LS-FPCA algorithm to estimate the panel data with interactive effects, which combines FPCA with the least squares method.

### 5.3.1 Functional PCA

Let us temporarily assume  $(\beta, \alpha_i)$  to be known in this subsection, and focus on a factor model (5.6):

$$y_{it}^\beta = \lambda'_i F_t + \xi_{it}, \quad (5.7)$$

where  $y_{it}^\beta$  is observable. Functional principal components analysis (FPCA) is a useful tool when the observed data are curves. Basic introduction to FPCA can be found in Jolliffe (2002, pp. 316–327). In the context of factor models it is easy to motivate and explain the method without an explicit reference to the spectrum decomposition of a reduced-rank covariance matrix. Suppose that we only observe the common components part  $\lambda'_i F_t$ . When  $F_t$  is a vector of smooth functions of  $t$ , then  $\lambda'_i F_t$  ( $i = 1, 2, \dots, N$ ) will be  $N$  curves. Given that  $F_t$  is a nonparametric function of  $t$ , it can be represented by basis functions.

Let  $\{\phi_j\}_{j \leq M}$  be a set of orthonormal basis functions such as cubic splines, polynomials, Fourier sequence, wavelets, etc, where  $M$  denotes the number of basis functions to be used. Larger  $M$  leads to more accurate approximation to  $F_t$ . Then there is an  $M \times r$  matrix  $B$  such that for large enough  $M$  and  $\phi(t)' = (\phi_1(t), \dots, \phi_M(t))$ ,

$$F_t' \approx (\phi_1(t), \dots, \phi_M(t)) B = \phi(t)' B.$$

Here  $B$  is an orthonormal coefficient matrix so that  $B'B = I_r$ . We take  $\{\phi_j\}$  to be a known set of basis functions and aim to estimate  $B$ . Let  $\widehat{B}$  be an estimator of  $B$  to be introduced later. Then  $F_t'$  can be estimated by  $\phi(t)' \widehat{B}$ .

Let us estimate  $B$  using the maximum likelihood method. Write

$$y_i^\beta = (y_i^\beta(t_i(1)), \dots, y_i^\beta(t_i(T_i)))', \quad T_i \times 1$$

and

$$F' = (F_{t_i(1)}, \dots, F_{t_i(T_i)}), \quad \xi_i' = (\xi_{i,t_i(1)}, \dots, \xi_{i,t_i(T_i)}),$$

then

$$y_i^\beta = F \lambda_i + \xi_i.$$

The above matrix  $F$  in fact depends on  $i$ . In addition, define a  $T_i \times M$  basis matrix

$$\Phi_i = \begin{pmatrix} \phi_1(t_i(1)) & \dots & \phi_M(t_i(1)) \\ \vdots & \ddots & \vdots \\ \phi_1(t_i(T_i)) & \dots & \phi_M(t_i(T_i)) \end{pmatrix},$$

which is known given the observation times  $(t_i(1), \dots, t_i(T_i))$  and the basis functions. From  $F \approx \Phi_i B$ , we have,

$$y_i^\beta \approx \Phi_i B \lambda_i + \xi_i$$

Suppose that  $\lambda_i \stackrel{iid}{\sim} N(0, \Sigma_\lambda)$  and  $\xi_{it} \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$  and  $\lambda_i$  and  $\xi_{it}$  are independent, then  $\Phi_i B \lambda_i + \xi_i$  is normally distributed with zero mean and variance  $\Phi_i B \Sigma_\lambda B' \Phi_i' + \sigma^2 I_{T_i}$  so the individual (negative) log-likelihood function based on  $y_i^\beta$  can be written as

$$l_i(B, \sigma^2, \Sigma_\lambda) = \log |\sigma^2 I_{T_i} + \Phi_i B \Sigma_\lambda B' \Phi_i'| + \text{tr} \left[ \frac{1}{T_i} y_i^\beta y_i^{\beta'} (\sigma^2 I_{T_i} + \Phi_i B \Sigma_\lambda B' \Phi_i')^{-1} \right].$$

We normalize  $\Sigma_\lambda$  to be a diagonal matrix. This is feasible; we can always find an orthogonal matrix  $R$  such that  $R \Sigma_\lambda R'$  is diagonal. We then define  $B^*$  to be  $BR'$  so that  $B^{*\prime} B^* = RB' BR' = RR' = I_r$ . Diagonality of  $\Sigma_\lambda$  together with  $BB' = I_r$  is sufficient to identify  $B$  and  $\Sigma_\lambda$ . This identification condition is analogous to (5.4). The log-likelihood function can be written as, for some constant  $C$ ,

$$\log L(B, \sigma^2, \Sigma_\lambda) = C - \frac{1}{2} \sum_{i=1}^N l_i(B, \sigma^2, \Sigma_\lambda). \quad (5.8)$$

Let  $\widehat{B}$  be the maximum likelihood estimator for  $B$ , which is the maximizer of (5.8). The common factor is then estimated by

$$\widehat{F}_t' = \phi(t)' \widehat{B}.$$

The above likelihood function is similar to the one considered by Paul and Peng (2009). They applied a Newton-Raphson iteration to maximize (5.8), and showed that the estimator  $\widehat{F}_t$  is consistent under mild conditions.

One can apply a leave-one-curve-out cross-validation procedure to choose the basis dimension  $M$  and the number of factors  $r$ . Specifically,  $M$  and  $r$  are chosen by solving

$$\min_{M,r} CV = \min_{M,r} \sum_{i=1}^N l_i(\widehat{B}^{-i}, \widehat{\sigma}^{2,-i}, \widehat{\Sigma}_\lambda^{-i}),$$

where  $(\widehat{B}^{-i}, \widehat{\sigma}^{2,-i}, \widehat{\Sigma}_\lambda^{-i})$  are the maximum likelihood estimators obtained by excluding the  $i$ -th individual.

### 5.3.2 LS-FPCA Iterations

Now consider the unbalanced panel data model with interactive effects and missing observations. As  $(\beta, \alpha_i)$  are unknown, we estimate them via least squares, and propose a simple LS-FPCA algorithm.

We first introduce some notation. Recall  $T = \max\{t_i(T_i)\}_{i \leq N} - \min\{t_i(T_i)\}_{i \leq N} + 1$ . Let  $W_i$  be a diagonal matrix of size  $T$  such that the  $t$ -th diagonal element equals 0 if  $y_{it}$  is missing and 1 otherwise. Adapting a method of Hwang and Schulman (1996), in the panel data context, we form groups based on time instead of variables. More specifically, we divide the  $T$  time periods into  $G$  groups so that within each group, each individual has either no observation or complete observations. Let  $\mathcal{T}_g$  denote the number of time periods in the  $g$ -th group, then  $\sum_{g=1}^G \mathcal{T}_g = T$ . In addition, the  $g$ -th group has complete observations on  $N_g$  individuals and missing observations on  $N - N_g$  individuals. For each group  $g$ , suppose it contains time periods  $(t_1, t_2, \dots, t_{\mathcal{T}_g}) \subset \{1, \dots, T\}$ , on which individuals  $(i_1, i_2, \dots, i_{N_g})$  have complete observations. Let  $P_g$  be a  $\mathcal{T}_g \times T$  matrix so that the rows of  $P_g$  equal the  $(t_1, t_2, \dots, t_{\mathcal{T}_g})$  rows of the  $T \times T$  identity matrix. Further, let  $Q_g$  be an  $N \times N_g$  matrix so that the columns of  $Q_g$  equal the  $(i_1, i_2, \dots, i_{N_g})$  columns of the  $N \times N$  identity matrix.

Take the missing pattern of Figure 5.1(a) as an example. Individuals 1–23 have missing observations at time periods  $t = 1 \dots 20$ , and individuals 29–50 have missing observations at time periods  $t = 31 \dots 50$ . We then divide all the 50 time periods into three groups, corresponding to time periods 1–20, 21–30, and 31–50. Hence  $\mathcal{T}_1 = 20, \mathcal{T}_2 = 10, \mathcal{T}_3 = 20$ . Then the rows of  $P_1, P_2, P_3$  are the first 20 rows, (21–30)-th rows and (31–50)-th rows of  $I_{50}$  respectively. The columns of  $Q_1, Q_2, Q_3$  are the (24–50)-th columns, all the columns, and first 28 columns of  $I_{50}$  respectively. For Figure 5.1(b), we divide the observations into just two groups. Group 1 contains the 20 time periods  $\{2, 7, 9 \dots 12, \dots, 49\}$ , and for these time periods, five individuals have complete observations. Hence  $N_1 = 5$  and  $\mathcal{T}_1 = 20$ . Group 2 contains all the remaining time periods on which all the individuals have complete data, so that  $N_2 = 50$  and  $\mathcal{T}_2 = 30$ .

Let  $\tilde{y}_{it} = y_{it} - \bar{y}_i$ ,  $\tilde{X}_{it} = X_{it} - \bar{X}_i$ , where  $\bar{y}_i = \frac{1}{T_i} \sum_{j=1}^{T_i} y_{i,t_i(j)}$  and  $\bar{X}_i = \frac{1}{T_i} \sum_{j=1}^{T_i} X_{i,t_i(j)}$ . For each  $k \leq \dim(\beta) \equiv K$ , let  $\tilde{X}_k = (\tilde{X}_{k,it})'_{i \leq N, t \leq T}$  be a  $T \times N$  matrix. Each column of  $\tilde{X}_k$  is a  $T$ -dimensional vector balanced so that missing observations are replaced with zeros, in the sense that the entries are either  $\tilde{X}_{k,it}$  or zero if  $\tilde{X}_{k,it}$  is missing. Our proposed LS-FPCA algorithm is described as follows.

#### LS-FPCA Algorithm:

- Initialize  $\hat{\beta} = \hat{\beta}_0$ .
- Apply FPCA to  $\tilde{y}_{it}^\beta = \tilde{y}_{it} - \tilde{X}'_{it} \hat{\beta}$ ; estimate  $\hat{F}$  and  $\hat{r}$  as in Section 5.3.1.
- Estimate the loading matrix by regressing  $\{\tilde{y}_i^\beta\}_{i \leq N}$  on  $\hat{F}$ :

$$\hat{\Lambda}'_i = (\hat{F}'_W \hat{F}_W)^{-1} \hat{F}'_W Z_i^\beta, \quad (5.9)$$

with  $\hat{F}_W = W_i \hat{F}$  and  $Z_i^\beta = W_i \tilde{y}_i^\beta$ .

(d) Update  $\hat{\beta}$  via regressing  $\tilde{y}_{it} - \hat{\Lambda}'_i \hat{F}_t$  on  $\tilde{X}_{it}$ :

$$\hat{\beta} = \left( \sum_{g=1}^G Z_g' Z_g \right)^{-1} \left( \sum_{g=1}^G Z_g' Y_g \right), \quad (5.10)$$

where  $Z_g = (\text{vec}(P_g \tilde{X}_1 Q_g), \dots, \text{vec}(P_g \tilde{X}_K Q_g))_{T_g N_g \times K}$  and  $Y_g = \text{vec}(P_g \tilde{w} Q_g)$ . Here  $\tilde{w} = (\tilde{w}_1, \dots, \tilde{w}_N)$  with each  $\tilde{w}_i = (\tilde{y}_{i1} - \hat{\lambda}'_1 \hat{F}_1, \dots, \tilde{y}_{iT} - \hat{\lambda}'_1 \hat{F}_T)'$  being a  $T \times 1$  vector, balanced so that missing observations are replaced with zeros.

(e) Iterate (b)–(d) until convergence.

When there are no missing observations, step (d) of the algorithm becomes the regular OLS that regresses  $(\tilde{y}_{11} - \hat{\Lambda}'_1 \hat{F}_1, \dots, \tilde{y}_{NT} - \hat{\Lambda}'_N \hat{F}_T)'$  on  $(\text{vec}(\tilde{X}_1), \dots, \text{vec}(\tilde{X}_K))$ . Note that  $(\tilde{y}_{11}, \dots, \tilde{y}_{NT})$  are already demeaned. Moreover, we can set the initial value as, for example,

$$\hat{\beta}_0 = \left( \sum_{g=1}^G Z_g' Z_g \right)^{-1} \left( \sum_{g=1}^G Z_g' \text{vec}(P_g \tilde{y} Q_g) \right), \quad (5.11)$$

where  $\tilde{y} = (\tilde{y}_1, \dots, \tilde{y}_N)$  with each  $\tilde{y}_i = (\tilde{y}_{i1}, \dots, \tilde{y}_{iT})'$  being a  $T \times 1$  vector, balanced so that missing observations are replaced with zeros. Such a choice is biased though, due to the endogeneity of  $\tilde{X}_{it}$  because it ignores the interactive effects. But starting values are not required to be consistent estimators, and the simulation results show that it is a good choice for initialization.

## 5.4 LS-EM-PCA ESTIMATION

The FPCA method described in Section 5.3 assumes deterministic common factors  $F_t$ . Because the basis representation works well only for smooth functions, once  $F_t$  is stochastic and no longer smooth in the time domain, FPCA will be inapplicable. In Section 5.4, we treat the factors to be non-smooth stochastic functions on the time domain, and employ the EM algorithm instead of FPCA. The EM algorithm has been commonly used to deal with missing data problems (Schafer 1997; McLachlan and Krishnan 1997; Meng and van Dyk 1997, etc). It was applied by Stock and Watson (1998) to the factor analysis.

Section 5.4.1 reviews the EM algorithm for factor analysis with missing data. In Section 5.4.2, we propose an LS-EM-PCA algorithm to estimate the unbalanced panel data model with interactive effects, which combines the EM algorithm with the least squares method.

### 5.4.1 EM Algorithm for Factor Models with Missing Data

Suppose  $\alpha$  and  $\beta$  are known. Recall the model

$$y_{it} - X'_{it}\beta - \alpha_i = y_{it}^\beta = \lambda'_i F_t + \xi_{it}. \quad (5.12)$$

In the E-step, missing observations  $y_{it}^\beta$  are replaced with an estimate of  $\lambda'_i F_t$  to obtain the “complete data” (or generated date). Hence the “complete data” contain individuals of the form

$$y_{it}^{*\beta} = \begin{cases} y_{it} - X'_{it}\beta - \alpha_i, & \text{if not missing} \\ \widehat{\lambda}'_i \widehat{F}_t & \text{if missing.} \end{cases} \quad (5.13)$$

for some estimator  $\widehat{\lambda}'_i \widehat{F}_t$ . In the M-step, the PCA is conducted on the “complete data” to estimate  $F_t$  and  $\lambda_i$ . To see how it works, under the Gaussian assumption  $\xi_{it} \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$ , the log-likelihood function of the “complete data”  $y^{*\beta} = \{y_{it}^{*\beta}\}$  is given by

$$l(y^{*\beta}, \Lambda, F) \propto -\frac{1}{NT} \sum_{i,t} \left( y_{it}^{*\beta} - \lambda'_i F_t \right)^2. \quad (5.14)$$

Maximizing (5.14) subject to the restriction (5.4) is a standard principal components problem (which gives rise to the name EM-PCA). At the  $j$ -th iteration, the columns of the  $T \times r$  matrix  $\widehat{F}^{(j)}$  are the eigenvectors corresponding to the  $r$  largest eigenvalues of the  $T \times T$  matrix  $(NT)^{-1} \sum_{i=1}^N y_i^{*\beta(j-1)} y_i^{*\beta(j-1)'} = T^{-1} \widehat{F}^{(j)'} \widehat{F}^{(j)}$ , and  $\widehat{\Lambda}_i^{(j)} = T^{-1} \widehat{F}^{(j)'} y_i^{*\beta(j-1)}$ . At the  $j$ -th iteration, we have

$$y_{it}^{*\beta(j)} = \begin{cases} y_{it} - X'_{it}\beta - \alpha_i, & \text{if not missing} \\ \widehat{\lambda}_i^{(j)} \widehat{F}_t^{(j)} & \text{if missing.} \end{cases}$$

The EM algorithm proceeds by iteratively maximizing the expected complete data likelihood with respect to  $(\Lambda, F_t)$ , until convergence.

### 5.4.2 LS-EM-PCA Iterations

In the unbalanced panel data model with missing data,  $(\beta, \alpha)$  need be estimated as well. Given  $(\beta, \alpha)$ , we can employ the EM algorithm to estimate  $(F, \Lambda)$ ; given  $(F, \Lambda)$ , we run LSDV to estimate  $(\beta, \alpha)$ . Hence our proposed LS-EM-PCA algorithm consists of two loops. The inner loop carries out the EM, while the outer loop estimates  $\beta$ . Note that when the EM algorithm is applied to factor analysis, the solution in the M-step is the principal components estimator of Stock and Watson (1998). Therefore we term this algorithm as LS-EM-PCA.

The iterations are described as follows.

**LS-EM-PCA Algorithm:**

- (a) Initialize  $(\widehat{\Lambda}, \widehat{F}) = (\widehat{\Lambda}_0, \widehat{F}_0)$ .
- (b) Estimate  $\widehat{\beta}$  via regressing  $\widetilde{y}_{it} - \widehat{\Lambda}' \widehat{F}_t$  on  $\widetilde{X}_{it}$  as in (5.10).
- (c) Apply EM algorithm to  $\widetilde{y}_{it} - \widetilde{X}'_{it} \widehat{\beta}$ ; update  $(\widehat{\Lambda}, \widehat{F})$  as in Section 5.4.1 (inner loop).
- (d) Iterate (b)-(c) until convergence (outer loop).

The initial value for  $(\Lambda, F)$  can be obtained by first carrying out the LS-FPCA algorithm. It is pointed out that, in principle, the generalized least squares (GLS) of Breitung and Tenhofen (2011) and Choi (2012) can be used in the estimation of factor and factor loadings, as well as the regression parameters.

## 5.5 DYNAMIC CASE

We now consider the dynamic panel data model:

$$\begin{aligned} y_{it} &= \sum_{j=1}^p \rho_j y_{i,t-j} + X'_{it} \beta + \varepsilon_{it}, \\ \varepsilon_{it} &= \alpha_i + \lambda'_i F_t + \xi_{it}, \\ i &= 1, \dots, N, \quad t = t_i = t_i(1), \dots, t_i(T_i), \end{aligned} \tag{5.15}$$

where the lagged variables  $y_{i,t-1}, \dots, y_{i,t-p}$  are included in the regressors. As in previous sections, the model contains unobserved interactive effects  $\lambda'_i F_t$ , and is still unbalanced due to possible missing observations. For notational simplicity, we just consider the case  $p = 1$ .

Given  $F$ , the model can be written similarly to (5.5) as:

$$M_F(y_{it} - \alpha_i) = \rho M_F y_{i,t-1} + M_F X'_{it} \beta + M_F \xi_{it}. \tag{5.16}$$

Conversely, given  $\rho, \beta$  and  $\alpha_i$ , one has a factor model

$$y_{it} - \rho y_{i,t-1} - X'_{it} \beta - \alpha_i \equiv y_{it}^\beta = \lambda'_i F_t + \xi_{it}, \tag{5.17}$$

An important issue is to estimate  $\rho$  in (5.16) even if  $(\alpha_i, \beta)$  are known. Because  $y_{i,t-1}$  can be endogenous due to the autocorrelation of  $\{\xi_{it}\}_{t=1}^\infty$ , OLS does not provide a consistent estimator of  $\rho$ , and we should apply the instrumental variables method instead. We employ  $\{X_{i,t_i(j)}\}_{t_i(j) \leq t-1}$  as instruments and estimate  $(\rho, \alpha_i, \beta)$  by the two-stage-least-squares (2SLS) from

$$y_{it} - \lambda'_i F_t \equiv y_{it}^F = \alpha_i + \rho y_{i,t-1} + X'_{it} \beta + \xi_{it}. \tag{5.18}$$

We iterate (5.17) and (5.18) as in previous sections. The use of  $\{X_{i,t_i(j)}\}_{t_i(j) \leq t-1}$  as IV's is widely adopted in panel data analysis (see Arellano and Honoré 2001; Hsiao 2003;

Arellano 2003; Baltagi 2008; and Sarafidis and Yamagata 2010). For an AR(1) dynamic model, a single missing observation corresponds to two missing observations in the static case because every observation appears twice in the model. The estimation of  $\rho$  is based on adjacent observations over  $t$ .

To formulate the 2SLS for (5.18), we modify the iteration stage (d) in the LS-FPCA algorithm (also (b) in the LS-EM-PCA) as follows. Given  $(\widehat{\lambda}_i, \widehat{F}_t)$ , let  $\tilde{w} = (\tilde{w}_{ti})_{T \times N}$  be a  $T \times N$  matrix so that each element  $\tilde{w}_{ti} = \tilde{y}_{it} - \widehat{\Lambda}_i' \widehat{F}_t$ ,  $t = 1, \dots, T$ ,  $i = 1, \dots, N$ . Each column of  $\tilde{w}$  is a  $T$ -dimensional vector with entries either  $\tilde{w}_{ti}$  or zero if  $\tilde{y}_{it}$  is missing. For the same grouping strategy of Section 5.3.2, denote  $Y_g = \text{vec}(P_g \tilde{w} Q_g)$  as a  $T_g N_g \times 1$  vector for the  $g$ -th group. Also let  $\tilde{w}_{-1}$  be a  $T \times N$  matrix with the  $i$ th column  $(0, \tilde{w}_{1,i}, \dots, \tilde{w}_{T-1,i})'$ , and  $\tilde{X}_{j,-1}$  be a  $T \times N$  matrix with the  $i$ th column  $(0, \tilde{X}_{i,1}, \dots, \tilde{X}_{i,T-1})'$ . Missing observations are replaced with zeros. Define  $Y_{g,-1} = \text{vec}(P_g \tilde{w}_{-1} Q_g)$ , and

$$\begin{aligned} Z_g &= (\text{vec}(P_g \tilde{X}_1 Q_g), \dots, \text{vec}(P_g \tilde{X}_K Q_g))_{T_g N_g \times K} \\ Z_{g,-1} &= (\text{vec}(P_g \tilde{X}_{1,-1} Q_g), \dots, \text{vec}(P_g \tilde{X}_{K,-1} Q_g))_{T_g N_g \times K}, \end{aligned}$$

For the regression model (5.18), the matrix of regressors is then given by  $X_g = (Y_{g,-1}, Z_g)$ , with the matrix of instruments  $H_g = (Z_{g,-1}, Z_g)$ . The 2SLS estimators of  $(\rho, \beta)$  can then be written as

$$(\widehat{\rho}, \widehat{\beta})' = \left( \sum_{g=1}^G X_g' H_g \left( H_g' H_g \right)^{-1} H_g' X_g \right)^{-1} \left( \sum_{g=1}^G X_g' H_g \left( H_g' H_g \right)^{-1} H_g' Y_g \right).$$

**Remark:** Suppose that we only employ  $x_{1,i,t-1}$  as the IV and apply the 2SLS, the estimator then becomes, for  $H_g = (\text{vec}(P_g \tilde{X}_{1,-1} Q_g), Z_g)$ ,

$$(\widehat{\rho}, \widehat{\beta})' = \left( \sum_{g=1}^G H_g' X_g \right)^{-1} \left( \sum_{g=1}^G H_g' Y_g \right),$$

the regular IV estimator with missing data.

## 5.6 MONTE CARLO SIMULATIONS

### 5.6.1 Data Generation and Implementation

We assess the finite sample performances of the proposed estimators, namely, the algorithms of LS-FPCA(static), LS-EM-PCA (static), IV-FPCA (dynamic) and IV-EM-PCA (dynamic), by simulations. Note that the LS-FPCA and IV-FPCA are designed for

smooth factor models, and LS-EM-PCA and IV-EM-PCA are for non-smooth factor models. In the simulation below, however, each method is computed for both smooth and stochastic factors for comparison.

Data are generated by

$$\begin{aligned} y_{it} &= \rho y_{i,t-1} + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \varepsilon_{it}, \\ \varepsilon_{it} &= \alpha_i + \lambda'_i F_t + \xi_{it}, \\ x_{1,it} &= \Gamma'_i F_t + S'_t \lambda_i + \kappa \lambda'_i F_t + u_{it}, \end{aligned}$$

where  $\rho = 0$  for the static models and 0.5 for the dynamic model;  $\beta_1 = 2, \beta_2 = 1, x_{2,it} \stackrel{i.i.d.}{\sim} U[0, 2], \alpha_i \stackrel{i.i.d.}{\sim} N[0, 4], \xi_{it} \stackrel{i.i.d.}{\sim} N[0, 1]$ , while  $\kappa = 1, \Gamma_i, S_t, \lambda_i \stackrel{i.i.d.}{\sim} U[0.5, 1.5], u_{it} \stackrel{i.i.d.}{\sim} N[0, 1]$ . Two procedures are used to generate the common factors, corresponding to the deterministic and stochastic ones as follows:

- (deterministic, smooth factors) Consider:

$$F_{1t} = \sin(\delta t/T), F_{2t} = t/T + \alpha \max\{t - t_0, 0\}/T.$$

with  $\delta = 5, \alpha = -0.5, t_0 = T/2$ . The sine function is used to represent a low frequency component, whereas the time trend in  $F_{2t}$  indicates a broken trend (see Figure 5.2 (a)).

- (stochastic, non-smooth factors) Consider the following VAR(1) process:

$$\begin{aligned} F_t &= A' F_{t-1} + \eta_{it}, \\ A &= \begin{pmatrix} a_{11} & 0 \\ 0 & a_{22} \end{pmatrix}, \end{aligned}$$

where  $\eta_{it} \stackrel{i.i.d.}{\sim} N[0, 1]$  and  $a_{11}, a_{22} \sim U[0.5, 1]$  (see Figure 5.2 (b)).

Finally, three missing patterns are designed as follows.

- Pattern 1: The first 40% time periods for the first 45% individuals are treated as unobserved while the last 40% time periods for the last 45% individuals are also treated as unobserved. This causes 36% missing data in total.
- Pattern 2: 90% of individuals have missing observations, and in the static case each has 40% observations missing at random (20% in the dynamic case). This causes 36% missing data in total.
- Pattern 3: 20% of data are missing randomly in the static case (10% for the dynamic case).

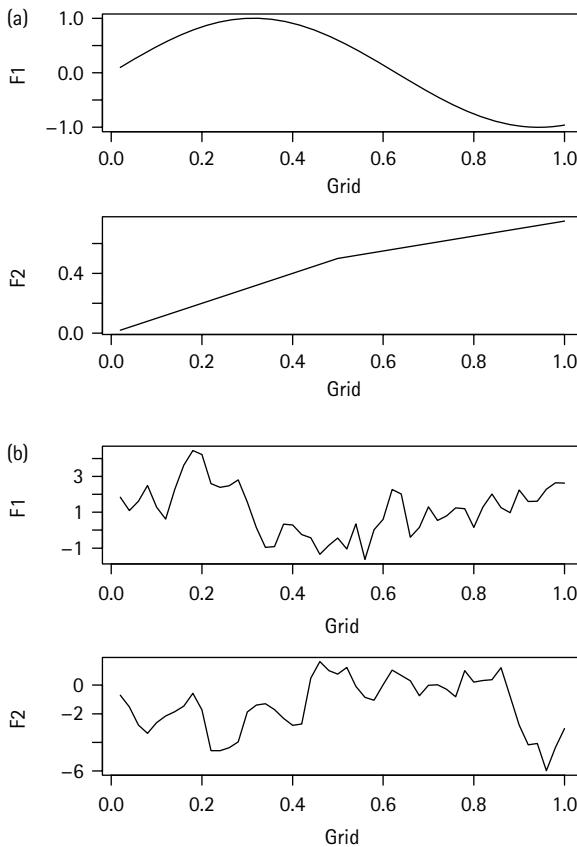


FIGURE 5.2 The factor process

### 5.6.2 Main Findings

We compare the proposed FPCA-based (LS-FPCA and IV-FPCA) algorithms with the EM-based (LS-EM-PCA and IV-EM-PCA) algorithms. These methods are also compared with OLS and LSDV. Here OLS pools together the observed data and ignores both interactive and individual effects. The LSDV method estimates  $\beta$  by (5.10) which removes the individual effects but not interactive effects.

Each simulation is replicated for 1000 times, and the results are provided in Tables 5.1, 5.2 (static) and Tables 5.3, 5.4 (dynamic). In the dynamic case,  $x_{1,i,t-1}$  is used as the IV. We summarize the major findings as follows.

- (a) The EM-type estimators outperform those of the FPCA-type no matter whether the factors are stochastic or not and regardless of the missing pattern.
- (b) The FPCA-type estimators are consistent for the case of smooth factors but not for the stochastic factors, which demonstrate that the basis representation approach of FPCA is inappropriate when factors are non-smooth on the time domain.

**Table 5.1 Static model with smooth factors**

			Missing Pattern 1				Missing Pattern 2				Missing Pattern 3			
			OLS	LSDV	LS-FPCA	LS-EM-PCA	OLS	LSDV	LS-FPCA	LS-EM-PCA	OLS	LSDV	LS-FPCA	LS-EM-PCA
N=50	$\hat{\beta}_1$	Mean	2.2611	2.2273	2.0779	2.0576	2.2619	2.2682	2.0593	2.0137	2.2663	2.2694	2.0513	2.0158
		Sd	0.0770	0.0199	0.0446	0.0397	0.0567	0.0223	0.0525	0.0316	0.0552	0.0178	0.0416	0.0252
T=50	$\hat{\beta}_2$	Mean	0.9980	1.0000	1.0006	0.9994	0.9973	0.9997	0.9996	1.0000	1.0019	0.9994	0.9988	0.9984
		Sd	0.0978	0.0477	0.0467	0.0482	0.0986	0.0485	0.0473	0.0490	0.0853	0.0423	0.0405	0.0415
N=100	$\hat{\beta}_1$	Mean	2.2570	2.2275	2.0846	2.0446	2.2609	2.2687	2.0480	2.0120	2.2615	2.2680	2.0447	2.0136
		Sd	0.0563	0.0145	0.0520	0.0232	0.0414	0.0179	0.0402	0.0203	0.0407	0.0126	0.0323	0.0169
T=50	$\hat{\beta}_2$	Mean	1.0013	1.0000	1.0002	0.9997	0.9976	1.0000	1.0007	1.0008	0.9983	1.0009	1.0010	1.0010
		Sd	0.0674	0.0329	0.0318	0.0330	0.0667	0.0344	0.0325	0.0333	0.0605	0.0291	0.0282	0.0292
N=50	$\hat{\beta}_1$	Mean	2.2597	2.2290	2.0662	2.0351	2.2643	2.2702	2.0405	2.0099	2.2697	2.2700	2.0352	2.0107
		Sd	0.0787	0.0171	0.0358	0.0213	0.0568	0.0176	0.0342	0.0194	0.0561	0.0152	0.0288	0.0160
T=100	$\hat{\beta}_2$	Mean	1.0010	0.9988	0.9987	0.9990	1.0023	1.0005	1.0002	1.0002	1.0026	1.0009	1.0004	1.0002
		Sd	0.0679	0.0328	0.0321	0.0341	0.0680	0.0324	0.0310	0.0326	0.0613	0.0298	0.0280	0.0286
N=100	$\hat{\beta}_1$	Mean	2.2618	2.2295	2.0674	2.0294	2.2629	2.2700	2.0327	2.0079	2.2645	2.2700	2.0295	2.0092
		Sd	0.0534	0.0119	0.0310	0.0149	0.0401	0.0133	0.0260	0.0129	0.0380	0.0109	0.0242	0.0114
T=100	$\hat{\beta}_2$	Mean	0.9975	1.0003	1.0002	0.9997	1.0028	1.0012	1.0007	1.0008	0.9980	0.9985	0.9985	0.9984
		Sd	0.0505	0.0223	0.0217	0.0220	0.0493	0.0239	0.0224	0.0232	0.0432	0.0209	0.0196	0.0198

**Table 5.2 Static model with stochastic factors**

		Missing Pattern 1				Missing Pattern 2				Missing Pattern 3				
		OLS	LSDV	LS-FPCA	LS-EM-PCA	OLS	LSDV	LS-FPCA	LS-EM-PCA	OLS	LSDV	LS-FPCA	LS-EM-PCA	
N=50	$\hat{\beta}_1$	Mean	2.4587	2.4602	2.4441	2.1079	2.4570	2.4616	2.4416	2.0260	2.4596	2.4632	2.4443	2.0528
		Sd	0.0303	0.0223	0.0222	0.0803	0.0277	0.0238	0.0244	0.0385	0.0248	0.0205	0.0210	0.0369
T=50	$\hat{\beta}_2$	Mean	0.9969	0.9995	1.0002	0.9998	0.9973	0.9997	1.0002	1.0001	1.0009	0.9987	0.9996	0.9997
		Sd	0.1007	0.0536	0.0534	0.0490	0.1030	0.0543	0.0541	0.0489	0.0891	0.0470	0.0467	0.0413
N=100	$\hat{\beta}_1$	Mean	2.4603	2.4626	2.4454	2.0917	2.4594	2.4642	2.4428	2.0184	2.4594	2.4639	2.4438	2.0501
		Sd	0.0237	0.0184	0.0191	0.0641	0.0227	0.0198	0.0220	0.0278	0.0215	0.0185	0.0196	0.0305
T=50	$\hat{\beta}_2$	Mean	1.0009	1.0000	1.0006	1.0002	0.9977	1.0004	1.0008	1.0006	0.9974	1.0008	1.0011	1.0012
		Sd	0.0705	0.0369	0.0371	0.0331	0.0699	0.0379	0.0372	0.0322	0.0638	0.0326	0.0321	0.0282
N=50	$\hat{\beta}_1$	Mean	2.4666	2.4686	2.4598	2.0923	2.4664	2.4699	2.4597	2.0165	2.4668	2.4699	2.4593	2.0358
		Sd	0.0241	0.0174	0.0161	0.0714	0.0212	0.0179	0.0170	0.0251	0.0214	0.0183	0.0165	0.0282
T=100	$\hat{\beta}_2$	Mean	1.0002	0.9974	0.9983	0.9992	1.0021	1.0005	1.0006	1.0004	1.0029	1.0011	1.0006	1.0001
		Sd	0.0716	0.0370	0.0363	0.0332	0.0717	0.0372	0.0367	0.0323	0.0641	0.0341	0.0334	0.0282
N=100	$\hat{\beta}_1$	Mean	2.4651	2.4680	2.4590	2.0879	2.4659	2.4698	2.4588	2.0111	2.4663	2.4699	2.4588	2.0355
		Sd	0.0196	0.0157	0.0145	0.0693	0.0183	0.0163	0.0155	0.0192	0.0187	0.0159	0.0145	0.0279
T=100	$\hat{\beta}_2$	Mean	0.9973	1.0000	1.0001	1.0003	1.0022	1.0005	1.0008	1.0011	0.9980	0.9988	0.9986	0.9982
		Sd	0.0525	0.0257	0.0254	0.0226	0.0519	0.0272	0.0269	0.0231	0.0451	0.0232	0.0224	0.0197

**Table 5.3 Dynamic model with smooth factors**

		Missing Pattern 1				Missing Pattern 2				Missing Pattern 3				
		OLS	LSDV	IV-FPCA	IV-EM-PCA	OLS	LSDV	IV-FPCA	IV-EM-PCA	OLS	LSDV	IV-FPCA	IV-EM-PCA	
N=50 T=50	$\hat{\rho}$	Mean	0.7058	0.5450	0.5207	0.5096	0.7060	0.5532	0.5210	0.5108	0.7044	0.5534	0.5200	0.5110
		Sd	0.0215	0.0055	0.0218	0.0172	0.0238	0.0071	0.0170	0.0183	0.0211	0.0050	0.0145	0.0142
	$\hat{\beta}_1$	Mean	1.6860	2.1134	2.0493	2.0193	1.6935	2.1378	2.0403	2.0197	1.6969	2.1365	2.0354	2.0166
		Sd	0.0813	0.0224	0.0464	0.0306	0.0774	0.0235	0.0373	0.0381	0.0695	0.0197	0.0311	0.0282
	$\hat{\beta}_2$	Mean	1.0021	1.0030	1.0036	1.0009	1.0041	1.0036	1.0031	1.0017	0.9991	1.0013	1.0017	1.0003
		Sd	0.0755	0.0460	0.0462	0.0489	0.0758	0.0454	0.0444	0.0483	0.0687	0.0409	0.0406	0.0432
N=100 T=50	$\hat{\rho}$	Mean	0.7081	0.5452	0.5200	0.5038	0.7085	0.5531	0.5184	0.5067	0.7066	0.5534	0.5180	0.5093
		Sd	0.0154	0.0042	0.0218	0.0110	0.0196	0.0064	0.0131	0.0105	0.0157	0.0039	0.0108	0.0084
	$\hat{\beta}_1$	Mean	1.6795	2.1138	2.0481	2.0109	1.6886	2.1380	2.0315	2.0103	1.6899	2.1369	2.0307	2.0137
		Sd	0.0577	0.0154	0.0407	0.0205	0.0612	0.0172	0.0267	0.0200	0.0531	0.0140	0.0223	0.0171
	$\hat{\beta}_2$	Mean	1.0020	1.0036	1.0040	1.0014	1.0021	1.0043	1.0046	1.0025	0.9981	1.0017	1.0021	1.0004
		Sd	0.0549	0.0320	0.0324	0.0335	0.0525	0.0307	0.0307	0.0317	0.0484	0.0293	0.0289	0.0300
N=50 T=100	$\hat{\rho}$	Mean	0.7205	0.5560	0.5199	0.5153	0.7208	0.5616	0.5169	0.5065	0.7201	0.5621	0.5159	0.5067
		Sd	0.0222	0.0044	0.0146	0.0125	0.0224	0.0051	0.0124	0.0112	0.0202	0.0039	0.0109	0.0087
	$\hat{\beta}_1$	Mean	1.6274	2.0921	2.0424	2.0312	1.6295	2.1059	2.0298	2.0104	1.6314	2.1049	2.0281	2.0102
		Sd	0.0779	0.0159	0.0308	0.0241	0.0733	0.0166	0.0259	0.0219	0.0669	0.0139	0.0219	0.0174
	$\hat{\beta}_2$	Mean	1.0042	1.0036	1.0033	1.0024	1.0050	1.0029	1.0029	1.0021	1.0008	1.0012	1.0013	1.0006
		Sd	0.0547	0.0326	0.0326	0.0348	0.0496	0.0337	0.0303	0.0322	0.0472	0.0288	0.0279	0.0292
N=100 T=100	$\hat{\rho}$	Mean	0.7210	0.5564	0.5207	0.5110	0.7206	0.5616	0.5146	0.5049	0.7220	0.5622	0.5131	0.5057
		Sd	0.0151	0.0031	0.0151	0.0070	0.0159	0.0039	0.0101	0.0067	0.0151	0.0029	0.0089	0.0058
	$\hat{\beta}_1$	Mean	1.6280	2.0919	2.0430	2.0227	1.6298	2.1060	2.0251	2.0078	1.6251	2.1053	2.0231	2.0088
		Sd	0.0555	0.0113	0.0313	0.0138	0.0526	0.0111	0.0186	0.0124	0.0505	0.0098	0.0166	0.0118
	$\hat{\beta}_2$	Mean	0.9972	1.0000	1.0002	0.9993	0.9993	1.0009	1.0006	0.9996	0.9998	1.0016	1.0012	1.0006
		Sd	0.0372	0.0228	0.0223	0.0230	0.0367	0.0228	0.0225	0.0230	0.0342	0.0205	0.0199	0.0203

**Table 5.4 Dynamic model with stochastic factors**

		Missing Pattern 1				Missing Pattern 2				Missing Pattern 3					
		OLS	LSDV	IV-FPCA	IV-EM-PCA	OLS	LSDV	IV-FPCA	IV-EM-PCA	OLS	LSDV	IV-FPCA	IV-EM-PCA		
N=50 T=50	$\hat{\rho}$	Mean	0.5514	0.5074	0.5104	0.5204	0.5521	0.5088	0.5103	0.5058	0.5498	0.5081	0.5102	0.5160	
		Sd	0.0154	0.0041	0.0053	0.0168	0.0158	0.0046	0.0054	0.0136	0.0136	0.0040	0.0047	0.0107	
	$\hat{\beta}_1$	Mean	2.3382	2.4462	2.4325	2.0764	2.3355	2.4425	2.4309	2.0157	2.3397	2.4449	2.4328	2.0226	
		Sd	0.0460	0.0200	0.0218	0.0588	0.0504	0.0199	0.0223	0.0385	0.0416	0.0186	0.0195	0.0423	
	$\hat{\beta}_2$	Mean	1.0047	1.0015	1.0025	1.0025	1.0063	1.0010	1.0022	1.0011	0.9995	1.0003	1.0012	1.0003	
		Sd	0.0957	0.0545	0.0544	0.0491	0.0943	0.0537	0.0524	0.0458	0.0870	0.0476	0.0470	0.0420	
	N=100 T=50	$\hat{\rho}$	Mean	0.5514	0.5074	0.5106	0.5181	0.5527	0.5089	0.5102	0.5041	0.5510	0.5081	0.5100	0.5155
		Sd	0.0137	0.0037	0.0045	0.0139	0.0151	0.0042	0.0049	0.0107	0.0131	0.0038	0.0041	0.0087	
N=100 T=100	$\hat{\beta}_1$	Mean	2.3356	2.4466	2.4320	2.0655	2.3318	2.4416	2.4291	2.0079	2.3358	2.4442	2.4321	2.0170	
		Sd	0.0400	0.0187	0.0192	0.0498	0.0462	0.0184	0.0212	0.0291	0.0415	0.0162	0.0175	0.0356	
	$\hat{\beta}_2$	Mean	1.0025	1.0015	1.0030	1.0028	1.0033	1.0029	1.0039	1.0025	0.9988	1.0008	1.0014	1.0011	
		Sd	0.0708	0.0373	0.0374	0.0341	0.0676	0.0352	0.0355	0.0311	0.0619	0.0340	0.0331	0.0292	
	$\hat{\rho}$	Mean	0.5483	0.5093	0.5107	0.5189	0.5489	0.5097	0.5107	0.5057	0.5479	0.5095	0.5107	0.5127	
		Sd	0.0122	0.0030	0.0032	0.0131	0.0128	0.0036	0.0034	0.0097	0.0113	0.0033	0.0028	0.0085	
	$\hat{\beta}_1$	Mean	2.3347	2.4437	2.4379	2.0689	2.3326	2.4419	2.4368	2.0125	2.3337	2.4421	2.4371	2.0184	
		Sd	0.0385	0.0153	0.0155	0.0513	0.0416	0.0164	0.0160	0.0270	0.0381	0.0154	0.0145	0.0283	
N=100 T=100	$\hat{\beta}_2$	Mean	1.0035	1.0017	1.0021	1.0032	1.0036	1.0015	1.0018	1.0017	1.0015	1.0005	1.0009	1.0009	
		Sd	0.0718	0.0380	0.0372	0.0341	0.0660	0.0358	0.0347	0.0312	0.0615	0.0330	0.0328	0.0287	
	$\hat{\rho}$	Mean	0.5487	0.5093	0.5108	0.5176	0.5492	0.5097	0.5107	0.5046	0.5480	0.5094	0.5106	0.5134	
		Sd	0.0105	0.0028	0.0028	0.0124	0.0113	0.0035	0.0032	0.0078	0.0100	0.0029	0.0026	0.0075	
	$\hat{\beta}_1$	Mean	2.3346	2.4433	2.4375	2.0612	2.3330	2.4413	2.4361	2.0088	2.3328	2.4433	2.4382	2.0204	
		Sd	0.0374	0.0125	0.0130	0.0483	0.0407	0.0147	0.0142	0.0214	0.0346	0.0124	0.0118	0.0258	
	$\hat{\beta}_2$	Mean	0.9975	0.9992	0.9998	0.9997	0.9994	1.0003	1.0003	0.9998	1.0005	1.0004	1.0008	1.0005	
		Sd	0.0484	0.0272	0.0267	0.0233	0.0465	0.0264	0.0260	0.0226	0.0436	0.0243	0.0231	0.0199	

- (c) In general, estimations with missing pattern 2 is more satisfactory than that of missing pattern 1.
- (d) Both OLS and LSDV are hardly consistent. The estimated  $\rho$  and  $\beta$  have large bias that does not decrease when sample size increases. In contrast, the bias of the proposed FPCA-type and EM-type estimators decays as the sample size becomes larger.

## 5.7 CONCLUSION

---

The missing data problem commonly exists in panel data. This chapter considers the estimation of panel models in the presence of interactive effects and missing observations. New algorithms are proposed to estimate the model parameters, and the common factors and factor loadings under various missing data patterns. When the common factors are smooth, the proposed LS-FPCA integrates the functional principal components analysis with OLS. When the factors are stochastic and non-smooth, we propose an LS-EM-PCA method that combines OLS with the EM algorithm to effectively deal with the missing observations in the panel data model. The method is applicable to dynamic panel data models. In this case, we use the lagged regressors as the instrumental variables.

Simulation studies show that the EM-type estimators are consistent and converge rapidly for both smooth and stochastic factors, while the FPCA-type estimators perform well only when the factors are smooth functions of the time.

This chapter focuses on the numerical estimation of panel data models, and no asymptotic analysis is provided. Proving consistency and further deriving the inferential theory of the proposed estimators can be technically difficult, which we leave as a future research topic.

## ACKNOWLEDGMENTS

---

Bai acknowledges financial supports from the NSF (SES-0962410) and Yang acknowledges financial supports from the NSFC (Grant No.71271096).

## REFERENCES

---

- Afifi, A. A., and Elashoff, R. M. (1966), “Missing Observations in Multivariate Statistics,” *Journal of American Statistical Association*, 61, 595–604.
- Ahn, S., Lee, Y., and Schmidt, P. (2013), “Panel Data Models with Multiple Time-Varying Individual Effects.” *Journal of Econometrics*, 174, 1–14.

- Antweiler, W. (2001), "Nested Random Effects Estimation in Unbalanced Panel Data, *Journal of Econometrics*, 101, 295–313.
- Arellano, M. (2003), *Panel Data Econometrics*, Oxford: Oxford University Press.
- Arellano, M. and Bond, S. R. (1991), "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an application to employment equations", *Review of Economic Studies*, 58, 277–297.
- Arellano, M. and Bover, O. (1995), "Another look at the instrumental variable estimation of error-components models," *Journal of Econometrics*, 68, 29–51.
- Arellano, M. and Honoré, B. (2001), "Panel Data Models: Some Recent Developments." In: J.J. Heckman and E. Leamer (eds.) *Handbook of Econometrics*, V5, Ch 53, 3229–3296.
- Bai, J. (2009), "Panel Data Models with Interactive Fixed Effects," *Econometrica*, 77(4), 1229–1279.
- Bai, J. (2010), "Likelihood Approach to Small  $T$  Dynamic Panel Models with Interactive Effects," Working Paper, Department of Economics, Columbia University.
- Bai, J. and Li, K. (2012), "Theory and methods of panel data models with interactive effects," unpublished manuscript, Department of Economics, Columbia University.
- Bai J. and S. Ng (2002), "Determining the Number of Factors in Approximate Factor Models", *Econometrica*, 70(1), 191–221.
- Baltagi, B. H. (2008), *Econometric Analysis of Panel Data*, Fourth edition, Chichester: John Wiley and Sons.
- Baltagi, B. H. and Chang, Y. J. (1994), "Incomplete Panels: A Comparative Study of Alternative Estimators for the Unbalanced One-way Error Component Regression Model, *Journal of Econometrics*, 62, 67–89.
- Baltagi, B. H. and Song, S. H. (2006), "Unbalanced Panel Data: A Survey, *Statistical Papers*, 47, 493–523.
- Baltagi, Badi, Garvin, H., Kerman, S. (1989), "Further Monte Carlo Evidence on Seemingly Unrelated Regressions with Unequal Number of Observations", *Annales d'Économie et de Statistique*, 14, 103–115.
- Baltagi, B. H., Song, S. H. and Jung, B. C. (2001), "The Unbalanced Nested Error Component Regression Model," *Journal of Econometrics*, 101, 357–381.
- Bańbura, M. and Modugno, M. (2010), "Maximum Likelihood Estimation of Factor Models on Data Sets with Arbitrary Pattern of Missing Data," Working Paper Series 1189, European Central Bank.
- Beale, E. M. L. and Little, R. J. L. (1975), "Missing Values in Multivariate Analysis," *Journal of the Royal Statistical Society (B)*, 37, 129–145.
- Biorn, E. (1981), "Estimating Economic Relations from Incomplete Cross-Section/ Time-Series Data", *Journal of Econometrics*, 16, 221–236.
- Blundell, R. and Bond, S. (1998), "Initial Conditions and Moment Restrictions in Dynamics Panel Data Models", *Journal of Econometrics*, 87, 115–143.
- Boente, G. and Fraiman, R. (2000), "Kernel-based Functional Principal Components," *Statistics Probability Letters*, 48, 335–345.
- Breitung, J. and J. Tenhofen (2011), "GLS Estimation of Dynamic Factor Models", *Journal of the American Statistical Association*, 106, 1150–1166.
- Cardot, H. (2000), "Nonparametric Estimation of Smoothed Principal Components Analysis of Sampled Noisy Functions," *Journal of Nonparametric Statistics*, 12, 503–538.
- Choi, I. (2012), "Efficient Estimation of Factor Models," *Econometric Theory*, 28, 274–308.

- Conniffe, D. (1985), "Estimating Regression Equations with Common Explanatory Variables but Unequal Numbers of Observations," *Journal of Econometrics*, 27, 179–196.
- Davis, P. (2001), "Estimating Multi-way Error Components Models with Unbalanced Data Structures Using Instrumental Variables," *Journal of Econometrics*, 106, 67–95.
- Hocking, R. R. and Smith, W. B. (1968), "Estimation of Parameters in the Multivariate Normal Distribution with Missing Observations," *Journal of American Statistical Association*, 63, 159–173.
- Hsiao, C. (2003), *Analysis of Panel Data*, Second edition, Cambridge: Cambridge University Press.
- Hwang, H. S. (1990), "Estimation of Linear SUR Model with Unequal Numbers of Observations," *Review of Economics and Statistics*, 72, 510–515.
- Hwang, H. S. and Schulman, C. (1996), "Estimation of SUR Model with Non-block missing Observations," *Annals of Economics and Statistics*, 44, 219–240.
- James, G. M., Hastie, T. J., and Sugar, C. A. (2000), "Principal Component Models for Sparse Functional Data," *Biometrika*, 87, 587–602.
- Jolliffe, I. T. (2002), *Principal Component Analysis*, New York: Springer.
- Kneip, A., Sickles, R. C., and Song, W. (2012), "A New Panel Data Treatment for Heterogeneity in Time Trends," *Econometric Theory*, 28, 590–628.
- McLachlan, G. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. New York: Wiley.
- Meng, X. and van Dyk, D. "The EM Algorithm-An Old Folk-Song Sung to a Fast New Tune," *Journal of the Royal Statistical Society, Ser. B*, 59, 511–567.
- Moon, R. and Weidner, M. (2010), "Dynamic Linear Panel Regression Models with Interactive Fixed Effects, Manuscript. University of South California.
- Paul, D. and Peng, J. (2009), "Consistency of Restricted Maximum Likelihood Estimators of Principal Components," *Annals of Statistics*, 37(3), 1229–1271.
- Peng, J. and Paul, D. (2009), "A Geometric Approach to Maximum Likelihood Estimation of the Functional Principal Components from Sparse Longitudinal Data," *Journal of Computational and Graphical Statistics*, 18(4), 995–1015.
- Pesaran, M.H. (2006). "Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure." *Econometrica*, 74, 967–1012.
- Rice, J. A. and Wu, C. (2001), "Nonparametric Mixed Effects Models for Unequally Sampled Noisy Curves," *Biometrics*, 57, 253–259.
- Sarafidis, V. and Yamagata, T. (2010), "Instrumental Variable Estimation of Dynamic Linear Panel Data Models with Defactored Regressors under Cross-sectional Dependence"; Working paper, University of York.
- Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*. Chapman and Hall/CRC, London.
- Schmidt, P. (1977), "Estimation of Seemingly Unrelated Regressions with Unequal Numbers of Observations," *Journal of Econometrics*, 5, 365–377.
- Stock, J. H. and Watson, M. W. (1998), "Diffusion indexes," NBER Working paper, no. 6702.
- Trawinski, I. M. and Bargmann, R. E. (1964), "Maximum Likelihood Estimation with Incomplete Multivariate Data," *Annals of Mathematical Statistics*, 35, 647–657.
- Wansbeek, T. and Kapteyn, A. (1989), "Estimation of the Error-Components Model with Incomplete Panels," *Journal of Econometrics*, 41, 341–361.
- Yao, F., Müller, H., and Wang, J. (2005), "Functional Data Analysis for Sparse Longitudinal Data," *Journal of the American Statistical Association*, 100, 577–590.

## CHAPTER 6

---

# PANEL DATA MODELS FOR DISCRETE CHOICE

---

WILLIAM GREENE

### 6.1 INTRODUCTION

---

We survey the intersection of two large areas of research in applied and theoretical econometrics. Panel data modeling broadly encompasses nearly all of modern microeconometrics and some of macroeconomics as well. Discrete choice is the gateway to and usually the default framework in discussions of nonlinear models in econometrics. We will select a few specific topics of interest: essentially modeling cross sectional heterogeneity in three foundational discrete choice settings: binary, ordered multinomial, unordered multinomial. A fourth, count data frameworks, is treated in detail in Cameron and Trivedi (2013).

We will examine some of the empirical models used in recent applications, mostly in parametric forms of panel data models. Formal development of the discrete outcome models described above can be found in numerous sources, such as Greene (2012) and Cameron and Trivedi (2005). We will focus on extensions of the models to panel data applications. The development here can contribute a departure point to the more specialized treatments such as Keane's (this volume) study of panel data discrete choice models of consumer demand or more theoretical discussions, such as Lee's (this volume) extensive development of dynamic multinomial logit models. Toolkits for practical application of most of the models noted here are built into familiar modern software such as Stata, SAS, R, NLOGIT, MatLab, and so on.

There are two basic threads of development of discrete choice models. *Random utility* based models emphasize the *choice* aspect of discrete choice. Discrete choices are the observable revelations of underlying preferences. For example, McFadden (1974) develops the random utility approach to multinomial qualitative choice. The fundamental building block is the binary choice model, which we associate with an agent's

revelation of their preference for one specific outcome over another. Ordered and unordered choice models build on this basic platform. The familiar estimation platforms, univariate probit and logit, ordered choice (see Greene and Hensher 2010), and multinomial logit for the former type and Poisson and negative binomial regressions for counts have been developed and extended in a vast literature. The extension of panel data models for heterogeneity and dynamic effects, which have been developed for linear regression in an equally vast literature, into these nonlinear settings is a bit narrower and is the subject of this chapter.

The second dimension of the treatment here is panel data modeling. The modern development of large, rich *longitudinal survey data sets*, such as the German Socioeconomic Panel (GSOEP), Household Income and Labor Dynamics in Australia (HILDA), Survey of Income and Program Participation (SIPP, US), British Household Panel Survey (BHPS), Medical Expenditure Panel Survey (MEPS, US), and European Community Household Panel Survey (ECHP) to name a few, has supported an ongoing interest in analysis of individual outcomes across households and within households through time. The BHPS, for example, now in its eighteenth wave, is long enough to have recorded a significant fraction of the life cycle of many family members. The National Longitudinal Survey (NLS, US) was begun in the 1960s and notes that for some purposes they have entered their second generation. Each of these surveys includes questions on discrete outcomes such as labor force participation, banking behavior, self-assessed health, subjective well-being, health care decisions, insurance purchase, and many others. The discrete choice models already noted are the natural platforms for analyzing these variables. For present purposes, a specific treatment of “panel data models” is motivated by interesting features of the population that can be studied in the context of longitudinal data, such as cross-sectional heterogeneity and dynamics in behavior and on estimation methods that differ from cross-section linear regression counterparts. We will narrow our focus to individual data. Contemporary applications include many examples in health economics: such as in Riphahn, Wambach, and Million’s (2003) study of insurance take-up and health care utilization using the GSOEP and Contoyannis, Rice, and Jones’s (2004) analysis of self-assessed health in the BHPS.

## 6.2 DISCRETE OUTCOME MODELS

---

We will denote the models of interest here as *discrete outcome models*. The data-generating process takes two specific forms, *random utility models* and *nonlinear regression models* for counts of events. In some applications, there is a bit of fuzziness of the boundary between these. Bhat and Pulugurta (1998) treat the number of vehicles owned, naturally a count, as a revelation of preferences for transport services (i.e., in a utility-based framework). For random utility, the departure point is the existence of

an individual preference structure that implies a utility index defined over states, or alternatives, that is,

$$U_{it,j} = U(\mathbf{x}_{it,j}, \mathbf{z}_i, A_i, \varepsilon_{it,j}).$$

We use  $\mathbf{x}_{it,j}$  to denote choice ( $j$ ) and choice situation ( $t$ ) varying attributes, such as a price,  $\mathbf{z}_i$  to denote invariant observable characteristics of the individual, such as gender, and  $A_i$ , generically, to denote unobserved heterogeneity (characteristics) of the chooser. Preferences are assumed to obey the familiar axioms—completeness, transitivity, and so on—we take the underlying microeconomic theory as given. In the econometric specification, “ $j$ ” indexes the alternative, “ $i$ ” indexes the individual and “ $t$ ” may index the particular choice situation in a set of  $T_i$  situations. In the cross-section case,  $T_i = 1 \dots$ . In panel data applications, the case  $T_i > 1$  will be of interest. The index “ $t$ ” is intended to provide for possible sequence of choices, such as consecutive observations in a longitudinal data setting or a stated choice experiment. The number of alternatives,  $J$ , may vary across both  $i$  and  $t$ —consider a stated choice experiment over travel mode or consumer brand choices in which individuals choose from possibly different available choice sets as the experiment progresses through time. Analysis of brand choices, for example, for ketchup, yogurt, and other consumer products based on the scanner data, is a prominent example from marketing research (see Allenby, Garrett, and Rossi 2010). With possibly some small loss of generality, we will assume that  $J$  is fixed throughout the discussion.

The number of choice situations,  $T$ , may vary across  $i$ . Most received theoretical treatments assume fixed (*balanced*)  $T$  largely for mathematical convenience, although many actual longitudinal data sets are *unbalanced*, that is, have variation in  $T_i$  across  $i$ . At some points this is a minor mathematical inconvenience—variation in  $T_i$  across  $i$  mandates a much more cumbersome notation than fixed  $T$  in most treatments. But, the variation in  $T_i$  can be substantive. If “unbalancedness” of the panel is the result of *endogenous attrition* in the context of the outcome model being studied, then a relative to the problem of *sample selection* becomes pertinent (see Heckman 1979 and a vast literature). The application to self-assessed health in the BHPS by Contoyannis, Jones, and Rice (2004) described below is an example. Wooldridge (2002) and Semykina and Wooldridge (2013) suggests procedures for modeling nonrandom attrition in binary choice and linear regression settings.

The data,  $\mathbf{x}_{it,j}$ , will include observable attributes of the outcomes, time varying characteristics of the chooser, such as age, and, possibly, previous outcomes;  $\mathbf{z}_i$  are time and choice invariant characteristics of the chooser, typically demographics such as gender or (in a stated preference experiment) income;  $\varepsilon_{it,j}$  is time varying and/or time invariant, unobserved, and random characteristics of the chooser. We will assume away at this point any need to consider the time series properties of  $\mathbf{x}_{it}$ —non-stationarity, for example. These are typically of no interest in longitudinal data applications. We do note that as the length of some panels such as the NLS, GSOEP, and the BHPS grow, the structural stability of the relationship under study might at least be questionable. Variables such as age and experience will appear non-stationary and mandate some

consideration of the nature of cross-period correlations. This consideration has also motivated a broader treatment of macroeconomic panel data such as the Penn World Tables. But, interest here is in individual, discrete outcomes for which these considerations are tangential or moot. The remaining element of the model is  $A_i$ , which will be used to indicate the presence of choice and time invariant, *unobservable heterogeneity*. As is common in other settings, the unobserved heterogeneity could be viewed as unobservable elements of  $\mathbf{z}_i$ , but it is more illuminating to isolate  $A_i$ .

The observation mechanism defined over the alternatives can be interpreted as a revelation of preferences;

$$y_{it} = G(U_{it,1}, U_{it,2}, \dots, U_{it,J})$$

The translation mechanism that maps underlying preferences to observed outcomes is part of the model. The most familiar (by far) application is the discrete choice over two alternatives, in which

$$y_{it} = G(U_{it,1}, U_{it,2}) = \mathbf{1}(U_{it,2} - U_{it,1} > 0). \quad (3)$$

Another common case is the *unordered multinomial choice* case in which  $G(\cdot)$  indexes the alternative with maximum utility.

$$\begin{aligned} y_{it} = G(U_{it,1}, U_{it,2}, \dots, U_{it,J}) &= j \text{ such that } U_{it,j} > U_{it,k} \forall j \neq k; \\ j, k &= 1, \dots, J \end{aligned}$$

(see, e.g., McFadden 1974). The convenience of the single outcome model comes with some loss of generality. For example, Van Dijk, Fok, and Paap (2007) examine a *rank ordered logit model* in which the observed outcome is the subject's vector of ranks (in their case, of six video games), as opposed to only the single most preferred choice. Multiple outcomes at each choice situation, such as this one, are somewhat unusual. Not much generality lost by maintaining the assumption of a scalar outcome—modification of the treatment to accommodate multiple outcomes will generally be straightforward. We can also consider a multivariate outcome in which more than one outcome is observed in each choice situation (see, e.g., Chakir and Parent 2009). The multivariate case is easily accommodated, as well. Finally, the *ordered multinomial choice* model is not one that describes utility maximization as such, but rather, a feature of the preference structure itself;  $G(\cdot)$  is defined over a single outcome, such that

$$\begin{aligned} y_{it} = G(U_{it,1}) &= j \text{ such that } U_{it,1} \in \text{ the } j^{\text{th}} \text{ interval of a partition of the real line,} \\ &(-\infty, \mu_0, \mu_1, \dots, \mu_J, \infty). \end{aligned}$$

The preceding has focused on random utility as an organizing principle. A second thread of analysis is models for counts. These are generally defined by the observed outcome and a discrete probability distribution

$$y_{it} = \#(\text{events for individual } i \text{ at time } t).$$

Note the inherently dynamic nature of the statement; in this context, “ $t$ ” means observed in the interval from the beginning to the end of a time period denoted  $t$ . Applications are typically normalized on the length of the observation window, such as the number of traffic incidents per day at given locations, or the number of messages that arrive at a switch per unit of time, or a physical dimension of the observation mechanism, such as the incidence of diabetes per thousand individuals. The “model” consists, again, of the observed data mechanism and a characterization of an underlying probability distribution ascribed to the rate of occurrence of events. The core model in this setting is a discrete process described by a distribution such as the Poisson or negative binomial distribution. A broader view might also count the number of events until some absorbing state is reached—for example, the number of periods that elapses until bankruptcy occurs, and so on. The model may also define treatments of sources of random variation, such as the negative binomial model or normal mixture models for counts that add a layer of unobservable heterogeneity into the Poisson platform. There is an intersection of the two types of models we have described. A *hurdle model* (see Mullahy 1987 and, e.g., Harris and Zhao’s (2007) analysis of smoking behavior) consists of a binary (utility-based) choice of whether to participate in an activity followed by an intensity equation or model that describes a count of events. Bago d’Uva (2006), for example, models health care usage using a latent class hurdle model and the BHPS data.

For purposes of developing the methodology of discrete outcome modeling in panel data settings, it is sufficient to work through the binary choice outcome in detail. Extensions to other choice models from this departure point are generally straightforward. However, we do note one important point at which this is decidedly not the case. A great deal has been written about semiparametric and nonparametric approaches to choice modeling. However, nearly all of this analysis has focused on binary choice models. The extension of these methods to multinomial choice, for example, is nearly nonexistent. Partly for this reason, and with respect to space limitations, with only an occasional exception, our attention will focus on parametric models. It also follows naturally that nearly all of the estimation machinery, both classical and “Bayesian” is grounded in likelihood-based methods.

### 6.3 INDIVIDUAL HETEROGENEITY IN A PANEL DATA MODEL OF BINARY CHOICE

---

After conventional estimation, in some cases, a so-called “*cluster correction*” (see Wooldridge 2003) is often used to adjust the estimated standard errors for effects

that would correspond to common unmeasured elements. But, the correction takes no account of heterogeneity in the estimation step. If the presence of unmeasured and unaccounted-for heterogeneity taints the estimator, then correcting the standard errors for “clustering” (or any other failure of the model assumptions) may be a moot point. This discussion will focus on accommodating heterogeneity in discrete choice modeling.

The binary choice model is the natural starting point in the analysis of “nonlinear panel data models.” Once some useful results are established, extensions to ordered choice models are generally straightforward and uncomplicated. There are only relatively narrow received treatments in unordered choice—we consider a few below. This leaves count data models which are treated conveniently later in discussions of nonlinear regression.

The base case is

$$\begin{aligned} y_{it} &= \mathbf{1}(U_{it,2} - U_{it,1} > 0) \\ U_{it,j} &= U(\mathbf{x}_{it,j}, \mathbf{z}_i, A_i, \varepsilon_{it,j}), \quad j = 1, 2. \end{aligned}$$

A linear utility specification (e.g., McFadden 1974) would be

$$U_{it,j} = U(\mathbf{x}_{it,j}, \mathbf{z}_i, A_i, \varepsilon_{it,j}) = \alpha_j + \beta'_j \mathbf{x}_{it,j} + \gamma' \mathbf{z}_i + \delta A_i + \varepsilon_{it,j},$$

where  $\varepsilon_{it,j}$  are independent and identically distributed across alternatives  $j$ , individuals,  $i$ , and period or choice situation,  $t$ . McFadden also assumed a specific distribution (type I extreme value) for  $\varepsilon_{it,j}$ . Subsequent researchers, including Manski (1975, 1985), Horowitz (1992), and Klein and Spady (1993) weakened the distribution assumptions. Matzkin (1991) suggested an alternative formulation, in which

$$U_{it,j} = U(\mathbf{x}_{it,j}, \mathbf{z}_i, A_i, \varepsilon_{it,j}) = V(\mathbf{x}_{it,j}, \mathbf{z}_i, A_i) + \varepsilon_{it,j}$$

with  $\varepsilon_{it,j}$  specified nonparametrically. In each of these cases, the question of what can be identified from observed data is central to the analysis. For McFadden’s model, for example, absent the complication of the unobserved  $A_i$ , all of the parameters shown are point identified, and probabilities and average partial effects can be estimated. Of course, the issue here is  $A_i$ , which is unobserved. Further fully parametric treatments (e.g., Train 2009) show how all parameters are identifiable. Under partially parametric approaches such as Horowitz (1992) or Klein and Spady (1993), parameters are identified up to scale (and location,  $\alpha$ ). This hampers computation of useful secondary results, such as probabilities and partial effects. Chesher and Smolinsky (2012) and Chesher and Rosen (2012a, b) and Chesher (2010, 2013) examine yet less parameterized cases in which point identification of interesting results such as marginal effects will be difficult. They consider specifications that lead only to set identification of aspects of preferences such as partial effects (see also Hahn 2010). Chernozhukov et al. (2013) also show that without some restrictions, average partial effects are not point identified in nonlinear models; they do indicate estimable sets for discrete covariates.

As Wooldridge (2010) notes, what these authors demonstrate is the large payoff to the palatable restrictions that we do impose in order to identify useful quantities in the parametric models that we estimate.

The generic model specializes in the binary case to

$$y_{it,j} = \mathbf{1}[U(\mathbf{x}_{it,j}, \mathbf{z}_i, A_i, \varepsilon_{it,j}) > 0].$$

The objective of estimation is to learn about features of the preferences, such as partial effects and probabilities attached to the outcomes as well as the superficial features of the model, which in the usual case would be a parameter vector. In the case of a probit model, for example, an overwhelming majority of treatment is devoted to estimation of  $\beta$  when the actual target is some measure of partial effect. This has been emphasized in some recent treatments, such as Wooldridge (2010) and Fernandez-Val (2009).

Combine the  $T_i$  observations on  $(\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT_i})$  in data matrix  $\mathbf{X}_i$ . The joint conditional density of  $y_{it}$  and  $A_i$  is

$$f(y_{i1}, y_{i2}, \dots, y_{it}, A_i | \mathbf{X}_i) = f(y_{i1}, y_{i2}, \dots, y_{it} | \mathbf{X}_i, A_i) f(A_i | \mathbf{X}_i).$$

A crucial ingredient of the estimation methodology is:

- *Conditional independence:* Conditioned on the observed data and the heterogeneity, the observed outcomes are independent. The joint density of the observed outcomes and the heterogeneity,  $A_i$ , can thus be written

$$\begin{aligned} f_{y_1, \dots, y_T}(y_{i1}, y_{i2}, \dots, y_{it} | \mathbf{X}_i, A_i) f_A(A_i | \mathbf{X}_i) \\ = \left[ \prod_{t=1}^{T_i} f_y(y_{it} | \mathbf{X}_i, A_i) \right] f_A(A_i | \mathbf{X}_i). \end{aligned}$$

Models of spatial interaction would violate this assumption (see Lee and Yu 2010 and Greene 2011a). The assumption will also be difficult to sustain when  $\mathbf{x}_{it}$  contains lagged values of  $y_{it}$ . The conditional log likelihood for a sample of  $n$  observations based on this assumption is

$$\log L = \sum_{i=1}^n \left\{ \left[ \sum_{t=1}^{T_i} \log f_y(y_{it} | A_i, \mathbf{X}_i) \right] + \log f_A(A_i | \mathbf{X}_i) \right\}.$$

If  $f_A(A_i | \mathbf{X}_i)$  actually involves  $\mathbf{X}_i$ , then this assumption is only a partial solution to setting up the estimation problem. It is difficult to construct a substantial application without this assumption. The challenge of developing models that include spatial correlation is the leading application (see Section 6.5 below).

The two leading cases are random and fixed effects. We will specialize to a linear utility function at this point,

$$U_{it} = \beta' \mathbf{x}_{it} + \gamma' \mathbf{z}_i + A_i + \varepsilon_{it}$$

and the usual observation mechanism

$$y_{it} = \mathbf{1}[U_{it} > 0].$$

We (semi) parameterize the data-generating process by assuming that there is a continuous probability distribution governing the random part of the model,  $\varepsilon_{it}$ , with distribution function  $F(\varepsilon_{it})$ . At least implicitly, we are directing our focus to cross-sectional variation. However, it is important to note possible unsystematic time variation in the process. The most general approach might be to loosen the specification of the model to  $F_t(\varepsilon_{it})$ . This would still require some statement of what would change over time and what would not—the heterogeneity carries across periods, for example. Time variation is usually not the main interest of the study. A common accommodation (again, see Wooldridge 2010) is a set of time dummy variables, so that

$$U_{it} = \beta' \mathbf{x}_{it} + \gamma' \mathbf{z}_i + \Sigma_t \delta_t d_{it} + A_i + \varepsilon_{it}.$$

Our interest is in estimating characteristics of the data generating process for  $y_{it}$ . Prediction of the outcome variable is considered elsewhere (e.g., Elliott and Leili 2013). We have also restricted our attention to features of the mean of the index function and mentioned scaling, or heteroscedasticity only in passing. (There has been recent research on less parametric estimators that are immune to heteroscedasticity. See, e.g., Chen and Khan 2003.) The semiparametric estimators suggested by Honoré and Kyriazidou (2000a, b) likewise consider explicitly the issue of heteroscedasticity. In the interest of brevity, we will leave this discussion for more detailed treatments of modeling discrete choices.

Two additional assumptions needed to continue are:

- *Random Sampling of the observation units:* All observation units  $i$  and  $t$  are generated and observed independently (within the overall framework of the data-generating process).
- *Independence of the random terms in the utility functions:* Conditioned on  $\mathbf{x}_{it}$ ,  $\mathbf{z}_i$ ,  $A_i$ , the unique random terms,  $\varepsilon_{it}$ , are statistically independent for all  $i, t$ .

The random sampling assumption is formed on the basis of all of the information that enters the analysis. Conceivably, the assumption could be violated, for example, in modeling choices made by participants in a social network or in models of spatial interaction. However, the apparatus described so far is wholly inadequate to deal with a modeling setting at that level of generality (see, e.g., Durlauf and Brock 2001a, b, 2002; Durlauf et al. 2010). Some progress has been made in modeling spatial correlation in discrete choices. However, the random effects framework has provided the only path to forward progress in this setting. The conditional independence assumption is crucial to the analysis.

### 6.3.1 Random Effects in a Static Model

The binary choice model with a common effect is

$$\begin{aligned} U_{it} &= \beta' \mathbf{x}_{it} + \gamma' \mathbf{z}_i + \Sigma_t \delta_t d_{it} + A_i + \varepsilon_{it}, \\ f_A(A_i | \mathbf{X}_i, \mathbf{z}_i) &= f_A(A_i), \\ y_{it} &= \mathbf{1}[U_{it} > 0]. \end{aligned}$$

Definitions of what constitutes a *random effects model* hinge on assumptions of the form of  $f_A(A_i | \mathbf{X}_i, \mathbf{z}_i)$ . For simplicity, we have made the broadest assumption, that the DGP of  $A_i$  is time invariant and independent of  $\mathbf{X}_i, \mathbf{z}_i$ . This implies that the conditional mean is free of the observed data;  $E[A_i | \mathbf{X}_i, \mathbf{z}_i] = E(A_i)$ . If there is a constant term in  $\mathbf{x}_{it}$ , then no generality is lost if we make the specific assumption  $E[A_i] = 0$  for all  $t$ . Whether the mean equals zero given all  $(\mathbf{X}_i, \mathbf{z}_i)$ , or equals zero given only the current (period  $t$ ) realization of  $\mathbf{x}_{it}$ , or specifically given only the past or only the future values of  $\mathbf{x}_{it}$  (none of which are testable) may have an influence on the estimation method employed (see, e.g., Wooldridge 2010, chapter 15). We also assume that  $\varepsilon_{it}$  are mutually independent and normally distributed for all  $i$  and  $t$ , which makes this a *random effects probit model*. Given the ubiquity of the logit model in cross-section settings, we will return below to the possibility of a *random effects logit* specification. The remaining question concerns the marginal (and, by assumption, conditional) distribution of  $A_i$ . For the present, motivated by the central limit theorem, we assume that  $A_i \sim N[0, \sigma_A^2]$ .

The log likelihood function for the parameters of interest is

$$\log L(\beta, \gamma, \delta | A_1, \dots, A_n) = \sum_{i=1}^n \log \left\{ \prod_{t=1}^{T_i} f_y(y_{it} | \mathbf{x}_{it}, t, A_i) \right\}.$$

The obstacle to estimation is the unobserved heterogeneity. The unconditional log likelihood is

$$\begin{aligned} \log L(\beta, \gamma, \delta) &= \sum_{i=1}^n \log \left\{ E_A \left[ \prod_{t=1}^{T_i} f_y(y_{it} | \mathbf{x}_{it}, A_i) \right] \right\} \\ &= \sum_{i=1}^n \log \left\{ \int_{-\infty}^{\infty} \left[ \prod_{t=1}^{T_i} f_y(y_{it} | \mathbf{x}_{it}, A_i) f_A(A_i) dA_i \right] \right\}. \end{aligned}$$

It will be convenient to specialize this to the random effects probit model. Write  $A_i = \sigma u_i$ , where  $u_i \sim N[0, 1]$ . The log likelihood becomes

$$\begin{aligned} \log L(\beta, \gamma, \delta, \sigma) &= \sum_{i=1}^n \log \left\{ \int_{-\infty}^{\infty} \left[ \prod_{t=1}^{T_i} \Phi[(2y_{it} - 1)(\alpha + \beta' \mathbf{x}_{it} \right. \right. \\ &\quad \left. \left. + \gamma' \mathbf{z}_i + \Sigma_t \delta_t d_{it} + \sigma u_i)] \phi(u_i) du_i \right] \right\}, \end{aligned}$$

where  $\Phi(\cdot)$  and  $\phi(\cdot)$  are the cdf and density of the standard normal distribution. (Note that we have exploited the symmetry of the normal distribution to combine the  $y_{it} = 0$  and  $y_{it} = 1$  terms.) To save some notation, for the present we will absorb the constant, time invariant variables and time dummy variables in  $\mathbf{x}_{it}$  and the corresponding parameters in  $\beta$  to obtain

$$\log L(\beta, \sigma) = \sum_{i=1}^n \log \left\{ \int_{-\infty}^{\infty} \left[ \prod_{t=1}^{T_i} \Phi[(2y_{it} - 1)(\beta' \mathbf{x}_{it} + \sigma u_i)] \phi(u_i) du_i \right] \right\}.$$

Two methods can be used in practice to obtain the maximum likelihood estimates of the parameters, Gauss-Hermite quadrature as developed by Butler and Moffitt (1982) and maximum simulated likelihood as analyzed in detail in Train (2003, 2009) and Greene (2012). The approximations to the log likelihood are

$$\log L_H(\beta, \sigma) = \sum_{i=1}^n \log \left\{ \sum_{h=1}^H w_h \left[ \prod_{t=1}^{T_i} \Phi[(2y_{it} - 1)(\beta' \mathbf{x}_{it} + \sigma W_h)] \right] \right\}$$

for the Butler and Moffitt approach, where  $(w, W)_h$ ,  $h = 1, \dots, H$  are the weights and nodes for an  $H$  point Hermite quadrature, and

$$\log L_S(\beta, \sigma) = \sum_{i=1}^n \log \left\{ \frac{1}{R} \sum_{r=1}^R \left[ \prod_{t=1}^{T_i} \Phi[(2y_{it} - 1)(\beta' \mathbf{x}_{it} + \sigma u_{ir})] \right] \right\},$$

for the maximum simulated likelihood approach, where  $u_{ir}$ ,  $r = 1, \dots, R$  are  $R$  pseudo-random draws from the standard normal distribution. Assuming that the data are well behaved and the approximations are sufficiently accurate, the likelihood satisfies the usual regularity conditions, and the MLE (or MSLE) is root- $n$  consistent, asymptotically normally distributed and invariant to one to one transformations of the parameters (for discussion of the additional assumptions needed to accommodate the use of the approximations to the log likelihood, see Train 2009). Bhat (1999) discusses the use of Halton sequences and other nonrandom methods of computing  $\log L_S$ . The quadrature method is widely used in contemporary software such as Stata (see Rabe-Hesketh, Skrondal, and Pickles 2005) SAS, and NLOGIT. Inference can be based on the usual trinity of procedures.

A *random effects logit model* would build off the same underlying utility function,

$$\begin{aligned} U_{it} &= \beta' \mathbf{x}_{it} + u_i + \varepsilon_{it}, \\ f_u(u_i) &= N[0, 1], f_\varepsilon(\varepsilon_{it}) = \frac{\exp(\varepsilon_{it})}{[1 + \exp(\varepsilon_{it})]^2} \\ y_{it} &= 1[U_{it} > 0]. \end{aligned}$$

The change in the earlier log likelihood is trivial—the normal cdf is replaced by the logistic (change “ $\Phi(\cdot)$ ” to “ $\Lambda(\cdot)$ ” (the cdf of the standard logistic distribution) in the

theory). It is more difficult to motivate the mixture of distributions in the model. The logistic model is usually specified in the interest of convenience of the functional form, while the random effect is the aggregate of all relevant omitted time invariant effects—hence the appeal to the central limit theorem. As noted, the modification of either of the practical approaches to estimation is trivial. A more orthodox approach would retain the logistic assumption for  $u_i$  as well as  $\varepsilon_{it}$ . It is not possible to adapt the quadrature method to this case as the Hermite polynomials are based on the normal distribution. But, it is trivial to modify the simulation estimator. In computing the simulated log likelihood function and any derivative functions, pseudo random normal draws are obtained by using  $u_{ir} = \Phi^{-1}(U_{ir})$ , where  $U_{ir}$  is either a pseudo-random  $U[0,1]$  draw, a Halton draw, or some other intelligent draw. To adapt the estimator to a logistic simulation, it would only be necessary to replace  $\Phi^{-1}(U_{ir})$  with  $\Lambda^{-1}(U_{ir}) = \log[U_{ir}/(1 - U_{ir})]$  (i.e., replace one line of computer code). The logit model becomes less natural as the model is extended in, for example, multiple equation directions and gives way to the probit model in nearly all recent applications.

The preceding is generic. The log likelihood function suggested above needs only to be changed to the appropriate density for the variable to adapt it to, for example, an ordered choice model or one of the models for count data. We will return briefly to this issue below.

### 6.3.1.1 Partial Effects

Partial effects in the presence of the heterogeneity are

$$\Delta(x, u) = \frac{\partial B(\beta'x + \sigma u)}{\partial x} = \beta B'(\beta'x + \sigma u),$$

where  $B(\cdot)$  is the function of interest, such as the probability, odds ratio, willingness to pay, or some other function of the latent index,  $\beta'x + \sigma u$ . The particular element of  $x$  might be a binary variable,  $D$ , with parameter  $\beta_D$ , in which case, the effect would be computed as  $B(\beta'x + \beta_D + \sigma u) - B(\beta'x + \sigma u)$ . If the index function includes a categorical variable such as education coded in levels such as  $ED_{low}$ ,  $ED_{hs}$ ,  $ED_{college}$ ,  $ED_{post}$ , the partial effects might be computed in the form of a transition matrix of effects,  $T$ , in which the  $ij^{th}$  element is

$$T_{from,to} = B(\beta'x + \beta_{to} + \sigma u) - B(\beta'x + \beta_{from} + \sigma u)$$

(for an application of this type of computation, see Contoyannis, Jones, and Rice 2004). For convenience, we will assume that  $\Delta(x, u)$  is computed appropriately for the application. The coefficients,  $\beta$  and  $\sigma$ , have been consistently estimated. The partial effect can be estimated directly at specific values of  $u$ , for example, its mean of zero. An *average partial effect* can also be computed. This would be

$$\Delta_x(x) = E_u \left[ \frac{\partial B(\beta'x + \sigma u)}{\partial x} \right] = \left[ \frac{\partial E_u[B(\beta'x + \sigma u)]}{\partial x} \right] = \left[ \frac{\partial [B_x(\beta'_x x)]}{\partial x} \right],$$

where  $B_x(\beta'_x \mathbf{x})$  is the expected value over  $u$  of the function of interest. The average partial effect will not equal the partial effect, as  $B_x(\cdot)$  need not equal  $B(\cdot)$ . Whether this average function is of interest is specific to the application. For the random effects probability model, we would usually begin with  $\text{Prob}(Y = 1|\mathbf{x}, u)$ . In this case, we can find  $B(\beta' \mathbf{x} + \sigma u) = \Phi(\beta' \mathbf{x} + \sigma u)$  while  $B_x(\mathbf{x}) = \Phi(\beta' \mathbf{x}/(1 + \sigma^2)^{1/2})$ . The simple partial effect is  $\partial \Phi(\beta' \mathbf{x} + \sigma u)/\partial \mathbf{x} = \phi(\beta' \mathbf{x} + \sigma u)\beta$  while the average partial effect is

$$\Delta_x(\mathbf{x}) = \frac{\partial \Phi\left(\frac{\beta' \mathbf{x}}{\sqrt{1+\sigma^2}}\right)}{\partial \mathbf{x}} = \frac{1}{\sqrt{1+\sigma^2}} \beta \phi\left(\frac{\beta' \mathbf{x}}{\sqrt{1+\sigma^2}}\right).$$

With estimates of  $\beta$  and  $\sigma$  in hand, it would be possible to compute the partial effects at specific values of  $u_i$ , such as zero. Whether this is an interesting value to use is questionable. However, it is also possible to obtain an estimate of the average partial effect, directly after estimation. Indeed, if at the outset, one simply ignores the presence of the heterogeneity, and uses maximum likelihood to estimate the parameters of the “*population averaged model*”

$$\text{Prob}(y = 1|\mathbf{x}) = \Phi(\beta'_x \mathbf{x}),$$

then the estimator consistently estimates  $\beta_x = \beta' \mathbf{x}/(1 + \sigma^2)^{1/2}$ . Thus, while conventional analysis does not estimate the parameters of the structural model, it does estimate something of interest, namely the parameters and partial effects of the population averaged model.

### 6.3.1.2 Alternative Models for the Random Effects

The random effects may enter the model in different forms. The so-called “generalized estimating equation” (GEE) (see Diggle, Liang, and Zeger 1994) approach to this analysis is difficult to motivate rigorously, but it is (loosely) generated by a seemingly unrelated regressions approach built around

$$y_{it} = \Phi(\beta' \mathbf{x}_{it}) + v_{it},$$

where the probability is also the regression function. A similar view is suggested by the *panel probit model* in Bertschuk and Lechner (1998),

$$U_{it} = \beta' \mathbf{x}_{it} + \varepsilon_{it},$$

$$\text{Cov}(\varepsilon_{it}, \varepsilon_{js}) = 1[i = j]\sigma_{ts}.$$

$$y_{it} = \mathbf{1}[U_{it} > 0].$$

Here, the SUR specification applies to the latent utilities, rather than the observed outcomes. The GEE estimator is estimated by a form of nonlinear generalized least squares. The terms in the log likelihood function for Bertschuk and Lechner’s model are  $T$ -variate normal probabilities. This necessitates computation of higher order

normal integrals. The authors devise a GMM estimator that avoids the burdensome calculations. Recent implementations of the GHK simulator and advances in computation capabilities do make the computations more reasonable (see Greene 2004a).

Heckman and Singer (1984) questioned the need for a full parametric specification of the distribution of  $u_i$ . (Their analysis was in the context of models for duration, but extends directly to this one.) A semiparametric, discrete specification based on their model could be written

$$u_i \subset (\alpha_1, \dots, \alpha_Q) \text{ with } \text{Prob}(u_i = \alpha_q) = \pi_q, q = 1, \dots, Q.$$

This gives rise to a “latent class” model, for which the log likelihood would be

$$\log L(\alpha, \beta, \pi) = \sum_{i=1}^n \log \left\{ \sum_{q=1}^Q \pi_q \left[ \prod_{t=1}^{T_i} \Phi[(2y_{it} - 1)(\alpha_q + \beta' \mathbf{x}_{it})] \right] \right\}.$$

This would be a partially semiparametric specification—it retains the fully parametric probit model as the platform. Note that this is a discrete counterpart to the continuous mixture model in Section 6.3.1.

The random effects model is, in broader terms, a *mixed model*. A more general statement of the *mixed model* would be

$$\begin{aligned} U_{it} &= (\beta + \mathbf{u}_i)' \mathbf{x}_{it} + \varepsilon_{it}, \\ f(\mathbf{u}_i | \mathbf{X}_i, \mathbf{z}_i) &= f(\mathbf{u}_i), \mathbf{u}_i \sim N[\mathbf{0}, \Sigma], \\ y_{it} &= \mathbf{1}[U_{it} > 0]. \end{aligned}$$

The extension here is that the entire parameter vector, not just the constant term, is heterogeneous. The mixture model used in recent applications is either continuous (see, e.g., Train 2009 and Rabe-Hesketh, Skrondal, and Pickles 2005) or discrete in the fashion suggested by Heckman and Singer (1984) (see Greene and Hensher 2010). Altonji and Matzkin (2005) considered other semiparametric specifications.

### 6.3.1.3 Specification Tests

It would be of interest to test for the presence of random effects against the null of the “pooled” model. That is, ultimately, a test of  $\sigma = 0$ . The test is complicated by the fact that the null value of  $\sigma$  lies on the boundary of the parameter space. See Breusch and Pagan (1980) and Chesher (1984) for discussion. In the random effects probit model, direct approaches based on the Wald or LR tests are available. The LM test has a peculiar feature; the score of the log likelihood is identically zero at  $\sigma = 0$ . Chesher (1984), Chesher and Lee (1986), and Cox and Hinkley (1974) suggest reparameterization of such models as a strategy for setting up the LM test. Greene and McKenzie

(2012) derived the appropriate statistic for the random effects probit model. The phenomenon would reappear in an ordered probit or ordered logit model as well. Their approach could be transported to those settings.

A second specification test of interest might be the distributional assumption. There is no natural residual based test such as the Bera and Jarque (1982) test for the linear regression. A test for the pooled (cross-section) probit model based essentially on Chesher and Irish's (1987) generalized residuals is suggested by Bera, Jarque, and Lee (1984). It is not clear how the test could be adapted to a random effects model, however, nor, in fact, whether it could be extended to other models such as ordered choice models. One possibility might be an information matrix test based on a robust (perhaps bootstrapped) covariance matrix (see Horowitz 1994).

#### 6.3.1.4 Other Discrete Choice Models

Application of the random effects models described above to an ordered choice model requires only a minor change in the assumed density of the observed outcome (see Greene and Hensher 2010, pp. 275–278). All other considerations are the same. The ordered probit model does contain an additional source of heterogeneity, in the thresholds. Ongoing development of the ordered choice methodology includes specifications of the thresholds, which may respond to observed effects (Pudney and Shields 2000; Lee and Kimhi 2005; Greene and Hensher 2010) and to unobserved random effects (Harris, Hollingsworth, and Greene 2012).

Random effects in count data models would build on a familiar specification in the cross-section form. For a Poisson regression, we would have

$$\text{Prob}(Y = y_{it} | \mathbf{x}_{it}, u_i) = \frac{\exp(-\lambda_{it}) \lambda_{it}^{y_{it}}}{y_{it}!}, \lambda_{it} = \exp(\beta' \mathbf{x}_{it} + \sigma u_i).$$

Since  $\lambda_{it}$  is the conditional mean, at one level, this is simply a nonlinear random effects regression model. However, maximum likelihood is the preferred estimator. If  $u_i$  is assumed to have a log-gamma distribution (see Hausman, Hall, and Griliches (HHG) 1984), then the unconditional model becomes a negative binomial (NB) regression. Recent applications have used a normal mixture approach (see, e.g., Riphahn, Wambach, and Million 2003). The normal model would be estimated by maximum simulated likelihood or by quadrature based on Butler and Moffitt (1982). See Greene (1995) for an application. A random effects negative binomial model would be obtained by applying the same methodology to the NB probabilities. HHG (1984) treat the NB model as a distinct specification rather than as the result of the mixed Poisson. The normal mixed NB model is discussed in Greene (2012).

There is an ambiguity in the mixed unordered multinomial choice model because it involves several utility functions. A fully specified random effects multinomial logit model would be

$$\text{Prob}(y_{it} = j) = \frac{\exp(\alpha_j + \beta' \mathbf{x}_{it,j} + u_{i,j})}{\sum_{j=1}^J \exp(\alpha_j + \beta' \mathbf{x}_{it,j} + u_{i,j})}.$$

A normalization is required since the probabilities sum to one—the constant and the random effect in the last utility function equal zero. An alternative specification would treat the random effect as a single choice invariant characteristic of the chooser, which would be the same in all utility functions. It would seem that this would be easily testable using the likelihood ratio statistic. However, this specification involves more than a simple parametric restriction. In the first specification, (we assume) the random effects are uncorrelated but this is also dubious because one of them must equal zero. In the second form, by construction, the utility functions are equicorrelated. This is a substantive change in the preference structure underlying the choices. Finally, the counterpart to the fully random parameters model is the mixed logit model,

$$\text{Prob}(y_{it} = j) = \frac{\exp(\alpha_{j,i} + (\beta + \mathbf{u}_i)' \mathbf{x}_{it,j})}{\sum_{j=1}^J \exp(\alpha_{j,i} + (\beta + \mathbf{u}_i)' \mathbf{x}_{it,j})}.$$

(The alternative specific constants are treated as random parameters as well.) See McFadden and Train (2000), Hensher, Rose, and Greene (2005), and Hensher and Greene (2003).

### 6.3.2 Fixed Effects in a Static Model

The single index model is

$$f(y_{it} | \mathbf{x}_{it}, \mathbf{z}_i, \alpha_i) = f(y_{it}, \beta' \mathbf{x}_{it} + \gamma' \mathbf{z}_i + \alpha_i) = f(y_{it}, a_{it}).$$

For empirical purposes, the model is recast with the unobserved effects treated as parameters to be estimated;

$$a_{it} = \beta' \mathbf{x}_{it} + \gamma' \mathbf{z}_i + \alpha_i c_{it},$$

where  $c_{it}$  is an element of a set of  $n$  group dummy variables. (Note: this is the estimation strategy. The model specification does not imply that the common effects are parameters in the same way that elements of  $\beta$  are. At this point,  $\mathbf{x}_{it}$  does not contain an overall constant term.) The leading cases in the received literature are the *fixed effects probit model*,

$$f(y_{it}, a_{it}) = \text{Prob}(y_{it} = 1 | a_{it}) = \Phi[(2y_{it} - 1)a_{it}],$$

where  $\Phi(w)$  is the standard normal cdf, and *fixed effects logit model*

$$f(y_{it}, a_{it}) = \Lambda[(2y_{it} - 1)a_{it}] = \exp[(2y_{it} - 1)a_{it}] / \{1 + \exp[(2y_{it} - 1)a_{it}]\}.$$

The fixed effects model is distinguished from the random effects model by relaxing the assumption that the unobserved heterogeneity,  $A_i$  is distributed independently of the observed covariates,  $(\mathbf{X}_i, \mathbf{z}_i)$ —that is,  $f_A[A_i | \mathbf{X}_i, \mathbf{z}_i] = f_A(A_i)$ . In the fixed effects case, the conditional distribution is not specified and may depend on  $\mathbf{X}_i$ . Other cases of interest

are the ordered choice models and the Poisson and negative binomial models for count data. We will examine the binary choice models first, then briefly consider the others. Chamberlain's (1980) proposed approach notwithstanding, fixed effects models have not provided an attractive framework for analysis of multinomial unordered choices in the received literature. For most of the discussion, we can leave the model in generic form and specialize when appropriate.

No specific assumption is made about the relationship between  $\alpha_i$  and  $x_{it}$ . The possibility that  $E[\alpha_i | x_{i1}, \dots, x_{iT}] = m(X_i)$  is not ruled out. If no restrictions are placed on the joint distribution of the unobservable  $\alpha_i$  and the observed  $X_i$ , then the random effects apparatus of the previous sections is unusable— $x_{it}$  becomes endogenous by dint of the omitted  $\alpha_i$ . Explicit treatment of  $\alpha_i$  is required for consistent estimation in the presence of random effects.

Any time invariant individual variables (TIVs),  $z_i$ , will lie in the column space of the unobservable  $\alpha_i$ . The familiar multicollinearity issue arises in the linear regression case and in nonlinear models. Coefficients  $\gamma$  cannot be identified without further restrictions (see Plümper and Troeger 2007, 2011; Greene 2011b; Breusch et al. 2011; and Hahn and Meinecke 2005). Consider a model with a single TIV,  $z_i$ . The log likelihood is

$$\log L = \sum_{i=1}^n \sum_{t=1}^{T_i} \log f(y_{it}, a_{it}).$$

The likelihood equations for  $\alpha_i$  and  $\gamma$  are

$$\begin{aligned} \frac{\partial \log L}{\partial \alpha_i} &= \sum_{t=1}^{T_i} \frac{\partial f(y_{it}, a_{it}) / \partial a_{it}}{f(y_{it}, a_{it})} \times 1 = \sum_{t=1}^{T_i} g_{a_{it}} = 0, \\ \frac{\partial \log L}{\partial \gamma} &= \sum_{i=1}^n \sum_{t=1}^{T_i} g_{a_{it}} z_i = \sum_{i=1}^n z_i \frac{\partial \log L}{\partial \alpha_i} = 0. \end{aligned}$$

This produces the singularity in the second derivatives matrix for the full set of parameters that is a counterpart to multicollinearity in the linear case. Gradient-based maximization methods will fail to converge because of the singularity of the weighting matrix, however formed. Bayesian methods (Lancaster 1999, 2000, 2001) will be able to identify the model parameters on the strength of informative priors. (For an example of Bayesian identification of individual effects on the strength of informative priors, see Koop, Osiewalski, and Steel 1997.) The GMM approach suggested byaisney and Lechner (2002) seems to provide a solution to the problem. The authors note, however,

Thus the coefficients of the time invariant regressors are identified provided there is at least one time varying regressor.... However, since this identification hinges on the local misspecification introduced by the Taylor series approximation, it seems preferable not to attempt an estimation of the coefficients of the time invariant variables, and to subsume the impact of the latter in the individual effect.

This would be an extreme example of *identification by the functional form of the model*. We assume that the model does not contain time invariant effects. It is worth noting that for purpose of analyzing modern longitudinal data sets, the inability to accommodate time invariant covariates is a vexing practical shortcoming of the fixed effects model. The hybrid formulations based on Mundlak's (1978) formulation or on *correlated random effects* in the next section present a useful approach that appears in many recent applications.

Strategies for estimation of models with fixed effects generally begin by seeking a way to avoid estimation of  $n$  effects parameters in the fully specified model (see, e.g., Fernandez-Val 2009). This turns on the existence of a sufficient statistic,  $S_i$  for the fixed effect such that the joint density,  $f(y_{it}, \dots, y_{iT} | S_i, X_i)$  does not involve  $\alpha_i$ . In the linear regression model,  $\Sigma_t y_{it}$  provides the statistic—the estimator based on the conditional distribution is the within groups linear least squares estimator. In all but a small few other cases (only two of any prominence in the contemporary literature), there is no sufficient statistic for  $\alpha_i$  in the log likelihood for the sample. In the Poisson regression, and in the binary logit model,  $\Sigma_t y_{it}$  provides the statistic. For the Poisson model, the marginal density is

$$f(y_{it}, a_{it}) = \frac{\exp(-\lambda_{it})\lambda_{it}^{y_{it}}}{y_{it}!}, \quad \lambda_{it} = \exp(\beta' \mathbf{x}_{it} + \alpha_i) = \exp(\alpha_i) \exp(\beta' \mathbf{x}_{it}).$$

The likelihood equation for  $\alpha_i$  is

$$\frac{\partial \log L}{\partial \alpha_i} = \left( \sum_{t=1}^{T_i} -\lambda_{it} \right) + \left( \sum_{t=1}^{T_i} y_{it} \right) = 0$$

which can be solved for

$$\alpha_i = \log \left( \frac{\sum_{t=1}^{T_i} y_{it}}{\sum_{t=1}^{T_i} \beta' \mathbf{x}_{it}} \right).$$

Note that there is no solution when  $y_{it}$  equals zero for all  $t$ . There need not be within group variation; the only requirement is that the sum be positive. Such observation groups must be dropped from the sample. The result for  $\alpha_i$  can be inserted into the log likelihood to form a concentrated log likelihood. The remaining analysis appears in HHG (1984). (HHG did not consider the case in which  $\sum_{i,t} y_{it} = 0$ , as in their data,  $y_{it}$  was always positive.)

Finally, for the binary logit model, the familiar result is

$$\begin{aligned} \text{Prob}(y_{i1}, y_{i2}, \dots, y_{iT_i}) &= f(y_{i1}, y_{i2}, \dots, y_{iT_i} | X_i, \Sigma_{t=1}^{T_i}) \\ &= \frac{\exp \left( \sum_{t=1}^{T_i} y_{it} (\beta' \mathbf{x}_{it}) \right)}{\sum_{\sum_t d_{it} = \sum_t y_{it}} \exp \left( \sum_{t=1}^{T_i} d_{it} (\beta' \mathbf{x}_{it}) \right)}, \end{aligned}$$

which is free of the fixed effects. The denominator in the probability is the sum over all  $\binom{T_i}{\sum_{t=1}^T y_{it}}$  configurations of the sequence of outcomes that sum to the same  $\sum_t y_{it}$ . This computation can, itself, be daunting—for example, if  $T_i = 20$  and  $\sum_t y_{it} = 10$ , there are  $20!/(10!)^2 = 184,756$  terms that all involve  $\beta$ . A recursive algorithm provided by Krailo and Pike (1984) greatly simplifies the calculations. (In an experiment with 500 individuals and  $T = 20$ , estimation of the model required about 0.25 seconds on an ordinary desktop computer.) Chamberlain (1980) details a counterpart of this method for a multinomial logit model. Borsch-Supan (1990) is an application.

In the probit model, which has attracted considerable interest, the practical implementation of the FEM requires estimation of the model with  $n$  dummy variables actually in the index function—there is no way to concentrate them out and no sufficient statistic. The complication of nonlinear models with possibly tens of thousands of coefficients to be estimated all at once has long been viewed as a substantive barrier to implementation of the model (see, e.g., Maddala 1983). The algorithm given in Greene (2004b, 2012) presents a solution to this practical problem. Fernandez-Val (2009) reports that he used this method to fit an FE probit model with 500,000 dummy variables. Thus, the physical complication is not a substantive obstacle in any problem of realistic dimensions. (In practical terms, the complication of fitting a model with 500,000+  $K$  coefficients would be a covariance matrix that would occupy nearly a terabyte of memory. Greene's algorithm exploits the fact that nearly the entire matrix is zeros to reduce the matrix storage requirements to linear in  $n$  rather than quadratic.)

The impediment to application of the fixed effects probit model is the *incidental parameters problem*. As has been widely documented in a long sequence of Monte Carlo studies and theoretical analyses, there is a persistent bias of  $O(1/T)$  in the maximum likelihood estimation of the parameters in many fixed effects model estimated by maximum likelihood. The problem was first reported in Neyman and Scott (1948), where it is shown that  $s^2$ , the MLE of the disturbance variance,  $\sigma^2$  in a fixed effects linear regression model has  $\text{plim } s^2 = \sigma^2(T - 1)/T$ . The obvious remedy, correcting for degrees of freedom, does not eliminate the vexing shortcoming of a perfectly well specified maximum likelihood estimator in other internally consistent model specifications. The problem persists in nonlinear settings where there is no counterpart “degrees of freedom correction” (for a detailed history, see Lancaster 2000). The extension of this result to other, nonlinear models has entered the orthodoxy of the field, though a precise result has actually been formally derived for only one case, the binomial logit model when  $T = 2$ , where it is shown that  $\text{plim } \hat{\beta}_{ML} = 2\beta$  (see, e.g., Abrevaya 1997 and Hsiao 2003). Although the regularity seems to be equally firm for the probit model and can be demonstrated with singular ease with a random number generator with any modern software, it has not been proved formally. Nor has a counterpart been found for any other  $T$ , for the unbalanced panel case, or for any other model. Other specific cases such as the ordered probit and logit models have

been persuasively demonstrated by Monte Carlo methods (see, e.g., Katz 2001 and Greene 2004b). The persistent finding is that the MLE for discrete choice models is biased away from zero. The result that does seem to persist is that when the incidental parameters problem arises, it does so with a proportional impact on some or all of the model parameters. The bias does not appear to depend substantively on the nature of the data support—it appears in the same form regardless of the process assumed to underlie the independent variables in the model. Rather, it is due to the presence of  $n$  additional estimation equations. We do note, once again, the generality of the bias, away from zero, appears to be peculiar to discrete outcome models.

Solutions to the incidental parameters problem in discrete choice cases—that is, consistent estimators of  $\beta$ —are of two forms. As discussed in Lancaster (2000), for a few specific cases, there exist sufficient statistics that will allow formation of a conditional density that is free of the fixed effects. The binary logit and Poisson regression cases are noted earlier.

Several recent applications have suggested a “*bias reduction*” approach. The central result as shown, for example, in Hahn and Newey (1994) and Hahn and Kuersteiner (2011) largely (again) for binary choice models is

$$\text{plim } \hat{\beta}_{ML} = \beta + B/T + O(1/T^2)$$

(see, as well, Arellano and Hahn 2007). That is, the unconditional MLE converges to a constant that is biased of  $O(1/T)$ . Three approaches have been suggested for eliminating  $B/T$ , a *penalized criterion* (modified log likelihood), *modified estimation (likelihood) equations*, and direct *bias correction* by estimating the bias, itself. In the first case, the direct log likelihood is augmented by a term in  $\beta$  whose maximizer is a good estimator of  $-B/T$  (see Carro and Trafirri 2011). In the second case, an estimator of  $-B/T$  is added to the MLE (see, e.g., Fernandez-Val 2009). The received theory has made some explicit use of the apparent proportionality result, that the bias in fixed effect discrete choice models, which are the only cases ever examined in detail, appears to be multiplicative, by a scalar of the form  $1 + b/T + O(1/T^2)$ . The effect seems to attach itself to scale estimation, not location estimators. The regression case noted earlier is obvious by construction. The binary choice case, though less so, does seem to be consistent with this. Write the model as  $y = 1[\beta'x + \alpha_i + \sigma w_{it} > 0]$ . The estimated parameters are  $\beta/\sigma$ , not  $\beta$ , where  $\sigma$  is typically normalized to 1 for identification. But, the multiplicative bias of the MLE does seem to affect the implicit “estimate” of the scale factor. The same result appears to be present in the MLE of the FE tobit model (see Greene 2004b). Fernandez-Val (2009) discusses this result at some length.

There is a loose end in the received results. The bias corrected estimators begin from the unconditional, brute force estimator that also estimates the fixed effects. However, this estimator, regardless of the distribution assumed (that will typically be the probit model), is incomplete. The estimator of  $\alpha_i$  is not identified when there is no within

group variation in  $y_i$ . For the probit model, the likelihood equation for  $\alpha_i$  is

$$\frac{\partial \log L}{\partial \alpha_i} = \sum_{t=1}^{T_i} \frac{(2y_{it} - 1)\phi[(2y_{it} - 1)(\beta' \mathbf{x}_{it} + \alpha_i)]}{\Phi[(2y_{it} - 1)(\beta' \mathbf{x}_{it} + \alpha_i)]} = 0.$$

If  $y_{it}$  equals one (zero) for all  $t$ , then the derivative is necessarily positive (negative) and cannot be equated to zero for any finite  $\alpha_i$ . In the “Chamberlain” estimator, groups for which  $y_{it}$  is always one or zero fall out of the estimation—they contribute  $\log(1.0) = 0.0$  to the log likelihood. Such groups must also be dropped for the unconditional estimator.

The starting point for consistent estimation of FE discrete choice models is the binary logit model. For the two period case, there are two obvious consistent estimators of  $\beta$ , the familiar textbook conditional estimator and one-half times the unconditional MLE. For more general (different  $T$ ) cases, the well-known estimator developed by Rasch (1960) and Chamberlain (1980), builds on the conditional joint distribution,  $\text{Prob}(y_{i1}, y_{i2}, \dots, y_{iT_i} | \Sigma_t y_{it}, \mathbf{X}_i)$ , which is free of the fixed effects. Two important shortcomings of the conditional approach are: (1) it does not provide estimators of any of the  $\alpha_i$  so it is not possible to compute probabilities or partial effects (see Wooldridge 2010, p. 622) and (2) it does not extend to other distributions or models. It does seem that there could be a remedy for (1). With a consistent estimator of  $\beta$  in hand, one could estimate individual terms of  $\alpha_i$  by solving the likelihood equation noted earlier for the probit model (at least for groups that have within group variation). The counterpart for the logit model is  $\Sigma_t [y_{it} - \Lambda(\beta' \mathbf{x}_{it} + \alpha_i)] = 0$ . A solution exists for  $\alpha_i$  for groups with variation over  $t$ . Each individual estimator is inconsistent as it is based on fixed  $T$  observations. Its asymptotic variance is  $O(1/T)$ . It remains to be established whether the estimators are systematically biased (upward or downward) when they are based on a consistent estimator of  $\beta$ . If not, it might pay to investigate whether the average over the useable groups provides useful information about  $E[\alpha_i]$ , which is what is needed to solve problem (1). The bias reduction estimators, to the extent that they solve the problem of estimation of  $\beta$ , may also help to solve this subsidiary problem. This was largely the finding of Hahn and Newey (1994). The conditional MLE in the binary logit model would appear to be a solution. This finding would be broadly consistent with Wooldridge’s arguments for the random effects pooled, or “population averaged” estimator.

The ordered choice cases are essentially the same as the binary cases as regards the conventional (brute force) estimator and the incidental parameters problem. There is no sufficient statistic for estimation of  $\beta$  in either case. However, the  $2\beta$  result for  $T = 2$  appears to extend to the ordered choice models. The broad nature of the result for  $T > 2$  would seem to carry over as well (see Greene and Hensher 2010). The ordered logit model provides an additional opportunity to manipulate the sample information. The base outcome probability for a fixed effects ordered logit model is

$$\text{Prob}(y_{it} = j | \mathbf{x}_{it}) = \Lambda(\mu_j - \beta' \mathbf{x}_{it} - \alpha_i) - \Lambda(\mu_{j-1} - \beta' \mathbf{x}_{it} - \alpha_i).$$

The implication is

$$\text{Prob}(y_{it} \geq j | \mathbf{x}_{it}) = \Lambda(\beta' \mathbf{x}_{it} + \alpha_i - \mu_j) = \Lambda(\beta' \mathbf{x}_{it} + \delta_i(j)).$$

Define the new variable  $D_{it}(j) = 1[y_{it} \geq j]$ ,  $j = 1, \dots, J$ . This defines  $J - 1$  binary fixed effects logit models, each with its own set of fixed effects, though they are the same save for the displacement by  $\mu_j$ . The Rasch/Chamberlain estimator can be used for each one. This does produce  $J - 1$  numerically different estimators of  $\beta$  that one might reconcile using a minimum distance estimator. The covariance matrices needed for the efficient weighting matrix are given in Brant (1990). An alternative estimator is based on the sums of outer products of the score vectors from the  $J - 1$  log likelihoods. Das and van Soest (1999) provide an application. Lee (2002) also provides an application.

Large sample bias corrected applications of the ordered choice models have been developed in Bester and Hansen (2009) and in Carro and Trafirri (2011). The methods employed limit attention to a three outcome case (low/medium/high). It is unclear if they can be extended to more general cases.

### 6.3.3 Correlated Random Effects

Mundlak (1978) suggested an approach between the questionable orthogonality assumptions of the random effects model and the frustrating limitations of the fixed effects specification,

$$\begin{aligned} y_{it} &= \beta' \mathbf{x}_{it} + \alpha_i + \varepsilon_{it} \\ \alpha_i &= \alpha + \gamma' \bar{\mathbf{x}}_i + w_i. \end{aligned}$$

Chamberlain (1980) proposed a less restrictive formulation,

$$\alpha_i = \alpha + \Sigma_t \gamma'_t \mathbf{x}_{it} + w_i.$$

This formulation is a bit cumbersome if the panel is not balanced—particularly if, as Wooldridge (2010) considers, the unbalancedness is due to endogenous attrition. The model examined by Plümper and Troeger (2007) is similar to Mundlak's;

$$\alpha_i = \alpha + \gamma' \mathbf{z}_i + w_i.$$

This is a “hierarchical model,” or multi (two) level model (see Bryk and Raudenbush 2002). In all of these cases, the assumption that  $E[w_i \mathbf{x}_{it}] = 0$  point identifies the parameters and the partial effects. The direct extension of this approach to nonlinear models such as the binary choice, ordered choice, and count data models converts them to random effects specifications that can be analyzed by conventional techniques. Whether the auxiliary equation should be interpreted as the conditional mean function in a structure or as a projection that, it is hoped, provides a good approximation to the

underlying structure is a minor consideration that nonetheless appears in the discussion. For example, Hahn, Ham, and Moon (2011) assume Mundlak's formulation as part of the structure at the outset, while Chamberlain (1980) would view that as restriction on the more general model.

The correlated random effects specification has a number of virtues for nonlinear panel data models. The practical appeal of a random effects vs. a full fixed effects approach is considerable. There are a number of conclusive results that can be obtained for the linear model that cannot be established for nonlinear models, such as Hausman's (1978) specification test for fixed vs. random effects. In the correlated random effects case, although the conditions needed to motivate Hausman's test are not met—the fixed effects estimator is not robust; it is not even consistent under either hypothesis—a variable addition test Wu (1973) is easily carried. In the Mundlak form, the difference between this version of the fixed effects model and the random effects model is the nonzero  $\gamma$ , which can be tested with a Wald test. Hahn, Ham, and Moon (2011) explored this approach in the context of panels in which there is very little within group variation and suggested an alternative statistic for the test. (The analysis of the data used in the World Health Report (WHO 2000) by Gravelle et al. (2002) would be a notable example.)

### 6.3.4 Attrition and Unbalanced Panels

Unbalanced panels may be more complicated than just a mathematical inconvenience. If the unbalanced panel results from attrition from what would otherwise be a balanced panel, and if the attrition is connected to the outcome variable, then the sample configuration is endogenous, and may taint the estimation process. Contoyannis, Jones, and Rice (2004) examine self-assessed health (SAH) in eight waves of the British Household Panel Survey. Their results suggest that individuals left the panel during the observation window in ways connected to the sequence of values of SAH. A number of authors, beginning with Verbeek and Nijman (1992) and Verbeek (2000) have suggested methods of detecting and correcting for endogenous attrition in panel data. Wooldridge (2002) proposes an “inverse probability weighting” procedure to weight observations in relation to their length of stay in the panel as a method of undoing the attrition bias. The method is refined in Wooldridge (2013) as part of an extension to a natural sample selection treatment.

## 6.4 DYNAMIC MODELS

---

An important benefit of panel data is the ability to study dynamic aspects of behavior in the model. The dynamic linear panel data regression

$$y_{it} = \beta' \mathbf{x}_{it} + \delta y_{i,t-1} + \alpha_i + \varepsilon_{it}$$

has been intensively studied since the field originated with Balestra and Nerlove (1966). Analysis of dynamic effects in discrete choice modeling has focused largely on binary choice. An empirical exception is Contoyannis, Jones, and Rice's (2004) ordered choice model for SAH. (Wooldridge (2005) also presents some more general theoretical results, e.g., for ordered choices.) For the binary case, the random effects treatment is untenable. The base case would be

$$y_{it} = 1[\beta' \mathbf{x}_{it} + \delta y_{i,t-1} + \gamma' \mathbf{z}_i + u_i + \varepsilon_{it} > 0].$$

Since the common effect appears in every period,  $u_i$  cannot be viewed as a random effect as it was earlier. However, in the conditional (on  $u_i$ ) log likelihood, as long as  $\varepsilon_{it}$  is independent across periods,  $y_{i,t-1}$  will be exogenous and the earlier treatment of random effects remains effective. A second complication is the “initial conditions problem” (Heckman 1981). The path of  $y_{it}$  will be determined at least partly (if not predominantly) by the value it took when the observation window opened. (The idea of initial conditions, itself, is confounded by the nature of the observation. It will rarely be the case that a process is observed from its beginning. Consider, for example, a model of insurance take-up or health status. Individuals have generally already participated in the process in periods before the observation begins. In order to proceed, it may be necessary to make some assumptions about the process, perhaps that it has reached an equilibrium at time  $t_0$  when it is first observed (see, e.g., Heckman 1981; Wooldridge 2002).) Arellano and Honoré (2001) consider this in detail as well.

Analysis of binary choice with lagged dependent variables, such as Lee (this volume) suggest that the incidental parameters problem is exacerbated by the lagged effects (see, e.g., Heckman 1981; Hahn and Kuersteiner 2002; Fernandez-Val 2009). Even under more restrictive assumptions, identification (and consistent estimation) of model parameters is complicated owing to the several sources of persistence in  $y_{it}$ , the heterogeneity itself and the state persistence induced by the lagged value. Analysis appears in Honoré and Kyriazidou (2000a, b), Chamberlain (1992), Hahn (2001), and Hahn and Moon (2006).

Semiparametric approaches to dynamics in panel data discrete choice have provided fairly limited guidance. Arellano and Honoré (2001) examine two main cases, one in which the model contains only current and lagged dependent variables and a second, three period model that has one regressor for which the second and third periods are equal. Lee (this volume) examines the multinomial logit model in similar terms. The results are suggestive, though perhaps more of methodological than practical interest. A practical approach is suggested by Heckman (1981), Hsiao (2003), Wooldridge (2010), and Semikyna and Wooldridge (2010). In a model of the form

$$y_{it} = 1[\beta' \mathbf{x}_{it} + \delta y_{i,t-1} + u_i + \varepsilon_{it} > 0],$$

the starting point,  $y_{i0}$ , is likely to be crucially important to the subsequent sequence of outcomes, particularly if  $T$  is small. We condition explicitly on the history;

$$\text{Prob}(y_{it} = 1 | \mathbf{X}_i, u_i, y_{i,t-1}, \dots, y_{i1}, y_{i0}) = f[y_{it}, (\beta' \mathbf{x}_{it} + \delta y_{i,t-1} + u_i)].$$

One might at this point take the initial outcome as exogenous and build up a likelihood,

$$f(y_{i1}, \dots, y_{iT} | \mathbf{X}_i, y_{i0}, u_i) = \prod_{t=1}^T f[(2y_{it} - 1)(\beta' \mathbf{x}_{it} + \delta y_{i,t-1} + u_i)],$$

then use the earlier methods to integrate  $u_i$  out of the function and proceed as in the familiar random effects fashion— $y_{i0}$  appears in the first term. The complication is that it is implausible to assume the common effect out of the starting point and have it appear suddenly at  $t = 1$ , even if the process (e.g., a labor force participation study that begins at graduation) begins at time 1. An approach suggested by Heckman (1981) and refined by Wooldridge (2005, 2010) is to form the joint distribution of the observed outcomes given  $(\mathbf{X}_i, y_{i0})$  and a plausible approximation to the marginal distribution  $f(u_i | y_{i0}, \mathbf{X}_i)$ . For example, if we depart from a probit model and use the Mundlak device to specify

$$u_i | y_{i0}, \mathbf{X}_i \sim N[\eta + \theta' \bar{\mathbf{x}}_i + \lambda y_{i0}, \sigma_w^2]$$

then

$$y_{it} = 1[\beta' \mathbf{x}_{it} + \delta y_{i,t-1} + \eta + \theta' \bar{\mathbf{x}}_i + \lambda y_{i0} + w_i + \varepsilon_{it} > 0].$$

(Some treatments, such as Chamberlain (1982), extend all of the rows of  $\mathbf{X}_i$  individually rather than use the group means. This creates a problem for unbalanced panels and, for a large model with even moderately large  $T$  creates an uncomfortably long list of right hand side variables. Recent treatments have usually used the projection onto the means instead.) Wooldridge (2010, p. 628) considers computation of average partial effects in this context. An application of these results to a dynamic random effects Poisson regression model appears in Wooldridge (2005). Contoyannis, Jones, and Rice (2004) specified a random effects dynamic ordered probit model, as

$$\begin{aligned} h_{it}^* &= \beta' \mathbf{x}_{it} + \gamma' \mathbf{h}_{i,t-1} + \alpha_i + \varepsilon_{it} \\ h_{it} &= j \text{ if } \mu_{j-1} < h_{it}^* \leq \mu_j \\ \alpha_i &= \eta + \alpha'_1 \mathbf{h}_{i0} + \alpha'_2 \mathbf{x}_i + w_i. \end{aligned}$$

This is precisely the application suggested above (with the Mundlak device). One exception concerns the treatment of the lagged outcome. Here, since the outcome variable is the label of the interval in which  $h_{it}^*$  falls,  $\mathbf{h}_{i,t}$  is a vector of  $J$  dummy variables for the  $J + 1$  possible outcomes (dropping one of them).

## 6.5 SPATIAL PANELS AND DISCRETE CHOICE

---

The final class of models noted is spatial regression models. Spatial regression has been well developed for the linear regression model. The linear model with *spatial auto-regression* is

$$\mathbf{y}_t = \mathbf{X}_t\beta + \lambda \mathbf{W}\mathbf{y}_t + \varepsilon_t,$$

where the data indicated are a sample of  $n$  observations at time  $t$ . The panel data counterpart will consist of  $T$  such samples. The matrix  $\mathbf{W}$  is the *spatial weight matrix*, or *contiguity matrix*. Nonzero elements  $w_{ij}$  define the two observations as neighbors. The relative magnitude of  $w_{ij}$  indicates how close the neighbors are.  $\mathbf{W}$  is defined by the analyst. Rows of  $\mathbf{W}$  are standardized to sum to one. The crucial parameter is the spatial auto-regression coefficient,  $\lambda$ . The transformation to the spatial moving average form is

$$\mathbf{y}_t = (\mathbf{I} - \lambda \mathbf{W})^{-1} \mathbf{X}_t \beta + (\mathbf{I} - \lambda \mathbf{W})^{-1} \varepsilon_t.$$

This is a generalized regression with disturbance covariance matrix  $\Omega = \sigma^2(\mathbf{I} - \lambda \mathbf{W})^{-1}(\mathbf{I} - \lambda \mathbf{W})^{-1'}$ . Some discussion of the model formulation may be found, for example, in Arbia (2006). An application to residential home sale prices is Bell and Bockstaal (2006). Extension of this linear model to panel data is developed at length in Lee and Yu (2010). An application to UK mental health expenditures appears in Moscone, Knapp, and Tosetti (2007).

Extensions of the spatial regression model to discrete choice are relatively scarce. A list of applications includes binary choice models Smirnov (2010), Pinske and Slade (1998), Bhat and Sener (2009), Klier and McMillen (2008), and Beron and Vijverberg (2004); a sample selection model applied to Alaskan trawlers by Flores Lagunes and Schnier (2012); an ordered probit analysis of accident severity by Kockelman and Wang (2009); a spatial multinomial probit model in Chakir and Parent (2009), and an environmental economics application to zero inflated counts by Rathbun and Fei (2006).

It is immediately apparent that if the spatial regression framework is applied to the underlying random utility specification in a discrete choice model that the density of the observable random vector,  $\mathbf{y}_t$  becomes intractable. In essence, the sample becomes one enormous fully auto-correlated observation. There is no transformation of the model that produces a tractable log likelihood. Each of the applications above develops a particular method of dealing with the issue. Smirnov, for example, separates the auto-correlation into “public” and “private” parts and assumes that the public part is small enough to discard. There is no generally applicable methodology in this setting on the level of the general treatment of simple dynamics and latent heterogeneity that

has connected the applications up to this point. We note, as well, that there are no received applications of spatial panel data to discrete choice models.

## REFERENCES

---

- Abrevaya, J., 1997. "The Equivalence of Two Estimators of the Fixed Effects Logit Model," *Economics Letters*, 55, 1, pp. 41–43.
- Allenby, G., J. Garrett, and P. Rossi, 2010. "A Model for Trade-Up and Change in Considered Brands," *Marketing Science*, 29, 1, pp. 40–56.
- Altonji, J., and R. Matzkin, 2005. "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors," *Econometrica*, 73, 3, pp. 1053–1102.
- Arbia, G., 2006. *Spatial Econometrics*, Springer, Berlin.
- Arellano, M., and J. Hahn, 2007. "Understanding Bias in Nonlinear Panel Models: Some Recent Developments," in R. Blundell, W. Newey, and T. Persson, eds., *Advances in Economics and Econometrics*, Ninth World Congress, Volume III, Cambridge University Press, Cambridge, pp. 381–409.
- Arellano, M., and B. Honoré, 2001. "Panel Data Models: Some Recent Developments," in J. Heckman and E. Leamer, eds., *Handbook of Econometrics*, Volume 5, Chapter 53, North-Holland, Amsterdam, 2001, pp. 3229–3296.
- Bago d'Uva, T., 2006. "Latent Class Models for Utilization of Health Care," *Health Economics* 15, 4, pp. 329–343.
- 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, pp. 585–612.
- Bell, K., and N. Bockstael, 2006. "Applying the Generalized Method of Moments Approach to Spatial Problems Involving Micro-Level Data," *Review of Economics and Statistics*, 82, 1, pp. 72–82.
- Bera, A., and C. Jarque, 1982. "Model Specification Tests: A Simultaneous Approach," *Journal of Econometrics*, 20, pp. 59–82.
- Bera, A., C. Jarque, and L. Lee, 1984. "Testing the Normality Assumption in Limited Dependent Variable Models," *International Economic Review*, 25, pp. 563–578.
- Beron, K., and W. Vijverberg, 2004. "Probit in a Spatial Context: A Monte Carlo Analysis," in L. Anselin, R. Florax, and S. Rey, eds., *Advances in Spatial Econometrics: Methodology, Tools and Applications*, Springer, New York, pp. 169–195.
- Bertschuk, I., and M. Lechner, 1998. "Convenient Estimators for the Panel Probit Model," *Journal of Econometrics*, 87, 2, pp. 329–372.
- Bester, C., and C. Hansen, 2009. "A Penalty Function Approach to Bias Reduction in Nonlinear Panel Models with Fixed Effects," *Journal of Business and Economic Statistics*, 27, 2, pp. 131–148.
- Bhat, C., 1999. "Quasi-Random Maximum Simulated Likelihood Estimation of the Mixed Multinomial Logit Model," Manuscript, Department of Civil Engineering, University of Texas, Austin.
- Bhat, C., and V. Pulugurta, 1998. "A Comparison of Two Alternative Behavioral Mechanisms for Car Ownership Decisions," *Transportation Research Part B*, 32, 1, pp. 61–75.

- Bhat, C., and I. Sener, 2009 "A Copula Based Closed Form Binary Logit Choice Model for Accommodating Spatial Correlation Across Observational Units," *Journal of Geographical Systems*, 11, pp. 243–272.
- Brant, R., 1990. "Assessing Proportionality in the Proportional Odds Model for Ordered Logistic Regression," *Biometrics*, 46, pp. 1171–1178.
- Breusch, T., and A. Pagan, 1980. "The LM Test and Its Applications to Model Specification in Econometrics," *Review of Economic Studies*, 47, pp. 239–254.
- Breusch, T., M. Ward, H. Nguyen, and T. Kompas, 2011. "On the Fixed-Effects Vector Decomposition," *Political Analysis*, 19, 2, pp. 123–134.
- Bryk, A., and S. Raudenbush, 2002. *Hierarchical Linear Models, Advanced Quantitative Techniques*, Sage, New York.
- Butler, J., and R. Moffitt, 1982. "A Computationally Efficient Quadrature Procedure for the One Factor Multinomial Probit Model," *Econometrica*, 50, pp. 761–764.
- Cameron, C., and P. Trivedi, 2005. *Microeometrics: Methods and Applications*, Cambridge University Press, Cambridge.
- Carro J., and A. Traferri, 2011. "State Dependence and Heterogeneity in Health Using a Bias Corrected Fixed Effects Estimator," *Journal of Applied Econometrics*, 26, pp. 1–27.
- Chakir, R., and O. Parent, 2009. "Determinants of Land Use Changes: A Spatial Multinomial Probit Approach," *Papers in Regional Science*, 88, 2, pp. 328–346.
- Chamberlain, G., 1980. "Analysis with Qualitative Data," *Review of Economic Studies*, 47, pp. 225–238.
- Chamberlain, G., 1982. "Multivariate Regression Models for Panel Data," *Journal of Econometrics*, 18, pp. 5–46.
- Chamberlain, G., 1984. "Panel Data," in Z. Griliches and M. Intriligator, eds., *Handbook of Econometrics*, Vol. 2, North Holland, Amsterdam, pp. 4–46.
- Chamberlain, G., 1992. "Binary Response Models for Panel Data: Identification and Information," Manuscript, Department of Economics, Harvard University.
- Chen, S., and S. Khan, 2003. "Rates of Convergence for Estimating Regression Coefficients in Heteroscedastic Discrete Response Models," *Journal of Econometrics*, 117, pp. 245–278.
- Chernozhukov, V., J. Hahn, I. Fernandez-Val, and W. Newey, 2013. "Average and Quantile Effects in Nonseparable Panel Models," *Econometrica*, 81, 2, pp. 535–580.
- Chesher, A., 1984. "Testing for Neglected Heterogeneity," *Econometrica*, 52, 4, pp. 865–872.
- Chesher, A., 2010 "Instrumental Variables Models for Discrete Outcomes," *Econometrica*, 78, pp. 575–601.
- Chesher, A., 2013. "Semiparametric Structural Models of Binary Response: Shape Restrictions and Partial Identification," *Econometric Theory*, 29, 2, pp. 231–266.
- Chesher, A., and M. Irish, 1987. "Residual Analysis in the Grouped Data and Censored Normal Linear Model," *Journal of Econometrics*, 34, pp. 33–62.
- Chesher, A., and L. Lee, 1986. "Specification Testing When Score Test Statistics are Identically Zero," *Journal of Econometrics*, 31, 2, pp. 121–149.
- Chesher, A., and A. Rosen, 2012a, "An Instrumental Variable Random Coefficients Model for Binary Outcomes," CeMMAP Working Paper CWP 34/12.
- Chesher, A., and A. Rosen, 2012b. "Simultaneous Equations for Discrete Outcomes: Coherence, Completeness and Identification," CEMMAP Working Paper CWP 21/12.

- Chesher, A., and K. Smolinsky, 2012. "IV Models of Ordered Choice," *Journal of Econometrics*, 166, pp. 33–48.
- Contoyannis, C., A. Jones, and N. Rice, 2004. "The Dynamics of Health in the British Household Panel Survey," *Journal of Applied Econometrics*, 19, 4, pp. 473–503.
- Cox, D., and D. Hinkley, 1974. *Theoretical Statistics*, Chapman and Hall, London.
- Das, M., and A. van Soest, 1999. "A Panel Data Model for Subjective Information on Household Income Growth," *Journal of Economic Behavior and Organization*, 40, pp. 409–426.
- Diggle, P., P. Liang, and S. Zeger, 1994. *Analysis of Longitudinal Data*, Oxford University Press, Oxford.
- Durlauf, S., and W. Brock, 2001a. "Discrete Choice with Social Interactions," *Review of Economic Studies*, 68, 2, pp. 235–260.
- Durlauf, S., and W. Brock, 2001b. "A Multinomial Choice Model with Neighborhood Effects," *American Economic Review*, 92, pp. 298–303.
- Durlauf, S., and W. Brock, 2002. "Identification of Binary Choice Models with Social Interactions," *Journal of Econometrics*, 140, 1, pp. 52–75.
- Durlauf, S., L. Blume, W. Brock, and Y. Ioannides, 2010. "Identification of Social Interactions," in J. Benhabib, A. Bisin, and M. Jackson, eds., *Handbook of Social Economics*, North Holland, Amsterdam.
- Elliott, G., and R. Leili, 2013. "Predicting Binary Outcomes," *Journal of Econometrics*, 174, 1, pp. 15–26.
- Fernandez-Val, I., 2009. "Fixed Effects Estimation of Structural Parameters and Marginal Effects in Panel Probit Models," *Journal of Econometrics*, 150, 1, pp. 71–85.
- Flores-Lagunes, A., and Schnier, K., 2012. "Sample Selection and Spatial Dependence," *Journal of Applied Econometrics*, 27, 2, pp. 173–204.
- Gravelle, H., R. Jacobs, A. Jones, and A. Street, 2002. "Comparing the Efficiency of National Health Systems: A Sensitivity Approach," Manuscript, University of York, Health Economics Unit.
- Greene, W., 1995. "Sample Selection in the Poisson Regression Model," Working Paper No. EC-95-6, Department of Economics, Stern School of Business, New York University.
- Greene, W., 2004a. "Convenient Estimators for the Panel Probit Model," *Empirical Economics*, 29, 1, pp. 21–47.
- Greene, W., 2004b, "The Behavior of the Fixed Effects Estimator in Nonlinear Models," *The Econometrics Journal*, 7, 1, pp. 98–119.
- Greene, W., 2011a. "Spatial Discrete Choice Models," Manuscript, Department of Economics, Stern School of Business, New York University, <http://people.stern.nyu.edu/wgreene/SpatialDiscreteChoiceModels.pdf>.
- Greene, W., 2011b. "Fixed Effects Vector Decomposition: A Magical Solution to the Problem of Time Invariant Variables in Fixed Effects Models?" *Political Analysis*, 19, 2, pp. 135–146.
- Greene, W., 2012. *Econometric Analysis*, 7th edn., Prentice Hall, Upper Saddle River, NJ.
- Greene, W., and D. Hensher, 2010. *Modeling Ordered Choices*, Cambridge University Press, Cambridge.
- Greene, W., and C. McKenzie, 2012. "LM Tests for Random Effects," Working Paper EC-12-14, Department of Economics, Stern School of Business, New York University.
- Hahn, J., 2001. "The Information Bound of a Dynamic Panel Logit Model with Fixed Effects," *Econometric Theory*, 17, pp. 913–932.
- Hahn, J., 2010, "Bounds on ATE with Discrete Outcomes," *Economics Letters*, 109, pp. 24–27.

- Hahn, J., and G. Kuersteiner, 2002. "Asymptotically Unbiased Inference for a Dynamic Panel Model with Fixed Effects When Both  $n$  and  $T$  are Large," *Econometrica*, 70, pp. 1639–1657.
- Hahn, J., and G. Kuersteiner, 2011. "Bias Reduction for Dynamic Nonlinear Panel Models with Fixed Effects," *Econometric Theory* 27, pp. 1152–1191.
- Hahn, J., and J. Meinecke, 2005. "Time Invariant Regressor in Nonlinear Panel Model with Fixed Effects," *Econometric Theory*, 21, pp. 455–469.
- Hahn, J., and H. Moon, 2006. "Reducing Bias of MLE in a Dynamic Panel Model," *Econometric Theory*, 22, pp. 499–512.
- Hahn, J., and W. Newey, 1994. "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models," *Econometrica* 72, pp. 1295–1319.
- Hahn, J., J. Ham, and H. Moon, 2011. "Test of Random vs. Fixed Effects with Small within Variation," *Economics Letters*, 112, pp. 293–297.
- Harris, M., and Y. Zhao, 2007. "Modeling Tobacco Consumption with a Zero Inflated Ordered Probit Model," *Journal of Econometrics*, 141, pp. 1073–1099.
- Harris, M., B. Hollingsworth, and W. Greene, 2012. "Inflated Measures of Self Assessed Health, Manuscript, School of Business, Curtin University.
- Hausman, J., 1978. "Specification Tests in Econometrics," *Econometrica*, 46, pp. 1251–1271.
- Hausman, J., B. Hall, and Z. Griliches, 1984. "Economic Models for Count Data with an Application to the Patents—R&D Relationship," *Econometrica*, 52, pp. 909–938.
- Heckman, J., 1979. "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 1979, pp. 153–161.
- Heckman, J., 1981. "Statistical Models for Discrete Panel Data," in C. Manski and D. McFadden, eds., *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, Cambridge, MA.
- Heckman, J., and B. Singer, 1984. "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*, 52, pp. 271–320.
- Hensher, D., and W. Greene, 2003. "The Mixed Logit Model: The State of Practice," *Transportation Research*, B, 30, pp. 133–176.
- Hensher, D., J. Rose, and W. Greene, 2006. *Applied Choice Analysis*, Cambridge University Press, Cambridge.
- Honoré, B., and E. Kyriazidou, 2000a. "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica* 68, 4, pp. 839–874.
- Honoré, B., and E. Kyriazidou, 2000b. "Estimation of Tobit-type Models with Individual Specific Effects," *Econometric Reviews* 19, pp. 341–366.
- Honoré, B., 2002, "Nonlinear Models with Panel Data," *Portuguese Economic Journal*, 1, 2, pp. 163–179.
- Horowitz, J., 1992. "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica*, 60, pp. 505–531.
- Horowitz, J., 1994. "Bootstrap-Based Critical Values for the Information Matrix Test," *Journal of Econometrics*, 61, pp. 395–411.
- Hsiao, C., 2003. *Analysis of Panel Data*, 2nd edn., Cambridge University Press, New York.
- Katz, E., 2001. "Bias in Conditional and Unconditional Fixed Effects Logit Estimation," *Political Analysis*, 9, 4, pp. 379–384.
- Klein, R., and R. Spady, 1993. "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica*, 61, pp. 387–421.

- Klier, T., and D. McMillen, 2008. "Clustering of Auto Supplier Plants in the United States: Generalized Method of Moments Spatial Logit for Large Samples," *Journal of Business and Economic Statistics*, 26, 4, pp. 460–471.
- Kockelman, K., and C. Wang, 2009. "Bayesian Inference for Ordered Response Data with a Dynamic Spatial Ordered Probit Model," Working Paper, Department of Civil and Environmental Engineering, Bucknell University.
- Koop, G., J. Osiewalski, and M. Steel, 1997. "Bayesian Efficiency Analysis through Individual Effects: Hospital Cost Frontiers," *Journal of Econometrics*, 76, pp. 77–106.
- Krailo, M., and M. Pike, 1984. "Conditional Multivariate Logistic Analysis of Stratified Case-Control Studies," *Applied Statistics*, 44, 1, pp. 95–103.
- Laisney, F., and M. Lechner, 2002. "Almost Consistent Estimation of Panel Probit Models with 'Small' Fixed Effects," ZEW Zentrum Discussion Paper No. 2002-64, <ftp://ftp.zew.de/pub/zew-docs/dp/dp0264.pdf>.
- Lancaster, T., 1999. "Panel Binary Choice with Fixed Effects," Discussion paper, Brown University.
- Lancaster, T., 2000. "The Incidental Parameter Problem since 1948," *Journal of Econometrics*, 95, pp. 391–413.
- Lancaster, T., 2001. "Orthogonal Parameters and Panel Data," Discussion paper, Brown University.
- Lee, L., and J. Yu, 2010. "Estimation of Spatial Panels," *Foundation and Trends in Econometrics*, 4, pp. 1–2.
- Lee, M., 2002. *Panel Data Econometrics*, Academic Press, New York.
- Lee, M., and A. Kimhi, 2005. "Simultaneous Equations in Ordered Discrete Responses with Regressor-Dependent Thresholds," *Econometrics Journal*, 8, 2, pp. 176–196.
- Maddala, G., 1983. *Limited Dependent and Qualitative Variables in Econometrics*, Cambridge University Press, Cambridge.
- Manski, C., 1975. "The Maximum Score Estimator of the Stochastic Utility Model of Choice," *Journal of Econometrics*, 3, pp. 205–228.
- Manski, C., 1985. "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics*, 27, pp. 313–333.
- Matzkin, R., 1991. "Semiparametric Estimation of Monotone and Concave Utility Functions for Polychotomous Choice Models," *Econometrica*, 59, 5, pp. 1315–1327.
- McFadden, D., 1974. "Conditional Logit Analysis of Qualitative Choice Behavior," in P. Zarembka, ed., *Frontiers in Econometrics*, Academic Press, New York.
- McFadden, D., and K. Train, 2000. "Mixed MNL Models for Discrete Choice," *Journal of Applied Econometrics*, 15, pp. 447–470.
- Moscone, F., M. Knapp, and E. Tosetti, 2007. "Mental Health Expenditures in England: A Spatial Panel Approach," *Journal of Health Economics*, 26, 4, pp. 842–864.
- Mullahy, J., 1987. "Specification and Testing of Some Modified Count Data Models," *Journal of Econometrics*, 33, pp. 341–365.
- Mundlak, Y., 1978. "On the Pooling of Time Series and Cross Sectional Data," *Econometrica*, 56, 1978, pp. 69–86.
- Neyman, J., and E. Scott, 1948. "Consistent Estimates Based on Partially Consistent Observations," *Econometrica*, 16, pp. 1–32.
- Pinske, J., and M. Slade, 1998. "Contracting in Space: An Application of Spatial Statistics to Discrete Choice Models," *Journal of Econometrics*, 85, pp. 125–154.

- Plümper, T., and V. Troeger, 2007. "Efficient Estimation of Time-Invariant and Rarely Changing Variables in Finite Sample Panel Analyses with Unit Fixed Effects," *Political Analysis*, 15, 2, pp. 124–139.
- Plümper, T., and V. Troeger, 2011. "Fixed-Effects Vector Decomposition: Properties, Reliability, and Instruments," *Political Analysis*, 19, 2, pp. 147–164.
- Pudney, S., and M. Shields, 2000. "Gender, Race, Pay and Promotion in the British Nursing Profession: Estimation of a Generalized Ordered Probit Model," *Journal of Applied Econometrics*, 15, 4, pp. 367–399.
- Rabe-Hesketh, S., A. Skrondal, and A. Pickles, 2005. "Maximum Likelihood Estimation of Limited and Discrete Dependent Variable Models with Nested Random Effects," *Journal of Econometrics*, 128, pp. 301–323.
- Rasch, G., 1960. "Probabilistic Models for Some Intelligence and Attainment Tests," *Denmark Paedogiska*, Copenhagen.
- Rathbun, S., and L. Fei, 2006. "A Spatial Zero-Inflated Poisson Regression Model for Oak Regeneration," *Environmental Ecology Statistics*, 13, pp. 409–426.
- Riphahn, R., A. Wambach, and A. Million, 2003. "Incentive Effects in the Demand for Health Care: A Bivariate Panel Count Data Estimation," *Journal of Applied Econometrics*, 18, 4, pp. 387–405.
- Semykina, A., and J. Wooldridge, 2013. "Estimation of Dynamic Panel Data Models with Sample Selection," *Journal of Applied Econometrics*, 28, 1, pp. 47–61.
- Smirnov, A., 2010. "Modeling Spatial Discrete Choice," *Regional Science and Urban Economics*, 40, 5, pp. 292–298.
- Train, K., 2003. *Discrete Choice Methods with Simulation*, Cambridge University Press, Cambridge.
- Train, K., 2009. *Discrete Choice Methods with Simulation*, 2nd edn., Cambridge University Press, Cambridge.
- Van Dijk, R., D. Fok, and R. Paap, 2007. "A Rank-Ordered Logit Model with Unobserved Heterogeneity in Ranking Capabilities," Econometric Institute, Erasmus University, Report 2007-07.
- Verbeek, M., 2000. *A Guide to Modern Econometrics*, Wiley, Chichester.
- Verbeek, M., and T. Nijman, 1992. "Testing for Selectivity Bias in Panel Data Models," *International Economic Review*, 33, 3, pp. 681–703.
- Wooldridge, J., 2002. "Inverse Probability Weighted M-Estimators for Sample Selection, Attrition, and Stratification," *Portuguese Economic Journal*, 1, pp. 117–139.
- Wooldridge, J., 2003. "Cluster-Sample Methods in Applied Econometrics," *American Economic Review*, 93, pp. 133–138.
- Wooldridge, J., 2005. "Simple Solutions to the Initial Conditions Problem in Dynamic Non-linear Panel Data Models with Unobserved Heterogeneity," *Journal of Applied Econometrics*, 20, pp. 39–54.
- Wooldridge, J., 2010. *Econometric Analysis of Cross Section and Panel Data*, 2nd edn., MIT Press, Cambridge, MA.
- Wooldridge, J., 2013. "Estimation of Dynamic Panel Data Models with Sample Selection," *Journal of Applied Econometrics*, 28, 1, pp. 47–61.
- World Health Organization, 2000. *The World Health Report, 2000, Health Systems: Improving Performance*. WHO, Geneva.
- Wu, D., 1973. "Alternative Tests of Independence between Stochastic Regressors and Disturbances," *Econometrica*, 41, pp. 733–750.

## CHAPTER 7

---

# PANEL CONDITIONAL AND MULTINOMIAL LOGIT ESTIMATORS

---

MYOUNG-JAE LEE

## 7.1 INTRODUCTION

---

MICRO panel data models for many individuals over only a few periods are often postulated to contain ‘incidental parameters’ which vary across individuals, along with ‘structural parameters’ which are common for all individuals. For instance, consider a panel linear model

$$y_{it} = x'_{it}\beta + \delta_i + u_{it}, \quad i = 1, \dots, N \text{ (large)} \quad \text{and} \quad t = 1, \dots, T \text{ (small)},$$

where  $y_{it}$  is a response variable of individual  $i$  at time  $t$ ,  $x_{it}$  is a  $k_x \times 1$  regressor vector,  $\beta$  is a parameter,  $\delta_i$  is a time-constant error (‘individual-specific effect’) and  $u_{it}$  is a time-varying error. Here  $\delta_i$  may be taken as a parameter to estimate, in which case  $\delta_i$  is an incidental parameter whereas  $\beta$  that is common to all individuals is a structural parameter (of interest).

Clearly,  $\delta_i$  cannot be consistently estimated in short panels, and thus it should be taken as a random variable just as  $y_{it}$ ,  $x_{it}$  and  $u_{it}$  are. In this case, the main concern has been that  $\delta_i$  might be the main source of  $x_{it}$  endogeneity; e.g.,  $\delta_i$  may represent genes or innate ability that influence  $y_{it}$ . In cross-section data, there are several approaches to deal with an endogenous  $x_{it}$  as reviewed in Lee (2012), but all of them require an instrument one way or another. In contrast, panel data do not necessarily need an instrument, as they provide a number of ways to get rid of  $\delta_i$ .

For the above linear model, as well known,  $\delta_i$  can be removed by first differencing:

$$\Delta y_{it} = \Delta x'_{it}\beta + \Delta u_{it} \quad \text{where} \quad \Delta y_{it} \equiv y_{it} - y_{i,t-1}.$$

If  $y_{it}$  is integer-valued with an exponential regression as in

$$E(y_{it}|\delta_i, x_{i1}, \dots, x_{iT}) = \exp(x'_{it}\beta + \delta_i + u_{it}),$$

then  $\delta_i$  can be removed with “dividing the model by  $y_{i,t-1}$ ” as can be seen in Kim (1988), Chamberlain (1992) and Wooldridge (1997). Other than these two approaches, there appeared yet another approach as follows to remove  $\delta_i$  by conditioning on ‘sufficient statistics’.

Given independent observations  $z_1, \dots, z_N$  from a density  $f(z;\theta)$ ,  $\Psi_N \equiv \Psi(z_1, \dots, z_N)$  is a ‘sufficient statistic’ for  $\theta$  if the distribution of  $(z_1, \dots, z_N)|\Psi_N$  does not depend on  $\theta$ . The sufficiency is known to be equivalent to the likelihood function to be written in the product form

$$\prod_i f(z_i; \theta) = g(\Psi_N, \theta) \times h(z_1, \dots, z_N) \quad \text{for some functions } g \text{ and } h.$$

Then, considering the maximum likelihood estimator (MLE) maximizing

$$\ln \prod_i f(z_i; \theta) = \ln g(\Psi_N, \theta) + \ln h(z_1, \dots, z_N)$$

with respect to  $\theta$ , since  $\ln h(z_1, \dots, z_N)$  drops out, we can see that  $\theta$  depends on  $(z_1, \dots, z_N)$  only through  $\Psi_N$ : all information in the data for  $\theta$  is contained in  $\Psi_N$ . Hence fixing  $\Psi_N$  in  $(z_1, \dots, z_N)|\Psi_N$  leaves no more information for  $\theta$  in the data  $(z_1, \dots, z_N)$ .

The idea of removing  $\delta_i$  by conditioning on a sufficient statistic in panel data has been applied to the exponential model under Poisson distributions and to binary response models under logistic distributions. This chapter reviews the latter: *panel conditional logit estimators (PCLE’s)* for binary responses, and their extensions. Since binary response models under logistic distributions are tightly specified, the room for improvement has been hard to come by—only once in several years. Nevertheless, the literature is filled with innovative ideas and valued highly in terms of the journals carrying the studies. Earlier reviews on PCLE can be found in Arellano and Honoré (2001), Lee (2002), Hsiao (2003), and Baltagi (2008).

The rest of this chapter is organized as follows. Section 7.2 examines static PCLE, and Section 7.3 reviews dynamic PCLE’s. Section 7.4 extends the static PCLE to ordered discrete responses by collapsing an ordered discrete response to multiple binary responses and then using minimum distance estimator (MDE). Section 7.5 examines static and dynamic PCLE’s for multinomial responses. Finally, Section 7.6 concludes.

Some words on notations. Let  $1[A] = 1$  if  $A$  holds and 0 otherwise, and

$$Y_i \equiv \begin{bmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{bmatrix} \quad \text{and} \quad X_i \equiv (x_{i1}, \dots, x_{iT}) = \begin{bmatrix} x_{i11} & \cdots & x_{iT1} \\ \vdots & & \vdots \\ x_{i1k_x} & \cdots & x_{iT k_x} \end{bmatrix} \implies X'_i \beta = \begin{bmatrix} x'_{i1} \beta \\ \vdots \\ x'_{iT} \beta \end{bmatrix}$$

where the entry below a matrix denotes its dimension. The independence between random vectors  $z_1$  and  $z_2$  given  $z_3$  will be denoted ' $z_1 \perp\!\!\!\perp z_2 | z_3$ '. As we will be assuming *iid* across  $i = 1, \dots, N$ , often the subscript  $i$  will be omitted.

In econometrics/statistics, often upper-case letters are used for random variables and the lower-case letters for their realized values, but this convention will not be followed as upper-case letters will be used for random matrices as in the last display. This may result in some abuse of notations (e.g.,  $y_{it}$  representing a random variable as well as a value it can take), but the meaning will be clear from the context. Unless otherwise noted, the time indexes in  $\sum_t$  and  $\prod_t$  run from 1 to  $T$ . ' $\rightsquigarrow$ ' denotes convergence in distribution.

## 7.2 STATIC PANEL CONDITIONAL LOGIT

This section examines static PCLE, i.e., PCLE with no lagged response as a regressor. Firstly, its likelihood function is derived for a general  $T$  to show the essential idea of PCLE: how to remove  $\delta_i$  by conditioning on its sufficient statistic  $\sum_t y_{it}$ . Secondly, two special cases  $T = 2$  and  $T = 3$  are examined for illustration. Thirdly, further remarks are provided.

### 7.2.1 Log-likelihood Function for Static PCLE

Consider a static panel binary-response logit model:

$$y_{it} = 1[x'_{it}\beta + \delta_i + u_{it} > 0], \quad i = 1, \dots, N, \quad t = 1, \dots, T;$$

$$(u_{i1}, \dots, u_{iT}) \text{ are iid logistic (i.e., } P(u_{it} \leq a) = \frac{\exp(a)}{1 + \exp(a)} \forall a)$$

and  $(u_{i1}, \dots, u_{iT}) \perp\!\!\!\perp (\delta_i, X_i)$ ;

$(Y_i, X_i)$  are observed and iid across  $i = 1, \dots, N$ .

The fact that  $y_{it}$  is determined only by  $(x_{it}, \delta_i, u_{it})$ , not by the other period regressors and errors, implies 'strict exogeneity', as the 'reduced form error'  $y_{it} - E(y_{it}|\delta_i, x_{it})$  is orthogonal, not just to  $x_{it}$ , but to all of  $(x_{i1}, \dots, x_{iT})$  due to  $E(y_{it}|\delta_i, x_{it}) = E(y_{it}|\delta_i, x_{i1}, \dots, x_{iT})$ .

The above assumptions imply that the joint probability for  $(Y_i = \Lambda) | (\delta_i, X_i)$  with  $\Lambda \equiv (\lambda_1, \dots, \lambda_T)' (\lambda_t = 0 \text{ or } 1 \forall t)$  is

$$P(Y_i = \Lambda | \delta_i, X_i) = \prod_t P(y_{it} = \lambda_t | \delta_i, X_i) = \prod_t P(y_{it} = \lambda_t | \delta_i, x_{it})$$

$$\begin{aligned}
 &= \prod_t \left\{ \frac{1}{1 + \exp(x'_{it}\beta + \delta_i)} \right\}^{1-\lambda_t} \left\{ \frac{\exp(x'_{it}\beta + \delta_i)}{1 + \exp(x'_{it}\beta + \delta_i)} \right\}^{\lambda_t} \\
 &= \prod_t \frac{\exp\{\lambda_t(x'_{it}\beta + \delta_i)\}}{1 + \exp(x'_{it}\beta + \delta_i)} = \frac{\exp(\delta_i \sum_t \lambda_t) \cdot \exp(\sum_t \lambda_t x'_{it}\beta)}{\prod_t \{1 + \exp(x'_{it}\beta + \delta_i)\}}. \quad (2.1)
 \end{aligned}$$

The special cases with  $\lambda_t = 1 \forall t$  (thus  $\sum_t \lambda_t = T$ ) and  $\lambda_t = 0 \forall t$  are, respectively,

$$\frac{\exp(\delta T) \cdot \exp(\sum_t x'_t \beta)}{\prod_t \{1 + \exp(x'_t \beta + \delta)\}} \quad \text{and} \quad \frac{1}{\prod_t \{1 + \exp(x'_t \beta + \delta)\}}.$$

If we write  $P(Y = \Lambda | \delta, X)$  just as  $P(\Lambda | \delta, X)$  and then replace  $\Lambda$  with  $Y$ , we get the joint likelihood (i.e., the probability of the random responses equal to the sample values):

$$P(Y | \delta, X) = \frac{\exp(\delta \sum_t y_t) \cdot \exp(\sum_t y_t x'_t \beta)}{\prod_t \{1 + \exp(x'_t \beta + \delta)\}}. \quad (2.2)$$

The probability of the sum of the random responses taking the sample value  $\sum_t y_t$  is

$$P\left(\sum_t y_t | \delta, X\right) = \sum_{\bar{\lambda}=\bar{y}} P(\lambda_1, \dots, \lambda_T | \delta, X) \quad (2.3)$$

where  $\sum_{\bar{\lambda}=\bar{y}} (\cdot)$  is the sum over all sequences  $\Lambda$  with

$$\bar{\lambda} \equiv \frac{1}{T} \sum_{t=1}^T \lambda_t = \frac{1}{T} \sum_t y_t \equiv \bar{y};$$

to avoid the cluttering notation ' $\sum_{\sum_t \lambda_t = \sum_t y_t}$ ', we use  $\sum_{\bar{\lambda}=\bar{y}}$ .

Substituting (2.1) into (2.3) and then using  $\sum_t \lambda_t = \sum_t y_t$ , (2.3) becomes

$$\begin{aligned}
 P\left(\sum_t y_t | \delta, X\right) &= \frac{\sum_{\bar{\lambda}=\bar{y}} \exp(\delta \sum_t \lambda_t) \exp(\sum_t \lambda_t x'_t \beta)}{\prod_t \{1 + \exp(x'_t \beta + \delta)\}} \\
 &= \frac{\exp(\delta \sum_t y_t) \sum_{\bar{\lambda}=\bar{y}} \exp(\sum_t \lambda_t x'_t \beta)}{\prod_t \{1 + \exp(x'_t \beta + \delta)\}}; \quad (2.4)
 \end{aligned}$$

e.g., with  $T = 2$  and  $\sum_t y_t = 1$ , as there are only two possibilities  $(0, 1)$  and  $(1, 0)$  for  $(\lambda_1, \lambda_2)$ ,

$$P\left(\sum_t y_t = 1 | \delta, X\right) = P(y_1 = 0, y_2 = 1 | \delta, X) + P(y_1 = 1, y_2 = 0 | \delta, X)$$

$$\begin{aligned}
&= \frac{\exp(x_2' \beta + \delta)}{\{1 + \exp(x_1' \beta + \delta)\}\{1 + \exp(x_2' \beta + \delta)\}} + \frac{\exp(x_1' \beta + \delta)}{\{1 + \exp(x_1' \beta + \delta)\}\{1 + \exp(x_2' \beta + \delta)\}} \\
&= \frac{\exp(\delta) \cdot \{\exp(x_1' \beta) + \exp(x_2' \beta)\}}{\{1 + \exp(x_1' \beta + \delta)\}\{1 + \exp(x_2' \beta + \delta)\}}.
\end{aligned}$$

Divide  $P(Y|\delta, X)$  in (2.2) by  $P(\sum_t y_t|\delta, X)$  in (2.4) to obtain

$$P\left(Y \mid \sum_t y_t, \delta, X\right) = P\left(Y \mid \sum_t y_t, X\right) = \frac{\exp(\sum_t y_t x_t' \beta)}{\sum_{\bar{\lambda}=\bar{y}} \exp(\sum_t \lambda_t x_t' \beta)};$$

the division removes two common terms,  $\exp(\delta \sum_t y_t)$  and  $\prod\{1 + \exp(x_t' \beta + \delta)\}$ . As the ratio is free of  $\delta$ ,  $\sum_t y_t$  is a sufficient statistic for  $\delta$  given  $X$ .

The sample conditional log-likelihood function to maximize for  $\beta$  is

$$\sum_{i=1}^N \ln \frac{\exp(\sum_t y_{it} x_{it}' \beta)}{\sum_{\bar{\lambda}=\bar{y}_i} \exp(\sum_t \lambda_t x_{it}' \beta)} \left\{ = \sum_{i=1}^N \ln \frac{\exp(Y_i' \cdot X_i' \beta)}{\sum_{\bar{\lambda}=\bar{y}_i} \exp(\Lambda' \cdot X_i' \beta)} \right\} \quad (2.5)$$

$$= \sum_{i=1}^N \ln \frac{\exp\{\sum_{t \geq 2} y_{it} (x_{it} - x_{i1})' \beta\}}{\sum_{\bar{\lambda}=\bar{y}_i} \exp\{\sum_{t \geq 2} \lambda_t (x_{it} - x_{i1})' \beta\}} \quad (2.6)$$

dividing the numerator and denominator by  $\exp(\sum_t y_{it} \cdot x_{i1}' \beta)$  and  $\exp(\sum_t \lambda_t \cdot x_{i1}' \beta)$ , respectively, that are the same—this is a normalization. Denoting the MLE as  $b_N$ ,  $\sqrt{N}(b_N - \beta)$  is asymptotically normal whose variance can be estimated in the “usual MLE way”: with  $-1$  times the inverse of the Hessian matrix, or with the inverse of the averaged outer-product of the score function.

Since

$$\sum_t y_t = 0 \iff (y_1 = 0, \dots, y_T = 0) \text{ and } \sum_t y_t = T \iff (y_1 = 1, \dots, y_T = 1),$$

when  $\sum_t y_{it} = 0$  or  $T$ ,  $\sum_{\bar{\lambda}=\bar{y}_i} \exp\{\sum_{t \geq 2} \lambda_t (x_{it} - x_{i1})' \beta\}$  has only one term with  $\Lambda = Y_i$ : the observations with  $\sum_t y_{it} = 0$  or  $T$  drop out, as their contribution to the log-likelihood function is  $\ln 1 = 0$ . Hence, maximizing (2.5) or (2.6) is equivalent to maximizing

$$\sum_{i=1}^N 1 \left[ \sum_t y_{it} \neq 0, T \right] \cdot \ln \frac{\exp\{\sum_{t \geq 2} y_{it} (x_{it} - x_{i1})' \beta\}}{\sum_{\bar{\lambda}=\bar{y}_i \neq 0,1} \exp\{\sum_{t \geq 2} \lambda_t (x_{it} - x_{i1})' \beta\}} \quad (2.7)$$

where the notation ‘ $\sum_{\bar{\lambda}=\bar{y}_i \neq 0,1}$ ’ means computing the sum  $\sum_{\bar{\lambda}=\bar{y}_i}$  only for the observations with  $\bar{y}_i \neq 0, 1 \iff \sum_t y_{it} \neq 0, T$ .

In (2.5) to (2.7), we saw different versions of sample maximand, with (2.5) looking the simplest. We can simply maximize (2.5), but the normalized version (2.6) shows better which parameters are identified; e.g., the intercept is not identified so long as it is time-constant because unity (the regressor for the intercept) is removed

in  $x_{it} - x_{i1}$ . Also going from (2.6) to (2.7) saves computation time, because the non-informative observations with no temporal change in  $y_{it}$ 's are removed. These points apply to other sample maximands as well to appear in the remainder of this chapter where there are a basic form, its normalized version and its informative-observations-only version. Regardless of which version is used, to implement the maximization, the sample maximand should be spelled out without the “implicit”  $\lambda_t$ 's, which can be best seen in  $T = 2$  and 3 special cases next.

### 7.2.2 Special Cases with Two and Three Periods

For  $T = 2$ , the sample maximand (2.7) is (with  $\sum_t y_{it} \neq 0, T \iff y_{i1} + y_{i2} = 1$ ),

$$\sum_i 1[y_{i1} + y_{i2} = 1] \ln \frac{\exp(y_{i2} \Delta x'_{i2} \beta)}{1 + \exp(\Delta x'_{i2} \beta)}$$

as  $\sum_{\bar{\lambda}=\bar{y}_i \neq 0,1} (\dots)$  contains only two terms for  $\Lambda = (1,0)$  and  $\Lambda = (0,1)$ :

$$\begin{aligned} \sum_{\bar{\lambda}=\bar{y}_i \neq 0,1} \exp \left\{ \sum_{t \geq 2} \lambda_t (x_{it} - x_{i1})' \beta \right\} &= \sum_{\Lambda=(1,0) \text{ or } (0,1)} \exp\{\lambda_2 (x_{i2} - x_{i1})' \beta\} \\ &= \exp\{0(x_{i2} - x_{i1})' \beta\} + \exp\{(x_{i2} - x_{i1})' \beta\} = 1 + \exp(\Delta x'_{i2} \beta). \end{aligned}$$

The PCLE with  $T = 2$  essentially becomes the cross-section logit with  $y_{i2}$  and  $\Delta x_{i2}$  being the binary response and regressors; hence, this PCLE is a “first-differencing” estimator.

For  $T = 3$ , (2.7) is

$$\sum_{i=1}^N 1 \left[ \sum_t y_{it} \neq 0, 3 \right] \cdot \ln \frac{\exp\{y_{i2}(x_{i2} - x_{i1})' \beta + y_{i3}(x_{i3} - x_{i1})' \beta\}}{\sum_{\bar{\lambda}=\bar{y}_i \neq 0,1} \exp\{\lambda_2(x_{i2} - x_{i1})' \beta + \lambda_3(x_{i3} - x_{i1})' \beta\}}.$$

For a further simplification, define  $S_{ia} \equiv 1[\sum_t y_{it} = a]$  for  $a = 1, 2$ . There are three possibilities for  $\Lambda$  given  $S_{i1} = 1$ :  $(1,0,0)$ ,  $(0,1,0)$  and  $(0,0,1)$ . Hence the probability of observing a particular  $Y$  given  $S_1 = 1$  is

$$\begin{aligned} &\frac{\exp(Y' \cdot X' \beta)}{\exp\{(1,0,0)X' \beta\} + \exp\{(0,1,0)X' \beta\} + \exp\{(0,0,1)X' \beta\}} \\ &= \frac{\exp(y_1 x'_1 \beta + y_2 x'_2 \beta + y_3 x'_3 \beta)}{\exp(x'_1 \beta) + \exp(x'_2 \beta) + \exp(x'_3 \beta)} = \frac{\exp\{y_2(x_2 - x_1)' \beta + y_3(x_3 - x_1)' \beta\}}{1 + \exp\{(x_2 - x_1)' \beta\} + \exp\{(x_3 - x_1)' \beta\}} \end{aligned} \tag{2.8}$$

dividing the numerator and the denominator by, respectively,  $\exp(\sum_t y_t \cdot x'_1 \beta)$  and  $\exp(x'_1 \beta)$  that are the same given  $S_1 = 1[\sum_t y_t = 1] = 1$ .

Doing analogously, the probability of observing a particular  $Y$  given  $S_2 = 1$  is

$$\begin{aligned} & \frac{\exp(Y' \cdot X'\beta)}{\exp\{(1, 1, 0)X'\beta\} + \exp\{(1, 0, 1)X'\beta\} + \exp\{(0, 1, 1)X'\beta\}} \\ &= \frac{\exp(y_2(x_2 - x_1)' \beta + y_3(x_3 - x_1)' \beta)}{\exp\{(x_2 - x_1)' \beta\} + \exp\{(x_3 - x_1)' \beta\} + \exp\{(x_2 - x_1 + x_3 - x_1)' \beta\}} \quad (2.9) \end{aligned}$$

normalizing with  $\exp(\sum_t y_t x'_1 \beta) = \exp(2x'_1 \beta)$ . Therefore, the sample maximand for  $T = 3$  is

$$\sum_i \{S_{i1} \ln(2.8) + S_{i2} \ln(2.9)\}. \quad (2.10)$$

For a general  $T$ , the static PCLE maximand (2.5) may look too difficult to compute due to the part  $\sum_{\bar{\lambda}=\bar{y}_i} (\cdot)$ . But there exists a fast recursive algorithm for  $\sum_{\bar{\lambda}=\bar{y}_i} (\cdot)$ . The algorithm was first noted by Howard (1972) for Cox (1972) ‘partial MLE’, and its applicability to PCLE is due to Kralo and Pike (1984). The popular econometric software STATA also uses this algorithm. Alternatively, when there are  $T > 3$  waves, (2.10) can be used for each three waves, and then MDE can be applied to combine the multiple sets of estimates; see the section on panel ordered logit to implement the idea. Instead of each three waves, we may use each two waves and then MDE.

## 7.2.3 Further Remarks

### 7.2.3.1 PCLE Literature

The idea of conditioning on a sum such as  $\sum_t y_t$  seems to have appeared first in Rasch (1961). The asymptotic distribution of conditional MLE was derived by Anderson (1970), who showed the static PCLE with  $T = 2$  and a single parameter as an example in Anderson (1970, p. 299). Chamberlain (1980, pp. 230–231) presented the static PCLE for a general  $T$  with multiple parameters, and noted its extensions to panel multinomial responses and cross-section models with a group structure.

So long as  $\delta_i$  is removed,  $\delta_i$  can be allowed to be related to  $X_i$  in an arbitrary fashion. Alternatively,  $\delta_i$  may be estimated as a fixed parameter when  $T$  is large, which appears to have prompted the name ‘fixed effect’ for  $\delta_i$ —a misnomer though. Usually, ‘fixed effect’ refers to  $\delta_i$  unrestricted in its relationship to  $X_i$ . Although not reviewed so far, the following theoretical studies on PCLE are notable; applied studies are omitted, as there are too many.

For  $T = 2$ , Chamberlain (2010) proved that, when  $(u_{i1}, u_{i2})$  are iid independently of  $(\delta_i, x_{i1}, x_{i2})$ , even if the support of  $x_{it}$  is unbounded, the semiparametric information bound for  $\beta$  is zero unless  $u_{it}$  is logistic; i.e.,  $\sqrt{N}$ -consistent estimator other than PCLE does not exist without further assumptions. Magnac (2004) sought to relax the ‘iid logistic’ assumption by having dependent non-logistic  $(u_{i1}, u_{i2})$ , while still allowing for the fixed effect  $\delta_i$  and maintaining the sufficiency of  $\sum_t y_{it}$ ; no application, however, seems to have been done so far. Thomas (2006) considered double fixed effects,

say  $\delta_i$  and  $\vartheta_i$ , to have  $\delta_i + t\vartheta_i$  in the model, which is then removed by conditioning on sufficient statistics  $\sum_{t=1}^T y_{it}$  and  $\sum_{t=1}^T ty_{it}$ . Halliday (2007) proposed a test for the sign of the state dependency in a dynamic model without specifying the regression functional form. Holderlein et al. (2011) considered a nonparametric regression function of  $x_{it}$  in the logistic model, and then sought to generalize the logit ‘link’ function à la Magnac (2004).

### 7.2.3.2 PCLE for Grouped Cross-Section Data

Often cross-section data have a group structure where group  $i$  has  $T_i$  members. For example, a group can be (i) a town, (ii) a matched group of individuals sharing similar characteristics (gender, age, etc.), or (iii) triples born to the same mother ( $T_i = 3$  in this case). In each group  $i$ , there are  $\sum_{t=1}^{T_i} y_{it}$ -many ones (and  $T_i - \sum_{t=1}^{T_i} y_{it}$  zeros), and the group members share the same unobserved trait  $\delta_i$  that may be related to  $x_{it}$ . *PCLE is applicable to grouped cross-section data*, because the time order plays no particular role in  $\sum_{\bar{\lambda}=\bar{y}_i} \exp(\sum_t \lambda_t x'_{it} \beta)$ .

In group-structure cross-section data, the normalization by  $x'_{i1} \beta$  can be done by any group member. In the town example, any town- $i$  member can be the first member. In a ‘matched case-control study’, often there are one ‘case’ ( $y = 1$ ) and multiple “controls” ( $y = 0$ ), and the single case subject is then the natural first member. For the triple example, the first born would be the natural first member.

### 7.2.3.3 Odds Ratio and Marginal Effect

As noted in Chamberlain (1984), the *logged odds ratio free of  $\delta$*  can be found from the PCLE because

$$\begin{aligned} & \ln \frac{P(y_t = 1|\delta, x_t = a_1)/P(y_t = 0|\delta, x_t = a_1)}{P(y_t = 1|\delta, x_t = a_0)/P(y_t = 0|\delta, x_t = a_0)} \\ &= \ln \left[ \left\{ \frac{\exp(a'_1 \beta + \delta)}{1 + \exp(a'_1 \beta + \delta)} / \frac{1}{1 + \exp(a'_1 \beta + \delta)} \right\} \right. \\ &\quad \left. / \left\{ \frac{\exp(a'_0 \beta + \delta)}{1 + \exp(a'_0 \beta + \delta)} / \frac{1}{1 + \exp(a'_0 \beta + \delta)} \right\} \right] \\ &= \ln \frac{\exp(a'_1 \beta + \delta)}{\exp(a'_0 \beta + \delta)} = (a_1 - a_0)' \beta, \end{aligned} \tag{2.11}$$

but PCLE cannot provide the ‘marginal effect’ of changing  $x_t$  from  $a_0$  to  $a_1$  because the marginal distribution  $F_\delta$  of  $\delta$  is not known in

$$\int \{P(y_t = 1|\delta, x_t = a_1) - P(y_t = 1|\delta, x_t = a_0)\} dF_\delta(\delta).$$

One way to get a marginal effect is using  $E_i(y_{it})$  where the subscript  $i$  denotes the expectation “within  $i$ ” using the time-series information of individual  $i$ . Observe

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T \frac{\exp(x'_{it}\beta)}{\exp(-\delta_i) + \exp(x'_{it}\beta)} &= \frac{1}{T} \sum_{t=1}^T \frac{\exp(x'_{it}\beta + \delta_i)}{1 + \exp(x'_{it}\beta + \delta_i)} \\ &\simeq E_i\{E_i(y_{it}|\delta_i, x_{it})\} = E_i(y_{it}) \simeq \bar{y}_i. \end{aligned}$$

Using the first and last expressions, set

$$\frac{1}{T} \sum_{t=1}^T \frac{\exp(x'_{it}b_N)}{\exp(-\delta_i) + \exp(x'_{it}b_N)} = \bar{y}_i$$

to solve this for the solution  $\hat{\delta}_i$ ;  $\hat{\delta}_i$  does not exist for the individuals with  $\bar{y}_i = 0$  or 1 as  $\hat{\delta}_i = -\infty$  or  $\infty$ , respectively. A  $\hat{\delta}_i$ -based marginal-effect estimator for  $x_t$  changing from  $a_0$  to  $a_1$  on the subpopulation  $\bar{y}_i \neq 0, 1$  is

$$\frac{1}{\sum_i 1[\bar{y}_i \neq 0, 1]} \sum_{i:\bar{y}_i \neq 0, 1} \left\{ \frac{\exp(a'_1 b_N)}{\exp(-\hat{\delta}_i) + \exp(a'_1 b_N)} - \frac{\exp(a'_0 b_N)}{\exp(-\hat{\delta}_i) + \exp(a'_0 b_N)} \right\}.$$

Note that  $\hat{\delta}_i$  includes time-constant variables, observed or not. In theory, this marginal effect estimator requires a large  $T$ , but it seems to work reasonably well even for a small  $T$ . The population version of the marginal effect estimator is

$$E\{E(y_{it}|\delta_i, x_{it} = a_1, \bar{y}_i \neq 0, 1) - E(y_{it}|\delta_i, x_{it} = a_0, \bar{y}_i \neq 0, 1) | \bar{y}_i \neq 0, 1\};$$

the outer expected value conditional on  $\bar{y}_i \neq 0, 1$  is obtained by integrating out  $\delta_i | (\bar{y}_i \neq 0, 1)$ .

#### 7.2.3.4 Time-Varying Parameters

Since any time-constant regressor is cancelled out in the regressor differences, only time-varying regressors are in  $x_{it}$  to appear in the form  $\Delta x_{it}$ . But if parameters change, then the corresponding time-constant regressors do not drop out of  $x_{it}$ . For PCLE, a time-constant regressor vector  $c_i$  with slopes  $\beta_{ct}$  appears in the form  $\beta'_{ct}c_i - \beta'_{c1}c_i = (\beta_{ct} - \beta_{c1})'c_i$ ; i.e., the parameter difference from the first period is identified. Often in practice, only the intercept is specified to be time-varying, and the intercept differences relative to the first period intercept are estimated with time dummies.

To see the time-varying intercept  $\beta_t$  identified by the time dummies, write  $\beta_t$  as

$$\beta_t = \beta_1 + (\beta_2 - \beta_1)1[t = 2] +, \dots, +(\beta_T - \beta_1)1[t = T] \quad \forall t.$$

The above  $\beta'_{ct}c_i - \beta'_{c1}c_i$  can be understood analogously:

$$\beta'_{ct}c_i = \beta'_{c1}c_i + (\beta_{c2} - \beta_{c1})'c_i1[t = 2] +, \dots, +(\beta_{cT} - \beta_{c1})'c_i1[t = T] \quad \forall t;$$

$\beta_{c1}$  is the slope for  $c_i$ ,  $\beta_{c2} - \beta_{c1}$  is the slope for the interaction term  $c_i 1[t = 2]$ , etc. That is,  $\beta_{c\tau} - \beta_{c1}$  can be estimated as the slope for  $c_i 1[t = \tau]$ . If all components of  $x_{it}$  are time-varying with slope  $\beta_t$ , then  $x'_{it}\beta_t - x'_{i1}\beta_1$  should replace  $(x_{it} - x_{i1})'\beta$ . Lee (2014) examined PCLE with time-varying parameters in detail for both static and dynamic models; as will be seen shortly, there are restrictions on parameters in dynamic models.

#### 7.2.3.5 Dynamics in Panel Logit Model

The dynamics allowed by PCLE is restrictive in two ways. Firstly,  $u_{i1}, \dots, u_{iT}$  are iid, and thus  $v_{it} = \delta_i + u_{it}$ ,  $t = 1, \dots, T$ , are allowed to be related only through  $\delta_i$ . This is plausible for cross-section data with a group structure where each subject in the group has the same relation with any other subject in the same group (i.e., ‘equi-correlation’ case), but not plausible otherwise because the serial correlation pattern of  $v_{it}$  is not allowed to change at all over time. Secondly,  $u_{it}$  is independent of  $(\delta_i, x_{i1}, \dots, x_{iT})$ , neither just of  $(\delta_i, x_{it})$  nor of  $(\delta_i, x_{i1}, \dots, x_{it})$ ; these three independences are of the type ‘strict exogeneity’, ‘contemporaneity’ and ‘predeterminedness’, respectively.

One disadvantage of strict exogeneity is that, by constraining  $u_{it}$  to be independent not just of the past and the present but also of the future regressors, we assume away economic agents who adjust the future  $x_{it}$  in view of the past  $u_{is}$ ,  $s < t$ . Another disadvantage is that  $y_{i,t-1}$  is not allowed in  $x_{it}$ : if  $y_{i,t-1}$  were in  $x_{it}$ , then  $u_{it}$  could not be independent of  $x_{i1}, \dots, x_{iT}$ , because  $y_{it}$  including  $u_{it}$  would appear in  $x_{i,t+1}$ . This shortcoming is overcome (partly) in dynamic PCLE to be discussed next.

### 7.3 DYNAMIC PANEL CONDITIONAL LOGIT

Consider a dynamic model with  $\alpha$  being an additional parameter:

$$y_{it} = 1[\alpha y_{i,t-1} + x'_{it}\beta + \delta_i + u_{it} > 0], \quad i = 1, \dots, N, \quad t = 1, \dots, T;$$

$(u_{i1}, \dots, u_{iT})$  are iid logistic and  $(u_{i1}, \dots, u_{iT}) \perp\!\!\!\perp (\delta_i, X_i, y_{i0})$ ;

$(y_{i0}, Y_i, X_i)$  are observed and iid across  $i = 1, \dots, N$ .

Here we expand the periods down to  $t = 0$  so that  $y_{i0}$  appears; for this dynamic model, the total number of periods needed is  $T + 1 \geq 3$ . Using the notation  $t = 0$  is not really necessary, but without  $t = 0$ , many notations used for the static model should be altered.

The dynamic model assumptions imply

$$P(y_t = 1|y_{t-1}, \dots, y_0, \delta, X) = P(y_t = 1|y_{t-1}, \delta, X_t) = \frac{\exp(\alpha y_{t-1} + x'_t \beta + \delta)}{1 + \exp(\alpha y_{t-1} + x'_t \beta + \delta)} \text{ and}$$

$$\ln \left\{ \frac{P(y_t = 1|y_{t-1} = 1, \delta, X)/P(y_t = 0|y_{t-1} = 1, \delta, X)}{P(y_t = 1|y_{t-1} = 0, \delta, X)/P(y_t = 0|y_{t-1} = 0, \delta, X)} \right\} = \ln \left\{ \frac{\exp(\alpha + x'_t \beta + \delta)}{\exp(x'_t \beta + \delta)} \right\} = \alpha :$$

$\alpha$  is the logged odds ratio for  $y_t$  as  $y_{t-1}$  changes from 0 to 1. The joint likelihood of  $Y$  given  $(y_0, \delta, X)$  is (compare to (2.1))

$$\begin{aligned} P(Y|y_0, \delta, X) &= \prod_t P(y_t|y_{t-1}, \dots, y_0, \delta, X) = \prod_t P(y_t|y_{t-1}, \delta, x_t) \\ &= \frac{\exp\{y_1(\alpha y_0 + x'_1 \beta + \delta)\}}{1 + \exp(\alpha y_0 + x'_1 \beta + \delta)} \cdots \frac{\exp\{y_T(\alpha y_{T-1} + x'_T \beta + \delta)\}}{1 + \exp(\alpha y_{T-1} + x'_T \beta + \delta)} \\ &= \frac{\exp(\alpha \sum_t y_{t-1} y_t + \sum_t y_t x'_t \beta + \delta \sum_t y_t)}{\prod_t \{1 + \exp(\alpha y_{t-1} + x'_t \beta + \delta)\}}. \end{aligned} \quad (3.1)$$

Unfortunately, removing  $\delta$  as in (2.1) to (2.5) does not work for (3.1). But the literature has seen three ideas to overcome this problem. The first (Chamberlain 1985) requiring at least four periods ( $T = 3$  and period 0) conditions on  $y_3$  under  $\beta = 0$  (no time-varying regressors). With  $(y_0, y_3, \sum_t y_t)$  fixed, the two middle-period responses  $y_1$  and  $y_2$  work just as in the two-period static PCLE, and the sample maximands with four periods and the general  $T + 1$  periods are  $(L_1)$  and  $(L'_1)$  below. The second (Honore and Kyriazidou 2000; HK) relaxes  $\beta = 0$  in the first approach by assuming that the last two period regressors are the same ( $x_2 = x_3$ ); the sample maximands with four periods are  $(L_2)$  and  $(L'_2)$ . The third (Bartolucci and Nigro 2010; BN) requires only three periods without conditioning on the last period response, but it adds a modifying term in the regression function; the maximand with a general  $T$  is  $(L_3)$ . BN also proposed a version requiring at least four periods with  $y_3$  conditioned on; the maximands with four periods and the general  $T + 1$  periods are  $(L_4)$  and  $(L'_4)$ . These three ideas are examined one by one in this section. The reader not interested in the ideas behind the estimators may skip to the sample maximands “ $L_\#$ ”.

### 7.3.1 Four Periods or More with No Regressor

With  $T = 3$  and  $\beta = 0$ , the likelihood function for  $(y_1, y_2, y_3)|(y_0, \delta)$  is

$$P(y_1, y_2, y_3|y_0, \delta) = \frac{\exp\{\alpha(y_0 y_1 + y_1 y_2 + y_2 y_3) + \delta(y_1 + y_2 + y_3)\}}{\{1 + \exp(\alpha y_0 + \delta)\}\{1 + \exp(\alpha y_1 + \delta)\}\{1 + \exp(\alpha y_2 + \delta)\}}. \quad (3.2)$$

Given  $(y_0, \delta)$ , consider  $(y_1 = 0, y_2 = 1, y_3)$  and  $(y_1 = 1, y_2 = 0, y_3)$ :

$$\begin{aligned} P(y_1 = 0, y_2 = 1, y_3 | y_0, \delta) &= \frac{\exp\{\alpha y_3 + \delta(1 + y_3)\}}{\{1 + \exp(\alpha y_0 + \delta)\}\{1 + \exp(\delta)\}\{1 + \exp(\alpha + \delta)\}}; \\ P(y_1 = 1, y_2 = 0, y_3 | y_0, \delta) &= \frac{\exp\{\alpha y_0 + \delta(1 + y_3)\}}{\{1 + \exp(\alpha y_0 + \delta)\}\{1 + \exp(\alpha + \delta)\}\{1 + \exp(\delta)\}}. \end{aligned} \quad (3.3)$$

Since the two denominators are the same, only the numerators matter in the ratio of the first probability to the sum of the two probabilities:

$$\frac{P(y_1 = 0, y_2 = 1, y_3 | y_0, \delta)}{P(y_1 = 1, y_2 = 0, y_3 | y_0, \delta) + P(y_1 = 0, y_2 = 1, y_3 | y_0, \delta)} = \frac{\exp\{\alpha y_3 + \delta(1 + y_3)\}}{\exp\{\alpha y_0 + \delta(1 + y_3)\} + \exp\{\alpha y_3 + \delta(1 + y_3)\}} = \frac{\exp\{\alpha(y_3 - y_0)\}}{1 + \exp\{\alpha(y_3 - y_0)\}}$$

which is free of  $\delta$ —this scenario will appear again. The ratio can be written also as

$$\frac{P(y_1 = 0, y_2 = 1, y_1 + y_2 = 1, y_3 | y_0, \delta)}{P(y_1 + y_2 = 1, y_3 | y_0, \delta)} = P(y_1 = 0, y_2 = 1 | y_1 + y_2 = 1, y_0, y_3, \delta).$$

Equating this to the preceding display gives the key expression

$$\begin{aligned} P(y_1 = 0, y_2 = 1 | y_1 + y_2 = 1, y_0, y_3, \delta) &= \frac{\exp\{\alpha(y_3 - y_0)\}}{1 + \exp\{\alpha(y_3 - y_0)\}} \\ \implies P(y_1 = 1, y_2 = 0 | y_1 + y_2 = 1, y_0, y_3, \delta) &= \frac{1}{1 + \exp\{\alpha(y_3 - y_0)\}}. \end{aligned}$$

Hence the sample log-likelihood function is

$$\sum_i 1[y_{i1} + y_{i2} = 1] \cdot \ln \left[ \frac{\exp\{y_{i2}\alpha(y_{i3} - y_{i0})\}}{1 + \exp\{\alpha(y_{i3} - y_{i0})\}} \right] \quad (\text{L}_1)$$

reminiscent of a binary response model for  $y_2$  with  $y_3 - y_0$  as a regressor for  $\alpha$ . Clearly,  $P(y_3 \neq y_0) > 0$  is necessary and we may thus attach  $1[y_{i3} \neq y_{i0}]$  to (L<sub>1</sub>).

For a general  $T$ , the maximand is (the last display can be derived from this)

$$\begin{aligned} \sum_i \ln P \left( y_{i1}, \dots, y_{iT-1} \mid \sum_{t=1}^{T-1} y_{it}, y_{i0}, y_{iT} \right) \\ = \sum_i \ln \left\{ \frac{\exp(\alpha \sum_t y_{i,t-1} y_{it})}{\sum_{\bar{\lambda}_{1,T-1}=\bar{y}} \exp(\alpha \sum_t \lambda_{t-1} \lambda_t)} \right\} \quad (\text{L}'_1) \end{aligned}$$

where  $\sum_{\bar{\lambda}_{1,T-1}=\bar{y}}$  denotes the sum over all sequences  $(y_0, \lambda_1, \dots, \lambda_{T-1}, y_T)'$  with  $\lambda_t = 0, 1$  for  $t = 1, \dots, T-1$  and  $\sum_{t=1}^{T-1} \lambda_t = \sum_{t=1}^{T-1} y_t$ . Although regressors are not allowed, time-constant ones are subsumed in  $\delta$ .

Bushway et al. (1999) used the above estimator with  $y_{it}$  having a police contact for criminal activity. They used  $N = 13,160$  males born in Philadelphia in 1958. Constructing artificial seven periods using three-year age bands (period 1 for ages 6–8, period 2 for ages 9–11, ...), they found  $\hat{\alpha}_1 = 1.58$  with  $t$ -value  $-38.6$ . Despite not controlling for any time-varying regressors, when  $\hat{\alpha}_1 = 1.58$  was divided by 1.8 to get a comparable number to probit, the resulting estimate was not much different from the

slope estimate of  $y_{i,t-1}$  in a dynamic panel probit controlling for time-varying regressors. Certainly, state dependence in crimes is an important issue for criminology, as much as state dependence in work or not is for economics and brand loyalty is for marketing.

If the duration in a certain state spans two adjacent periods (e.g., an unemployment duration starting before the end of  $t - 1$  and continuing after the beginning of  $t$ , and  $y_t = 1$  if unemployed anytime in period  $t$ ), then this automatically results in the first-order dependence of  $y_t$ . This gives a motivation to check for a *second-order dynamic model*. A second-order dynamic logit model with no regressor in Chamberlain (1985, p. 16) is

$$P(y_{it} = 1 | y_{i,t-1}, y_{i,t-2}, \dots, y_{i0}, \alpha_{1i}, \delta_i) = \frac{\exp(\alpha_{1i}y_{it-1} + \alpha_2y_{i,t-2} + \delta_i)}{1 + \exp(\alpha_{1i}y_{it-1} + \alpha_2y_{i,t-2} + \delta_i)} \quad (3.4)$$

where  $\alpha_{1i}$  is allowed to vary across  $i$  as  $\alpha_{1i}$  becomes removed along with  $\delta_i$  eventually; only  $\alpha_2$  is to be estimated.

For the second-order dynamic model, at least six waves are needed ( $T = 5$  along with  $y_0$ ), and the sample maximand for  $\alpha_2$  with six waves is

$$\sum_i 1[y_{i2} + y_{i3} = 1, y_{i1} = y_{i4}] \cdot \ln \left[ \frac{\exp\{y_{i3}\alpha_2(y_{i5} - y_{i0})\}}{1 + \exp\{\alpha_2(y_{i5} - y_{i0})\}} \right] \quad (3.5)$$

reminiscent of logit with  $y_3$  as the response and  $y_5 - y_0$  as the regressor for  $\alpha_2$ . Clearly  $P(y_5 \neq y_0) > 0$  is necessary, and thus we may attach  $1[y_{i5} \neq y_{i0}]$  to the maximand.

For a general  $T$ ,

$$\begin{aligned} P\left(y_2, y_3, \dots, y_{T-2} | \sum_{t=2}^{T-2} y_t, \sum_{t=2}^{T-1} y_{t-1}y_t, y_0, y_1, y_{T-1}, y_T\right) \\ = \frac{\exp\left(\alpha_2 \sum_{t=2}^T y_{t-2}y_t\right)}{\sum_{\tilde{\lambda}_{2,T-2}=\bar{y}, \bar{y}_1} \exp\left(\alpha_2 \sum_{t=2}^T \lambda_{t-2}\lambda_t\right)} \end{aligned}$$

where  $\sum_{\tilde{\lambda}_{2,T-2}=\bar{y}, \bar{y}_1}$  denotes the sum over all sequences  $(y_0, y_1, \lambda_2, \dots, \lambda_{T-2}, y_{T-1}, y_T)'$  such that  $\lambda_t = 0, 1$  for  $t = 2, \dots, T - 2$ ,  $\sum_{t=2}^{T-2} \lambda_t = \sum_{t=2}^{T-2} y_t$  and  $\sum_{t=2}^{T-1} \lambda_{t-1}\lambda_t = \sum_{t=2}^{T-1} y_{t-1}y_t$ . Taking  $\sum_i \ln(\cdot)$  on this yields the sample maximand.

An empirical application appeared in Corcoran and Hill (1985) on whether the current unemployment  $y_t$  depends on the past unemployment  $y_{t-1}$  or not. Corcoran and Hill (1985) used PSID for 1972–1976 (five waves) with  $N = 1251$  to find a significant result:  $\hat{\alpha}_1 = 1.29$  with standard error 0.36. To see if this state dependence is spurious or not (i.e., due to the aforementioned problem of unemployment duration spanning two adjacent periods), they then used PSID for 1969–1976 to find an insignificant result:  $\hat{\alpha}_2 = -0.25$  with standard error 0.43. They concluded no true state dependence, contrary to most studies on state dependence. Of course, the conclusion is subject to

the ‘no relevant time-varying regressor’ assumption, which is unlikely to hold. Also the “effective number” of observations was too small in their application, essentially relying on fewer than 100 individuals.

Magnac (2000) proposed a multinomial generalization of (L<sub>1</sub>) with  $y_{it}$  representing a multiple choice on labor market states: stable employment, temporary employment, paid training, unemployment and schooling. Using French panel data with  $N = 5454$  over 1989–1992, Magnac (2000) constructed Markov-chain logged odds ratio matrices by estimating the  $y_{i,t-1}$ ’s slope indexed by the ‘source state’  $y_{i,t-1} = j$  and ‘destination state’  $y_{it} = l$  while allowing for arbitrary  $\delta_{il}$ ’s to find, overall, a substantial state dependence. Magnac (2000) also considered dependence of second order ( $y_{it}$  depending on  $y_{i,t-1}$  and  $y_{i,t-2}$ ) or higher.

### 7.3.2 Four Periods with the Same Last Two-Period Regressors

Analogous to (3.2) is the likelihood function for  $(y_1, y_2, y_3) | (y_0, \delta, X)$  with  $T = 3$ :

$$\frac{\exp\{\alpha(y_0y_1 + y_1y_2 + y_2y_3) + (y_1x_1 + y_2x_2 + y_3x_3)'\beta + \delta(y_1 + y_2 + y_3)\}}{\{1 + \exp(\alpha y_0 + x'_1\beta + \delta)\}\{1 + \exp(\alpha y_1 + x'_2\beta + \delta)\}\{1 + \exp(\alpha y_2 + x'_3\beta + \delta)\}}.$$

From this,

$$\begin{aligned} & P(y_1 = 0, y_2 = 1, y_3 | y_0, \delta, X) \\ &= \frac{\exp\{\alpha y_3 + (x_2 + y_3x_3)'\beta + \delta(1 + y_3)\}}{\{1 + \exp(\alpha y_0 + x'_1\beta + \delta)\}\{1 + \exp(x'_2\beta + \delta)\}\{1 + \exp(\alpha + x'_3\beta + \delta)\}}; \\ & P(y_1 = 1, y_2 = 0, y_3 | y_0, \delta, X) \\ &= \frac{\exp\{\alpha y_0 + (x_1 + y_3x_3)'\beta + \delta(1 + y_3)\}}{\{1 + \exp(\alpha y_0 + x'_1\beta + \delta)\}\{1 + \exp(\alpha + x'_2\beta + \delta)\}\{1 + \exp(x'_3\beta + \delta)\}}. \end{aligned}$$

If  $x_2 = x_3$ , then the two denominators are the same, under which we can proceed analogously to the no regressor case.

The ratio of the first probability to the sum of the two probabilities given  $x_2 = x_3$  is

$$\begin{aligned} & \frac{P(y_1 = 0, y_2 = 1, y_3 | y_0, \delta, X, x_2 = x_3)}{P(y_1 = 1, y_2 = 0, y_3 | y_0, \delta, X, x_2 = x_3) + P(y_1 = 0, y_2 = 1, y_3 | y_0, \delta, X, x_2 = x_3)} \\ &= P(y_1 = 0, y_2 = 1 | y_0, y_3, y_1 + y_2 = 1, \delta, X, x_2 = x_3) \\ &= \frac{\exp\{\alpha y_3 + (x_2 + y_3x_3)'\beta + \delta(1 + y_3)\}}{\exp\{\alpha y_0 + (x_1 + y_3x_3)'\beta + \delta(1 + y_3)\} + \exp\{\alpha y_3 + (x_2 + y_3x_3)'\beta + \delta(1 + y_3)\}} \\ &= \frac{\exp\{\alpha(y_3 - y_0) + (x_2 - x_1)'\beta\}}{1 + \exp\{\alpha(y_3 - y_0) + (x_2 - x_1)'\beta\}} \quad (\text{free of } \delta). \end{aligned}$$

When  $x_{it}$  is discretely distributed, the sample log-likelihood function is

$$\sum_i 1[y_{i1} + y_{i2} = 1] \cdot 1[x_{i2} = x_{i3}] \cdot \ln \left[ \frac{\exp[y_{i2}\{\alpha(y_{i3} - y_{i0}) + (x_{i2} - x_{i1})'\beta\}]}{1 + \exp\{\alpha(y_{i3} - y_{i0}) + (x_{i2} - x_{i1})'\beta\}} \right] \quad (\text{L}_2)$$

where  $y_{i2}$  plays the role of a binary response, and ' $x_{i2} = x_{i3}$ ' rules out the time dummies for  $t = 2, 3$  as well as age in  $x_{it}$ —a rather restrictive feature for micro panel data. The resulting estimator  $g_N$  is  $\sqrt{N}$ -consistent for  $\gamma \equiv (\alpha, \beta')$  and asymptotically normal. The asymptotic variance can be estimated in the usual MLE way. Differently from the no-regressor case, the observations with  $y_3 = y_0$  can be useful for  $\beta$ , although not for  $\alpha$ .

When  $x_{it}$  is continuously distributed, the maximand is

$$\begin{aligned} & \sum_i 1[y_{i1} + y_{i2} = 1] \cdot K\left(\frac{x_{i2} - x_{i3}}{h}\right) \\ & \cdot \ln \left[ \frac{\exp[y_{i2}\{\alpha(y_{i3} - y_{i0}) + (x_{i2} - x_{i1})'\beta\}]}{1 + \exp\{\alpha(y_{i3} - y_{i0}) + (x_{i2} - x_{i1})'\beta\}} \right] \end{aligned} \quad (\text{L}'_2)$$

where  $K(\cdot)$  is a kernel (e.g., the  $N(0, 1)$  density) and  $h$  is a bandwidth; the density for  $x_2 - x_3$  should be positive at zero. The resulting estimator is  $(Nh^{k_x})^{1/2}$ -consistent and asymptotically normal with the variance estimable as follows.

The asymptotic variance matrix is  $J^{-1}VJ^{-1}$ : defining  $x_{i23} \equiv x_{i2} - x_{i3}$ ,  $z_i \equiv (y_{i3} - y_{i0}, x'_{i2} - x'_{i1})'$  and  $\gamma \equiv (\alpha, \beta')$  and denoting the density of  $x_{23}$  as  $f_{23}(\cdot)$ ,

$$\begin{aligned} J & \equiv f_{23}(0) \cdot E \left\{ 1[y_1 \neq y_2] \frac{\exp(z'\gamma)}{\{1 + \exp(z'\gamma)\}^2} zz' | x_{23} = 0 \right\}, \\ V & \equiv \int K(v)^2 dv \cdot f_{23}(0) \cdot E \left[ 1[y_1 \neq y_2] \left\{ y_2 - \frac{\exp(z'\gamma)}{1 + \exp(z'\gamma)} \right\}^2 zz' | x_{23} = 0 \right]. \end{aligned}$$

With the estimator  $g_N \rightarrow^P \gamma$ , consistent estimators of  $J$  and  $V$  are, respectively,

$$J_N \equiv \frac{1}{Nh^{k_x}} \sum_i K\left(\frac{x_{i23}}{h}\right) \cdot 1[y_{i1} \neq y_{i2}] \frac{\exp(z'_i g_N)}{\{1 + \exp(z'_i g_N)\}^2} z_i z'_i,$$

$$V_N \equiv \frac{1}{Nh^{k_x}} \sum_i K\left(\frac{x_{i23}}{h}\right)^2 \cdot 1[y_{i1} \neq y_{i2}] \left\{ y_{i2} - \frac{\exp(z'_i g_N)}{1 + \exp(z'_i g_N)} \right\}^2 z_i z'_i.$$

An application of the estimator can be seen, e.g., in Chintagunta et al. (2001) for two major yogurt brand choices with  $N = 737$  and  $\sum_i T_i = 5618$  (unbalanced panel with each purchase occasion serving as each period) using scanner data from Sioux Falls, South Dakota, over 1986–1988. Other applications appeared in Lee and Tae (2005) for female working status using Korean panel data “KLIPS” over 1998–2001 with  $N = 3882$ , and Biewen (2009) for poverty status using German panel data “GSOEP”

over 2000–2006 with  $N = 3952$ . HK showed an extension to panel multinomial choice where the current choice depends on the previous choice, which will be examined later.

HK also looked at the second-order dependence of the form (3.4) with  $x'_{it}\beta$  added. HK stated that six waves ( $T = 5$  along with  $y_0$ ) are enough to estimate  $\alpha_2$  and  $\beta$  with sequences satisfying  $x_3 = x_4 = x_5$  as well as

$$y_2 + y_3 = 1 \text{ and } (y_0 \neq y_1, y_1 = y_4 = y_5) \text{ or } (y_4 \neq y_5, y_0 = y_1 = y_4). \quad (3.6)$$

‘ $x_3 = x_4 = x_5$ ’ is for  $\beta$ , and (3.6) implies the no-regressor-case condition in (3.5) that is

$$y_2 + y_3 = 1, y_1 = y_4 \text{ and } y_0 \neq y_5. \quad (3.7)$$

But (3.7) does not imply (3.6), which is somewhat counter-intuitive.

When there are more than four waves, there are at least three ways to proceed as explained in Lee (2002). The first is the “genuine conditional approach”: HK (p. 852) proposed a sample maximand using five waves that conditions on  $x_2 = x_3 = x_4$  and  $\sum_t y_t$  to remove  $\delta$ ; this, however, slows down the convergence rate when  $x_t$  is continuously distributed. The second is using each four waves one by one. For instance, with five waves, the first four waves (0, 1, 2, 3) give one set of estimates and the second (1, 2, 3, 4) gives another set; the time-constant parameter estimates can be then combined using MDE. This kind of MDE will be explained in Section 7.4; see also Lee (2002, 2010) and the references therein for MDE. The third is maximizing a single sample maximand that is the sum of sub-maximands corresponding to sub-waves, as was suggested in HK (p. 852). Among these three approaches, the MDE would be the most practical and efficient: the first approach has a slower convergence rate than the MDE, and the third makes the sum of the score functions zero, not necessarily the individual score functions.

### 7.3.3 Three Periods or More With Regressors

#### 7.3.3.1 Three Periods or More without $y_T$ Conditioned on

BN presented a dynamic PCLE without  $x_2 = x_3$ . They assumed

$$P(y_t = 1 | y_{t-1}, \dots, y_0, \delta, X) = P(y_t = 1 | y_{t-1}, \delta, X) = \frac{\exp\{\alpha y_{t-1} + x'_t \beta + \delta + e_t^*(\delta, X)\}}{1 + \exp\{\alpha y_{t-1} + x'_t \beta + \delta + e_t^*(\delta, X)\}}$$

instead of the panel logit model, where  $e_t^*(\delta, X)$  is defined recursively from  $t = T$  backward as (with an additional parameter  $\beta_T^*$ )

$$e_T^*(\delta, X) \equiv x'_T \beta_T^* \text{ and } e_t^*(\delta, X) \equiv \ln \frac{1 + \exp\{\alpha + x'_{t+1} \beta + \delta + e_{t+1}^*(\delta, X)\}}{1 + \exp\{x'_{t+1} \beta + \delta + e_{t+1}^*(\delta, X)\}}, \quad t < T.$$

In the model,  $y_t$  depends on the future  $x_t$ 's through  $e_t^*(\delta, X)$ . As  $y_T$  is not conditioned on, three waves ( $T = 2$  along with  $y_0$ ) are enough, differently from HK. Despite

$e_t^*(\delta, X)$ ,  $\alpha$  still equals the logged odds ratio of  $y_t$  with  $y_{t-1}$  changing from 0 to 1:

$$\begin{aligned} & \ln \left\{ \frac{P(y_t = 1|y_{t-1} = 1, \delta, X)/P(y_t = 0|y_{t-1} = 1, \delta, X)}{P(y_t = 1|y_{t-1} = 0, \delta, X)/P(y_t = 0|y_{t-1} = 0, \delta, X)} \right\} \\ &= \ln \left\{ \frac{\exp\{\alpha + x'_t \beta + \delta + e_t^*(\delta, X)\}}{\exp\{x'_t \beta + \delta + e_t^*(\delta, X)\}} \right\} = \alpha. \end{aligned}$$

Unfortunately, the steps from the above model assumption to the sample log-likelihood just below are not shown in BN despite its importance; they can be seen in Lee (2014) in a more general model with time-varying parameters.

The sample log-likelihood function for a general  $T$  is

$$\begin{aligned} & \sum_i \ln \left[ \frac{\exp\{\alpha \sum_t y_{i,t-1} y_{it} + \sum_t y_{it} x'_{it} \beta + y_{iT} x'_{iT} \beta_T^*\}}{\sum_{\bar{\lambda}_{1T}=\bar{y}} \exp\{\alpha \sum_t \lambda_{t-1} \lambda_t + \sum_t \lambda_t x'_{it} \beta + \lambda_{iT} x'_{iT} \beta_T^*\}} \right] \quad (\text{L}_3) \\ &= \sum_i \ln \left[ \sum_t y_{it} \neq 0, T \right] \\ & \ln \left[ \frac{\exp\{\alpha \sum_t y_{i,t-1} y_{i,t} + \sum_{t \geq 2} y_{it} (x_{it} - x_{i1})' \beta + y_{iT} x'_{iT} \beta_T^*\}}{\sum_{\bar{\lambda}_{1T}=\bar{y}} \exp\{\alpha \sum_t \lambda_{t-1} \lambda_t + \sum_{t \geq 2} \lambda_t (x_{it} - x_{i1})' \beta + \lambda_{iT} x'_{iT} \beta_T^*\}} \right] \end{aligned}$$

dividing the numerator and denominator by  $\exp(\sum_t y_{it} \cdot x'_{i1} \beta) = \exp(\sum_t \lambda_t \cdot x'_{i1} \beta)$ , where  $\sum_{\bar{\lambda}_{1T}=\bar{y}} (\cdot)$  denotes the sum over all sequences  $(y_0, \lambda_1, \dots, \lambda_T)'$  such that  $\lambda_t = 0, 1$  for  $t = 1, \dots, T$ , and  $\sum_t \lambda_t = \sum_t y_t$ .

(L<sub>3</sub>) is to be maximized for  $(\alpha, \beta, \beta_T^*)$  and the resulting estimator is  $\sqrt{N}$ -consistent and asymptotically normal; the variance can be estimated in the usual MLE way. *Differently from HK, no restriction such as  $x_{T-1} = x_T$  is needed* for this estimator. This allows age and the time dummies for the last two periods, which is critical in micro-econometrics; also, as was already mentioned, three waves are enough.

### 7.3.3.2 Four Periods or More with $y_T$ Conditioned on

The motivation to add  $e_t^*(\delta, X)$  in the above panel model is that it results in a PCLE with the maximand consisting only of lagged responses,  $x_{it}$ 's and  $e_T^*(\delta, X)$ —no  $\delta$ , nor  $e_t^*(\delta, X)$ 's. But the presence of  $e_t^*(\delta, X)$  in  $P(y_t = 1|y_{t-1}, \delta, X)$  is hard to justify. Also, the assumptions about  $e_T^*(\delta, X)$  are implausible because: (i)  $e_T^*(\delta, X)$  does not depend on  $\delta$  while  $e_t^*(\delta, X)$  does; (ii) only  $x_T$ , not the other  $x_t$ 's, is relevant for  $e_T^*(\delta, X)$ ; (iii)  $e_T^*(\delta, X)$  is linear in parameters while  $e_t^*(\delta, X)$ 's are not; and (iv) when  $\alpha = 0$ ,  $e_t^*(\delta, X) = 0 \forall t < T$  but  $e_T^*(\delta, X) \neq 0$ , which means that the dynamic model does not reduce to the static model when  $\alpha = 0$ . This heavy dependence on the last period is reminiscent of “identification at infinity.”

To avoid these shortcoming, BN suggested another dynamic PCLE *conditioning extra on  $y_T$* , which then removes  $e_T^*(\delta, X)$  from the maximand; with  $\beta_T^*$  no longer estimated,

this version reduces to the static PCLE when  $\alpha = 0$ . This dynamic PCLE with at least four waves and  $y_T$  conditioned on is well aligned with the other preceding dynamic PCLE's, and it is presented next.

When  $T = 4$ , there are only two informative sequences  $(y_0, 0, 1, y_3)$  and  $(y_0, 1, 0, y_3)$  as in the other PCLE's, and the sample maximand is (recall (L<sub>3</sub>) without  $x'_{iT}\beta_T^*$ )

$$\begin{aligned} & \sum_i 1[y_{i1} + y_{i2} = 1] \cdot \ln \left[ \frac{\exp(\alpha \sum_t y_{i,t-1} y_{it} + y_{i2}(x_{i2} - x_{i1})' \beta)}{\sum_{\bar{\lambda}_{1,2}=\bar{y}} \exp\{\alpha \sum_t \lambda_{t-1} \lambda_t + \lambda_2(x_{i2} - x_{i1})' \beta\}} \right] \\ &= \sum_i 1[y_{i1} + y_{i2} = 1] \cdot \ln \left[ \frac{\exp\{\alpha(y_{i0}y_{i1} + y_{i2}y_{i3}) + y_{i2}(x_{i2} - x_{i1})' \beta\}}{\exp\{\alpha y_{i3} + (x_{i2} - x_{i1})' \beta\} + \exp(\alpha y_{i0})} \right] \\ &= \sum_i 1[y_{i1} + y_{i2} = 1] \cdot \ln \left[ \frac{\exp\{\alpha y_{i0}(y_{i1} - 1) + \alpha y_{i2}y_{i3} + y_{i2}(x_{i2} - x_{i1})' \beta\}}{1 + \exp\{\alpha(y_{i3} - y_{i0}) + (x_{i2} - x_{i1})' \beta\}} \right] \\ &= \sum_i 1[y_{i1} + y_{i2} = 1] \cdot \ln \left[ \frac{\exp[y_{i2}\{\alpha(y_{i3} - y_{i0}) + (x_{i2} - x_{i1})' \beta\}]}{1 + \exp\{\alpha(y_{i3} - y_{i0}) + (x_{i2} - x_{i1})' \beta\}} \right] \end{aligned} \quad (\text{L}_4)$$

where the first equality uses  $(\sum_t \lambda_{t-1} \lambda_t = y_{i3}, \lambda_2 = 1)$  for  $(y_{i0}, 0, 1, y_{i3})$ , and  $(\sum_t \lambda_{t-1} \lambda_t = y_{i0}, \lambda_2 = 0)$  for  $(y_{i0}, 1, 0, y_{i3})$ ; the third uses  $y_{i1} - 1 = -y_{i2}$ . (L<sub>4</sub>) equals (L<sub>2</sub>) without the restriction  $x_2 = x_3$ . For a general  $T$ , the log-likelihood function is

$$\sum_i 1 \left[ \sum_t y_{it} \neq 0, T \right] \cdot \ln \left[ \frac{\exp\{\alpha \sum_t y_{i,t-1} y_{it} + \sum_{t=2}^{T-1} y_{it}(x_{it} - x_{i1})' \beta\}}{\sum_{\bar{\lambda}_{1,T-1}=\bar{y}} \exp\{\alpha \sum_t \lambda_{t-1} \lambda_t + \sum_{t=2}^{T-1} \lambda_t(x_{it} - x_{i1})' \beta\}} \right]. \quad (\text{L}'_4)$$

### 7.3.3.3 Three Periods or More Using an Estimator for $\delta_i$

As an alternative way of having the dynamic PCLE reduce to the static PCLE when  $\alpha = 0$  despite  $T = 2$  and  $y_0$  (i.e., only three waves), Bartolucci and Nigro (2012) proposed to approximate the logarithm of the denominator in (3.1) around  $(\alpha, \beta, \delta_i) = (0, \bar{\beta}, \bar{\delta}_i)$ :

$$\begin{aligned} & \sum_t \ln\{1 + \exp(\alpha y_{i,t-1} + x'_{it}\beta + \delta_i)\} \\ & \simeq \sum_t \left[ \ln\{1 + \exp(x'_{it}\bar{\beta} + \bar{\delta}_i)\} + \frac{\exp(x'_{it}\bar{\beta} + \bar{\delta}_i)}{1 + \exp(x'_{it}\bar{\beta} + \bar{\delta}_i)} \{y_{i,t-1}\alpha + x'_{it}(\beta - \bar{\beta}) + \delta_i - \bar{\delta}_i\} \right] \\ &= \sum_t [\ln\{1 + \exp(x'_{it}\bar{\beta} + \bar{\delta}_i)\} + \bar{q}_{it}(\delta_i - \bar{\delta}_i) + \bar{q}_{it}x'_{it}(\beta - \bar{\beta})] \\ & \quad + \alpha \bar{q}_{i1}y_{i0} + \alpha \sum_{t=2}^T \bar{q}_{it}y_{i,t-1} \\ \text{where } & \bar{q}_{it} \equiv \frac{\exp(x'_{it}\bar{\beta} + \bar{\delta}_i)}{1 + \exp(x'_{it}\bar{\beta} + \bar{\delta}_i)}. \end{aligned}$$

Taking  $\exp(\cdot)$  on this gives

$$\begin{aligned} \prod_t \{1 + \exp(\alpha y_{i,t-1} + x'_{it}\beta + \delta)\} &\simeq \exp\left(\alpha \bar{q}_{i1} y_{i0} + \alpha \sum_{t=2}^T \bar{q}_{it} y_{i,t-1}\right) \\ &\cdot \prod_t \{1 + \exp(x'_{it}\bar{\beta} + \bar{\delta}_i)\} \exp\{\bar{q}_{it}(\delta_i - \bar{\delta}_i) + \bar{q}_{it}x'_{it}(\beta - \bar{\beta})\}. \end{aligned}$$

The main “headache” in dynamic PCLE relative to static PCLE is  $(y_1, \dots, y_{T-1})$  in the denominator of the likelihood function (3.1), which makes removing  $\delta$  by conditioning impossible. But, in the last display, only  $\alpha \sum_{t=2}^T \bar{q}_{it} y_{i,t-1}$  depends on  $(y_{i1}, \dots, y_{iT-1})$ . Hence, moving this term upward to the numerator of (3.1) with a minus sign attached, there remains no term in the denominator of (3.1) that depends on  $(y_{i1}, \dots, y_{iT-1})$  to result in

$$P\left(Y \mid \sum_t y_t, y_0, \delta, X\right) = \frac{\exp\left\{\alpha \sum_t y_{t-1} y_t + \sum_t y_t x'_t \beta - \alpha \sum_{t=2}^T \bar{q}_{it} y_{i,t-1}\right\}}{\sum_{\bar{\lambda}_{1T}=\bar{y}} \exp\left\{\alpha \sum_t \lambda_{t-1} \lambda_t + \sum_t \lambda_t x'_t \beta - \alpha \sum_{t=2}^T \bar{q}_{it} \lambda_{t-1}\right\}}.$$

To implement this estimator, Bartolucci and Nigro (2012) proposed a two-stage procedure: the first stage is obtaining  $\bar{\beta}$  using the static PCLE and then  $\bar{\delta}_i$  as the static PCLE maximizer only for each individual’s time-series with  $\bar{\beta}$  plugged in; the second stage is maximizing the last display for  $(\alpha, \beta)$ . The shortcomings of this approach are that it is not clear how good the above approximation is, and that solving for  $\bar{\delta}_i$  with each individual’s time-series requires  $T \rightarrow \infty$  in principle for the consistency.

## 7.4 PANEL CONDITIONAL ORDERED LOGIT

Consider a panel ordered logit model:

$$y_{it} = \sum_{r=1}^{R-1} 1[x'_{it}\beta + \delta_i + u_{it} \geq \tau_r] \quad \text{where } \tau_1 < \tau_2 < \dots < \tau_{R-1}$$

and  $u_{i1}, \dots, u_{iT}$  are iid logistic independently of  $(\delta_i, x_{i1}, x_{i2}, \dots, x_{iT})$ .

Here  $y_{it}$  takes  $R$  ordered categories  $(0, 1, \dots, R-1)$  with  $R \geq 3$ . Differently from panel logit for binary responses,  $\delta$  cannot be removed by conditioning on  $\sum_t y_t$ . One can easily see this by trying to remove  $\delta$  in the simplest case  $T = 2$  and  $R = 3$ .

*Although there is no genuine PCLE for ordered responses, it is possible to apply PCLE by collapsing an ordered discrete response to binary responses.* For example, if  $y_{it}$  takes 0, 1, 2, then collapse  $y_{it}$  into binary responses: 0 to 0 and (1, 2) to 1, or (0, 1) to 0 and 2 to 1. We can then estimate the parameters for each binary response model and impose the restriction with MDE that the same parameters are estimated by the multiple PCLE’s.

This is explained in this section using a three-category and two-wave case first and then four categories; the general  $R$ -category case can be inferred from these.

For more generality, instead of the above basic model, we allow the intercept and the thresholds to change over time—when the intercept is time-varying, thresholds are likely time-varying as well because the intercept includes a threshold due to a location normalization:

$$y_{it} = \sum_{r=1}^{R-1} 1[\psi_t + x'_{it}\beta + \delta_i + u_{it} \geq \tau_{rt}], \quad t = 1, 2.$$

where  $\psi_t + x'_{it}\beta$  is now the regression function with the regressors  $(1, x'_{it})'$ .

### 7.4.1 Three Categories

When  $R = 3$ , the two collapsed panel binary response models are

- (i) : 0 to 0 and (1, 2) to 1,  $y_{it} = 1[\psi_t - \tau_{1t} + x'_{it}\beta + \delta_i + u_{it} \geq 0]$ ,
- (ii) : (0, 1) to 0 and 2 to 1,  $y_{it} = 1[\psi_t - \tau_{2t} + x'_{it}\beta + \delta_i + u_{it} \geq 0]$ .

Since  $\{\Delta(\psi_2 - \tau_{12}), \beta\}$  are identified in (i) and  $\{\Delta(\psi_2 - \tau_{22}), \beta\}$  are identified in (ii), define ( $\beta$ 's carry a subscript for a reason to become clear shortly)

$$\gamma_1 \equiv \{\Delta(\psi_2 - \tau_{12}), \beta'_1\}', \quad \gamma_2 \equiv \{\Delta(\psi_2 - \tau_{22}), \beta'_2\}' \quad \text{and} \quad \gamma \equiv (\gamma'_1, \gamma'_2)'.$$

Applying PCLE to (i) and (ii) separately, we estimate the “reduced form (RF)” parameters  $\gamma$  while the “structural form (SF)” parameters are  $\gamma_o \equiv \{\Delta(\psi_2 - \tau_{12}), \Delta(\psi_2 - \tau_{22}), \beta'\}'$ , and  $\gamma$  is linked to  $\gamma_o$  by

$$\begin{bmatrix} \Delta(\psi_2 - \tau_{12}) \\ \beta_1 \\ \Delta(\psi_2 - \tau_{22}) \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & I_{k_x} \\ 1 & 0 & 0 \\ 0 & 0 & I_{k_x} \end{bmatrix} \begin{bmatrix} \Delta(\psi_2 - \tau_{12}) \\ \Delta(\psi_2 - \tau_{22}) \\ \beta \end{bmatrix}.$$

Denoting the PCLE for  $\gamma$  as  $\hat{\gamma}$  and the middle matrix with  $I_{k_x}$  as  $L$ , MDE estimates  $\gamma_o$  by minimizing

$$N \cdot (\hat{\gamma} - Lg_o)' W_N^{-1} (\hat{\gamma} - Lg_o)$$

with respect to  $g_o$ , where  $W_N$  is a consistent estimator for the asymptotic variance of  $\sqrt{N}(\hat{\gamma} - \gamma)$ . To obtain  $W_N$ , let  $s_{i1}$  and  $s_{i2}$  denote the score function estimators of the

PCLE's for (i) and (ii), respectively. Then we can use

$$W_N \equiv \frac{1}{N} \sum_i \eta_i \eta'_i \text{ where } \eta_i \equiv (\eta'_{i1}, \eta'_{i2})',$$

$$\eta_{i1} \equiv \left( \frac{1}{N} \sum_i s_{i1} s'_{i1} \right)^{-1} s_{i1} \text{ and } \eta_{i2} \equiv \left( \frac{1}{N} \sum_i s_{i2} s'_{i2} \right)^{-1} s_{i2}.$$

The solution  $g_{mde}$  to the minimization problem and its asymptotic distribution are

$$g_{mde} = (L' W_N^{-1} L)^{-1} L' W_N^{-1} \hat{\gamma} \text{ and } \sqrt{N}(g_{mde} - \gamma_o) \rightsquigarrow N\{0, (L' W^{-1} L)^{-1}\}$$

where  $W \equiv E(\eta \eta')$ . The MDE minimand evaluated at  $g_{mde}$  can be used as a test statistic for the over-identification condition  $H_0 : \beta_1 = \beta_2 = \beta$  because under the  $H_0$

$$N \cdot (\hat{\gamma} - L g_{mde})' W_N^{-1} (\hat{\gamma} - L g_{mde}) \rightsquigarrow \chi_{k_x}^2.$$

### 7.4.2 Four Categories

Turning to  $R = 4$ , the collapsed binary response models are

- (i) : 0 to 0 and (1, 2, 3) to 1,  $y_{it} = 1[\psi_t - \tau_{1t} + x'_{it}\beta + \delta_i + u_{it} \geq 0]$ ,
- (ii) : (0, 1) to 0 and (2, 3) to 1,  $y_{it} = 1[\psi_t - \tau_{2t} + x'_{it}\beta + \delta_i + u_{it} \geq 0]$ ,
- (iii) : (0, 1, 2) to 0 and 3 to 1,  $y_{it} = 1[\psi_t - \tau_{3t} + x'_{it}\beta + \delta_i + u_{it} \geq 0]$ .

The identified parameters are

$$\gamma_1 \equiv \{\Delta(\psi_2 - \tau_{12}), \beta'_1\}', \quad \gamma_2 \equiv \{\Delta(\psi_2 - \tau_{22}), \beta'_2\}', \quad \gamma_3 \equiv \{\Delta(\psi_2 - \tau_{32}), \beta'_3\}'$$

With  $\gamma \equiv (\gamma'_1, \gamma'_2, \gamma'_3)'$  and  $\gamma_o \equiv \{\Delta(\psi_2 - \tau_{12}), \Delta(\psi_2 - \tau_{22}), \Delta(\psi_2 - \tau_{32}), \beta'\}'$ , the restriction linking  $\gamma$  to  $\gamma_o$  is

$$\begin{bmatrix} \Delta(\psi_2 - \tau_{12}) \\ \beta_1 \\ \Delta(\psi_2 - \tau_{22}) \\ \beta_2 \\ \Delta(\psi_2 - \tau_{32}) \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & I_{k_x} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & I_{k_x} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & I_{k_x} \end{bmatrix} \begin{bmatrix} \Delta(\psi_2 - \tau_{12}) \\ \Delta(\psi_2 - \tau_{22}) \\ \Delta(\psi_2 - \tau_{32}) \\ \beta \end{bmatrix}.$$

Let  $s_{i1}$ ,  $s_{i2}$  and  $s_{i3}$  denote the score function estimators of the PCLE's for the binary models (i), (ii) and (iii), respectively. Then

$$W_N = \frac{1}{N} \sum_i \eta_i \eta'_i \text{ where } \eta_i \equiv (\eta'_{i1}, \eta'_{i2}, \eta'_{i3})' \text{ and } \eta_{ij} \equiv \left( \frac{1}{N} \sum_i s_{ij} s'_{ij} \right)^{-1} s_{ij}, j = 1, 2, 3.$$

The MDE  $g_{mde}$  and its asymptotic distribution are

$$g_{mde} = (L' W_N^{-1} L)^{-1} L' W_N^{-1} \hat{\gamma} \quad \text{and} \quad \sqrt{N}(g_{mde} - \gamma_0) \rightsquigarrow N\{0, (L' W^{-1} L)^{-1}\}$$

where  $W = E(\eta\eta')$ . The test statistic for  $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta$  is

$$N \cdot (\hat{\gamma} - Lg_{mde})' W_N^{-1} (\hat{\gamma} - Lg_{mde}) \rightsquigarrow \chi_{2k_x}^2.$$

The above analysis can be generalized allowing for  $T \geq 3$  and time-varying slopes. Yet another generalization is regressor-dependent thresholds: if  $\tau_r$ 's are regressor-dependent as in Lee and Kimhi (2005), then subtracting a different threshold from the original regression function alters the slope parameters differently, which implies different slopes across the binary response models.

### 7.4.3 PCLE with More than Enough Waves

The above MDE can be applied with little change to PCLE with “more-than-enough waves.” For instance, for the dynamic PCLE with no regressor, if there are five waves (one more than the minimum four) 0 to 4, then 0 to 3 can be used to obtain one estimator for  $\alpha$  and 1 to 4 can be used to obtain another estimator for  $\alpha$ . The two estimators can be combined with MDE to obtain a single estimator that is essentially a weighted average of the two estimators with the higher weight given to the one with the smaller variance. Analogous procedures hold for the other dynamic PCLE's that require at least three or four waves. When MDE is applied in this context, only the time-constant parameters should be restricted to be the same in the MDE, as illustrated in the following.

Suppose two binary PCLE's have been done. If the intercept is allowed to change over time but the slope is time-constant, then the appropriate restriction for MDE is

$$\begin{bmatrix} \psi_1 \\ \beta_1 \\ \psi_2 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & I_{k_x} \\ 0 & 1 & 0 \\ 0 & 0 & I_{k_x} \end{bmatrix} \begin{bmatrix} \psi_1 \\ \psi_2 \\ \beta \end{bmatrix}$$

where  $\psi_1$  is the intercept from the first PCLE and  $\psi_2$  is the intercept from the second PCLE. Thus the restriction is binding only on the slope  $\beta$ .

Recall  $(x_{it} - x_{i1})'\beta$  in (2.6) where the time-constant intercept drops out. Since the PCLE intercepts  $\psi_1$  and  $\psi_2$  in the last display are in fact intercept changes relative to the base period,  $\psi_1$  and  $\psi_2$  may not be estimated at all if the intercept is believed to be time-constant. In this case, the MDE restriction to be used is not the last display, but only

$$\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} I_{k_x} \\ I_{k_x} \end{bmatrix} \cdot \beta.$$

## 7.5 PANEL CONDITIONAL MULTINOMIAL LOGIT (PCML)

---

This section examines panel conditional multinomial logit estimator (PCML) that was proposed by Chamberlain (1980, p. 231). Because the panel conditional multinomial logit model has one more dimension ( $J$  alternatives) in addition to  $N$  and  $T$ , notation in this section differs somewhat from that in the other sections. To ease exposition, we deal with the simplest case (3 alternatives and 2 waves) in detail first, and then the general case with  $J$  and  $T$ .

### 7.5.1 Conditional Likelihood for Three Alternatives and Two Periods

Set  $T = 2$  and consider three unordered alternatives ( $j = a, b, c$ ) to choose from. Define

$$\gamma_{ijt} = 1 \text{ if } i \text{ chooses } j \text{ at time } t \text{ and 0 otherwise } \quad \forall i, j, t,$$

$$Y_i \equiv (y_{ia1}, y_{ib1}, y_{ic1}, y_{ia2}, y_{ib2}, y_{ic2})'_{6 \times 1};$$

e.g., if person  $i$  chooses  $b$  at  $t = 1$  and  $c$  at  $t = 2$ , then  $Y_i = (0, 1, 0, 0, 0, 1)'$ .

Suppose that alternative  $j$  at time  $t$  gives satisfaction (or utility)  $s_{ijt}$  to person  $i$ , and

$$s_{ijt} = m'_{it} \alpha_j + x'_{ijt} \beta + \delta_{ij} + u_{ijt},$$

where  $m_{it}$  and  $x_{ijt}$  are regressors with  $m_{it}$  being the alternative-constant regressors and  $x_{ijt}$  being the alternative-varying regressors, and  $\delta_{ij}$  and  $u_{ijt}$  are error terms. One chooses the alternative giving the maximum satisfaction, and consequently, the choice depends only on the location-normalized satisfactions  $s_{ibt} - s_{iat}$  and  $s_{ict} - s_{iat}$ , with alternative  $a$  as the base.

Let  $k_m$  and  $k_x$  denote the row dimension of  $m_{it}$  and  $x_{ijt}$ , respectively. Define

$$\begin{aligned} w_{ijt} &\equiv \begin{bmatrix} m_{it} \\ x_{ijt} - x_{iat} \end{bmatrix}_{(k_m+k_x) \times 1}, \quad \gamma_j \equiv \begin{bmatrix} \alpha_j - \alpha_a \\ \beta \end{bmatrix}, \quad j = a, b, c, \\ w_{it} &\equiv \text{diag}(w_{iat}, w_{ibt}, w_{ict}), \quad \gamma \equiv (\gamma'_a, \gamma'_b, \gamma'_c)'_{3(k_m+k_x) \times 1}. \end{aligned}$$

From this, we obtain

$$\begin{aligned} w'_{ijt} \gamma_j &= m'_{it}(\alpha_j - \alpha_a) + (x_{ijt} - x_{iat})' \beta \quad (\implies w'_{iat} \gamma_a = 0) \\ \gamma' w_{it} &= (\gamma'_a, \gamma'_b, \gamma'_c) \cdot \text{diag}(w_{iat}, w_{ibt}, w_{ict}) \\ &= (w'_{iat} \gamma_a, w'_{ibt} \gamma_b, w'_{ict} \gamma_c) = (0, w'_{ibt} \gamma_b, w'_{ict} \gamma_c); \end{aligned} \tag{5.1}$$

$$\begin{aligned} s_{ijt} - s_{iat} &= m_{it}(\alpha_j - \alpha_a) + (x_{ijt} - x_{iat})'\beta + \delta_{ij} - \delta_{ia} + u_{ijt} - u_{iat} \\ &= w'_{ijt}\gamma_j + \delta_{ij} - \delta_{ia} + u_{ijt} - u_{iat}. \end{aligned} \quad (5.2)$$

Assume that  $(u_{ia1}, u_{ib1}, u_{ic1}, u_{ia2}, u_{ib2}, u_{ic2})$  are independent given

$$W_i \equiv (m_{i1}, m_{i2}, x_{ij1}, x_{ij2}, j = a, b, c) \text{ and } \delta_{ia}, \delta_{ib}, \delta_{ic},$$

which is equivalent to  $Y_i = (y_{ia1}, y_{ib1}, y_{ic1}, y_{ia2}, y_{ib2}, y_{ic2})'$  being independent given these variables. Although the independence is restrictive, still relations are permitted through the conditioning variables in this display.

Omitting  $i$ , assume a multinomial logit model

$$P(y_{jt} = 1 | W, \delta_a, \delta_b, \delta_c) = \frac{\exp(m'_t\alpha_j + x'_{jt}\beta + \delta_j)}{\sum_{j=a,b,c} \exp(m'_t\alpha_j + x'_{jt}\beta + \delta_j)} \quad (5.3)$$

which follows from  $(u_{a1}, u_{b1}, u_{c1}, u_{a2}, u_{b2}, u_{c2})$  iid with the Type-1 extreme value distribution (i.e.,  $P(u_{jt} \leq q) = \exp\{-\exp(-q)\}$ ). After the normalization by alternative  $a$  (i.e., dividing the numerator and denominator by  $\exp(m'_t\alpha_a + x'_{at}\beta + \delta_a)$ ), (5.3) becomes (recall (5.2))

$$P(y_{jt} = 1 | W, \delta_a, \delta_b, \delta_c) = \frac{\exp(w'_{jt}\gamma_j + \delta_j - \delta_a)}{\sum_j \exp(w'_{jt}\gamma_j + \delta_j - \delta_a)}; \text{ note } \exp(w'_{at}\gamma_a + \delta_a - \delta_a) = 1.$$

The likelihood function for  $Y = (y_{a1}, y_{b1}, y_{c1}, y_{a2}, y_{b2}, y_c)'$  is (note  $\sum_j y_{jt} = 1 \forall t$ )

$$\begin{aligned} P(Y | W, \delta_a, \delta_b, \delta_c) &= \prod_j \left( \frac{\exp(w'_{j1}\gamma_j + \delta_j - \delta_a)}{\sum_j \exp(w'_{j1}\gamma_j + \delta_j - \delta_a)} \right)^{y_{j1}} \left( \frac{\exp(w'_{j2}\gamma_j + \delta_j - \delta_a)}{\sum_j \exp(w'_{j2}\gamma_j + \delta_j - \delta_a)} \right)^{y_{j2}} \\ &= \frac{\exp \left\{ \sum_j y_{j1}(w'_{j1}\gamma_j + \delta_j - \delta_a) + \sum_j y_{j2}(w'_{j2}\gamma_j + \delta_j - \delta_a) \right\}}{\sum_j \exp(w'_{j1}\gamma_j + \delta_j - \delta_a) \cdot \sum_j \exp(w'_{j2}\gamma_j + \delta_j - \delta_a)} \\ &= \frac{\exp \left\{ \sum_j y_{j1}w'_{j1}\gamma_j + \sum_j y_{j2}w'_{j2}\gamma_j + \sum_j (y_{j1} + y_{j2})(\delta_j - \delta_a) \right\}}{\sum_j \exp(w'_{j1}\gamma_j + \delta_j - \delta_a) \cdot \sum_j \exp(w'_{j2}\gamma_j + \delta_j - \delta_a)}. \end{aligned} \quad (5.4)$$

Hence  $y_{j1} + y_{j2}, j = a, b, c$ , are candidates to condition on to remove  $\delta_j - \delta_a, j = a, b, c$ .

Observe

$$\begin{aligned} P(y_{j1} + y_{j2}, j = a, b, c | W, \delta_a, \delta_b, \delta_c) &= \sum_{\bar{\lambda}_j = \bar{y}_j \forall j} \frac{\exp \left\{ \sum_j \lambda_{j1}w'_{j1}\gamma_j + \sum_j \lambda_{j2}w'_{j2}\gamma_j + \sum_j (\lambda_{j1} + \lambda_{j2})(\delta_j - \delta_a) \right\}}{\sum_j \exp(w'_{j1}\gamma_j + \delta_j - \delta_a) \cdot \sum_j \exp(w'_{j2}\gamma_j + \delta_j - \delta_a)} \\ &= \frac{\exp \left\{ \sum_j (y_{j1} + y_{j2})(\delta_j - \delta_a) \right\} \cdot \sum_{\bar{\lambda}_j = \bar{y}_j \forall j} \exp \left( \sum_j \lambda_{j1}w'_{j1}\gamma_j + \sum_j \lambda_{j2}w'_{j2}\gamma_j \right)}{\sum_j \exp(w'_{j1}\gamma_j + \delta_j - \delta_a) \cdot \sum_j \exp(w'_{j2}\gamma_j + \delta_j - \delta_a)} \end{aligned} \quad (5.5)$$

where  $\sum_{\bar{\lambda}_j=\bar{y}_j} \forall j$  stands for the sum over the sequences  $(\lambda_{a1}, \lambda_{b1}, \lambda_{c1}, \lambda_{a2}, \lambda_{b2}, \lambda_{c2})'$  such that  $\lambda_{jt} = 0, 1 \forall j$  and  $t$ ,  $\sum_j \lambda_{jt} = 1 \forall t$ , and  $\lambda_{j1} + \lambda_{j2} = y_{j1} + y_{j2} \forall j$ .

Hence, dividing (5.4) by (5.5) renders

$$\begin{aligned} P(Y | y_{j1} + y_{j2}, \delta_j, j = a, b, c, W) \\ = \frac{\exp\left(\sum_j y_{j1} w'_{j1} \gamma_j + \sum_j y_{j2} w'_{j2} \gamma_j\right)}{\sum_{\bar{\lambda}_j=\bar{y}_j} \exp\left(\sum_j \lambda_{j1} w'_{j1} \gamma_j + \sum_j \lambda_{j2} w'_{j2} \gamma_j\right)} \end{aligned}$$

and the sample maximand is

$$\begin{aligned} \sum_i \ln \left\{ \frac{\exp\left(\sum_j y_{ij1} w'_{ij1} \gamma_j + \sum_j y_{ij2} w'_{ij2} \gamma_j\right)}{\sum_{\bar{\lambda}_j=\bar{y}_j} \exp\left(\sum_j \lambda_{j1} w'_{ij1} \gamma_j + \sum_j \lambda_{j2} w'_{ij2} \gamma_j\right)} \right\} \\ = \sum_i \ln \left\{ \frac{\exp\{Y'_i (\gamma' w_{i1}, \gamma' w_{i2})'\}}{\sum_{\bar{\lambda}_j=\bar{y}_j} \exp\{\Lambda'(\gamma' w_{i1}, \gamma' w_{i2})'\}} \right\}. \end{aligned} \quad (5.6)$$

The values that  $y_{j1} + y_{j2}, j = a, b, c$  can take jointly are restricted: e.g., if  $y_{a1} + y_{a2} = 1$  and  $y_{b1} + y_{b2} = 1$ , then necessarily  $y_{c1} + y_{c2} = 0$ .

Even for the simplest case with three alternatives and two waves, the likelihood function (5.6) is hard to vision. Hence we will take a detailed look in the following. As will be seen in (5.7) below, it is convenient to group the observations depending on the type of moves in  $Y$  and construct the sample maximand consisting of the types of moves.

Consider three types of alternative changes across two periods:

$$(a, b) \text{ or } (b, a), \quad (a, c) \text{ or } (c, a), \quad \text{and} \quad (b, c) \text{ or } (c, b).$$

We examine  $(a, b)$  or  $(b, a)$  first. In terms of  $Y$ ,  $(a, b)$  and  $(b, a)$  are, respectively,

$$(1, 0, 0, \quad 0, 1, 0) \quad \text{and} \quad (0, 1, 0, \quad 1, 0, 0).$$

Before proceeding further, define  $y_{it}$  as ( $y_{it}$  is not a dummy)

$$y_{it} = j \quad \text{if } i \text{ chooses } j \text{ at time } t \quad (\text{i.e., } y_{it} = \sum_j j \times y_{ijt}).$$

Omitting the conditioning variables and recalling the  $1 \times 3$  vector  $\gamma' w_t = (0, w'_{bt} \gamma_b, w'_{ct} \gamma_c)$  in (5.1), the probability of observing  $(a, b)$  given  $(a, b)$  or  $(b, a)$  can be found with (5.6):

$$\begin{aligned} & P\{y_1 = a, y_2 = b | (y_1 = a, y_2 = b) \text{ or } (y_1 = b, y_2 = a)\} \\ &= \frac{\exp\{(1, 0, 0, 0, 1, 0) \cdot (\gamma' w_1, \gamma' w_2)'\}}{\exp\{(1, 0, 0, 0, 1, 0) \cdot (\gamma' w_1, \gamma' w_2)'\} + \exp\{(0, 1, 0, 1, 0, 0) \cdot (\gamma' w_1, \gamma' w_2)'\}} \\ &= \frac{\exp(w'_{b2} \gamma_b)}{\exp(w'_{b2} \gamma_b) + \exp(w'_{b1} \gamma_b)} \equiv P_{ab} \end{aligned}$$

as (5.1) gives  $(1, 0, 0)(\gamma' w_t)' = w_{at}' \gamma_a = 0$ . Also,

$$P\{y_1 = b, y_2 = a \mid (y_1 = a, y_2 = b) \text{ or } (y_1 = b, y_2 = a)\} = 1 - P_{ab} \equiv P_{ba}.$$

Doing analogously for  $(a, c)$  or  $(c, a)$ , we get

$$\begin{aligned} & P\{y_1 = a, y_2 = c \mid (y_1 = a, y_2 = c) \text{ or } (y_1 = c, y_2 = a)\} \\ &= \frac{\exp\{(1, 0, 0, 0, 0, 1) \cdot (\gamma' w_1, \gamma' w_2)'\}}{\exp\{(1, 0, 0, 0, 0, 1) \cdot (\gamma' w_1, \gamma' w_2)'\} + \exp\{(0, 0, 1, 1, 0, 0) \cdot (\gamma' w_1, \gamma' w_2)'\}} \\ &= \frac{\exp(w_{c2}' \gamma_c)}{\exp(w_{c2}' \gamma_c) + \exp(w_{c1}' \gamma_c)} \equiv P_{ac}; \\ & P\{y_1 = c, y_2 = a \mid (y_1 = a, y_2 = c) \text{ or } (y_1 = c, y_2 = a)\} = 1 - P_{ac} \equiv P_{ca}. \end{aligned}$$

As for  $(b, c)$  or  $(c, b)$ ,

$$\begin{aligned} & P\{y_1 = b, y_2 = c \mid (y_1 = b, y_2 = c) \text{ or } (y_1 = c, y_2 = b)\} \\ &= \frac{\exp\{(0, 1, 0, 0, 0, 1) \cdot (\gamma' w_1, \gamma' w_2)'\}}{\exp\{(0, 1, 0, 0, 0, 1) \cdot (\gamma' w_1, \gamma' w_2)'\} + \exp\{(0, 0, 1, 0, 1, 0) \cdot (\gamma' w_1, \gamma' w_2)'\}} \\ &= \frac{\exp(w_{b1}' \gamma_b + w_{c2}' \gamma_c)}{\exp(w_{b1}' \gamma_b + w_{c2}' \gamma_c) + \exp(w_{c1}' \gamma_c + w_{b2}' \gamma_b)} \equiv P_{bc}; \\ & P\{y_1 = c, y_2 = b \mid (y_1 = b, y_2 = c) \text{ or } (y_1 = c, y_2 = b)\} = 1 - P_{bc} \equiv P_{cb}. \end{aligned}$$

Therefore, the sample maximand is

$$\begin{aligned} & \sum_i \{y_{ia1} y_{ib2} \ln P_{ab} + y_{ib1} y_{ia2} \ln P_{ba} + y_{ia1} y_{ic2} \ln P_{ac} \\ & \quad + y_{ic1} y_{ia2} \ln P_{ca} + y_{ib1} y_{ic2} \ln P_{bc} + y_{ic1} y_{ib2} \ln P_{cb}\} \end{aligned} \tag{5.7}$$

with the identified parameters  $\zeta \equiv (\alpha'_b - \alpha'_a, \alpha'_c - \alpha'_a, \beta')'$ .

### 7.5.2 General Cases

For a general  $T$ , use  $1, 2, \dots, J$  to list the alternatives, not  $a, b, c$ . Let

$$Y_i \equiv (y_{i11}, \dots, y_{ij1}, \dots, y_{iT1}, \dots, y_{iTJ})'_{JT \times 1}.$$

The identified parameters are

$$\zeta \equiv (\alpha'_2 - \alpha'_1, \dots, \alpha'_J - \alpha'_1, \beta')'.$$

Define

$$w_{ijt} \equiv \begin{bmatrix} w_{ijt} \\ (k_m + k_x) \times 1 \end{bmatrix} \equiv \begin{bmatrix} m_{it} \\ x_{ijt} - x_{i1t} \end{bmatrix}, \quad \begin{bmatrix} \gamma_j \\ (k_m + k_x) \times 1 \end{bmatrix} \equiv \begin{bmatrix} \alpha_j - \alpha_1 \\ \beta \end{bmatrix}, \quad \forall j,$$

$$\begin{matrix} w_{it} \\ J(k_m+k_x) \times J \end{matrix} \equiv \text{diag}(w_{i1t}, \dots, w_{iJt}), \quad \begin{matrix} \gamma \\ J(k_m+k_x) \times 1 \end{matrix} \equiv (\gamma'_1, \dots, \gamma'_J)'.$$

From this, we get

$$\begin{matrix} \gamma' w_{it} \\ 1 \times J \end{matrix} = (0, w'_{i2t}\gamma_2, \dots, w'_{iJt}\gamma_J) \quad \text{as } w'_{i1t}\gamma_1 = 0 \forall t.$$

The conditional log-likelihood generalizing (5.6) is

$$\frac{\exp(\sum_t \sum_j y_{jt} w'_{jt}\gamma_j)}{\sum_{\bar{\lambda}_j=\bar{y}_j \forall j} \exp(\sum_t \sum_j \lambda_{jt} w'_{jt}\gamma_j)} = \frac{\exp\{Y'(\gamma' w_1, \gamma' w_2, \dots, \gamma' w_T)'\}}{\sum_{\bar{\lambda}_j=\bar{y}_j \forall j} \exp\{\Lambda'(\gamma' w_1, \gamma' w_2, \dots, \gamma' w_T)'\}} \quad (5.8)$$

where  $\sum_{\bar{\lambda}_j=\bar{y}_j \forall j}$  stands for the sum over the sequences  $\Lambda=(\lambda_{11}, \dots, \lambda_{J1}, \dots, \lambda_{1T}, \dots, \lambda_{JT})'$  such that  $\lambda_{jt} = 0, 1 \forall j$  and  $t$ ,  $\sum_j \lambda_{jt} = 1 \forall t$  and  $\sum_t \lambda_{jt} = \sum_t y_{jt} \forall j$ ; the sufficient statistics for  $\delta_j - \delta_1$ ,  $j = 2, \dots, J$ , are  $\sum_t y_{1t}, \dots, \sum_t y_{Jt}$ . The non-informative observations are those with  $\sum_t y_{jt} = T$  for some  $j$  (i.e., sticking to alternative  $j$  throughout the entire periods).

To better understand (5.8), consider  $J = 4$  and  $T = 2$  that is “one alternative” more complicated than the simplest case  $J = 3$  and  $T = 2$ . With four alternatives  $(a, b, c, d)$  and  $T = 2$ —we revert back to alphabets instead of  $(1, 2, 3, 4)$  to avoid confusion—consider 6 =  $\binom{4}{2}$  groups of observations that differ on the values that  $\sum_t y_{jt}$ ,  $j = a, b, c, d$ , can take over two periods subject to  $\sum_j \sum_t y_{jt} = 2$  (excluding the non-informative cases  $\sum_t y_{jt} = 2$  for some  $j$ ):

$$\sum_t y_{at} = 0, \sum_t y_{bt} = 0, \sum_t y_{ct} = 1, \sum_t y_{dt} = 1 : (c, d) \text{ or } (d, c);$$

$$\sum_t y_{at} = 0, \sum_t y_{bt} = 1, \sum_t y_{ct} = 0, \sum_t y_{dt} = 1 : (b, d) \text{ or } (d, b);$$

$$\sum_t y_{at} = 0, \sum_t y_{bt} = 1, \sum_t y_{ct} = 1, \sum_t y_{dt} = 0 : (b, c) \text{ or } (c, b);$$

$$\sum_t y_{at} = 1, \sum_t y_{bt} = 0, \sum_t y_{ct} = 0, \sum_t y_{dt} = 1 : (a, d) \text{ or } (d, a);$$

$$\sum_t y_{at} = 1, \sum_t y_{bt} = 0, \sum_t y_{ct} = 1, \sum_t y_{dt} = 0 : (a, c) \text{ or } (c, a);$$

$$\sum_t y_{at} = 1, \sum_t y_{bt} = 1, \sum_t y_{ct} = 0, \sum_t y_{dt} = 0 : (a, b) \text{ or } (b, a).$$

For instance, if subject  $i$  has  $Y_i = (0, 0, 1, 0, 0, 0, 0, 1)'$  (choosing  $c$  at  $t = 1$  and  $d$  at  $t = 2$ ), then he/she falls in the first group, and his/her likelihood contribution is

$$\frac{\exp\{(0, 0, 1, 0, 0, 0, 0, 1)(\gamma' w_{i1}, \gamma' w_{i2})'\}}{\exp\{(0, 0, 1, 0, 0, 0, 0, 1)(\gamma' w_{i1}, \gamma' w_{i2})'\} + \exp\{(0, 0, 0, 1, 0, 0, 1, 0)(\gamma' w_{i1}, \gamma' w_{i2})'\}}.$$

Now suppose  $J = 3$  with alternatives  $(a, b, c)$  and  $T = 3$ . In this case, we should consider all possibilities for  $\sum_t y_{jt}$ ,  $j = a, b, c$ , over three periods subject to  $\sum_j \sum_t y_{jt} = 3$  (except the non-informative cases  $\sum_t y_{jt} = 3$  for some  $j$ ):

$$\sum_t y_{at} = 0, \sum_t y_{bt} = 1, \sum_t y_{ct} = 2 : (b, c, c), (c, b, c), (c, c, b);$$

$$\sum_t y_{at} = 0, \sum_t y_{bt} = 2, \sum_t y_{ct} = 1 : (c, b, b), (b, c, b), (b, b, c);$$

$$\sum_t y_{at} = 1, \sum_t y_{bt} = 0, \sum_t y_{ct} = 2 : (a, c, c), (c, a, c), (c, c, a);$$

$$\sum_t y_{at} = 1, \sum_t y_{bt} = 1, \sum_t y_{ct} = 1 : (a, b, c), (a, c, b), (b, a, c), (b, c, a), (c, a, b), (c, b, a);$$

$$\sum_t y_{at} = 1, \sum_t y_{bt} = 2, \sum_t y_{ct} = 0 : (a, b, b), (b, a, b), (b, b, a);$$

$$\sum_t y_{at} = 2, \sum_t y_{bt} = 0, \sum_t y_{ct} = 1 : (a, a, c), (a, c, a), (c, a, a);$$

$$\sum_t y_{at} = 2, \sum_t y_{bt} = 1, \sum_t y_{ct} = 0 : (a, a, b), (a, b, a), (b, a, a).$$

For instance, if subject  $i$  has  $Y_i = (0, 1, 0, 0, 0, 1, 0, 0, 1)'$  (choosing  $b$  at  $t = 1$  and  $c$  at  $t = 2, 3$ ), then he/she falls in the first group, and his/her likelihood contribution is

$$\begin{aligned} & \exp\{(0, 1, 0, 0, 0, 1, 0, 0, 1)(\gamma' w_{i1}, \gamma' w_{i2}, \gamma' w_{i3})'\}/S \quad \text{where} \\ & S \equiv \exp\{(0, 1, 0, 0, 0, 1, 0, 0, 1)(\gamma' w_{i1}, \gamma' w_{i2}, \gamma' w_{i3})'\} \\ & + \exp\{(0, 0, 1, 0, 1, 0, 0, 0, 1)(\gamma' w_{i1}, \gamma' w_{i2}, \gamma' w_{i3})'\} \\ & + \exp\{(0, 0, 1, 0, 0, 1, 0, 1, 0)(\gamma' w_{i1}, \gamma' w_{i2}, \gamma' w_{i3})'\}. \end{aligned}$$

### 7.5.3 PCML Model Variations and Applied Studies

Recall  $s_{ijt} = m'_{it}\alpha_j + x'_{ijt}\beta + \delta_{ij} + u_{ijt}$  that we have been using so far. In the literature of PCML, variations of this model, some simpler and some more complicated, have been used, and we review those in the following.

One application that used the same  $s_{ijt}$  model is Börsch-Supan (1990) with four housing alternatives: renting/owning a small/large residential unit. He used 880 families from PSID over 1977–1981 ( $T = 5$ ), and

$$m_{it} = (\text{age}_{it}, \text{age}_{it}^2, \text{income}_{it}, \text{family-size}_{it}, \text{work}_{it})' \quad \text{and} \quad x_{ijt} = \text{price}_{ijt};$$

$\delta_{ij}$  can be an individual characteristic such as being claustrophobic for a small unit. In multinomial choices, getting information on the alternative-varying variables for

the non-chosen alternatives is hard, and Börsch-Supan (1990) imputed the price information.

HK provided a dynamic PCML with  $y_{it}$  depending on  $y_{i,t-1}$  by an intercept shift  $\gamma_{hj}$ :

$$P(y_{it} = j | y_{i,t-1} = h, \delta\text{'s all current \& past regressors}) = \frac{\exp(x'_{ijt}\beta_j + \gamma_{hj} + \delta_{ij})}{\sum_{j'} \exp(x'_{ij't}\beta_{j'} + \gamma_{hj'} + \delta_{ij'})}$$

where  $\beta_j$ 's and  $\gamma_{hj}$ 's are to be estimated under the restriction that the regressors in the periods  $s+1$  and  $t+1$  be the same when the observations with  $y_{is} + y_{it} = 1$  are used. HK considered only  $x_{ijt}$ , not  $m_{it}$ ; although  $x_{ijt}$  can be thought of as including  $m_{it}$ , it is better to separate them, as they can carry different types of parameters as in  $m'_{it}\alpha_j + x'_{ijt}\beta_j$ .

The dynamic PCML of HK was applied by Chintagunta et al. (2001) for three yogurt brand choice as part of a sensitivity analysis for their main binary choice problem. They used only three regressors (all alternative-variant): brand price, whether the brand was displayed or not and whether the brand was advertised or not. Chintagunta et al. used

$$s_{ijt} = x'_{ijt}\beta_j + \gamma_j 1[y_{i,t-1} = j] + \delta_{ij} + u_{ijt}$$

to find that  $\gamma_j$  changes much as  $j$  changes.

Egger and et al. (2007) considered  $\beta_{hj}$  in finding the effects of trade and outsourcing on employment in different sectors, where the preceding state is  $h$ ; i.e., the effect of  $x_{ijt}$  on  $s_{ijt}$  depends on the previous choice  $h$  as well as on the current choice  $j$ . Egger and et al. (2007) used both  $x_{ijt}$  and  $m_{it}$ , but this generalization was done only for  $x_{ijt}$  and the intercept:

$$\begin{aligned} P(y_{it} = j | y_{i,t-1} = h, \delta\text{'s, all current and past regressors}) \\ = \frac{\exp(m'_{it}\alpha_j + x'_{ijt}\beta_{hj} + \gamma_{hj} + \delta_{ij})}{\sum_{j'} \exp(m'_{it}\alpha_{j'} + x'_{ij't}\beta_{hj'} + \gamma_{hj'} + \delta_{ij'})}. \end{aligned}$$

## 7.6 CONCLUSIONS

This chapter reviewed panel conditional logit estimators (PCLE). Although the so-called panel ‘random effect’ estimators with an individual-specific effect unrelated to regressors are as popular as PCLE in practice, PCLE has the main advantage of allowing for an arbitrary relation between the individual-specific effect and regressors. Also, PCLE converges well in computation unlike many other estimators. Hence, despite its shortcomings such as not being able to identify the parameters for time-constant regressors and the difficulty in obtaining the marginal effect, PCLE will remain an important and widely used estimator in econometrics.

## ACKNOWLEDGMENTS

---

I am grateful to the comments by Jin-young Choi, an anonymous reviewer, and the seminar participants of the Melbourne Institute, University of Auckland, University of New South Wales, University of Adelaide and the Institute of Economics of Academia Sinica.

## REFERENCES

---

- Anderson, E.B., 1970, Asymptotic properties of conditional maximum likelihood estimators, *Journal of the Royal Statistical Society (Series B)* 32, 283–301.
- Arellano, M. and B. Honoré, 2001, Panel data models: some recent developments, in *Handbook of Econometrics* 5, edited by J.J. Heckman and E. Leamer, North-Holland.
- Baltagi, B.H., 2008, *Econometric analysis of panel data*, 4th ed., Wiley.
- Bartolucci, F. and V. Nigro, 2010, A dynamic model for binary panel data with unobserved heterogeneity admitting a  $\sqrt{n}$ -consistent conditional estimator, *Econometrica* 78, 719–733.
- Bartolucci, F. and V. Nigro, 2012, Pseudo conditional maximum likelihood estimation of the dynamic logit model for binary panel data, *Journal of Econometrics* 170, 102–116.
- Biewen, M., 2009, Measuring state dependence in individual poverty histories when there is feedback to employment status and household composition, *Journal of Applied Econometrics* 24, 1095–1116.
- Börsch-Supan, A., 1990, Panel data analysis of housing choices, *Regional Science and Urban Economics* 20, 65–82.
- Bushway, S., R. Brame and R. Paternoster, 1999, Assessing stability and change in criminal offending: a comparison of random effects, semiparametric, and fixed effects modeling strategies, *Journal of Quantitative Criminology* 15, 23–61.
- Chamberlain, G., 1980, Analysis of covariance with qualitative data, *Review of Economic Studies* 47, 225–238.
- Chamberlain, G., 1984, Panel data, in *Handbook of Econometrics* 2, edited by Z. Griliches and M. Intriligator, North-Holland.
- Chamberlain, G., 1985, Heterogeneity, omitted variable bias and duration dependence, in *Longitudinal Analyses of Labor Market Data*, edited by J.J. Heckman and B. Singer, Academic Press.
- Chamberlain, G., 1992, Comment: Sequential moment restrictions in panel data, *Journal of Business and Economic Statistics* 10, 20–26.
- Chamberlain, G., 2010, Binary response models for panel data: identification and information, *Econometrica* 78, 159–168.
- Chintagunta, P., E. Kyriazidou and J. Perktold, 2001, Panel data analysis of household brand choices, *Journal of Econometrics* 103, 111–153.
- Corcoran, M. and M.S. Hill, 1985, Reoccurrence of unemployment among young adult men, *Journal of Human Resources* 20, 165–183.

- Cox, D.R., 1972, Regression models and life tables, *Journal of the Royal Statistical Society (Series B)* 34, 187–220.
- Egger, P., M. Pfaffermayr and A. Weber, 2007, Sectoral adjustment of employment to shifts in outsourcing and trade: evidence from a dynamic fixed effects multinomial logit model, *Journal of Applied Econometrics* 22, 559–580.
- Halliday, T.J., 2007, Testing for state dependence with time-variant transition probabilities, *Econometric Reviews* 26, 685–703.
- Hoderlein, S., E. Mammen and K. Yu, 2011, Non-parametric models in binary choice fixed effects panel data, *Econometrics Journal* 14, 351–367.
- Honoré, B.E. and E. Kyriazidou, 2000, Panel data discrete choice models with lagged dependent variables, *Econometrica* 68, 839–874.
- Howard, S., 1972, Remark on the paper by D. R. Cox ‘Regression models and life-tables’, *Journal of the Royal Statistical Society (Series B)* 34, 187–220.
- Hsiao, C., 2003, Analysis of panel data, 2nd ed., Cambridge University Press.
- Kim, Joo Hyung, 1988, A method of moments estimator to circumvent the incidental parameters problem in short panels, Ph.D. Thesis, University of Wisconsin-Madison.
- Krailo, M.D. and M. C. Pike, 1984, Algorithm AS 196: conditional multivariate logistic analysis of stratified case-control studies, *Journal of the Royal Statistical Society Series C (Applied Statistics)* 33, 95–103.
- Lee, M.J., 2002, Panel data econometrics: methods-of-moments and limited dependent variables, Academic Press.
- Lee, M.J., 2010, Micro-econometrics: methods of moments and limited dependent variables, Springer.
- Lee, M.J., 2012, Semiparametric estimators for limited dependent variable (LDV) models with endogenous regressors, *Econometric Reviews* 31, 171–214.
- Lee, M.J., 2014, Panel conditional and multinomial logit with time-varying parameters, unpublished paper.
- Lee, M.J. and A. Kimhi, 2005, Simultaneous equations in ordered discrete responses with regressor-dependent thresholds, *Econometrics Journal* 8, 176–196.
- Lee, M.J. and Y.H. Tae, 2005, Analysis of labor-participation behavior of Korean women with dynamic probit and conditional logit, *Oxford Bulletin of Economics and Statistics* 67, 71–91.
- Magnac, T., 2000, Subsidized training and youth employment: distinguishing unobserved heterogeneity from state dependence in labor market histories, *Economic Journal* 110, 805–837.
- Magnac, T., 2004, Panel binary variables and sufficiency: generalizing conditional logit, *Econometrica* 72, 1859–1876.
- Rasch, G., 1961, On general law and the meaning of measurement in psychology, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* 4, 321–333.
- Thomas, A., 2006, Consistent estimation of binary-choice panel data models with heterogeneous linear trends, *Econometrics Journal* 9, 177–195.
- Wooldridge, J.M., 1997, Multiplicative panel data models without the strict exogeneity assumption, *Econometric Theory* 13, 667–678.

## CHAPTER 8

---

# COUNT PANEL DATA

---

A. COLIN CAMERON AND PRAVIN K. TRIVEDI

### 8.1 INTRODUCTION

---

THIS chapter surveys panel data methods for a count dependent variable that takes nonnegative integer values, such as number of doctor visits. The focus is on short panels, with  $T$  fixed and  $n \rightarrow \infty$ , as the literature has concentrated on this case.

The simplest panel models specify the conditional mean to be of exponential form, and specify the conditional distribution to be Poisson or, in some settings, a particular variant of the negative binomial. Then it can be possible to consistently estimate slope parameters provided only that the conditional mean is correctly specified, and to obtain standard errors that are robust to possible misspecification of the distribution. This is directly analogous to panel linear regression under normality where consistent estimation and robust inference are possible under much weaker assumptions than normality. In particular, it is possible to consistently estimate the slope parameters in a fixed effects version of the Poisson model, even in a short panel.

Richer models account for special features of count data. In particular, the Poisson is inadequate in modelling the conditional distribution as it is a one parameter distribution that imposes variance-mean equality. In most applications the conditional variance exceeds the conditional mean. Richer parametric models are negative binomial models and finite mixture models. Furthermore, even for a given parametric model there can be a bunching or excess of zeros, leading to modified count models—hurdle models and with-zeros models. These considerations are especially important for applications that need to model the conditional distribution, not just the conditional mean. For example, interest may lie in predicting the probability of an excessive number of doctor visits.

Section 8.2 briefly reviews standard cross-section models for count data, the building block for section 8.3 that presents standard static models for panel counts with focus on short panels. Section 8.4 presents extension to the dynamic case, where the

current count depends on lagged realizations of the count. Again the emphasis is on short panels, and the Arellano-Bond estimator for linear dynamic models with fixed effects can be adapted to count data. Section 8.5 considers extensions that address more complicated features of count data.

## 8.2 MODELS FOR CROSS-SECTION COUNT DATA

The main cross-section data models for counts are the Poisson and negative binomial models, hurdle and zero-inflated variants of these models, and latent class or finite mixture models.

### 8.2.1 Poisson Quasi-MLE

The Poisson regression model specifies that  $y_i$  given  $\mathbf{x}_i$  is Poisson distributed with density

$$f(y_i|\mathbf{x}_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \dots \quad (1)$$

and mean parameter

$$\mathbb{E}[y_i|\mathbf{x}_i] = \mu_i = \exp(\mathbf{x}'_i \beta). \quad (2)$$

The exponential form in (2) ensures that  $\mu_i > 0$ . It also permits  $\beta$  to be interpreted as a semi-elasticity, since  $\beta = [\partial \mathbb{E}[y_i|\mathbf{x}_i]/\partial \mathbf{x}_i]/\mathbb{E}[y_i|\mathbf{x}_i]$ . In the statistics literature the model is often called a log-linear model, since the logarithm of the conditional mean is linear in the parameters:  $\ln \mathbb{E}[y_i|\mathbf{x}_i] = \mathbf{x}'_i \beta$ .

Given independent observations, the log-likelihood is  $\ln L(\beta) = \sum_{i=1}^n \{y_i \mathbf{x}'_i \beta - \exp(\mathbf{x}'_i \beta) - \ln y_i\}$ . The Poisson MLE  $\hat{\beta}_P$  solves the first-order conditions

$$\sum_{i=1}^n (y_i - \exp(\mathbf{x}'_i \beta)) \mathbf{x}_i = \mathbf{0}. \quad (3)$$

These first-order conditions imply that the essential condition for consistency of the Poisson MLE is that  $\mathbb{E}[y_i|\mathbf{x}_i] = \exp(\mathbf{x}'_i \beta)$ , i.e., that the conditional mean is correctly specified—the data need not be Poisson distributed.

The Poisson quasi-MLE is then asymptotically normally distributed with mean  $\beta$  and variance-covariance matrix

$$\mathbb{V}[\hat{\beta}_P] = \left( \sum_{i=1}^n \mu_i \mathbf{x}_i \mathbf{x}'_i \right)^{-1} \left( \sum_{i=1}^n \mathbb{V}[y_i|\mathbf{x}_i] \mathbf{x}_i \mathbf{x}'_i \right) \left( \sum_{i=1}^n \mu_i \mathbf{x}_i \mathbf{x}'_i \right)^{-1}, \quad (4)$$

where  $\mu_i = \exp(\mathbf{x}'_i \beta)$ . This can be consistently estimated using a heteroskedasticity-robust estimate

$$\widehat{V}[\widehat{\beta}_P] = \left( \sum_{i=1}^n \widehat{\mu}_i \mathbf{x}_i \mathbf{x}'_i \right)^{-1} \left( \sum_{i=1}^n (y_i - \widehat{\mu}_i)^2 \mathbf{x}_i \mathbf{x}'_i \right) \left( \sum_{i=1}^n \widehat{\mu}_i \mathbf{x}_i \mathbf{x}'_i \right)^{-1}. \quad (5)$$

A property of the Poisson distribution is that the variance equals the mean. Then  $V[y_i | \mathbf{x}_i] = \mu_i$ , so (4) simplifies to  $V[\widehat{\beta}_P] = (\sum_{i=1}^n \mu_i \mathbf{x}_i \mathbf{x}'_i)^{-1}$ . In practice for most count data, the conditional variance exceeds the conditional mean, a feature called overdispersion. Then using standard errors based on  $V[\widehat{\beta}_P] = (\sum_{i=1}^n \mu_i \mathbf{x}_i \mathbf{x}'_i)^{-1}$ , the default in most Poisson regression packages, can greatly underestimate the true standard errors; one should use (5).

The robustness of the Poisson quasi-MLE to distributional misspecification, provided the conditional mean is correctly specified, means that Poisson regression can also be applied to continuous nonnegative data. In particular, OLS regression of  $\ln y$  on  $\mathbf{x}$  cannot be performed if  $y = 0$  and leads to a retransformation problem if we wish to predict  $y$ . Poisson regression of  $y$  on  $\mathbf{x}$  (with exponential conditional mean) does not have these problems.

The robustness to distributional misspecification is shared with the linear regression model with assumed normal errors. More generally, this holds for models with specified density in the linear exponential family, i.e.,  $f(y|\mu) = \exp\{a(\mu) + b(y) + c(\mu)y\}$ . In the statistics literature this class is known as generalized linear models (GLM). It includes the normal, Poisson, geometric, gamma, Bernoulli, and binomial.

## 8.2.2 Parametric Models

In practice, the Poisson distribution is too limited as it is a one-parameter distribution, depending on only the mean  $\mu$ . In particular, the distribution restricts the variance to equal the mean, but count data used in economic applications generally are overdispersed.

The standard generalization of the Poisson is the negative binomial (NB) model, most often the NB2 variant that specifies the variance to equal  $\mu + \alpha\mu^2$ . Then

$$f(y_i|\mu_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + 1)\Gamma(\alpha^{-1})} \left( \frac{\alpha^{-1}}{\alpha^{-1} + \mu_i} \right)^{\alpha^{-1}} \times \left( \frac{\mu_i}{\alpha^{-1} + \mu_i} \right)^{y_i}, \quad \alpha > 0, y_i = 0, 1, 2, \dots \quad (6)$$

This reduces to the Poisson for  $\alpha \rightarrow 0$ . Specifying  $\mu_i = \exp(\mathbf{x}'_i \beta)$ , the MLE solves for  $\beta$  and  $\alpha$  the first-order conditions

$$\begin{aligned} \sum_{i=1}^n \frac{y_i - \mu_i}{1 + \alpha \mu_i} \mathbf{x}_i &= \mathbf{0} \\ \sum_{i=1}^n \left\{ \frac{1}{\alpha^2} \left( \ln(1 + \alpha \mu_i) - \sum_{j=0}^{y_i-1} \frac{1}{(j + \alpha^{-1})} \right) + \frac{y_i - \mu_i}{\alpha(1 + \alpha \mu_i)} \right\} &= 0. \end{aligned} \quad (7)$$

As for the Poisson, the NB2 MLE for  $\beta$  is consistent provided  $E[y_i|\mathbf{x}_i] = \exp(\mathbf{x}'_i \beta)$ .

A range of alternative NB models can be generated by specifying  $V[y|\mathbf{x}] = \mu + \alpha \mu^p$ , where  $p$  is specified or is an additional parameter to be estimated. The most common alternative model is the NB1 that sets  $p = 1$ , so the conditional variance is a multiple of the mean. For these variants the quasi-MLE is no longer consistent—the distribution needs to be correctly specified. Yet another variation parameterizes  $\alpha$  to depend on regressors.

The NB models are parameterized to have the same conditional mean as the Poisson. In theory the NB MLE is more efficient than the Poisson QMLE if the NB model is correctly specified, though in practice the efficiency gains are often small. The main reason for using the NB is in settings where the desire is to fit the distribution, not just the conditional mean. For example, interest may lie in predicting the probability of ten or more doctor visits. And a fully parametric model such as the NB may be necessary if the count is incompletely observed, due to truncation, censoring or interval-recording (e.g., counts recorded as 0, 1, 2, 3–5, more than 5).

Both the Poisson and NB models are inadequate if zero counts do not come from the same process as positive counts. Then there are two commonly used modified count models, based on different behavioral models. Let  $f_2(y)$  denote the latent count density. A hurdle or two-part model specifies that positive counts are observed only after a threshold is crossed, with probability  $1 - f_1(0)$ . Then we observe  $f(0) = f_1(0)$  and, for  $y > 0$ ,  $f(y) = f_2(y)(1 - f_1(0))/(1 - f_2(0))$ . A zero-inflated or with-zeros model treats some zero counts as coming from a distinct process due to, for example, never participating in the activity or mismeasurement. In that case  $f_2(0)$ , the probability of zero counts from the baseline density, is inflated by adding a probability of, say,  $\pi$ . Then we have  $f(0) = \pi + (1 - \pi)f_2(0)$  and, for  $y > 0$ ,  $f(y) = (1 - \pi)f_2(y)$ .

A final standard adaptation of cross-section count models is a latent class or finite mixtures model. Then  $y$  is a draw from an additive mixture of  $C$  distinct populations with component (subpopulation) densities  $f_1(y), \dots, f_C(y)$ , in proportions  $\pi_1, \dots, \pi_C$ , where  $\pi_j \geq 0$ ,  $j = 1, \dots, C$ , and  $\sum_{j=1}^C \pi_j = 1$ . The mixture density is then  $f(y) = \sum_{j=1}^C \pi_j f_j(y)$ . Usually the  $\pi_j$  are not parameterized to depend on regressors, the  $f_j(y)$  are Poisson or NB models with regressors, and often  $C = 2$  is adequate.

## 8.3 STATIC PANEL COUNT MODELS

---

The standard methods for linear regression with data from short panels—pooled OLS and FGLS, random effects, and fixed effects—extend to Poisson regression and, to a lesser extent, to NB regression. Discussion of other panel count models is deferred to Section 8.5.

### 8.3.1 Individual Effects in Count Models

Fixed and random effects models for short panels introduce an individual-specific effect. For count models, with conditional mean restricted to be positive, the effect is multiplicative in the conditional mean, rather than additive. Then

$$\mu_{it} \equiv E[y_{it} | \mathbf{x}_{it}, \alpha_i] = \alpha_i \lambda_{it} = \alpha_i \exp(\mathbf{x}'_{it} \beta), \quad i = 1, \dots, n, \quad t = 1, \dots, T, \quad (8)$$

where the last equality specifies an exponential functional form. Note that the intercept is merged into  $\alpha_i$ , so that now the regressors  $\mathbf{x}_{it}$  do not include an intercept.

In this case the model can also be expressed as

$$\mu_{it} \equiv \exp(\delta_i + \mathbf{x}'_{it} \beta), \quad (9)$$

where  $\delta_i = \ln \alpha_i$ . For the usual case of an exponential conditional mean, the individual-specific effect can be interpreted as either a multiplicative effect or as an intercept shifter. If there is reason to specify a conditional mean that is not of exponential form, then a multiplicative effects model may be specified, with  $\mu_{it} \equiv \alpha_i g(\mathbf{x}'_{it} \beta)$ , or an intercept shift model may be used, with  $\mu_{it} \equiv g(\delta_i + \mathbf{x}'_{it} \beta)$ .

Unlike the linear model, consistent estimation of  $\beta$  here does not identify the marginal effect. The marginal effect given (8) is

$$ME_{itj} \equiv \frac{\partial E[y_{it} | \mathbf{x}_{it}, \alpha_i]}{\partial x_{itj}} = \alpha_i \exp(\mathbf{x}'_{it} \beta) \beta_j = \beta_j E[y_{it} | \mathbf{x}_{it}, \alpha_i], \quad (10)$$

which depends on the unknown  $\alpha_i$ . Instead, the slope coefficient  $\beta_j$  can be interpreted as a semi-elasticity, giving the proportionate increase in  $E[y_{it} | \mathbf{x}_{it}, \alpha_i]$  associated with a one-unit change in  $x_{itj}$ . For example, if  $\beta_j = .06$ , then a one-unit change in  $x_j$  is associated with a 6% increase in  $y_{it}$ , after controlling for both regressors and the unobserved individual effect  $\alpha_i$ .

### 8.3.2 Pooled or Population-Averaged Models

Before estimating models with individual-specific effects, namely fixed and random effects models, we consider pooled regression. A pooled Poisson model bases estimation on the marginal distributions of the individual counts  $y_{it}$ , rather than on the joint distribution of the counts  $y_{i1}, \dots, y_{iT}$  for the  $i^{th}$  individual.

The pooled Poisson QMLE is obtained by standard Poisson regression of  $y_{it}$  on an intercept and  $\mathbf{x}_{it}$ . Define  $\mathbf{z}'_{it} = [1 \ \mathbf{x}'_{it}]$  and  $\gamma' = [\delta \ \beta']$ , so  $\exp(\mathbf{z}'_{it}\gamma) = \exp(\delta + \mathbf{x}'_{it}\beta)$ . Then the first-order conditions are

$$\sum_{i=1}^n \sum_{t=1}^T (y_{it} - \exp(\mathbf{z}'_{it}\gamma)) \mathbf{z}_{it} = \mathbf{0}. \quad (11)$$

The estimator is consistent if

$$E[y_{it}|\mathbf{x}_{it}] = \exp(\delta + \mathbf{x}'_{it}\beta) = \alpha \exp(\mathbf{x}'_{it}\beta), \quad (12)$$

i.e., if the conditional mean is correctly specified. Default standard errors are likely to be incorrect, however, as they assume that  $y_{it}$  is equidispersed and is uncorrelated over time for individual  $i$ . Instead it is standard in short panels to use cluster-robust standard errors, with clustering on the individual, based on the variance matrix estimate

$$\left[ \sum_{i=1}^n \sum_{t=1}^T \widehat{\mu}_{it} \mathbf{z}_{it} \mathbf{z}'_{it} \right]^{-1} \sum_{i=1}^n \sum_{t=1}^T \sum_{s=1}^T \widehat{u}_{it} \widehat{u}_{is} \mathbf{z}_{it} \mathbf{z}'_{is} \left[ \sum_{i=1}^n \sum_{t=1}^T \widehat{\mu}_{it} \mathbf{z}_{it} \mathbf{z}'_{it} \right]^{-1}, \quad (13)$$

where  $\widehat{\mu}_{it} = \exp(\mathbf{z}'_{it}\widehat{\gamma})$ , and  $\widehat{u}_{it} = y_{it} - \exp(\mathbf{z}'_{it}\widehat{\gamma})$ .

The multiplicative effects model (8) for  $E[y_{it}|\mathbf{x}_{it}, \alpha_i]$  leads to condition (12) for  $E[y_{it}|\mathbf{x}_{it}]$  if  $\alpha_i$  is independent of  $\mathbf{x}_{it}$  and  $\alpha = E_{\alpha_i}[\alpha_i]$ . This condition holds in a random effects model, see below, but not in a fixed effects model. The statistics literature refers to the pooled estimator as the population-averaged estimator, since (12) is assumed to hold after averaging out any individual-specific effects. The term "marginal analysis," meaning marginal with respect to  $\alpha_i$ , is also used.

More efficient pooled estimation is possible by making assumptions about the correlation between  $y_{it}$  and  $y_{is}$ ,  $s \neq t$ , conditional on regressors  $\mathbf{X}_i = [\mathbf{x}'_{i1} \cdots \mathbf{x}'_{iT}]'$ . Let  $\mu_i(\gamma) = [\mu_{i1} \cdots \mu_{iT}]'$ , where  $\mu_{it} = \exp(\mathbf{z}'_{it}\gamma)$  and let  $\Sigma_i$  be a model for  $V[y_i|\mathbf{X}_i]$  with  $ts^{th}$  entry  $Cov[y_{it}, y_{is}|\mathbf{X}_i]$ . For example, if we assume data are equicorrelated, so  $Cor[y_{it}, y_{is}|\mathbf{X}_i] = \rho$  for all  $s \neq t$ , and that data are overdispersed with variance  $\sigma_{it}^2$ , then  $\Sigma_{i,ts} \equiv Cov[y_{it}, y_{is}|\mathbf{X}_i] = \rho \sigma_{it} \sigma_{is}$ . An alternative model permits more flexible correlation for the first  $m$  lags, with  $Cor[y_{it}, y_{i,t-k}|\mathbf{X}_i] = \rho_k$ , where  $\rho_k = 0$  for  $|k| > m$ . Such assumptions enable estimation by more efficient feasible nonlinear generalized least squares. The first-order conditions for  $\gamma$  are

$$\sum_{i=1}^n \frac{\partial \mu'_i(\gamma)}{\partial \gamma} \widehat{\Sigma}_i^{-1} (\mathbf{y}_i - \mu_i(\gamma)) = \mathbf{0}, \quad (14)$$

where  $\widehat{\Sigma}_i$  is obtained from initial first-stage pooled Poisson estimation of  $\beta$  and consistent estimation of any other parameters that determine  $\Sigma_i$ .

The statistics literature calls this estimator the Poisson generalized estimating equations (GEE) estimator. The variance model  $\Sigma_i$  is called a working matrix, as it is possible to obtain a cluster-robust estimate of the asymptotic variance matrix robust to misspecification of  $\Sigma_i$ , provided  $n \rightarrow \infty$ . Key references are Zeger and Liang (1986) and Liang and Zeger (1986). Liang, Zeger, and Qaqish (1992) consider generalized GEE estimators that jointly estimate the regression and correlation parameters. Bränäs and Johansson (1996) allow for time-varying random effects  $\alpha_{it}$  and estimation by generalized method of moments (GMM).

The preceding pooled estimators rely on correct specification of the conditional mean  $E[y_{it}|\mathbf{x}_{it}]$ . In richer parametric models, such as a hurdle model or an NB model other than NB2, stronger assumptions are needed for estimator consistency. The log-likelihood for pooled ML estimation is based for individual  $i$  on  $\prod_{t=1}^T f(y_{it}|\mathbf{x}_{it})$ , the product of the marginal densities, rather than the joint density  $f(\mathbf{y}_i|\mathbf{X}_i)$ . Consistent estimation generally requires that the marginal density  $f(y_{it}|\mathbf{x}_{it})$  be correctly specified. Since  $y_{it}$  is in fact correlated over  $t$ , there is an efficiency loss. Furthermore, inference should be based on cluster-robust standard errors, possible given  $n \rightarrow \infty$ .

### 8.3.3 Random Effects Models and Extensions

A random effects (RE) model is an individual effects model with the individual effect  $\alpha_i$  (or  $\delta_i$ ) assumed to be distributed independently of the regressors. Let  $f(y_{it}|\mathbf{x}_{it}, \alpha_i)$  denote the density for the  $it^{th}$  observation, conditional on both  $\alpha_i$  and the regressors. Then the joint density for the  $i^{th}$  observation, conditional on the regressors, is

$$f(\mathbf{y}_i|\mathbf{X}_i) = \int_0^\infty \left[ \prod_{t=1}^T f(y_{it}|\alpha_i, \mathbf{x}_{it}) \right] g(\alpha_i|\eta) d\alpha_i, \quad (15)$$

where  $g(\alpha_i|\eta)$  is the specified density of  $\alpha_i$ . In some special cases there is an explicit solution for the integral (15). Even if there is no explicit solution, Gaussian quadrature numerical methods work well since the integral is only one-dimensional, or estimation can be by maximum simulated likelihood.

The Poisson random effects model is obtained by supposing  $y_{it}$  is Poisson distributed, conditional on  $\mathbf{x}_{it}$  and  $\alpha_i$ , with mean  $\alpha_i \lambda_{it}$ , and additionally that  $\alpha_i$  is gamma distributed with mean 1, a normalization, and variance  $1/\gamma$ . Then, integrating out  $\alpha_i$ , the conditional mean  $E[y_{it}|\mathbf{x}_{it}] = \lambda_{it}$ , the conditional variance  $V[y_{it}|\mathbf{x}_{it}] = \lambda_{it} + \lambda_{it}^2/\gamma$ , and there is a closed form solution to (15), with

$$f(\mathbf{y}_i|\mathbf{X}_i) = \left[ \prod_t \frac{\lambda_{it}^{y_{it}}}{y_{it}!} \right] \times \left( \frac{\gamma}{\sum_t \lambda_{it} + \gamma} \right)^\gamma \\ \times \left( \sum_t \lambda_{it} + \gamma \right)^{-\sum_t y_{it}} \frac{\Gamma(\sum_t y_{it} + \gamma)}{\Gamma(\gamma)}. \quad (16)$$

For exponential conditional mean the ML first-order conditions for  $\hat{\beta}$  are

$$\sum_{i=1}^n \sum_{t=1}^T \mathbf{x}_{it} \left( y_{it} - \lambda_{it} \frac{\bar{y}_i + \gamma/T}{\bar{\lambda}_i + \gamma/T} \right) = 0, \quad (17)$$

where  $\bar{y}_i = \frac{1}{T} \sum_{t=1}^T y_{it}$  and  $\bar{\lambda}_i = \frac{1}{T} \sum_{t=1}^T \lambda_{it}$ . A sufficient condition for consistency is that  $E[y_{it}|\mathbf{X}_i] = \lambda_{it}$ . The model has the same conditional mean as the pooled Poisson, but leads to more efficient estimation if in fact overdispersion is of the NB2 form.

The Poisson random effects model was proposed by Hausman, Hall, and Griliches (1984). They also presented a random effects version of the NB2 model, with  $y_{it}$  specified to be i.i.d. NB2 with parameters  $\alpha_i \lambda_{it}$  and  $\phi_i$ , where  $\lambda_{it} = \exp(\mathbf{x}'_{it}\beta)$ . Conditional on  $\lambda_{it}$ ,  $\alpha_i$ , and  $\phi_i$ ,  $y_{it}$  has mean  $\alpha_i \lambda_{it} / \phi_i$  and variance  $(\alpha_i \lambda_{it} / \phi_i) \times (1 + \alpha_i / \phi_i)$ . A closed form solution to (15) is obtained by assuming that  $(1 + \alpha_i / \phi_i)^{-1}$  is a beta-distributed random variable with parameters  $(a, b)$ .

The preceding examples specify a distribution for  $\alpha_i$  that leads to a closed-form solution to (15). This is analogous to specifying a natural conjugate prior in a Bayesian setting. Such examples are few, and in general there is no closed form solution to (15). Furthermore, the most obvious choice of distribution for the multiplicative effect  $\alpha_i$  is the lognormal, equivalent to assuming that  $\delta_i$  in  $\exp(\delta_i + \mathbf{x}'_{it}\beta)$  is normally distributed. Since there is then no closed form solution to (15), Gaussian quadrature or maximum simulated likelihood methods are used. If  $\delta_i \sim N[\delta, \sigma_\delta^2]$  then  $E[y_{it}|\mathbf{x}_{it}] = \exp(\delta + \sigma_\delta^2/2)\lambda_{it}$ , a rescaling of the conditional mean in the Poisson-gamma random effects model. This is absorbed in the intercept if  $\lambda_{it} = \exp(\mathbf{x}'_{it}\beta)$ .

More generally, slope coefficients in addition to the intercept may vary across individuals. A random coefficients model with exponential conditional mean specifies  $E[y_{it}|\mathbf{x}_{it}, \delta_i, \beta_i] = \exp(\delta_i + \mathbf{x}'_{it}\beta_i)$ . Assuming  $\delta_i \sim N[\delta, \sigma_\delta^2]$  and  $\beta_i \sim N[\beta, \Sigma_\beta]$  implies  $\delta_i + \mathbf{x}'_{it}\beta_i \sim N[\delta + \mathbf{x}'_{it}\beta, \sigma_\delta^2 + \mathbf{x}'_{it}\Sigma_\beta\mathbf{x}_{it}]$ . The conditional mean is considerably more complicated, with  $E[y_{it}|\mathbf{x}_{it}] = \exp\{\delta + \mathbf{x}'_{it}\beta + (\sigma_\delta^2 + \mathbf{x}'_{it}\Sigma_\beta\mathbf{x}_{it})/2\}$ . This model falls in the class of generalized linear latent and mixed models; see Skrondal and Rabe-Hesketh (2004). Numerical integration methods are more challenging as the likelihood now involves multi-dimensional integrals.

One approach is to use Bayesian Markov chain Monte Carlo (MCMC) methods. Chib, Greenberg, and Winkelmann (1998) consider the following model. Assume  $y_{it}|\gamma_{it}$  is Poisson distributed with mean  $\exp(\gamma_{it})$ , where  $\gamma_{it} = \mathbf{x}'_{it}\beta + \mathbf{w}'_{it}\alpha_i$  and  $\alpha_i \sim N[\alpha, \Sigma_\alpha]$ . The RE model is the specialization  $\mathbf{w}'_{it}\alpha_i = \alpha_i$ , and the random coefficients

model sets  $w_{it} = x_{it}$ , though they argue that  $x_{it}$  and  $w_{it}$  should share no common variables to avoid identification and computational problems. Data augmentation is used to add  $\gamma_{it}$  as parameters leading to augmented posterior  $p(\beta, \eta, \Sigma, \gamma | y, X)$ . A Gibbs sampler is used where draws from  $p(\gamma | \beta, \eta, \Sigma, y, X)$  use the Metropolis-Hastings algorithm, while draws from the other full conditionals  $p(\beta | \eta, \Sigma, \gamma, y, X)$ ,  $p(\eta | \beta, \Sigma, \gamma, y, X)$ , and  $p(\Sigma | \beta, \eta, \gamma, y, X)$  are straightforward if independent normal priors for  $\beta$  and  $\eta$  and a Wishart prior for  $\Sigma^{-1}$  are specified.

Another generalization of the RE model is to model the time-invariant individual effect to depend on the average of individual effects, an approach proposed for linear regression by Mundlak (1978) and Chamberlain (1982). The conditionally correlated random (CCR) effects model specifies that  $\alpha_i$  in (8) can be modelled as

$$\alpha_i = \exp(\bar{x}'_i \lambda + \varepsilon_i), \quad (18)$$

where  $\bar{x}_i$  denotes the time-average of the time-varying exogenous variables and  $\varepsilon_i$  may be interpreted as unobserved heterogeneity that is uncorrelated with the regressors. Substituting into (8) yields

$$E[y_{it} | x_{i1}, \dots, x_{iT}, \alpha_i] = \exp(x'_{it} \beta + \bar{x}'_i \lambda + \varepsilon_i). \quad (19)$$

This can be estimated as an RE model, with  $\bar{x}_i$  as an additional regressor.

### 8.3.4 Fixed Effects Models

Fixed effects (FE) models treat the individual effect  $\alpha_i$  in (8) as being random and potentially correlated with the regressors  $X_i$ . In the linear regression model with additive errors the individual effect can be eliminated by mean-differencing or by first-differencing. In the nonlinear model (8),  $\alpha_i$  can be eliminated by quasi-differencing as follows.

Assume that the regressors  $x_{it}$  are strictly exogenous, after conditioning on  $\alpha_i$ , so that

$$E[y_{it} | x_{i1}, \dots, x_{iT}, \alpha_i] \equiv E[y_{it} | X_i, \alpha_i] = \alpha_i \lambda_{it}. \quad (20)$$

This is a stronger condition than (8) which conditions only on  $x_{it}$  and  $\alpha_i$ . Averaging over time for individual  $i$ , it follows that  $E[\bar{y}_i | X_i, \alpha_i] = \alpha_i \bar{\lambda}_i$ , where  $\bar{y}_i = \frac{1}{T} \sum_{t=1}^T y_{it}$  and  $\bar{\lambda}_i = \frac{1}{T} \sum_{t=1}^T \lambda_{it}$ . Subtracting from (20) yields

$$E[(y_{it} - (\lambda_{it}/\bar{\lambda}_i)\bar{y}_i) | \alpha_i, X_i] = 0, \quad (21)$$

and hence by the law of iterated expectations

$$E[x_{it} \left( y_{it} - \frac{\lambda_{it}}{\bar{\lambda}_i} \bar{y}_i \right)] = 0. \quad (22)$$

Given assumption (20),  $\beta$  can be consistently estimated by the method of moments estimator that solves the sample moment conditions corresponding to (22):

$$\sum_{i=1}^n \sum_{t=1}^T \mathbf{x}_{it} \left( y_{it} - \frac{\bar{y}_i}{\lambda_i} \lambda_{it} \right) = 0. \quad (23)$$

For short panels a panel-robust estimate of the variance matrix of  $\hat{\beta}$  can be obtained using standard GMM results.

Wooldridge (1990) covers this moment-based approach in more detail and gives more efficient GMM estimators when additionally the variance is specified to be of the form  $\psi_i \alpha_i \lambda_{it}$ . Chamberlain (1992a) gives semi-parametric efficiency bounds for models using only specified first moment of form (8). Attainment of these bounds is theoretically possible but practically difficult, as it requires high-dimensional nonparametric regressions.

Remarkably, the method of moments estimator defined in (23) coincides with the Poisson fixed effects estimator in the special case that  $\lambda_{it} = \exp(\mathbf{x}'_{it} \beta)$ . This estimator in turn can be derived in two ways.

First, suppose we assume that  $y_{it} | \mathbf{x}_{it}, \alpha_i$  is independently distributed over  $i$  and  $t$  as Poisson with mean  $\alpha_i \lambda_{it}$ . Then maximizing the log likelihood function  $\sum_{i=1}^n \sum_{t=1}^T \ln f(y_{it} | \mathbf{x}_{it}, \alpha_i)$  with respect to both  $\beta$  and  $\alpha_1, \dots, \alpha_n$  leads to first-order conditions for  $\beta$  that after some algebra can be expressed as (23), while  $\hat{\alpha}_i = \bar{y}_i / \hat{\lambda}_i$ . This result, given in Blundell, Griffith, and Windmeijer (1997, 2002) and Lancaster (1997) is analogous to that for MLE in the linear regression model under normality with fixed effects—in principle the  $\alpha_i$  introduce an incidental parameters problem, but in these specific models this does not lead to inconsistent estimation of  $\beta$ , even if  $T$  is small.

Second, consider the Poisson conditional MLE that additionally conditions on  $T\bar{y}_i = \sum_{t=1}^T y_{it}$ . Then some algebra reveals that  $\alpha_i$  drops out of the conditional log-likelihood function  $\sum_{i=1}^n \sum_{t=1}^T \ln f(y_{it} | \mathbf{x}_{it}, \alpha_i, T\bar{y}_i)$ , and that maximization with respect to  $\beta$  leads to the first-order conditions (23). This is the original derivation of the Poisson FE estimator due to Palmgren (1981) and Hausman, Hall, and Griliches (1984). Again, a similar result holds for the linear regression model under normality.

In general it is not possible to obtain consistent estimates of  $\beta$  in a fixed effects model with data from a short panel, due to too many incidental parameters  $\alpha_1, \dots, \alpha_n$ . The three leading exceptions are regression with additive errors, regression with multiplicative errors, including the Poisson, and the logit model. Hausman, Hall, and Griliches (1984) additionally proposed a fixed effects estimator for the NB1 model, but Guimarães (2008) shows that this model places a very strong restriction on the relationship between  $\alpha_i$  and the NB1 overdispersion parameter. One consequence, pointed out by Allison and Waterman (2002), is that the coefficients of time-invariant regressors are identified in this model.

One alternative is to estimate a regular NB model, such as NB2, with a full set of individual dummies. While this leads to inconsistent estimation of  $\beta$  in short panels due to the incidental parameters problem, Allison and Waterman (2002) and Greene

(2004) present simulations that suggest that this inconsistency may not be too large for moderately small  $T$ ; see also Fernández-Val (2009) who provides theory for the probit model. A second alternative is to use the conditionally correlated random effects model presented in (18).

The distinction between fixed and random effects is fundamentally important, as pooled and random effects estimators are inconsistent if in fact the data are generated by the individual-specific effects model (8) with  $\alpha_i$  correlated with  $x_{it}$ . Let  $\beta_1$  denote the subcomponent of  $\beta$  that is identified in the fixed effects model (i.e. the coefficient of time-varying regressors), or a subset of this, and let  $\hat{\beta}_{1,RE}$  and  $\tilde{\beta}_{1,FE}$  denote, respectively, the corresponding RE and FE estimators. The Hausman test statistic is

$$H = (\hat{\beta}_{1,RE} - \tilde{\beta}_{1,FE})' [\hat{V}[\tilde{\beta}_{1,FE} - \hat{\beta}_{1,RE}]]^{-1} (\hat{\beta}_{1,RE} - \tilde{\beta}_{1,FE}). \quad (24)$$

If  $H < \chi^2_\alpha(\dim(\beta_1))$  then at significance level  $\alpha$  we do not reject the null hypothesis that the individual specific effects are uncorrelated with regressors. In that case there is no need for fixed effects estimation.

This test requires an estimate of  $V[\tilde{\beta}_{1,FE} - \hat{\beta}_{1,RE}]$ . This reduces to  $V[\tilde{\beta}_{1,FE}] - V[\hat{\beta}_{1,RE}]$ , greatly simplifying analysis, under the assumption that the RE estimator is fully efficient under the null hypothesis. But it is very unlikely that this additional restriction is met. Instead in short panels one can do a panel bootstrap that resamples over individuals. In the  $b^{th}$  resample compute  $\tilde{\beta}_{1,FE}^{(b)} - \hat{\beta}_{1,RE}^{(b)}$  and, given  $B$  bootstraps, compute the variance of these  $B$  differences.

## 8.4 DYNAMIC PANEL COUNT MODELS

An individual-specific effect  $\alpha_i$  induces dependence over time in  $y_{it}$ . An alternative way to introduce dependence over time is a dynamic model that specifies the distribution of  $y_{it}$  to depend directly on lagged values of  $y_{it}$ .

### 8.4.1 Specifications for Dynamic Models

We begin by considering the Poisson model in the time series case, with  $y_t$  Poisson distributed with mean that is a function of  $y_{t-1}$  and regressors  $x_t$ . Then a much wider range of specifications for a dynamic model have been proposed than in the linear case; Cameron and Trivedi (2013, chapter 7) provide a survey.

One obvious time series model, called exponential feedback, is that  $y_t$  is Poisson with mean  $\exp(\rho y_{t-1} + x'_t \beta)$ , but this model is explosive if  $\rho > 0$ . An alternative is to specify the mean to be  $\exp(\rho \ln y_{t-1} + x'_t \beta)$ , but this model implies that if  $y_{t-1} = 0$ , then  $y_t$  necessarily equals zero. A linear feedback model specifies the mean to equal  $\rho y_{t-1} +$

$\exp(\mathbf{x}'_t \beta)$ . This model arises from a Poisson integer-valued autoregressive model of order 1 (INAR(1)), a special case of the more general class of INARMA models.

For panel data we allow for both dynamics and the presence of an individual specific effect. Define the conditional mean to be

$$\mu_{it} = E[y_{it} | \mathbf{X}_i^{(t)}, \mathbf{Y}_i^{(t-1)}, \alpha_i], \quad (25)$$

where  $\mathbf{X}_i^{(t)} = \{\mathbf{x}_{it}, \mathbf{x}_{i,t-1}, \dots\}$  and  $\mathbf{Y}_i^{(t-1)} = \{y_{i,t-1}, y_{i,t-2}, \dots\}$ . Blundell, Griffith, and Windmeijer (1997, 2002) discuss various forms for  $\mu_{it}$  and emphasize the linear feedback model

$$\mu_{it} = \rho y_{i,t-1} + \alpha_i \exp(\mathbf{x}'_{it} \beta), \quad (26)$$

where for simplicity we consider models where  $\mu_{it}$  depends on just the first lag of  $y_{it}$ . The exponential feedback model instead specifies

$$\mu_{it} = \alpha_i \exp(\rho y_{i,t-1} + \mathbf{x}'_{it} \beta). \quad (27)$$

Yet another model, proposed by Crepon and Duguet (1997), is that

$$\mu_{it} = h(y_{i,t-1}, \gamma) \alpha_i \exp(\mathbf{x}'_{it} \beta), \quad (28)$$

where the function  $h(y_{it-1}, \gamma)$  parameterizes the dependence of  $\mu_{it}$  on lagged values of  $y_{it}$ . A simple example is  $h(y_{i,t-1}, \gamma) = \exp(\gamma \mathbf{1}[y_{i,t-1} > 0])$  where  $\mathbf{1}[\cdot]$  is the indicator function. More generally a spline-type specification with a set of dummies determined by ranges taken by  $y_{it-1}$  might be specified.

### 8.4.2 Pooled Dynamic Models

Pooled dynamic models assume that all regression coefficients are the same across individuals, so that there are no individual-specific fixed or random effects. Then one can directly apply the wide range of methods suggested for time series data, even for small  $T$  provided  $n \rightarrow \infty$ . This approach is given in Diggle et al. (2002, chapter 10), who use autoregressive models that directly include  $y_{i,t-k}$  as regressors. Brännäs (1995) briefly discusses a generalization of the INAR(1) time series model to longitudinal data.

Under weak exogeneity of regressors, which requires that there is no serial correlation in  $(y_{it} - \mu_{it})$ , the models can be estimated by nonlinear least squares, GEE, method of moments, or GMM based on the sample moment condition  $\sum_i \sum_t \mathbf{z}_{it} (y_{it} - \mu_{it})$ , where  $\mathbf{z}_{it}$  can include  $y_{i,t-1}$  and  $\mathbf{x}_{it}$  and, if desired, additional lags in these variables.

This approach leads to inconsistent estimation if fixed effects are present. But inclusion of lagged values of  $y_{it}$  as a regressor may be sufficient to control for correlation between  $y_{it}$  and lagged  $y_{it}$ , so that there is no need to additionally include individual-specific effects.

### 8.4.3 Random Effects Dynamic Models

A random effects dynamic model is an extension of the static RE model that includes lagged  $y_{it}$  as regressors. However, the log-likelihood will depend on initial condition  $y_{i0}$ , this condition will not disappear asymptotically in a short panel, and most importantly it will be correlated with the random effect  $\alpha_i$  (even if  $\alpha_i$  is uncorrelated with  $\mathbf{x}_{it}$ ). So it is important to control for the initial condition.

Heckman (1981) writes the joint distribution of  $y_{i0}, y_{i1}, \dots, y_{iT}, \alpha_i | \mathbf{x}_{it}$  as

$$f(y_{i0}, y_{i1}, \dots, y_{iT}, \alpha_i | \mathbf{X}_i) = f(y_{i1}, \dots, y_{iT} | \mathbf{X}_i, y_{i0}, \alpha_i) f(y_{i0} | \mathbf{X}_i, \alpha_i) f(\alpha_i | \mathbf{X}_i). \quad (29)$$

Implementation requires specification of the functional forms  $f(y_{i0} | \mathbf{X}_i, \alpha_i)$  and  $f(\alpha_i | \mathbf{X}_i)$  and, most likely, numerical integration; see Stewart (2007).

Wooldridge (2005) instead proposed a conditional approach, for a class of nonlinear dynamic panel models that includes the Poisson model, based on the decomposition

$$f(y_{i1}, \dots, y_{iT}, \alpha_i | \mathbf{X}_i, y_{i0}) = f(y_{i1}, \dots, y_{iT} | \mathbf{X}_i, y_{i0}, \alpha_i) f(\alpha_i | y_{i0}, \mathbf{X}_i). \quad (30)$$

This simpler approach conditions on  $y_{i0}$  rather than modelling the distribution of  $y_{i0}$ . Then the standard random effects conditional ML approach identifies the parameters of interest. One possible model for  $f(\alpha_i | y_{i0}, \mathbf{X}_i)$  is the CCR model in (18) with  $y_{i0}$  added as a regressor, so

$$\alpha_i = \exp(\delta_0 y_{i0} + \bar{\mathbf{x}}_i' \lambda + \varepsilon_i), \quad (31)$$

where  $\bar{\mathbf{x}}_i$  denotes the time-average of the time-varying exogenous variables, and  $\varepsilon_i$  is an i.i.d. random variable. Then the model (30)–(31) can be estimated using RE model software commands. Note that in a model with just one lag of  $y_{i,t-1}$  as a regressor, identification in the CCR model requires three periods of data  $(y_{i0}, y_{i1}, y_{i2})$ .

### 8.4.4 Fixed Effects Dynamic Models

The Poisson FE estimator eliminates fixed effects under the assumption that  $E[y_{it} | \mathbf{X}_i] = \alpha_i \lambda_{it}$ ; see (20). This assumption rules out predetermined regressors. To allow for predetermined regressors that may be correlated with past shocks, we make the weaker assumption that regressors are weakly exogenous, so

$$E[y_{it} | \mathbf{X}_i^{(t)}] = E[y_{it} | \mathbf{x}_{it}, \dots, \mathbf{x}_{i1}] = \alpha_i \lambda_{it}, \quad (32)$$

where now conditioning is only on current and past regressors. Then, defining  $u_{it} = y_{it} - \alpha_i \lambda_{it}$ ,  $E[u_{it} | \mathbf{x}_{is}] = 0$  for  $s \leq t$ , so future shocks are indeed uncorrelated with current  $\mathbf{x}$ , but there is no restriction that  $E[u_{it} | \mathbf{x}_{is}] = 0$  for  $s > t$ .

For dynamic models, lagged dependent variables also appear as regressors, and we assume

$$E[y_{it} | \mathbf{X}_i^{(t)}, \mathbf{Y}_i^{(t-1)}] = E[y_{it} | \mathbf{x}_{it}, \dots, \mathbf{x}_{i1}, y_{i,t-1}, \dots, y_{i1}] = \alpha_i \lambda_{it}, \quad (33)$$

where conditioning is now also on past values of  $y_{it}$ . (For the linear feedback model defined in (26),  $\alpha_i \lambda_{it}$  in (33) is replaced by  $\rho y_{i,t-1} + \alpha_i \lambda_{it}$ .)

If regressors are predetermined then the Poisson FE estimator is inconsistent, since quasi-differencing subtracts  $(\lambda_{it}/\bar{\lambda}_i)\bar{y}_i$  from  $y_{it}$ , see (21), but  $\bar{y}_i$  includes future values  $y_{is}$ ,  $s > t$ . This problem is analogous to the inconsistency (or Nickell bias) of the within or mean-differenced fixed effects estimator in the linear dynamic model.

Instead GMM estimation is based on alternative differencing procedures that eliminate  $\alpha_i$  under the weaker assumption (33). These generalize the use of first differences in linear dynamic models with fixed effects. Chamberlain (1992b) proposed eliminating the fixed effects  $\alpha_i$  by the transformation

$$q_{it}(\theta) = \frac{\lambda_{i,t-1}}{\lambda_{it}} y_{it} - y_{i,t-1}, \quad (34)$$

where  $\lambda_{it} = \lambda_{it}(\theta)$ . Wooldridge (1997) instead proposed eliminating the fixed effects using

$$q_{it}(\theta) = \frac{y_{i,t-1}}{\lambda_{i,t-1}} - \frac{y_{it}}{\lambda_{it}}. \quad (35)$$

For either specification of  $q_{it}(\theta)$  it can be shown that, given assumption (33),

$$\mathbb{E}[q_{it}(\theta)|\mathbf{z}_{it}] = 0, \quad (36)$$

where  $\mathbf{z}_{it}$  can be drawn from  $\mathbf{x}_{i,t-1}, \mathbf{x}_{i,t-2}, \dots$  and, if lags up to  $y_{i,t-p}$  appear as regressors,  $\mathbf{z}_{it}$  can also be drawn from  $y_{i,t-p-1}, y_{i,t-p-2}, \dots$  Often  $p = 1$ , so  $y_{i,t-2}, y_{i,t-3}, \dots$  are available as instruments.

In the just-identified case in which there are as many instruments as parameters, the method of moments estimator solves

$$\sum_{i=1}^n \sum_{t=1}^T \mathbf{z}_{it} q_{it}(\theta) = \mathbf{0}. \quad (37)$$

In general there are more instruments  $\mathbf{z}_{it}$  than regressors and the GMM estimator of  $\beta$  minimizes

$$\left( \sum_{i=1}^n \sum_{t=1}^T \mathbf{z}_{it} q_{it}(\theta) \right)' \mathbf{W}_n \left( \sum_{i=1}^n \sum_{t=1}^T \mathbf{z}_{it} q_{it}(\theta) \right). \quad (38)$$

Given efficient two-step GMM estimation, model adequacy can be tested using an over-identifying restrictions test.

It is also important to test for serial correlation in  $q_{it}(\theta)$ , using  $q_{it}(\hat{\theta})$ , as correct model specification requires that  $\text{Cor}[q_{it}(\theta), q_{is}(\theta)] = 0$  for  $|t - s| > 1$ . Blundell, Griffith, and Windmeijer (1997) adapt serial correlation tests proposed by Arellano and Bond (1991) for the linear model. Crepon and Duguet (1997) and Brännäs and Johansson (1996) apply serial correlation tests in the GMM framework.

Windmeijer (2008) provides a broad survey of GMM methods for the Poisson panel model, including the current setting. Two-step GMM estimated coefficients and standard errors can be biased in finite samples. Windmeijer (2008) proposes an extension of the variance matrix estimate of Windmeijer (2005) to nonlinear models. In a Monte Carlo exercise with predetermined regressor he shows that this leads to improved finite sample inference, as does the Newey and Windmeijer (2009) method applied to the continuous updating estimator variant of GMM.

Blundell, Griffith, and Windmeijer (2002) proposed an alternative transformation, the mean-scaling transformation

$$q_{it}(\theta) = y_{it} - \frac{\bar{y}_{i0}}{\lambda_{i0}}\lambda_{it}, \quad (39)$$

where  $\bar{y}_{i0}$  is the presample mean value of  $y_i$  and the instruments are  $(x_{it} - x_{i0})$ . This estimator is especially useful if data on the dependent variable are available farther back in time than data on the explanatory variables. The transformation leads to inconsistent estimation, but in a simulation this inconsistency is shown to be small and efficiency is considerably improved.

The GMM methods of this section can be adapted to estimate FE models with endogenous regressors. Suppose the conditional mean of  $y_{it}$  with exogenous regressors is

$$\mu_{it} = \alpha_i \exp(x'_{it}\beta). \quad (40)$$

Due to endogeneity of regressor(s), however,  $E[y_{it} - \mu_{it}|x_{it}] \neq 0$ , so the standard Poisson FE estimator is inconsistent. Windmeijer (2000) shows that in the panel case, the individual-specific fixed effects  $\alpha_i$  can only be eliminated if a multiplicative errors specification is assumed and if the Wooldridge transformation is used. Then nonlinear IV or GMM estimation is based on  $q_{it}(\theta)$  defined in (35), where the instruments  $z_{it}$  satisfy  $E[(y_{it} - \mu_{it})/\mu_{it}|z_{it}] = 0$  and  $z_{it}$  can be drawn from  $x_{i,t-2}, x_{i,t-3}, \dots$

## 8.5 EXTENSIONS

---

In this section we survey recent developments that extend to the panel setting complications for counts that were introduced briefly in section 8.2.2 on cross-section data models. We consider panel versions of hurdle models, latent class models, and dynamic latent class models.

### 8.5.1 Hurdle Models

The panel count models covered in previous sections specify the same stochastic process for zero counts and for positive counts. Both the hurdle model and the

zero-inflated model relax this restriction. Here we focus on panel versions of the hurdle or two-part model; similar issues arise for the zero-inflated model.

We specify a two-part data generating process. The split between zeros and positives is determined by a Bernoulli distribution with probabilities of, respectively,  $f_1(0|\mathbf{z}_{it})$  and  $1 - f_1(0|\mathbf{z}_{it})$ . The distribution of positives is determined by a truncated-at-zero variant of the count distribution  $f_2(y_{it}|\mathbf{x}_{it})$ . Then

$$f(y_{it}|\mathbf{x}_{it}, \mathbf{z}_{it}) = \begin{cases} f_1(0|\mathbf{z}_{it}) & \text{if } y_{it} = 0 \\ (1 - f_1(0|\mathbf{z}_{it})) \frac{f_2(y_{it}|\mathbf{x}_{it})}{1 - f_2(0|\mathbf{x}_{it})} & \text{if } y_{it} \geq 1, \end{cases} \quad (41)$$

which specializes to the standard model only if  $f_1(0|\mathbf{z}_{it}) = f_2(0|\mathbf{x}_{it})$ , and  $\mathbf{z}_{it} = \mathbf{x}_{it}$ . In principle,  $\mathbf{z}_{it}$  and  $\mathbf{x}_{it}$  may have distinct and/or overlapping elements, though in practice they are often the same. This model can handle both excess zeros in the count distribution  $f_2(y_{it}|\mathbf{x}_{it})$ , if  $f_1(0) > f_2(0)$ , and too few zeros if  $f_2(0) > f_1(0)$ .

This model is simply a pooled version of the standard cross-section hurdle model. Its implementation involves no new principles if the cross-section assumptions are maintained, though cluster-robust standard errors analogous to those in (13) for pooled Poisson should be used. Because the two parts of the model are functionally independent, maximum likelihood estimation can be implemented by separately maximizing the two terms in the likelihood. A binary logit specification is usually used to model the positive outcome, and a Poisson or negative binomial specification is used for  $f_2(y_{it}|\mathbf{x}_{it})$ .

A random effects variant of this model introduces individual-specific effects, so

$$f(y_{it}|\mathbf{x}_{it}, \mathbf{z}_{it}, \alpha_{1i}, \alpha_{2i}) = \begin{cases} f_1(y_{it}|\mathbf{z}_{it}, \alpha_{1i}) & \text{if } y_{it} = 0 \\ (1 - f_1(0|\mathbf{z}_{it}, \alpha_{1i})) \frac{f_2(y_{it}|\mathbf{x}_{it}, \alpha_{2i})}{1 - f_2(0|\mathbf{x}_{it}, \alpha_{2i})} & \text{if } y_{it} \geq 1, \end{cases} \quad (42)$$

where  $\alpha_{1i}$  and  $\alpha_{2i}$  are individual-specific effects for the first and second part of the model, respectively. Under the assumption of exogeneity of  $\mathbf{x}_{it}$  and  $\mathbf{z}_{it}$ , and given the bivariate density of  $(\alpha_{1i}, \alpha_{2i})$  denoted by  $h(\alpha_{1i}, \alpha_{2i})$ , the marginal distribution of  $y_{it}$  is given by

$$\int \int f(y_{it}|\mathbf{x}_{it}, \mathbf{z}_{it}, \alpha_{1i}, \alpha_{2i}) h(\alpha_{1i}, \alpha_{2i}) d\alpha_{1i} d\alpha_{2i}. \quad (43)$$

This calculation can be expected to be awkward to implement numerically. First, the likelihood no longer splits into two pieces that can be maximized individually. Second, it seems plausible that the individual-specific effects in the two distributions should not be independent. In some cases the assumption of a bivariate normal distribution is appropriate, perhaps after transformation such as for  $\ln \alpha_{2i}$  rather than  $\alpha_{2i}$ . Experience with even simpler problems of the same type suggest that more work is needed on the computational aspects of this problem; see Olsen and Schafer (2001).

Consistent estimation of a fixed effects variant of this model in a short panel is not possible. Conditional likelihood estimation is potentially feasible for some special

choices of  $f_1(\cdot)$ , but a sufficient statistic for  $\alpha_{2i}$  in a zero-truncated model is not available. Given  $T$  sufficiently large, individual-specific dummy variables may be added to the model. Then the profile likelihood approach (Dhaene and Jochmans 2011) is potentially appealing, but there is no clear guidance from the literature. Yet another approach is to specify a conditionally correlated random effects model, introduced in (18).

### 8.5.2 Latent Class Models

Latent class models, or finite mixture models (FMM), have been used effectively in cross-section analysis of count data. They are generally appealing because they offer additional flexibility within a parametric framework. In this section we consider their extension to panel counts.

The key idea underlying latent class modeling is that an unknown distribution may be parsimoniously approximated by a mixture of parametric distributions with a finite and small number of mixture components. For example, a mixture of Poissons may be used to approximate an unknown distribution of event counts. Such models can provide an effective way of handling both excess zeros and overdispersion in count models (Deb and Trivedi 2002).

The general expression for a panel latent class model in which all parameters are assumed to vary across latent classes is

$$f(y_{it} | \mathbf{x}_{it}, \theta_1, \dots, \theta_C, \pi_1, \dots, \pi_{C-1}) = \sum_{j=1}^C \pi_j f_j(y_{it} | \mathbf{x}_{it}, \theta_j), \quad (44)$$

where  $0 \leq \pi_j \leq 1$ ,  $\pi_1 > \pi_2 \dots > \pi_C$ ,  $\sum_j \pi_j = 1$ ,  $\mathbf{x}_{it}$  is a vector of  $K$  exogenous variables, and  $\theta_j$  denotes the vector of unknown parameters in the  $j^{th}$  component. For simplicity the component probabilities  $\pi_j$  in (44) are time-invariant and individual-invariant, an assumption that is relaxed below.

The estimation objective is to obtain consistent estimates of  $(\pi_j, \theta_j)$ ,  $j = 1, \dots, C$ , where  $C$  also should be determined from the data. For simplicity the analysis below concentrates on modeling issues that are specific to panel models, avoiding some general issues that arise in identification and estimation of all latent class models, and just a two-component mixture is considered, so we assume  $C = 2$  is adequate.

We begin by considering a pooled panel latent class model. Introduce an unobserved variable  $d_{it}$  that equals  $j$  if individual  $i$  in period  $t$  is in the  $j^{th}$  latent class, and let

$$f(y_{it} | d_{it} = j, \mathbf{x}_{it}) = P(\mu_{it}^{(j)}), \quad j = 1, 2,$$

where  $P(\mu_{it}^{(j)})$  is the density of a Poisson distribution with mean  $\mu_{it}^{(j)} = \exp(\mathbf{x}'_{it}\beta_j)$ . For the case  $C = 2$ ,  $d_{it}$  is a Bernoulli random variable. Let  $\Pr[d_{it} = 1] = \pi$ , for the moment

constant over  $i$  and  $t$ . Then the joint density of  $(d_{it}, y_{it})$  is

$$f(y_{it}, d_{it} | \mathbf{x}_{it}, \beta_1, \beta_2, \pi) = \left[ \pi \mathbb{P}(\mu_{it}^{(1)}) \right]^{1[d_{it}=1]} \left[ (1-\pi) \mathbb{P}(\mu_{it}^{(2)}) \right]^{1[d_{it}=2]}, \quad (45)$$

where  $1[A] = 1$  if event  $A$  occurs and equals 0 otherwise. The marginal density of  $y_{it}$  is

$$f(y_{it} | \mathbf{x}_{it}, \beta_1, \beta_2, \pi) = \pi \mathbb{P}(\mu_{it}^{(1)}) + (1-\pi) \mathbb{P}(\mu_{it}^{(2)}). \quad (46)$$

Let  $\theta = (\beta_1, \beta_2)$ . Under the assumption that the observations are independent across individuals and over time, the complete-data (joint) likelihood, conditioning on both  $y_{it}$  and  $d_{it}$  for all  $i$  and  $t$ , is

$$L^c(\theta, \pi) = \prod_{i=1}^n \prod_{t=1}^T \left[ \pi \mathbb{P}(\mu_{it}^{(1)}) \right]^{1[d_{it}=1]} \left[ (1-\pi) \mathbb{P}(\mu_{it}^{(2)}) \right]^{1[d_{it}=2]}, \quad (47)$$

and the marginal likelihood, conditioning on  $y_{it}$  and  $d_{it}$  for all  $i$  and  $t$ , is

$$L^m(\theta, \pi) = \prod_{i=1}^n \prod_{t=1}^T \left[ \left\{ \pi \mathbb{P}(\mu_{it}^{(1)}) + (1-\pi) \mathbb{P}(\mu_{it}^{(2)}) \right\} \right]. \quad (48)$$

ML estimation may be based on an EM algorithm applied to (47) or, more directly, a gradient-based algorithm applied to (48). These expressions, especially the marginal likelihood, have been used to estimate a pooled panel model; see Bago d'Uva (2005). When following this pooled approach, the modeling issues involved are essentially the same as those for the cross-section latent class models.

In a panel with sufficiently long time series dimension, there is some motivation for considering transitions between classes (states). One way to do so is to allow the mixture proportion  $\pi$ , assumed constant in the above exposition, to vary over time (and possibly individuals). This can be done by specifying  $\pi$  as a function of some time-varying regressors. Let  $\Pr[d_{it} = 1 | \mathbf{z}_{it}] = F(\mathbf{z}'_{it} \gamma)$  where  $F$  denotes a suitable c.d.f. such as logit or probit, and  $\theta = (\gamma, \beta_1, \beta_2)$ . Then the complete-data likelihood is

$$L^c(\theta, \gamma) = \prod_{i=1}^n \prod_{t=1}^T \left[ F(\mathbf{z}'_{it} \gamma) \mathbb{P}(\mu_{it}^{(1)}) \right]^{1[d_{it}=1]} \left[ (1 - F(\mathbf{z}'_{it} \gamma)) \mathbb{P}(\mu_{it}^{(2)}) \right]^{1[d_{it}=2]}. \quad (49)$$

This specification was used by Hyppolite and Trivedi (2012).

If, in the interests of parsimonious specification,  $C$  is kept low, then some latent class components may still show substantial within-class heterogeneity. This provides motivation for adding individual-specific effects to improve the fit of the model. A multiplicative random effects model variant of (47), with individual-specific effects  $\alpha_i$  and Poisson component means  $\alpha_i \lambda_{it}^{(j)}$  where  $\lambda_{it}^{(j)} = \exp(\mathbf{x}'_{it} \beta_j)$ , has the following form:

$$L^c(\theta, \pi | \alpha_1, \dots, \alpha_n) = \prod_{i=1}^n \prod_{t=1}^T \left[ \pi \mathbb{P}(\alpha_i \lambda_{it}^{(1)}) \right]^{1[d_{it}=1]} \left[ (1-\pi) \mathbb{P}(\alpha_i \lambda_{it}^{(2)}) \right]^{1[d_{it}=2]}. \quad (50)$$

Assuming that the parametric distribution  $g(\alpha_i|\eta)$  for the individual-specific effects is the same for both latent classes, the individual-specific effects can be integrated out, analytically or numerically, yielding the likelihood function

$$L^c(\theta, \pi, \eta) = \prod_{i=1}^n \int \left\{ \prod_{t=1}^T \left[ \pi P(\alpha_i \lambda_{it}^{(1)}) \right]^{1[d_{it}=1]} \left[ (1-\pi) P(\alpha_i \lambda_{it}^{(2)}) \right]^{1[d_{it}=2]} \right\} \\ \times g(\alpha_i|\eta) d\alpha_i. \quad (51)$$

For a popular specification of  $g(\alpha_i|\eta)$  such as the gamma, the integral will be a mixture of two negative binomial distributions (Deb and Trivedi 2002); for the log-normal specification there is no closed form but a suitable numerical approximation can be used. Under the more flexible assumption that the two classes have different distributions for the  $\alpha_i$ , likelihood estimation is potentially more complicated. More generally the slope coefficients may also vary across individuals; Greene and Hensher (2013) estimate a latent class model with random coefficients for cross-section multinomial data.

The fixed effects model is very popular in econometric studies as it allows the individual specific effects to be correlated with the regressors. Until recently, however, there has been no attempt to combine finite mixtures and fixed effects. In a recent paper, Deb and Trivedi (2013) take the first steps in this direction. They use the conditional likelihood approach to eliminate the incidental parameters  $\alpha_i$  from the likelihood. The resulting likelihood is a complete-data form likelihood which is maximized using an EM algorithm.

For the one-component Poisson panel model with  $y_{it} \sim P(\alpha_i \lambda_{it})$  and  $\lambda_{it} = \exp(\mathbf{x}'_{it}\beta)$ , the incidental parameters can be concentrated out of the likelihood using the first-order conditions with respect to  $\alpha_i$ . Then  $\hat{\alpha}_i = \sum_t y_{it} / \sum_t \lambda_{it}$ , leading to the following concentrated likelihood function, ignoring terms not involving  $\beta$ :

$$\ln L_{\text{conc}}(\beta) \propto \sum_{i=1}^n \sum_{t=1}^T \left[ y_{it} \ln \lambda_{it} - y_{it} \ln \left( \sum_{s=1}^T \lambda_{is} \right) \right]. \quad (52)$$

We wish to extend this conditional maximum likelihood approach to Poisson finite mixture models. The above conditioning approach will not work for the mixture of Poissons because in this case a sufficient statistic for the  $\alpha_i$  is not available. However, the approach can work if the incidental parameters  $\alpha_i$  are first concentrated out of the mixture components and the mixture is expressed in terms of the concentrated components. Denote by  $s_i$  the sufficient statistic for  $\alpha_i$ . Then the mixture representation after conditioning on  $s_i$  is

$$f(y_{it}|\mathbf{x}_{it}, s_i, \theta_1, \dots, \theta_C, \pi_1, \dots, \pi_{C-1}) = \sum_{j=1}^C \pi_j f(y_{it}|\mathbf{x}_{it}, s_i, \theta_j). \quad (53)$$

Specializing to the Poisson mixture, the complete-data concentrated likelihood is

$$L_{conc}(\theta, \pi_1, \dots, \pi_C) = \prod_{i=1}^n \prod_{t=1}^T \sum_{j=1}^C \left[ \pi_j P\left( \frac{\sum_{s=1}^T y_{is}}{\sum_{s=1}^T \lambda_{is}^{(j)}} \times \lambda_{it}^{(j)} \right) \right]^{1[d_{it}=j]} . \quad (54)$$

Because in this case the sufficient statistic  $s_i = \sum_t y_{it} / \sum_t \lambda_{it}^{(j)}$  depends on model parameters and not just on data, the EM algorithm needs to be applied to the full-data likelihood. For a Monte Carlo evaluation and an empirical application, see Deb and Trivedi (2013).

### 8.5.3 Dynamic Latent Class Models

Dynamics can be introduced into the general model (44) in several ways. One way is to introduce dynamics into the component densities, such as using  $f_j(y_{it}|x_{it}, y_{i,t-1}, \theta_j)$ . Böckenholt (1999) does this using a pooled version of the Poisson with conditional mean  $\rho_j y_{i,t-1} + \exp(x'_{it} \beta_j)$  for the  $j^{th}$  component.

Alternatively, dynamics can be introduced through the latent class membership probabilities. Specifically, the class to which an individual belongs may evolve over time, depending on class membership in the previous period. If we assume that all unobserved past information useful in predicting class membership is contained in the most recent class membership, the process that determines the class of an individual can be characterized by a first-order Markov Chain.

First assume that the chains for each individual are characterized by the same time-homogeneous transition matrix and the same initial probability vector. Let  $p_{kl}$  denote the probability that an individual in state  $k$  switches to state  $l$  in the next period, where there are two possible states in a two-component model. Then

$$p_{kl} = \Pr[d_{it} = l | d_{i,t-1} = k], \quad k, l = 1, 2, \quad (55)$$

and, since  $\sum_{l=1}^2 p_{kl} = 1$ , there are two free parameters, say  $p_{11}$  and  $p_{21}$ . The corresponding transition matrix is

$$\mathbf{P} = \begin{bmatrix} p_{11} & 1 - p_{11} \\ p_{21} & 1 - p_{21} \end{bmatrix}. \quad (56)$$

We additionally need to specify the probabilities at initial time  $t = 1$ , and let

$$\pi = \Pr[d_{i1} = 1], \quad (57)$$

and  $1 - \pi = \Pr[d_{i1} = 2]$ . The bivariate discrete-time process  $(d_{it}, y_{it})$ , where  $d_{it}$  is an unobserved Markov chain and  $y_{it}|(d_{it}, \mu_{it}^{(1)}, \mu_{it}^{(2)})$  is independent, is known as a hidden Markov model. The hidden Markov model is a mixture whose mixing distribution is a

Markov chain. The joint density for  $(d_{it}, y_{it})$  is given by

$$f(y_{it}, d_{it} | \mathbf{x}_{it}, \theta, \mathcal{I}_{i,t-1}) = \left( \Pr[d_{it} = 1 | \mathcal{I}_{i,t-1}] \times P(\mu_{it}^{(1)}) \right)^{1[d_{it}=1]} \\ \times \left( \Pr[d_{it} = 2 | \mathcal{I}_{i,t-1}] \times P(\mu_{it}^{(2)}) \right)^{1[d_{it}=2]}, \quad (58)$$

where  $\theta = (\beta_1, \beta_2, p_{11}, p_{12}, \pi)$  and  $\mathcal{I}_{it}$  denotes information about individual  $i$  available up to time  $t$ . Because time dependence is modeled as a first-order Markov chain, the probability of being in a given state at a given point in time now depends on the previous history of the bivariate process. The difference between (58) and (45) is that in (58) the whole history of the process matters.

The corresponding marginal density is then

$$f(y_{it} | \mathbf{x}_{it}, \theta, \mathcal{I}_{i,t-1}) = \sum_{j=1}^2 \Pr(d_{it} = j | \mathcal{I}_{i,t-1}) P(\mu_{it}^{(j)}). \quad (59)$$

Again the difference between (46) and (59) is that the history of the process enters the marginal density of the hidden Markov model.

The exact expression for the full-data likelihood depends upon whether or not the path of each individual chain is observed. For a detailed discussion of different assumptions, as well as estimation algorithms, see Hyppolite and Trivedi (2012).

The restriction that the transition matrix is time invariant can be relaxed. A more flexible model results if we parametrize the transition matrix. One such model specifies

$$\mathbf{P}_{it} = \begin{bmatrix} F(\mathbf{z}'_{it}\gamma) & 1 - F(\mathbf{z}'_{it}\gamma) \\ F(\mathbf{z}'_{it}\gamma + \lambda) & 1 - F(\mathbf{z}'_{it}\gamma + \lambda) \end{bmatrix},$$

where  $F$  is a suitable c.d.f. such as that for the logit or the probit. The complete-data likelihood and the marginal likelihood for this model are obtained the same way as for previous models. For details see Hyppolite and Trivedi (2012).

## 8.6 CONCLUSION

---

The methods of Sections 8.2 to 8.4 are well established, and many of the methods have been integrated into the leading econometrics packages. Many econometrics textbooks provide discussion of count data models, though the treatment of panel counts is generally brief. The specialized monograph of Cameron and Trivedi (2013) provides a more comprehensive presentation. The richer parametric models of Section 8.5 seek to model features of the data not well captured in some applications by the simpler panel models. These richer models are computationally more demanding, and eliminating fixed effects in these models is challenging.

## ACKNOWLEDGMENTS

---

This chapter has benefitted considerably from the comments of two referees.

## REFERENCES

---

- Allison P.D., and R.P. Waterman (2002), “Fixed-Effects Negative Binomial Regression Models,” *Sociological Methodology*, 32, 247–265.
- Arellano, M., and S. Bond (1991), “Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations,” *Review of Economic Studies*, 58, 277–298.
- Bago d’Uva, T. (2005), “Latent Class Models for Use of Primary Care: Evidence from a British Panel,” *Health Economics*, 14, 873–892.
- Blundell, R., R. Griffith, and F. Windmeijer (1997), “Individual Effects and Dynamics in Count Data,” manuscript, University College London.
- Blundell, R., R. Griffith, and F. Windmeijer (2002), “Individual Effects and Dynamics in Count Data,” *Journal of Econometrics*, 108, 113–131.
- Böckenholt, U. (1999), “Mixed INAR(1) Poisson Regression Models: Analyzing Heterogeneity and Serial Dependencies in Longitudinal Count Data,” *Journal of Econometrics*, 89, 317–338.
- Brännäs, K. (1995), “Explanatory Variables in the AR(1) Model,” Umea Economic Studies No. 381, University of Umea.
- Brännäs, K., and P. Johansson (1996), “Panel Data Regression for Counts,” *Statistical Papers*, 37, 191–213.
- Cameron, A.C., and P.K. Trivedi (2013), *Regression Analysis of Count Data*, Second Edition, Econometric Society Monograph No. 53, Cambridge, Cambridge University Press.
- Chamberlain, G. (1982), “Multivariate Regression Models for Panel Data,” *Journal of Econometrics*, 18, 5–46.
- Chamberlain, G. (1992a), “Efficiency Bounds for Semiparametric Regression,” *Econometrica*, 60, 567–596.
- Chamberlain, G. (1992b), “Comment: Sequential Moment Restrictions in Panel Data,” *Journal of Business and Economic Statistics*, 10, 20–26.
- Chib, S., E. Greenberg, and R. Winkelmann (1998), “Posterior Simulation and Bayes Factors in Panel Count Data Models,” *Journal of Econometrics*, 86, 33–54.
- Crepion, B., and E. Duguet (1997), “Estimating the Innovation Function from Patent Numbers: GMM on Count Data,” *Journal of Applied Econometrics*, 12, 243–264.
- Deb, P., and P.K. Trivedi (2002), “The Structure of Demand for Health Care: Latent Class versus Two-part Models,” *Journal of Health Economics*, 21, 601–625.
- Deb, P., and P.K. Trivedi (2013), “Finite Mixture for Panels with Fixed Effects,” *Journal of Econometric Methods*, 2, 35–51.
- Dhaene, G., and K. Jochmans (2011), “Profile-Score Adjustments for Nonlinear Fixed-effect Models,” Paper presented at the 17th International Panel Data Conference, Montreal.

- Diggle, P.J., P. Heagerty, K.-Y. Liang, and S.L. Zeger (2002), *Analysis of Longitudinal Data*, Second Edition, Oxford, Oxford University Press.
- Fernández-Val, I. (2009), “Fixed Effects Estimation of Structural Parameters and Marginal Effects in Panel Probit Models,” *Journal of Econometrics*, 150, 71–85.
- Greene, W.H. (2004), “The Behaviour of the Maximum Likelihood Estimator of Limited Dependent Variable Models in the Presence of Fixed Effects,” *Econometrics Journal*, 7, 98–119.
- Greene, W.H., and D.A. Hensher (2013), “Revealing Additional Dimensions of Preference Heterogeneity in a Latent Class Mixed Multinomial Logit Model,” *Applied Economics*, 45, 1897–1902.
- Guimarães, P. (2008), “The Fixed Effects Negative Binomial Model Revisited,” *Economics Letters*, 99, 63–66.
- 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, 909–938.
- Heckman, J.J. (1981), “The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time-Discrete Data Stochastic Process,” in C. Manski and D. McFadden, eds., *Structural Analysis of Discrete Data with Econometric Applications*, Cambridge (MA), MIT Press.
- Hypolite, J., and P.K. Trivedi (2012), “Alternative Approaches for Econometric Analysis of Panel Count Data Using Dynamic Latent Class Models (with Application to Doctor Visits Data),” *Health Economics*, 21, S101–128.
- Lancaster, T. (1997), “Orthogonal Parameters and Panel Data,” Working Paper No. 1997–32, Department of Economics, Brown University.
- Liang, K.-Y., and S.L. Zeger (1986), “Longitudinal Data Analysis Using Generalized Linear Models,” *Biometrika*, 73, 13–22.
- Liang K.-Y., S.L. Zeger, and B. Qaqish (1992), “Multivariate Regression Analyses for Categorical Data,” *Journal of the Royal Statistical Society (B)*, 54, 3–40.
- Mundlak, Y. (1978), “On the Pooling of Time Series and Cross Section Data,” *Econometrica*, 46, 69–85.
- Newey, W.K., and F. Windmeijer (2009), “GMM with Many Weak Moment Conditions,” *Econometrica*, 77, 687–719.
- Olsen, M.K., and J.L. Schafer (2001), “The Two-Part Random Effects Model for Semicontinuous Longitudinal Data,” *Journal of American Statistical Association*, 96, 730–745.
- Palmgren, J. (1981), “The Fisher Information Matrix for Log-Linear Models Arguing Conditionally in the Observed Explanatory Variables,” *Biometrika*, 68, 563–566.
- Skrondal, A., and S. Rabe-Hesketh (2004), *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*, London, Chapman and Hall.
- Stewart, M. (2007), “The Interrelated Dynamics of Unemployment and Low-wage Employment,” *Journal of Applied Econometrics*, 22, 511–531.
- Windmeijer F. (2000), “Moment Conditions for Fixed Effects Count Data Models with Endogenous Regressors,” *Economics Letters*, 68, 21–24.
- Windmeijer F. (2005), “A Finite Sample Correction for the Variance of Linear Efficient Two-step GMM Estimators,” *Journal of Econometrics*, 126, 25–517.
- Windmeijer, F. (2008), “GMM for Panel Count Data Models,” *Advanced Studies in Theoretical and Applied Econometrics*, 46, 603–624.

- Wooldridge, J.M. (1990), "Distribution-Free Estimation of Some Nonlinear Panel Data Models," Working Paper No. 564, Department of Economics, Massachusetts Institute of Technology.
- Wooldridge, J.M. (1997), "Multiplicative Panel Data Models without the Strict Exogeneity Assumption," *Econometric Theory*, 13, 667–679.
- Wooldridge, J.M. (2005), "Simple Solutions to the Initial Conditions Problem in Dynamic, Nonlinear Panel Data Models with Unobserved Heterogeneity," *Journal of Applied Econometrics*, 20, 39–54.
- Zeger, S.L., and K.-Y. Liang (1986), "Longitudinal Data Analysis for Discrete and Continuous Outcomes," *Biometrics*, 42, 121–130.

## CHAPTER 9

---

# TREATMENT EFFECTS AND PANEL DATA

---

MICHAEL LECHNER

### 9.1 INTRODUCTION

---

In the last three decades two rapidly developing fields had an immense effect on how microeconometric empirical studies are conducted in our days. On the one hand, the literature on panel econometrics clarified how the increasing availability of panel data sets could improve the estimation of econometric models by exploiting the fact that repeated observations from a unit of the population are available. This led to more precise and more robust estimation strategies. Many of these methods made it into our standard econometric textbooks and became part of the standard econometric curriculum. This handbook, as well as the recently published 3rd edition of the *Econometrics of Panel Data* (Mátyás and Sevestre 2008), give a good account of the latest (as well as less new) developments in this field.

On the other hand, the so-called treatment effects literature exploded over the last two decades as well. It is a major achievement of the econometric treatment effects literature to clarify under which conditions causal effects are nonparametrically identified. This also led to a much better understanding of how to choose appropriate research designs and of “what we are really estimating.” This is particularly so for the case when effects are heterogeneous, which is the prominent case in that literature. Furthermore, this literature, which is not yet as mature as the panel econometrics one, puts emphasis on identifying causal effects under as weak as possible (and plausible) conditions, which limits the role of tightly specified statistical parametric models. To the contrary, non- and semiparametric methods are emphasized. Angrist and Pischke (2010) give a good account of these ideas and show how they influence the way microeconometric studies are done, while Heckman, LaLonde, and Smith (1999), and Imbens and Wooldridge (2009) provide rich surveys of the econometric methods. Angrist

and Pischke (2009) give a (graduate) textbook treatment of this topic,<sup>1</sup> which also received top journal space in the “Forum on Estimation of Treatment Effects” in the *Journal of Economic Literature* (e.g., Deaton 2010; Heckman 2010; Imbens 2010) and the Symposium on “Con out of Econometrics” by the *Journal of Economic Perspectives* (2010).

The first part of this chapter shows how panel data can be used to improve the credibility of methods usually used in the (static) treatment effects literature. These improvements come mainly from the availability of outcome variables measured prior to treatment. However, in addition to improving the credibility of static causal models, panel data are essential for estimating causal effects obtained from dynamic causal models, which is the main theme of the second part of this chapter. This chapter also shows that such dynamic causal effects have only weak links to parameters usually appearing in the dynamic panel data model literature (e.g., Arellano and Bond 1991).

This survey has many omissions indeed. As the panel econometrics literature, as well as the literature on treatment effects, is huge, we had to omit several important topics to stick with the space constraints of such a handbook. First of all, all the semiparametric panel data literature is completely ignored. The interested reader is referred to the chapter by Bo Honoré in this handbook. Duration models are another important omission from this chapter, for reasons of lack of space and not because of lack of relevance, although, for example, in the work of Abbring and van den Berg (2003) there is a clear link between panel data and the identification and estimation of causal effects. Furthermore, we ignore the extensive literature on distributional treatment effects (e.g., Firpo 2007) as well as a substantial part of the more structural dynamics treatment literature (see, e.g., Abbring and Heckman 2007; Heckman and Navarro-Lozano 2007). Finally, recent developments on testing, as well as developing instrumental variable assumptions are ignored as well (for a recent example, see Klein 2010).

Even for the subjects not omitted, this chapter will neither review the whole panel literature nor the whole treatment effects literature. Instead it focuses on parts of the treatment literature where panel data are particularly helpful and important. “Importance” and “helpfulness,” of course, entirely depend on which empirical subject is analyzed, in addition to some degree of subjective judgment. Therefore, this chapter takes an applied perspective in the sense of prominently using an empirical example to exemplify ideas and of treating formal assumptions and properties rather informally (and relating the reader to the corresponding papers in the literature instead). In the same thrust, we exemplify the main ideas in a very simple linear regression setting.

The main empirical example is the evaluation of active labor market programs, which will now be introduced. This literature tries to answer the question of whether the unemployed benefit from participating in some public financed training or employment program. Of course, the effects of interest of such programs have many dimensions, usually including individual reemployment chances and earnings.<sup>2</sup> Many of such studies are based on reasonably large administrative data sets that allow the observation of individuals before, during, and after participating in a program. Thus, usually the empirical analysis is based on panel data. The econometric methods developed in the treatment effects literature are used in the respective empirical

analyses, because their main advantages mentioned above are deemed to be important. Furthermore, a diverse set of different identification and estimation strategies is employed.<sup>3</sup>

In the next section, the static treatment effects model is introduced. We consider three approaches that figure prominently in the applied literature. Starting with matching and regression type methods, we continue with differences-in-differences methods and instrumental variable estimation. In Section 9.3, we discuss matching and regression type methods within a dynamic treatment effects framework. Finally, Section 9.4 concludes and the Appendix contains some derivations omitted from the main body of the text.

## 9.2 THE STATIC TREATMENT MODEL

---

### 9.2.1 Notation

This chapter features the most simple static treatment effects model, namely a model in which the treatment is binary. This simplification allows us to concentrate on the key issues without an overly complex technical apparatus hiding the main insights relevant for empirical work even if more general models are used in some of the applied work. In this model, as in any other treatment effects model, we are interested in the effect of a *ceteris paribus* change of the treatment (e.g., participating in a program),  $D$ , on the outcomes, (e.g., earnings)  $Y$ . To denote the values of the outcomes that would occur if  $D = 1$  or  $D = 0$ , respectively, we define so-called potential outcomes  $Y^d$  (i.e.  $Y^1$  and  $Y^0$ ) (Rubin 1974).<sup>4</sup> By construction, both potential outcomes can never be observed simultaneously. The link between observable and potential outcomes is given by the following observation rule, i.e.  $Y = DY^1 + (1 - D)Y^0$ . The observation rule directly implies that  $E(Y_t^{d_1} | D_1 = d_1) = E(Y_t | D_1 = d_1)$ .

Other variables that play a role as control variables are denoted by  $X$  (e.g., past education or the labor market history), while instrumental variables are denoted by  $Z$ . Depending on the context,  $X$  may be scalars or vectors of random variables. If not mentioned explicitly otherwise,  $X$  and  $Z$  are assumed not to be influenced by the treatment and are in this sense exogenous.

With respect to timing, we assume that the treatment occurs after period 0 (i.e.,  $t = 0$ ) and before period 1. In other words, in the static model the treatment variable  $D_t$  equals 0 prior to period 1. In period 1, it switches from 0 to 1 with positive probability. Otherwise,  $D_t$  is time constant. Thus, fixing/knowing  $D_1$  is like fixing/knowing all  $D_t$ . There may be further pre- and post-treatment periods, denoted by  $t$ . Thus, the random variables consist of several elements over time, which will be stacked on top of each other, for example,  $Y = (\dots, Y_0, Y_1, \dots)'$ . If not mentioned otherwise, all the random variables, describing a larger population of interest (unemployed individuals in our example), and functions thereof are assumed to have as many moments as required

for the particular analysis. Finally, assume that we observe data from  $D$  and  $Y$ , as well as  $X$  and  $Z$  if needed. The data is obtained from  $N$  independent draws from the population described by these random variables. In other words,  $(d_i, y_i, x_i, z_i)$  are i.i.d. in the cross-sectional dimension “ $i$ ” but may be arbitrarily correlated over time.

It is usually the goal of a treatment effect analysis to uncover causal effects aggregated over specific subpopulations while allowing individual causal effects to vary across observations in a general way.<sup>5</sup> In this chapter, we consider the average treatment effect (ATE,  $\gamma_t$ ), the average treatment effect for the treated (ATET),  $\gamma_t(1)$  and non-treated (ATENT),  $\gamma_t(0)$ , as well as the local average treatment effect (LATE,  $\gamma_t(z)$ ), all in a particular period  $t$ .<sup>6</sup> These effects are defined as follows:

$$\begin{aligned}\gamma_t &= E(Y_t^1 - Y_t^0); && \text{(ATE)} \\ \gamma_t(d_1) &= E(Y_t^1 - Y_t^0 | D_1 = d_1), & \forall d_1 \in \{0, 1\}; & \text{(ATET, ATENT)} \\ \gamma_t(z) &= E(Y_t^1 - Y_t^0 | \text{Complier}(z)); & \forall t \in \{\dots, 0, 1, \dots\}. & \text{(LATE)}\end{aligned}$$

In our empirical example, the ATE is the relevant object of estimation when interest is in the expected effect of the program for a randomly chosen unemployed, while ATET will be the expected effect of a program for a randomly chosen participant, and ATENT for a non-participant.<sup>7</sup> Note that ATE can be directly derived from ATET and ATENT, because  $\gamma_t = \gamma_t(1)P(D_1 = 1) + \gamma_t(0)[1 - P(D_1 = 1)]$ . Therefore, we focus most of the discussion only on the ATET, since the arguments for the ATENT are symmetric and ATE can always be obtained by a combination of the two. Note that one of the two terms that appear in the sum that defines ATET and ATENT can be expressed in terms of variables that have sample counterparts, that is,  $E(Y_t^{d_1} | D_1 = d_1) = E(Y_t | D_1 = d_1)$  (and can therefore consistently be estimated by the respective sample mean). However, this is not so for the second term,  $E(Y_t^{d_1} | D_1 = 1 - d_1)$ , called the counterfactual, which requires further assumptions for identification, that is, assumptions required to express the counterfactual in terms of the observable variables  $Y$ ,  $X$ , and  $D$  (and  $Z$ ).

Finally, the LATE parameter, introduced by Imbens and Angrist (1994), measures the mean program effect for a randomly drawn unemployed from a so-called complier population. A complier population is characterized by the fact that for every member a change in the value of the instrument leads to a change in the value of the treatment. Thus, in the empirical example, a complier is a person who would have a different program participation status when faced with different values of the instrument.<sup>8</sup> While the data is informative about who is a participant and who is a non-participant, it is silent about who is a complier. Usually, we expect all these effects to be zero in the pre-treatment periods,  $t \leq 0$ .<sup>9</sup>

In the following, we consider several identification and estimation strategies that are popular in empirical studies trying to uncover causal effects and discuss the value of the availability of panel studies for these strategies. All these strategies provide potential solutions when a direct comparison of the means of  $\gamma_t$  for observations with  $d_{1i} = 1$

(treated) and  $d_{1i} = 0$  (controls) will be confounded by some other observable or unobservable variable that jointly influences the potential outcomes ( $Y_t^d$ ) and the treatment ( $D_1$ ).

To illustrate some of the ideas and to simplify a comparison with standard (linear) panel data methods, we specify simple linear models for the conditional expectations of the potential outcomes. These models will not be the most general possible. In particular, we abstract among other things from effect heterogeneity (implying  $\gamma_t = \gamma_t(1) = \gamma_t(0)$ ), which plays a key role in the treatment effects literature. However, keeping this parametric example simple allows obtaining additional intuition for most major ideas discussed in this survey.<sup>10</sup>

$$\begin{aligned} E(Y_t^0 | X_0 = x_0, D_0 = 0, D_1 = d_1, \dots, D_T = d_T) &= \alpha_t + x_0 \beta_t + d_1 \delta_t, \\ E(Y_t^1 | X_0 = x_0, D_0 = 0, D_1 = d_1, \dots, D_T = d_T) &= \underbrace{\alpha_t + x_0 \beta_t + d_1 \delta_t}_{E(Y_t^0 | X_0 = x_0, D_1 = d_1)} + \gamma_t; \\ \forall x_0 \in X_0, \forall t &= \{\dots, 0, 1, \dots, T\}. \end{aligned}$$

$T$  denotes the final period of data used. This (simple) specification captures the idea that the different groups (defined by treatment status) may exhibit the same effect of the treatment (allowed to be variable over time), but may have different levels in the potential outcomes. If subjects self-select or are selected into the treatments,  $d_1 \delta_t$  may be termed the time-varying conditional-on- $X$  selection effect. The value of  $\delta_t$  is inherently linked to the nature of the “selection process.” For example, if treatment is assigned in a random experiment, then  $\delta_t$  equals zero. If differences of average outcomes between treated and control individuals result from differences in  $x_0$  only, then, again,  $\delta_t$  equals zero.

Using the observation rule, this model leads to the following conditional expectation for  $Y_t$ :

$$E(Y_t | X_0 = x_0, D_t = d_t) = \alpha_t + x_0 \beta_t + d_t(\gamma_t + \delta_t); \quad \forall x_0 \in X_0, \forall t = \{\dots, 0, 1, \dots, T\}.$$

From this equation it is obvious that additional assumptions are necessary in order to obtain consistent estimates of the treatment effect,  $\gamma_t$ , because it is confounded by the selection effect,  $\delta_t$ .

## 9.2.2 Selection on Observables: The Conditional Independence Assumption

### 9.2.2.1 Nonparametric Identification

In this section we analyze the case when information on background variables  $X$  is rich enough such that the potential outcomes are unconfounded (conditionally independent of  $D_1$ ) given  $X$  (conditional independence assumptions, CIA). This is formalized as<sup>11</sup>

$$Y_t^0, Y_t^1 \perp\!\!\!\perp D_1 | X_0 = x_0, \quad x_0 \in \chi_0, \quad \forall t > 0.$$

This assumption states that the potential outcomes are independent (denoted by  $\perp\!\!\!\perp$ ) of treatment in period 1 conditional on  $X_0$  for the values of  $x_0$  in  $\chi_0$ . In addition, assume that there is common support, that is,  $0 < P(D_1|X_0 = x_0) < 1$ , and that  $X_0$  is not influenced by  $D_1$ .<sup>12</sup> These assumptions imply that  $E(Y_t^{d_1}|D_1 = 1 - d_1) = E[E(Y_t|X_0 = x_0, D_1 = d_1)|D_1 = 1 - d_1]$  so that ATE, ATET, and ATENT are identified, that is, they can be expressed in terms of random variables for which realizations are available for all sampled members of the population:

$$\begin{aligned}\gamma_t(1) &= E(Y_t|D_1 = 1) - E[E(Y_t|X_0 = x_0, D_1 = 0)|D_1 = 1], \\ \gamma_t(0) &= E[E(Y_t|X_0 = x_0, D_1 = 1)|D_1 = 0] - E(Y_t|D_1 = 0), \quad \forall x_0 \in \chi_0, \quad \forall t > 0.\end{aligned}$$

For the linear model outlined above, note that as indicated already in the previous section, CIA implies that  $\delta_t = 0$  for  $t > 0$ . Thus, the treatment effects can easily be obtained by standard regression methods.

Why are panel data helpful in this essentially static setting? First, having further post-treatment time periods available allows estimating the dynamics of the effects. Second, whether this selection-on-observable assumption is plausible depends on the particular pre-treatment information available, as the data needs to contain all variables that jointly influence the treatment and the post-treatment potential outcomes. Third, assuming some homogeneity over time, we may argue that the CIA also holds for outcomes prior to the treatment. If this is true, and if the treatment effect is zero for those pre-treatment periods ( $\gamma_t = \gamma_t(1) = \gamma_t(0) = 0, t \leq 0$ ), we may conduct placebo tests (pre-program tests in the language of Heckman and Hotz 1989). If we find statistically significant nonzero effects, this will be an indication that the CIA does not hold prior to treatment. This in turn may indicate that it does not hold in the post-treatment periods either.<sup>13</sup>

Understanding the outcome dynamics in the empirical example is important because many programs have initial negative effects, so-called lock-in effects (e.g., due to a reduced job search while participating in a program). Positive effects, if any, appear only later (see, e.g., Lechner, Miquel, and Wunsch 2011). The issue about using pre-treatment variables to control for confounding may be even more important. In the case of labor market evaluations, Lechner and Wunsch (2013), for example, show the importance of controlling for variables capturing an informative individual labor market history to avoid biased estimation. It is probably true for many applications that key elements of  $X$  are pre-treatment outcome variables ( $X_0 = (Y_{-\tau}, \dots, Y_0, \tilde{X}_0)$ ; the last element in this vector denotes some other exogenous confounders. One reason for this may be that they contain the same or similar unobservables as the post-treatment variables and that such unobservables are likely to be correlated with  $D_1$ . Finally, placebo tests may or may not be an appropriate tool in practice. For example, in many countries unemployment is a requirement to become eligible for the programs of the active

labor market policy. In such case, the sample will be selected such that everybody is unemployed in  $t = 0$  (otherwise there would be no common support). Then, if we estimate an effect for the pre-treatment period  $t = 0$  in a placebo experiment, we will always find a zero effect, at least for the outcome variable unemployment. Thus the test has no power for this variable in this period. As outcomes are likely to be correlated over time, and as different outcome variables, like earnings and various employment indicators, are also correlated in the cross-sectional dimension, the test may generally lack power in such situations.

### 9.2.2.2 Estimation

For our “toy-linear” model, the CIA implies:

$$E(Y_t|X_0 = x_0, D_t = d_t) = \alpha_t + x_0\beta_t + d_t\gamma_t; \quad \forall x_0 \in X_0, \forall t = \{1, \dots, T\}.$$

Thus, the treatment effect,  $\gamma_t$ , is consistently estimated by a cross-sectional regression (in the post-treatment periods) in which the observable outcome,  $Y_t$ , is regressed on  $X_0$  and  $D_1$ . Remember that the vector of confounding variables here includes functions of past outcomes as well as other exogenous variables for which the realized values are known in period 0. Indeed, there is no gain by using panel data methods in this model with linear confounding correction, time varying coefficients, and a treatment effect that does not vary with confounders and treatment, be it fixed or random effects. Cross-sectional regressions for post-treatment periods ( $t > 0$ ) are consistent estimators of the treatment effects. Of course, if some of the coefficients are constant over time, then panel data methods may lead to more efficient estimates. Similar arguments will be valid if conditional expectations of the potential outcomes are nonlinear functions of the confounders, or if the effects vary with the confounders.

However, estimating a parametric model is unnecessarily restrictive, since the identifying assumptions provide nonparametric identification of the mean causal effects. Therefore, it is not surprising that the literature emphasized methods that do not require the parametric assumptions (and the implied restrictions on effect heterogeneity). To estimate the effects nonparametrically, we need a nonparametric regression of  $P(D_1 = 1|X_0 = x_0)$  for weighting-type estimators (Hirano, Imbens, and Ridder 2003). Alternatively, for regression based estimators a nonparametric regression of  $E(Y_t|X_0 = x_0, D_1 = 0)$  for the ATET, of  $E(Y_t|X_0 = x_0, D_1 = 1)$  for the ATENT, or both for the ATE is required (Imbens, Newey, and Ridder 2007). At least one of those nonparametric regressions is also needed for many other nonparametric methods, like for most versions of matching (Rubin 1979). This is the case because in the selection-on-observables framework all methods, whether parametric or nonparametric, are explicitly or implicitly based on adjusting the distribution of  $X_0$  in the  $D_1 = 1$  and  $D_1 = 0$  subsamples such that the adjusted distribution of the confounders is very similar for treated and non-treated. If this is successful, using the same adjustment for the outcome variables gives the desired mean causal effects. The higher the dimension of

$X_0$ , the more difficult it is to create this kind of comparability in all dimensions of  $X_0$ , and thus the curse of dimensionality comes into its damaging play.

The results of Rosenbaum and Rubin (1983) reduce this problem in some sense, as they show that it is sufficient to make those subpopulations comparable with respect to a one-dimensional random variable, instead of the high-dimensional  $X_0$ . This one-dimensional random variable is the so-called propensity score,  $p(X_0)$ , which is the probability of treatment given the confounders, that is,  $p(x_0) := P(D_1 = 1|X_0 = x_0)$ . The methods used most in empirical work are semiparametric in the sense that the propensity score is estimated by (flexible) parametric models. Then this score is used for either weighting, regression-type adjustments, or matching estimation. Since there is nothing specific to panel data when using these methods in this context, we will refer the reader to the excellent surveys by Imbens (2004) and Imbens and Wooldridge (2009). Several Monte Carlo studies compare the performance of the various estimators, like Frölich (2004), Busso, DiNardo, and McCrary (2009a, 2009b) and the very extensive study of Huber, Lechner, and Wunsch (2013). The latter compares more than hundred different estimators using what they call an “empirical Monte Carlo” study, which is a Monte Carlo design that shares many features with real empirical studies. In the latter study, a particular radius matching estimator with bias adjustment showed some superior large and small sample properties.

## 9.2.3 Selection on Unobservables I: Difference-in-Difference Methods

### 9.2.3.1 Semiparametric Identification

Whereas matching-type methods discussed in the previous section may not necessarily require data from different periods, such data are essential for difference-in-difference (DiD) methods. The basic idea of the DiD concept is to have (at least) four different subsamples available for the empirical analysis: one group that has already been subject to the treatment (observed in  $t > 0$ ), one group that will be subject to the treatment in the future (observed in  $t \leq 0$ ), and another two groups not subject to the treatment that are observed in the same periods as the two treatment groups. If the treatment has no effect in period 0 and if the outcomes of the treatment and non-treatment groups develop in the same fashion over time (usually called either “common-trend” or “bias stability” assumption), then, conceptionally, we may either (i) use period “0” to estimate the bias of any estimator based on selection-on-observables (since the true effect is 0 in period 0) and use this estimate to purge the similar estimate in  $t > 0$  from this bias, or (ii) use the change of the outcome variables of the non-treatment group over time together with the pre-treatment outcomes of the future treated to estimate what would have happened to the treated group in  $t > 0$  had they not been treated.

These ideas are indeed old and can at least be traced back to a paper by Snow (1855). He was interested in whether cholera was transmitted by (bad) air or (bad) water.

Snow (1855) used a change in the water supply in one district of London, namely the switch from polluted water taken from the Thames in the center of London to a supply of cleaner water taken upriver, to isolate the effect of the water quality from other confounders. In our days there are many applications of these methods, mainly in applied microeconomics. They are also well explained in most modern econometric textbooks (see, e.g., the excellent discussions in Angrist and Pischke 2009). Since these methods are also contained in several excellent surveys on treatment effects (e.g., Blundell and Costa Dias 2009; Imbens and Wooldridge 2009), I keep this section brief and reiterate a few panel data related points that appeared in my recent survey on DiD estimation (Lechner 2011a).

The common-trend assumption,

$$\begin{aligned} & E(Y_1^0 | X_0 = x_0, D_1 = 1) - E(Y_0^0 | X_0 = x_0, D_1 = 1) \\ &= E(Y_1^0 | X_0 = x_0, D_1 = 0) - E(Y_0^0 | X_0 = x_0, D_1 = 0) \\ &= E(Y_1^0 | X_0 = x_0) - E(Y_0^0 | X_0 = x_0), \quad \forall x_0 \in \chi_0, \end{aligned}$$

together with the assumptions that (i)  $D_1$  has no effect prior to treatment, i.e.  $\gamma_t(d) = 0, t \leq 0$ , (ii) the covariates are not influenced by the treatment, and that (iii) there is the necessary common support,  $\gamma_t(1)$ , are identified.<sup>14</sup> Note that identification is not nonparametric (as in the previous section) in the sense that the validity of the common-trend assumption depends on the chosen transformation (unit of measurement) of the outcome variables. In other words, if the common-trend assumption is deemed correct, for example, for earnings it will be violated for nonlinear transformations of earnings, like log-earnings (at least for non-trivial cases). As it is usually difficult to explain why such an assumption should only be valid for a particular functional form, this is a limitation of this method (see the generalization of Athey and Imbens 2006, which is however more difficult to apply).<sup>15</sup>

Going back to our “toy” model and forming the differences for the non-participation potential outcomes, we obtain the following expressions:

$$\begin{aligned} E(Y_t^0 - Y_\tau^0 | X_0 = x_0, D_1 = 1) &= (\alpha_t - \alpha_\tau) + x_0(\beta_t - \beta_\tau) + (\delta_t - \delta_\tau), \\ E(Y_t^0 - Y_\tau^0 | X_0 = x_0, D_1 = 0) &= (\alpha_t - \alpha_\tau) + x_0(\beta_t - \beta_\tau); \quad \tau \in \{\dots, 0\}, t \in \{1, \dots, T\}. \end{aligned}$$

Thus, the required condition for the common-trend assumption to hold is that the impact of the selection effect is time constant  $(\delta_t - \delta_\tau) = 0$ , at least for the (minimum of) two periods used for estimation. This may seem to be somewhat more general than in the case of CIA, which required the absence of post-treatment confounding (i.e.,  $\delta_t = 0$ ). However, since obviously  $\delta_t = 0$  does not imply  $(\delta_t - \delta_\tau) = 0$ , the two methods are not nested. On top of this, DiD also requires the absence of pre-treatment effects ( $\gamma_\tau = 0, \tau \leq 0$ ).<sup>16</sup> Under these assumptions, we obtain the following conditional expectations for the observable outcome variables:

$$E(Y_t | X_0 = x_0, D_1 = d_1, \dots, D_T = d_T) = \alpha_t + x_0\beta_t + \underline{1}(t > 0)d_t\gamma_t + d_1\delta.$$

$\underline{1}(\cdot)$  denotes the indicator function, which is one if its argument is true. Thus, the treatment effects can be recovered by regression methods.

### 9.2.3.2 Estimation

The name of the estimation strategy is already indicative of the underlying estimation principle in general. If the common-trend assumption holds conditional on  $X_0$ , then the estimate of the effect conditional on  $X$  can be obtained by forming the differences of the pre- and post-treatment periods' outcomes of the treated and subtracting the differences of the pre- and post-treatment periods' outcomes of the non-treated. In the (virtual) second step the conditional-on- $X$  effects are averaged with weights implied by the distribution of  $X$  among the treated.

For our simple linear model, this leads to the specification given above. Clearly, the treatment effects cannot be estimated with one period of data alone because of the presence of the selection term,  $d_1\delta$ , which was absent in the model of the section discussing the selection-on-observables only. Within a post-treatment cross-section this selection effect cannot be distinguished from the causal effect  $\underline{1}(t > 0)d_t\gamma_t$ . Further note that panel data are not necessary for estimating the parameters of this equation. Repeated cross-sections of at least one pre-treatment and one post-treatment period are sufficient as long as they contain also the information about the (past and future) treatment status and confounders.

If the linear model is not deemed to be appropriate for modeling the conditional expectations, then there are some nonlinear and/or less parametric methods available, many of which are discussed in Lechner (2011a). Therefore, for the sake of brevity the interested reader is referred to that paper.

### 9.2.3.3 The Value of Panel Data Compared to Repeated Cross-Sections

In the previous sections, we saw that panel data allowed us to (i) follow the outcome dynamics, (ii) compute more informative control variables, and (iii) check the credibility of the identifying assumptions with placebo tests. While (ii) always requires panel data, for (i) and (iii) it is only essential to have data from additional periods (so that repeated cross-sections are sufficient). The same is true for DiD.

If panel data are available, the linear DiD estimator can be estimated by fixed effects methods.<sup>17</sup> One consequence of basing the estimator on individual differences over time is that all influences of time constant confounding factors that are additively separable from the remaining part of the conditional expectations of the potential outcomes are removed by the DiD-type of differencing. Therefore, it is not surprising that adding fixed individual effects instead of the treatment group dummy  $d$  in the regression formulation leads to the same quantity to be estimated (e.g., Angrist and Pischke 2009). This way it becomes obvious, as it was for our “toy”-model, that the usual advantages attributed to fixed effects models, like controlling of time constant endogeneity and selectivity within a linear setting, are also advantages of the difference-in-difference approach.

Furthermore, from the point of view of identification, a substantial advantage of panel data is that matching estimation based on conditioning on pre-treatment outcomes is feasible as well. This is an important issue because it appears to be a natural requirement for a “good” comparison group to have similar pre-treatment means of the outcome variables (because it is likely that pre-treatment outcomes are correlated with post-treatment outcomes as well as selection, either directly, or because the unobservables that influence those three quantities are correlated).<sup>18</sup> This conditioning is not possible with repeated cross-sections, since we do not observe pre- and post-treatment outcomes of the same individuals.

The corresponding matching-type assumptions for the case when lagged outcome variables are available (and used) imply the following:

$$E(Y_t^0 | Y_0 = y_0, X_0 = x_0, D_1 = 1) = E(Y_t^0 | Y_0 = y_0, X_0 = x_0, D_1 = 0), \quad \forall t > 0.$$

Imbens and Wooldridge (2009) observe that the common-trend assumption and this matching-type assumption impose different identifying restrictions on the data which are not nested and must be rationalized based on substantive knowledge about the selection process, that is, only one of them can be true. Angrist and Krueger (1999) elaborate on this issue based on regression models and come to the same conclusions.

The advantage of the DiD method, as mentioned before, is that it allows for time constant confounding unobservables ( $\delta_t \neq 0$ ) while requiring common-trends ( $\delta_t = \delta_\tau$ ), whereas matching does not require common-trends ( $\delta_t \neq \delta_\tau$ ) but assumes that conditional on pre-treatment outcomes confounding unobservables are irrelevant ( $\delta_t = 0$ ). As  $\delta_t, \delta_\tau$  capture the effects of variables jointly influencing selection, as well as outcomes, their interpretation depends on the conditioning sets used. For example, if the selection process is entirely governed by  $x_0$  and  $y_\tau$ , then controlling for those variables implies  $\delta_t = 0$ . In this case matching may be used and there is no need for any assumptions concerning the selection process in period  $\tau$ . More generally, one may argue that conditioning on the past outcome variables already controls for the part of the unobservables that manifested themselves in the lagged outcome variables.

One may try to combine the positive features of both methods by including pre-treatment outcomes among the covariates in a DiD framework. This is however identical to matching: taking the difference while keeping the pre-treatment part of that difference constant at the individual level in any comparison (i.e., the treated and matched control observations have the same pre-treatment level) is equivalent to just ignoring the differencing of DiD and to focus on the post-treatment variables alone. Thus, such a procedure implicitly requires the matching assumptions. In other words, assuming common-trends conditional on the start of the trend (which means it has to be the same starting point for treated and controls) is practically identical to assuming no confounding (i.e., that the matching assumptions hold) conditional on past outcomes.

Thus, Imbens and Wooldridge’s (2009, p. 70) conclusion about the usefulness of DiD in panel data compared to matching is negative: “As a practical matter, the DiD

approach appears less attractive than the unconfoundedness-based approach in the context of panel data. It is difficult to see how making treated and control units comparable on lagged outcomes will make the causal interpretation of their difference less credible, as suggested by the DiD assumptions.” However, Chabé-Ferret (2012) gives several examples in which a difference-in-difference strategy leads to a consistent estimator while matching conditional on past outcomes may be biased. However, even for those examples given, the assumptions necessary for the consistency of DiD require substantive knowledge on how the selection bias impacts the potential outcomes, which are similar to our toy-model. He also shows simulations that indicate that for the case when the assumptions for matching on lagged outcomes, as well as for DiD, are not exactly fulfilled, both estimators are biased, but matching appears to be more robust than DiD. He concludes that for the cases for which one or the other set of assumptions is not clearly preferred on theoretical grounds, results from both estimation strategies should be presented.

## 9.2.4 Selection on Unobservables II: Instrumental Variables

### 9.2.4.1 Nonparametric Identification

Either when selection-on-observables or differences-in-differences approaches are not credible, or when the instrument-specific LATE parameter is the more interesting parameter compared to the ATE, ATET, or ATENT,<sup>19</sup> then instrumental variable estimation may be the method of choice. The seminal paper by Imbens and Angrist (1994) increased considerably our understanding of which kind of causal effect is estimated by 2SLS when effects are heterogeneous. This literature was further extended by Heckman (1997), Vytlacil (2002), and Heckman and Vytlacil (2005) for continuous instruments, as well as Abadie (2003) and Frölich (2007) for ways to deal with covariates. These papers also clarify that with heterogeneous effects the IV assumptions have to be strengthened somewhat. In other words, on top of the assumption that the instrument has no effect on the outcomes other than by changing the treatment (exclusion restriction, no direct effect assumption), the assumption that a change in the instrument affects the treatment only in one direction (i.e., it either increases or decreases treatment probability for all), the so-called monotonicity assumption, is required as well.

As before, the key question for this chapter is about the role of panel data in IV estimation. As before, the first benefit panel data provide is that observing more post-treatment outcomes allows uncovering how the effects of the treatment in period 1 evolve over time. Second, instruments may not be valid unconditionally and current period control variables may not help as they might already be affected by the treatment. In this case observing more pre-treatment variables may be very helpful. In our example of active labor market policy evaluation, Frölich and Lechner (2010) used an

instrument that measured on which side of a regional borderer within a local labor market an unemployed lived. The rationale for this instrument was that this fact mattered for their program participation probability but not (directly) for their labor market success. The concern in the paper was that they might have chosen one or the other side of the border by considerations that could involve other characteristics, like tax rates and past labor market success. These factors may be however related to outcomes via different channels than program participation thus violating the exclusion restriction. With panel data, we are able to condition on such past events and thus improve the credibility of the instrument. The third benefit one might derive from panel data is that past values of some variables that are not time constant may provide instruments. A word of caution is in order in this instance, because there are a couple of empirical papers that use lagged outcomes as instruments without giving the explicit reasoning that would justify doing so. This is somewhat at odds with the arguments made in the previous section about the value of lagged outcomes as a confounding control variable, because by definition a confounder has a direct effect on outcomes thus violating one of the key assumptions required for consistent IV estimation. In other words, as it is likely that those lagged outcome variables depend on the same unobservable than the current period outcome variables do, one needs very explicit arguments why this should not matter with respect to the exclusion restriction in the particular study at hand. The fourth benefit of panel data, namely placebo tests, is that it may allow estimating effects for periods in which the true effect is known to be zero (and the instrument is valid as well), thus providing some empirical evidence on the credibility of the instrument.

#### *9.2.4.2 Estimation*

The easiest way to conceptualize the linear model is to follow exactly the same steps as for selection-on-observables, and to assume that one of the confounders that are contained in the required conditioning set  $X_0$  is unobservable. Thus, the linear model for the observable outcomes derived above cannot be a basis for consistent estimation by regression methods. Note that by the definition of a confounder as a variable jointly correlated with treatments and outcomes, this leads to the endogeneity of  $D_t$  in the regression formulated in terms of observable variables. In this case, and if a valid instrument is available, the panel econometric IV methods for linear models, described for example in Baltagi (2008) and Biorn and Krishnakumar (2008), may be applied to obtain estimates that are consistent under the linearity and homogeneity assumptions discussed in the previous section.

For non- or semiparametric estimation similar problems concerning the dimension of the confounders in case of selection-on-observables occur. Frölich (2007) showed that the IV estimate is a ratio of estimators that would be consistent under a no-confounding assumption of the relation of the instrument and the outcome. In fact, IV estimates can be obtained by dividing the effect of the instrument  $Z$  on the outcome  $Y_t$  by the effect of  $Z$  on  $D_1$  each time controlling for variables,  $X_0$ , that are jointly related to the instrument and to the outcome or the treatment.<sup>20</sup> Since these are similar

estimation strategies as described in the section on selection-on-observables the same tools for reducing the dimension are available. The only difference is, of course, that the propensity score in this case is the probability of the binary instrument (instead of the treatment) being one given the confounders, that is,  $p^z(x_0) := P(D_1 = 1|X = x_0)$ .

## 9.3 A DYNAMIC TREATMENT MODEL

---

### 9.3.1 Motivation and Basic Structure of the Model

The static treatment model, which is widely used in micro econometrics even when panel data are used, allows for dynamics in the sense that the effects of *the* treatment,  $D_1$ , are allowed to vary over time, and that variables measured in pre-treatment periods were used to tackle the confounding problem in different ways. The treatment itself, however, was not allowed to change over time more than once (from period 0 to period 1). In this section, we present a model that allows for more treatment dynamics. For such a model, the availability of panel data is essential.

Robins (1986) suggested an explicitly dynamic causal framework based on potential outcomes. It allows the definition of causal effects of dynamic interventions and clarifies the resulting endogeneity and selectivity problems. Identification is achieved by sequential selection-on-observable assumptions (see Abbring 2003, for a comprehensive summary).<sup>21</sup> His approach was subsequently applied in epidemiology and biostatistics (e.g., Robins 1989, 1997, 1999; Robins, Greenland, and Hu 1999, for discrete treatments; Gill and Robins 2001, for continuous treatments; and many other applications by various authors) to define and estimate the effect of time-varying treatments in discrete time. It is common in that literature to estimate the effects by parametric models usually based on the so-called G-computation algorithm as suggested by Robins (1986).

Lechner and Miquel (2010, LM10 hereinafter) extend Robins's (1986) framework to different causal parameters. Since the assumptions used in LM10 are similar to the selection-on-observables or conditional independence assumptions (CIA) of the static model, Lechner (2009b) proposed dynamic extensions of the matching and inverse-probability-weighting estimators discussed above, which are more robust than parametric models. The applications of this approach in economics are limited so far.<sup>22</sup> One reason is that this approach, in particular in its semiparametric and nonparametric form requires larger and more informative data than required for estimating causal effects in a static treatment effects model.

Below, the definitions of the dynamic causal model as well as the identification results derived by Robins (1986) and LM10 are briefly reviewed.<sup>23</sup> To ease the notational burden, we use a three-periods-binary-treatment model to discuss the most relevant issues that distinguish the dynamic from the static model.<sup>24</sup> Using again our

labor market program evaluation example for illustration, suppose that, as before, there is an initial pre-treatment period,  $D_0 = 0$ , plus two subsequent periods in which different treatment states (participation in a program) are realized. Denote the history of variables up to period  $t$  by a bar below that variable, that is,  $\underline{d}_2 = (0, d_1, d_2)$ .<sup>25</sup> Therefore, in this setting all treatment combinations are fully described by the four sequences  $(0, 0)$ ,  $(1, 0)$ ,  $(0, 1)$ , and  $(1, 1)$ . The potential outcomes are indexed by these treatment combinations,  $Y_t^{\underline{d}_1}$  ( $t \geq 1$ ) or  $Y_t^{\underline{d}_2}$  ( $t \geq 2$ ). They are measured at the end of each period, whereas treatment status is measured at the beginning of each period. For each sequence length of length of one or two periods (plus the initial period), one of the respective potential outcomes is observable:

$$\begin{aligned} Y_t &= D_1 Y_t^1 + (1 - D_1) Y_t^0, \quad \forall t \geq 1; \\ Y_t &= D_1 Y_t^1 + (1 - D_1) Y_t^0 = D_1 D_2 Y_t^{11} + (1 - D_1) D_2 Y_t^{01} \\ &\quad + D_1 (1 - D_2) Y_t^{10} + (1 - D_1) (1 - D_2) Y_t^{00}; \quad \forall t \geq 2. \end{aligned}$$

Finally, note that the confounders,  $X_t$ , will be explicitly considered to be time varying and may contain functions of  $Y_t$ . Like the outcomes they are observable at the end of each period.

As for the static model, the causal effect of the sequences is formalized using averages of potential outcomes. The following expression defines the causal effect (for period  $t$ ) of a sequence of treatments up to period 1 or 2,  $\underline{d}_{\tau}^k$ , compared to an alternative sequence of the same or a different length,  $\underline{d}_{\tau}'$ , for a population defined by one of those sequences or a third sequence,  $\underline{d}_{\tau}^j$ :

$$\begin{aligned} \gamma_{\tau}^{\underline{d}_{\tau}^k, \underline{d}_{\tau}'^j}(\underline{d}_{\tau}^j) &= E(Y_t^{\underline{d}_{\tau}^k} | \underline{D}_{\tilde{\tau}} = \underline{d}_{\tilde{\tau}}^j) - E(Y_t^{\underline{d}_{\tau}'^j} | \underline{D}_{\tilde{\tau}} = \underline{d}_{\tilde{\tau}}^j), \\ 0 \leq \tilde{\tau}; 1 \leq \tau, \tau' \leq 2, \tilde{\tau} &\leq \tau', \tau; \tilde{\tau}, \tau', \tau \leq t; \\ k \neq l, k &\in (1, \dots, 2^{\tau}), l \in (1, \dots, 2^{\tau'}), j \in (1, \dots, 2^{\tau}). \end{aligned}$$

The treatment sequences indexed by  $k$ ,  $l$ , and  $j$  may correspond to  $d_1 = 0$  or  $d_1 = 1$  if  $\tau$  (or  $\tau'$ ) denotes period 1, or to the longer sequences  $(d_1, d_2) = (0, 0), (0, 1), (1, 0)$ , or  $(1, 1)$  if  $\tau$  (or  $\tau'$ ) equals two. LM10 call  $\gamma_{\tau}^{\underline{d}_{\tau}^k, \underline{d}_{\tau}'^j}$  the dynamic average treatment effect (DATE). Accordingly,  $\gamma_{\tau}^{\underline{d}_{\tau}^k, \underline{d}_{\tau}'^j}(\underline{d}_{\tau}^k)$  is termed DATE on the treated (DATET). There are also cases in-between, like  $\gamma_{\tau}^{\underline{d}_{\tau}^k, \underline{d}_{\tau}'^j}(\underline{d}_{\tau}^l)$ , for which the conditioning set is defined by a sequence shorter than the one defining the causal contrast. Finally, note that the effects are symmetric for the same population ( $\gamma_{\tau}^{\underline{d}_{\tau}^k, \underline{d}_{\tau}'^j}(\underline{d}_{\tau}^k) = -\gamma_{\tau}^{\underline{d}_{\tau}'^j, \underline{d}_{\tau}^k}(\underline{d}_{\tau}^k)$ , but  $\gamma_{\tau}^{\underline{d}_{\tau}^k, \underline{d}_{\tau}'^j}(\underline{d}_{\tau}^k) \neq \gamma_{\tau}^{\underline{d}_{\tau}^k, \underline{d}_{\tau}'^j}(\underline{d}_{\tau}'^j)$ ). This feature, however, does not restrict effect heterogeneity.

Let us now postulate a simple linear model that serves the same purpose as in the case of the static model:

$$E(Y_t^0 | X_0 = x_0, D_1 = d_1) = \alpha_{1,t} + x_0 \beta_{1,t} + d_1 \delta_{1,t},$$

$$E(Y_t^1 | X_0 = x_0, D_1 = d_1) = \alpha_{1,t} + \underbrace{x_0 \beta_{1,t} + d_1 \delta_{1,t}}_{E(Y_t^0 | X_0 = x_0, D_1 = d_1)} + \gamma_t^1, \quad \forall x_0 \in \chi_0,$$

$$\forall t = \{\dots, 0, 1, \dots, T\}.$$

Therefore, for the “observable” outcomes the observation rule implies the following:

$$E(Y_t | X_0 = x_0, D_1 = d_1) = \alpha_{1,t} + x_0 \beta_{1,t} + d_1 (\gamma_t^1 + \delta_{1,t}).$$

Note that this part of the specification that relates to the effects of the treatments in period 1 only is specified exactly as for the static model to ease comparison (with the exception of not conditioning on all  $D_t$ , which has a different meaning in the dynamic than in the static model). Therefore, it is also clear that identification of  $\gamma_{1,t}$  is exactly as for the static model discussed in the previous section. Therefore, from now on we concentrate on sequences that include treatment status in period 2 as well.

The key features that the toy model for dynamic treatments is supposed to capture the impact of confounders already influenced by the treatment in period 1 as well as the selection effects that come from selecting into  $D_1$  in period 1 and into  $D_2$  in period 2. The following specifications contain these features while keeping all other complications to a minimum:

$$E(Y_t^{00} | \underline{X}_1 = \underline{x}_1, \underline{D}_T = \underline{d}_T) = \alpha_{2,t} + x_0 \beta_{2,t} + y_1 \phi_t + d_1 d_2 \lambda_t^{11}$$

$$+ (1 - d_1) d_2 \lambda_t^{01} + d_1 (1 - d_2) \lambda_t^{10};$$

$$E(Y_t^{d_1', d_2'} | \underline{X}_1 = \underline{x}_1, \underline{D}_T = \underline{d}_T) = \underbrace{\alpha_{2,t} + x_0 \beta_{2,t} + y_1 \phi_t + d_1 d_2 \lambda_t^{11}}_{E(Y_t^{00} | \underline{X}_1 = \underline{x}_1, \underline{D}_T = \underline{d}_T)}$$

$$+ (1 - d_1) d_2 \lambda_t^{01} + d_1 (1 - d_2) \lambda_t^{10}$$

$$+ \gamma_t^{d_1' d_2'}; \quad \forall x_0, y_1, d_2, d_1.$$

Note that the coefficients  $\lambda_t^{d_1, d_2}$  denote the subsample specific selection effects. Next, we derive the conditional expectation of the observable outcome for this case using the observation rules given above:

$$E(Y_t | \underline{X}_1 = \underline{x}_1, \underline{D}_T = \underline{d}_T) = E(Y_t | \underline{X}_1 = \underline{x}_1, \underline{D}_2 = \underline{d}_2) = \alpha_{2,t} + x_0 \beta_{2,t} + y_1 \phi_t$$

$$+ d_1 d_2 (\gamma_t^{11} + \lambda_t^{11}) + (1 - d_1) d_2 (\gamma_t^{01} + \lambda_t^{01}) + d_1 (1 - d_2) (\gamma_t^{10} + \lambda_t^{10}).$$

In a similar fashion as before, this equation shows that the treatment effects are not identified without further assumptions that concern the selection effects,  $\lambda_t^{d_1 d_2}$ . It is also important to note that this model is not a typical dynamic panel data model, as the conditioning is not on  $y_{t-1}$ , but on  $y_1$  (i.e., it does not depend on  $t$ ).

### 9.3.2 Identification

The weak dynamic conditional independence assumption (W-DCIA) postulates that the variables that jointly influence selection at each stage of the sequence as well as the outcomes are observable in the time period corresponding to that stage:

- (a)  $Y_t^{00}, Y_t^{10}, Y_t^{01}, Y_t^{11} \coprod D_1 | X_0 = x_0,$
- (b)  $Y_t^{00}, Y_t^{10}, Y_t^{01}, Y_t^{11} \coprod D_2 = d_2 | D_1 = d_1, X_1 = x_1, \forall x_1 \in \chi_1, \forall t \geq 1.$

$\underline{\chi}_1 = (\chi_0, \chi_1)$  denotes the support of  $X_0$  and  $X_1$ . Part (a) of W-DCIA states that the potential outcomes are independent of treatment choice in period 1 ( $D_1$ ) conditional on  $X_0$ . This is the standard version of the static CIA. Part (b) states that conditional on the treatment in period 1 and on the confounding variables of periods 0 and 1,  $\underline{X}_1$ , potential outcomes are independent of participation in period 2 ( $D_2$ ).

In the Appendix, it is shown that using this assumption and the observation rule gives us the relation between shorter and longer sequences of potential outcomes (which also provides the link between the static and the dynamic models):

$$\begin{aligned} E(Y_t^{d_1} | D_1 = d_1) &= E(Y_t^{d_1 1} | D_1 = d_1, D_2 = 1) p^{D_2}(d_1) \\ &\quad + E(Y_t^{d_1 0} | D_1 = d_1, D_2 = 0) [1 - p^{D_2}(d_1)]; \\ E(Y_t^{1-d_1} | D_1 = d_1) &= \underset{X_0 | D_1 = d_1}{E} [E(Y_t^{(1-d_1)1} | X_0 = x_0, D_1 = 1 - d_1, D_2 = 1) p^{D_2|X}(x_0, 1 - d_1)] \\ &\quad + \underset{X_0 | D_1 = d_1}{E} [E(Y_t^{(1-d_1)0} | X_0 = x_0, D_1 = 1 - d_1, D_2 = 0) \\ &\quad [1 - p^{D_2|X}(x_0, 1 - d_1)]]; \\ p^{D_2}(d_1) &:= P(D_2 = 1 | D_1 = d_1); \quad p^{D_2|X}(x_0, d_1) := P(D_2 = 1 | X_0 = x_0, D_1 = d_1). \end{aligned}$$

Thus, the expectation of the outcomes of the shorter sequences is a weighted average of the expectation of the two longer sequences that have the same first period treatment as the shorter sequence.

To see whether the W-DCIA is plausible in our example, the question is which variables influence program participation in each period as well as subsequent labor market outcomes and whether such variables are observable. If the answer to the latter question is yes (and if there is common support, i.e., there are individuals with the same observable characteristics that are observed in both treatment sequences of interest), then there is identification, even if some or all conditioning variables in period 2 are influenced by the labor market and program participation outcomes of period 1. LM10 show that, for example, quantities that are for subpopulations defined by treatment status in period 1 or 0 only, like  $E(Y_2^{11})$ ,  $E(Y_2^{11} | D_1 = 0)$ , and  $E(Y_2^{11} | D_1 = 1)$ , are identified. Mean potential outcomes for subpopulations defined by treatment status in period 1 and 2 are only identified if the sequences coincide in the first period

(e.g.,  $E[Y_2^{11} | \underline{D}_2 = (1, 0)]$ ). However,  $E[Y_2^{11} | \underline{D}_2 = (0, 0)]$  or  $E[Y_2^{11} | \underline{D}_2 = (0, 1)]$  are not identified. Thus,  $\gamma_t^{\underline{d}_t^k, \underline{d}_t^l}$  and  $\gamma_t^{\underline{d}_t^k, \underline{d}_t^l}(\underline{d}_1^j)$  are identified  $\forall \underline{d}_1^k, \underline{d}_2^k, \underline{d}_1^l, \underline{d}_2^l, \underline{d}_1^j, \underline{d}_2^j \in \{0, 1\}$ , but  $\gamma_2^{\underline{d}_2^k, \underline{d}_2^l}(\underline{d}_2^j)$  is not identified if  $\underline{d}_1^l \neq \underline{d}_2^k$ ,  $\underline{d}_1^l \neq \underline{d}_2^j$ , or  $\underline{d}_2^k \neq \underline{d}_2^j$ . The relevant distinction between the populations defined by participation states in period 1 and subsequent periods is that in period 1, treatment choice is random conditional on exogenous variables, which is the result of the initial condition stating that  $D_0 = 0$  holds for everybody. However, in the second period, randomization into these treatments is conditional on variables already influenced by the first part of the treatment. W-DCIA has an appeal for applied work as a natural extension of the static framework. However, W-DCIA is not strong enough to identify the classical treatment effects on the treated which would define the population of interest using one of the complete sequences (for all three periods), if the sequences of interest differ in period 1.

Let us now consider identification in our linear example. Note that part (b) of the W-DCIA implies that  $E(Y_t^{d_1^l, d_2^l} | \underline{X}_1 = \underline{x}_1, \underline{D}_T = \underline{d}_T) = E(Y_t^{d_1^l, d_2^l} | \underline{X}_1 = \underline{x}_1, D_1 = d_1)$ , thus  $\lambda_t^{11} = \lambda_t^{10} (= \lambda_t^1)$  and  $\lambda_t^{01} = \lambda_t^{00} (= 0)$ , thus we have:

$$\begin{aligned} E(Y_t^{00} | \underline{X}_1 = \underline{x}_1, \underline{D}_T = \underline{d}_T) &= E(Y_t^{00} | \underline{X}_1 = \underline{x}_1, D_1 = d_1) = \alpha_{2,t} + x_0\beta_{2,t} + y_1\phi_t + d_1\lambda_t^1; \\ E(Y_t^{d_1^l, d_2^l} | \underline{X}_1 = \underline{x}_1, \underline{D}_T = \underline{d}_T) &= E(Y_t^{d_1^l, d_2^l} | \underline{X}_1 = \underline{x}_1, D_1 = d_1) \\ &= \alpha_{2,t} + x_0\beta_{2,t} + y_1\phi_t + d_1\lambda_t^1 + \gamma_t^{d_1^l, d_2^l}. \\ E(Y_t | \underline{X}_1 = \underline{x}_1, \underline{D}_T = \underline{d}_T) &= \alpha_{2,t} + x_0\beta_{2,t} + y_1\phi_t + d_1 d_2 (\gamma_t^{11} + \lambda_t^1) + \\ &\quad + (1 - d_1) d_2 \gamma_t^{01} + d_1 (1 - d_2) (\gamma_t^{10} + \lambda_t^1). \end{aligned}$$

Thus,  $\gamma_t^{01}$  as well as  $\alpha_{2,t}, \beta_{2,t}, \phi_t$  are identified. Furthermore since part (a) of the W-DCIA implies that  $\delta_{1,t} = 0$ , the effects of the treatment of the first period,  $\alpha_{1,t}, \beta_{1,t}, \gamma_t^1$ , are identified as well. However, there is still a selection effect,  $\lambda_t^1$ , that hinders full identification of the causal effects (i.e., only  $(\gamma_t^{11} + \lambda_t^1)$  and  $(\gamma_t^{10} + \lambda_t^1)$  are identified from this regression).

However, this is already enough because the treatment effects are linked in the sense that  $\gamma_t^1$ , by definition, must be a weighted average of  $\gamma_t^{11}$  and  $\gamma_t^{10}$ . Indeed the Appendix shows that  $\gamma_t^1 = (\gamma_t^{11} - \gamma_t^{10}) \frac{E}{X_0 | D_1 = 0} [P(D_2 = 1 | X_0 = x_0, D_1 = 1)] - \gamma_t^{01} P(D_2 = 1 | D_1 = 0) + \gamma_t^{10}$  holds in this model, which provides identification (because (i)  $\frac{E}{X_0 | D_1 = 0} [P(D_2 = 1 | X_0 = x_0, D_1 = 1)]$  and  $P(D_2 = 1 | D_1 = 0)$  are identified; (ii)  $\gamma_t^{11} - \gamma_t^{10}$  can be expressed in terms of identified terms and  $\gamma_t^{10}$ ; (iii) thus  $(\gamma_t^{11} + \lambda_t^1)$  and  $(\gamma_t^{10} + \lambda_t^1)$  depend only on two further unknowns and have, for non-trivial cases, a solution).

For the general model, LM10 show that to identify all treatment parameters, W-DCIA must be strengthened by essentially imposing that the confounding variables used to control selection into the treatment of the second period are not influenced by

the selection into the first-period treatment. This can be summarized by an independence condition like  $Y_t^{d_2} \perp\!\!\!\perp D_2 | X_1 = \underline{x}_1$  (LM10 call this the strong dynamic conditional independence assumption, *S-DCIA*). Note that the conditioning set includes the outcome variables from the first period. This is the usual conditional independence assumption used in the multiple static treatment framework (with four treatments; see Imbens 2000; Lechner 2001). In other words, when the control variables (including the outcome variables in period 1) are not influenced by the previous treatments, the dynamic problem collapses to a static problem of four treatments with selection on observables. An example of such a situation would be an assignment to two subsequent training programs that was made already before the first program began and for which there is no chance to drop out once assigned to both programs.

Any attempt of nonparametrically estimating these effects faces the same problem that distributional adjustments based on a potentially high-dimensional vector of characteristics and intermediate outcomes are required. However, as before for the static case, propensity scores are available to allow the construction of semiparametric estimators (see LM10 for details).

### 9.3.3 Estimation

Lechner (2008, L08 hereinafter) shows that for the model using W-DCIA these propensity scores are convenient tools for constructing sequential propensity score matching and reweighting estimators.<sup>26</sup> Using such propensity scores, the following identification result based on W-DCIA is the key ingredient for building appropriate estimators:

$$\begin{aligned} E(Y_2^{d_2^k} | D_1 = d_1^j) &= E_{p^{d_1^k}(x_0)} \{ E_{p^{d_2^k|d_1^k}(\underline{x}_1)} [E(Y_2 | D_2 = \underline{d}_2^k, p^{d_2^k|d_1^k, d_1^k}(\underline{x}_1)) | D_1 \\ &= d_1^k, p^{d_1^k}(x_0)] | D_1 = d_1^j\}, p^{d_2^k|d_1^k, d_1^j}(X_1) := [p^{d_2^k|d_1^k}(\underline{X}_1), p^{d_1^j}(X_0)], \\ &\forall d_1^k, d_2^k, d_1^j, d_1 \in \{0, 1\}, \end{aligned}$$

where  $p^{d_2^k|d_1^k}(\underline{x}_1) := p^{d_2^k|d_1^k} P(D_2 = \underline{d}_2^k | D_1 = d_1^k, \underline{X}_1 = \underline{x}_1)$  and  $p^{d_1^j}(x_0) = P(D_1 = d_1^j | X_0 = x_0)$  are the respective participation probabilities. To learn the counterfactual outcome for the population participating in  $d_1^j$  (the target population) had they participated in the sequence  $\underline{d}_2^k$ , characteristics (and thus outcomes) of observations with  $\underline{d}_2^k$  must be reweighted to make them comparable to the characteristics of the observations in the target population ( $d_1^j$ ). The dynamic, sequential structure of the causal model restricts the possible ways to do so. Intuitively, for the members of the target population, observations that share the first element of the sequence of interest ( $d_1^k$ ) should be reweighted such that they have the same distribution of  $p^{d_1^k}(X_0)$  as the target population. Call this artificially created group comparison group one. Yet, to estimate the effect of the

full sequence, the outcomes of observations that share  $\underline{d}_2^k$  instead of  $d_1^k$  are required. Thus, an artificial subpopulation of observations in  $\underline{d}_2^k$  that has the same distribution of characteristics of  $p_{\underline{d}_1^k}(X_0)$  and  $p_{\underline{d}_2^k|d_1^k}(\underline{X}_1)$  as the artificially created comparison group 1 is required. The same principle applies for dynamic average treatment effects in the population (DATE).

All proposed estimators in L08 have the same structure: they are computed as weighted means of the outcome variables observed in the subsample  $D_2 = \underline{d}_2^k$ . The weights,  $w(\cdot)$ , depend on the specific effects of interest and are functions of the balancing scores.

$$\begin{aligned}\widehat{E(Y_2^{d_2^k}|D_1=d_1^j)} &= \sum_{i \in \underline{d}_2^k} w_i^{d_2^k, d_1^j} (\underline{p}^{d_2^k|d_1^k, d_1^k}(\underline{x}_{1,i}), d_1^k) y_i; \quad w_i^{d_2^k, d_1^j} \geq 0; \quad \sum_{i \in \underline{d}_2^k} w_i^{d_2^k, d_1^j} = 1 \\ (9.1) \quad \widehat{E(Y_2^{d_2^k})} &= \sum_{i \in \underline{d}_2^k} w_i^{d_2^k} (\underline{p}^{d_2^k|d_1^k, d_1^k}(\underline{x}_{1,i}), d_1^k) y_i; \quad w_i^{d_2^k} \geq 0; \quad \sum_{i \in \underline{d}_2^k} w_i^{d_2^k} = 1.\end{aligned}\quad (9.2)$$

It remains to add a note on estimation of our linear toy model. Following the considerations in the identification part, the model consists of two linear regressions based on the following two equations:

$$\begin{aligned}E(Y_t|X_0=x_0, D_1=d_1) &= \alpha_{1,t} + x_0 \beta_{1,t} + d_1 \gamma_t^1; \\ E(Y_t|\underline{X}_1=\underline{x}_1, \underline{D}_2=\underline{d}_2) &= \alpha_{2,t} + x_0 \beta_{2,t} + y_1 \phi_t + d_1 d_2 (\gamma_t^{11} + \lambda_t^1) \\ &\quad + (1-d_1) d_2 \gamma_t^{01} + d_1 (1-d_2) (\gamma_t^{10} + \lambda_t^1).\end{aligned}$$

In a second step the estimated coefficients together with the link between the one-period and two-period treatment effects are used to uncover the causal effects. Since these effects are assumed to be homogeneous, the W-DCIA is sufficient to identify all relevant quantities of this model. It is important to note, though, that the outlined procedure is very different from estimating a classical dynamic or static linear or nonlinear panel data model.

## 9.4 CONCLUDING REMARKS

---

For many empirical applications, panel data are essential for the credible identification and precise estimation of causal effects. The first part of this chapter, which discussed matching and instrumental variable estimation in the static treatment model, showed how the additional information provided by panel data could be used to measure pre-treatment variables that improve the credibility of those strategies. Furthermore, if several post-treatment periods were available, more interesting effects capturing the outcome dynamics could be estimated. The latter was also true for the so-called

difference-in-difference approach, although the use of the pre-treatment outcomes differ for this approach: lagged outcomes do not appear in the conditioning set but were used instead to form pre-treatment/post-treatment outcome differences. Thus, the latter approach is not robust to nonlinear transformations of the outcome variables while the former two approaches are robust to such transformations. Another difference between IV, matching, and difference-in-difference approaches is that for the latter panel data are not strictly necessary as repeated cross-sections will do. Finally, for all approaches based on a static treatment framework panel data may allow for so-called placebo tests (i.e., estimating effects for periods for which it is known that they should be zero). Such tests are another tool of improving the credibility of the chosen identifying assumptions.

The second part of this chapter showed how panel data could be used to identify and estimate causal parameters derived from dynamic treatment effect models, an area that did not yet receive much attention in econometrics. Therefore, the results on non-parametric identification and non- or semiparametric estimation are mainly limited to the case of imposing sequential selection-on-observable assumptions, a case which is popular in other fields as well, like epidemiology. It is a perhaps a surprising insight from this analysis that the parameters usually estimated by linear parametric panel data models and the causal parameters derived from the dynamic treatment models are only loosely related.

There are still many open ends in this literature. For example, in the dynamic treatment models instrumental variable estimation seems to be rather unexplored, while for the static models we just start to understand when it makes more sense to use lagged outcome variables as covariates instead of forming differences and apply a difference-in-difference approach instead. In conclusion, we can expect the intersection of the literatures on panel data and treatment effects to produce many interesting research papers in the near future.

## APPENDIX: RELATION OF DIFFERENT POTENTIAL OUTCOMES IN THE DYNAMIC TREATMENT MODEL

---

In this appendix, we provide the derivations that lead to the link of the potential outcomes of sequences of length one and two. First, define the following short-cut notation for the conditional selection probabilities.

$$p^{D_2}(d_1) := P(D_2 = 1|D_1 = d_1); \quad p^{D_2|X}(x_0, d_1) := P(D_2 = 1|X_0 = x_0, D_1 = d_1).$$

Next, we use the observation rule to establish the desired relation:

$$E[D_1 Y_t^1 + (1 - D_1) Y_t^0 | D_1 = 1] = E[Y_t^1 | D_1 = 1] = E[D_2 Y_t^{11} + (1 - D_2) Y_t^{10} | D_1 = 1].$$

Using iterated expectations, for the general case we obtain the following expression:

$$\begin{aligned} E(Y_t^{d_1} | D_1 = d_1) &= E(Y_t^{d_1} | D_1 = d_1, D_2 = 1)p^{D_2}(d_1) + E(Y_t^{d_1} | D_1 = d_1, D_2 = 0) \\ [1 - p^{D_2}(d_1)] &= E(Y_t^{d_11} | D_1 = d_1, D_2 = 1)p^{D_2}(d_1) \\ &+ E(Y_t^{d_10} | D_1 = d_1, D_2 = 0)[1 - p^{D_2}(d_1)]. \end{aligned}$$

Next, we want to establish a similar link for the mean counterfactual  $E(Y_t^{d_1} | D_1 = 1 - d_1)$ , which requires the use of part (a) of the W-DCIA.

$$\begin{aligned} E(Y_t^{1-d_1} | D_1 = d_1, X_0 = x_0) &\stackrel{W-DCIA\ a)}{=} E(Y_t^{1-d_1} | X_0 = x_0, D_1 = 1 - d_1) \Rightarrow \\ E(Y_t^{1-d_1} | D_1 = d_1) &= \underset{X_0|D_1=d_1}{E} E(Y_t^{1-d_1} | X_0 = x_0, D_1 = 1 - d_1) = \\ &= \underset{X_0|D_1=d_1}{E} [E(Y_t^{(1-d_1)1} | X_0 = x_0, D_1 = 1 - d_1, D_2 = 1)p^{D_2|X}(x_0, 1 - d_1)] \\ &+ \underset{X_0|D_1=d_1}{E} [E(Y_t^{(1-d_1)0} | X_0 = x_0, D_1 = 1 - d_1, D_2 = 0)[1 - p^{D_2|X}(x_0, 1 - d_1)]] \end{aligned}$$

This formula can now be used to connect the treatment effects as well:

$$\begin{aligned} \gamma_t^1(1) &= E(Y_t^1 | D_1 = 1) - E(Y_t^0 | D_1 = 1) = \\ &= E(Y_t^{11} | D_1 = 1, D_2 = 1)p^{D_2}(1) + E(Y_t^{10} | D_1 = 1, D_2 = 0)[1 - p^{D_2}(1)] - \\ &- \underset{X_0|D_1=1}{E} [E(Y_t^{01} | X_0 = x_0, D_1 = 0, D_2 = 1)p^{D_2|X}(x_0, 0)] \\ &- \underset{X_0|D_1=1}{E} [E(Y_t^{00} | X_0 = x_0, D_1 = 0, D_2 = 0)[1 - p^{D_2|X}(x_0, 0)]]; \end{aligned}$$

$$\begin{aligned} \gamma_t^1(0) &= \underset{X_0|D_1=0}{E} \left[ E(Y_t^{11} | X_0 = x_0, D_1 = 1, D_2 = 1)p^{D_2|X}(x_0, 1) \right] + \\ &+ \underset{X_0|D_1=0}{E} \left[ E(Y_t^{10} | X_0 = x_0, D_1 = 1, D_2 = 0) \left[ 1 - p^{D_2|X}(x_0, 1) \right] \right] - \\ &- \left[ E(Y_t^{01} | D_1 = 0, D_2 = 1)p^{D_2}(0) + E(Y_t^{00} | D_1 = 0, D_2 = 0) \left[ 1 - p^{D_2}(0) \right] \right]. \end{aligned}$$

Finally, we consider the special case of the dynamic linear toy model postulated for  $\gamma_t^1(0)$ :

$$\begin{aligned} \gamma_t^1 = \gamma_t^1(1) = \gamma_t^1(0) &= \underset{X_0|D_1=0}{E} \left[ E(Y_t^{11} | X_0 = x_0, D_1 = 1, D_2 = 1)p^{D_2|X}(x_0, 1) \right] + \\ &+ \underset{X_0|D_1=0}{E} \left[ E(Y_t^{10} | X_0 = x_0, D_1 = 1, D_2 = 0) \left[ 1 - p^{D_2|X}(x_0, 1) \right] \right] - \\ &- \left[ E(Y_t^{01} | D_1 = 0, D_2 = 1)p^{D_2}(0) + E(Y_t^{00} | D_1 = 0, D_2 = 0) \left[ 1 - p^{D_2}(0) \right] \right] = \end{aligned}$$

$$\begin{aligned}
&= \underset{X_0|D_1=0}{E} \left[ E(Y_t^{11} - Y_t^{10}|X_0 = x_0, D_1 = 1, D_2 = 1) p^{D_2|X}(x_0, 1) \right. \\
&\quad \left. + E(Y_t^{10}|X_0 = x_0, D_1 = 1, D_2 = 0) \right] - \\
&\quad - \left[ E(Y_t^{01} - Y_t^{00}|D_1 = 0, D_2 = 1) p^{D_2}(0) + E(Y_t^{00}|D_1 = 0, D_2 = 0) \right] = \\
&= \underset{X_0|D_1=0}{E} \left[ E(Y_t^{11} - Y_t^{10}|X_0 = x_0, D_1 = 1, D_2 = 1) p^{D_2|X}(x_0, 1) \right] - \\
&\quad - \left[ E(Y_t^{01} - Y_t^{00}|D_1 = 0, D_2 = 1) p^{D_2}(0) \right] + \\
&\quad + \underset{X_0|D_1=0}{E} \left[ E(Y_t^{10}|X_0 = x_0, D_1 = 1, D_2 = 0) \right] - E(Y_t^{00}|D_1 = 0, D_2 = 0) = \\
&= \underset{X_0|D_1=0}{E} \left[ E(Y_t^{11} - Y_t^{10}|X_0 = x_0, D_1 = 0, D_2 = 1) p^{D_2|X}(x_0, 1) \right] - \\
&\quad - \left[ E(Y_t^{01} - Y_t^{00}|D_1 = 0, D_2 = 1) p^{D_2}(0) \right] + \\
&\quad + \underset{X_0|D_1=0}{E} \left[ E(Y_t^{10} - Y_t^{00}|X_0 = x_0, D_1 = 1, D_2 = 0) \right] = \\
&= (\gamma_t^{11} - \gamma_t^{10}) \underset{X_0|D_1=0}{E} \left[ p^{D_2|X}(x_0, 1) \right] - \gamma_t^{01} p^{D_2}(0) + \gamma_t^{10}.
\end{aligned}$$

## ACKNOWLEDGMENTS

---

I am also affiliated with CEPR and PSI, London, CESIfo, Munich, IAB, Nuremberg, and IZA, Bonn. I thank Hugo Bodory for carefully reading the manuscript and Martin Huber as well as two anonymous referees for helpful comments and suggestions. The usual disclaimer applies.

## NOTES

---

1. This book also has a chapter on panel data but for the special case of difference-in-difference estimation, which so far provided the main formal link between treatment effects and panel data.
2. The meta study by Card, Kluve, and Weber (2010) gives a comprehensive overview of recent studies in that field, and Lechner, Miquel, and Wunsch (2011) provide a recent prototypical example for Germany.
3. For a different field applying treatment effects models with panel data, see, e.g., Lechner (2009a), who analyzes the impact of individual sports activity on labor market outcomes using the German GSOEP panel data.
4. Capital letters denote random variables, while small letter denote either their realizations or fixed values.

5. Again, by considering only averages the focus is on the simplest case, but conceptionally most of the considerations below carry over to quantile treatment effects, or other objects for which the knowledge of the marginal distribution of the potential outcome is sufficient. Furthermore, we also do not discuss a large range of other parameters that relate, for example, to continuous instruments (see Heckman and Vytlacil 2005, for an extensive discussion of such parameters).
6. By some inconsistency of notation,  $\gamma_t(d_1)$  refers to a specific population defined by the value of  $d_1$ , while  $\gamma_t(z)$  refers to some population that is implied by the use of a specific instrument  $z$  (see below).
7. Note that this way of coding  $D$  implies that we measure the effect of the intervention that occurred in period 1 only. Analyzing further changes in  $D$  is relegated to the section on dynamic treatment models.
8. Indeed, there may be two such groups with become either more or less likely to receive treatment for an identical change of the instrument. Usually, one of those groups, called defiers, is assumed to be absent (see Imbens and Angrist 2004).
9. If this is not true, for example, due to changing behavior in anticipation of treatment, then it is sometimes possible to adjust the calendar date of the treatment (i.e., period 0) just prior to the first period when such reaction could be expected.
10. Note that these simple linear specifications sometimes allow specialized identification and estimation strategies exploiting these parametric features. As this is not the purpose of this example, such cases will be ignored in the discussions below.
11.  $A \perp\!\!\!\perp B/C$  means that *each element* of the vector of random variables  $B$  is independent of each element of the random vector  $A$  conditional on the random vector  $C$  taking values of  $c$  in the sense of Dawid (1979).
12. See the excellent survey by Imbens (2004), who extensively discusses this case.
13. To be precise, such tests can be informative about confounders that are simultaneously related to the current treatment and the past and current outcomes.
14. This literature usually attempts only to identify effects for the treated. Although identifying effects for non-treated would technically just involve a redefinition of the treatment, this setting is usually unattractive in empirical studies, because it requires three treated groups one of which becomes non-treated from period 0 to period 1.
15. See Lechner (2011a) for more discussion on how to deal with nonlinearities in this approach.
16. Note that although this is not required by CIA in general, once pre-treatment outcomes are used as covariates they must be exogenous. Of course, this is only plausible in the absence of pre-treatment effects.
17. The remaining part of this section follows closely section 3.2.8 of Lechner (2011a).
18. Note that although such an intuition of controlling for more information is plausible in many applications, it is easy to create an example with a larger and a smaller conditioning set for which CIA holds in the *smaller* but not in the larger set.
19. For example, Frölich and Lechner (2010) analyze the effects of active labor market programs and argue that the compliers that relate to their instruments are close to a population that would join the programs if they were marginally extended. In fact, for the policy question about the effects of extending the programs estimating such a parameter would be more interesting than estimating the ATE, the ATET, or the ATENT.

20. Note that although the same notation  $X$  is used here for both variables, usually these “instrument confounders” may be different from the “treatment confounders” that are required under the CIA.
21. Until now identification of dynamic treatment models by instrumental variable methods and generalized difference-in-difference methods is a rather unexplored area, although there are some results in Miquel (2002, 2003), that await further research. Therefore, this section focuses entirely only on the sequential-selection-on-observable approach proposed by Robins (1986) in his seminal paper. Alternative reduced form approaches have been suggested, for example, by Fitzenberger, Osikominu, and Paul (2010).
22. Exceptions are Lechner and Wiesler (2013), who analyze the effects of the timing and order of Austrian active labor market programs and LM10 who analyze the effects of the German active labor market policies. A further exception is Ding and Lehrer (2003), who use this framework and related work by Miquel (2002, 2003) to evaluate a sequentially randomized class size study using difference-in-difference-type estimation methods. Lechner (2008) discusses practical issues when using this approach for labor market evaluations.
23. The dynamic potential outcome framework is also useful to compare concepts of causality used in microeconomics and time series econometrics (see Lechner 2011b, for details).
24. As before, there may be more periods available to measure pre- or post-treatment outcomes though.
25. Therefore, the first element of this sequence,  $d_0$ , is mainly ignored in the notation as it does not vary.
26. Of course, other static matching-type estimators (e.g., Huber, Lechner, and Wunsch 2013) can be adapted to the dynamic context in a similar way.

## REFERENCES

---

- Abadie, A. (2003), “Semiparametric Instrumental Variable Estimation of Treatment Response Models,” *Journal of Econometrics*, 113, 231–263.
- Abbring, J. H. (2003), “Dynamic Econometric Program Evaluation,” IZA, Discussion Paper, 804.
- Abbring, J. H., and J. J. Heckman (2007), “Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation,” in J.J. Heckman and E. E. Leamer (eds.), *Handbook of Econometrics*, Vol. 6B, 5145–5303. Amsterdam: Elsevier.
- Abbring, J. H., and G. J. van den Berg (2003), “The Nonparametric Identification of Treatment Effects in Duration Models,” *Econometrica*, 71, 1491–1517.
- Angrist, J. D., and A. B. Krueger (1999), “Empirical Strategies in Labor Economics,” in O. Ashenfelter and D. Card (eds.), *Handbook of Labor Economics*, Vol. III A, Ch. 23, 1277–1366. Amsterdam: Elsevier.
- Angrist, J. D., and J.-S. Pischke (2009), *Mostly Harmless Econometrics*, New York: Princeton University Press.
- Angrist, J. D., and J.-S. Pischke (2010), “The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics,” *Journal of Economic Perspectives*, 24, 3–30.

- Arellano, M., and S. Bond (1991), "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *Review of Economic Studies*, 58, 277–297.
- Athey, S., and G. W. Imbens (2006), "Identification and Inference in Nonlinear Difference-in-Difference Models," *Econometrica*, 74, 431–497.
- Baltagi, B. (2008), *Econometric Analysis of Panel Data*. Chichester: Wiley.
- Biorn, E., and J. Krishnakumar (2008), "Measurement Errors and Simultaneity," in L. Mátyás and P. Sevestre (eds.), *The Econometrics of Panel Data*, Chapter 10, 323–367. Berlin, Heidelberg: Springer.
- Blundell, R., and M. Costa Dias (2009), "Alternative Approaches to Evaluation in Empirical Microeconomics," *Journal of Human Resources*, 44, 565–640.
- Busso, M., J. DiNardo, and J. McCrary (2009a), "Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects," *Journal of Business and Economic Statistics* (forthcoming).
- Busso, M., J. DiNardo, and J. McCrary (2009b), "New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators," IZA Discussion Paper 3998.
- Card, D., J. Klueve, and A. Weber (2010), "Active Labour Market Policy Evaluations: A Meta-Analysis," *Economic Journal*, 120, F452–F477.
- Chabé-Ferret, S. (2012), "Matching vs. Differencing when Estimating Treatment Effects with Panel Data: The Example of the Effect of Job Training Programs on Earnings," Toulouse School of Economics Working Paper, 12–356.
- Dawid, A. P. (1979), "Conditional Independence in Statistical Theory," *Journal of the Royal Statistical Society B*, 41, 1–31.
- Deaton, A. (2010), "Instruments, Randomization and Learning about Development," *Journal of Economic Literature*, 48, 424–455.
- Ding, W., and S. F. Lehrer (2003), "Estimating Dynamic Treatment Effects from Project STAR," mimeo.
- Firpo, S. (2007), "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica*, 75, 259–276.
- Fitzenberger, B., Osikominu, A., and M. Paul (2010), "The Heterogeneous Effects of Training Incidence and Duration on Labor Market Transitions," IZA Discussion Paper No. 5269.
- Frölich, M. (2004), "Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators," *Review of Economics and Statistics*, 86, 77–90.
- Frölich, M. (2007), "Nonparametric IV Estimation of Local Average Treatment Effects with Covariates," *Journal of Econometrics*, 139, 35–75.
- Frölich, M., and M. Lechner (2010), "Exploiting Regional Treatment Intensity for the Evaluation of Labour Market Policies," *Journal of the American Statistical Association*, 105, 1014–1029.
- Gill, R. D., and J. M. Robins (2001), "Causal Inference for Complex Longitudinal Data: The Continuous Case," *The Annals of Statistics*, 1–27.
- Heckman, J. J. (1997), "Instrumental Variables," *Journal of Human Resources*, 32, 441–462.
- Heckman, J. J. (2010), "Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy," *Journal of Economic Literature*, 48, 356–398.
- Heckman, J. J., and V. J. Hotz (1989), "Choosing among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," *Journal of the American Statistical Association*, 84, 862–880.

- Heckman, J., and S. Navarro-Lozano (2007), "Dynamic Discrete Choice and Dynamic Treatment Effects," *Journal of Econometrics*, 136, 341–396.
- Heckman, J. J., and E. Vytlacil (2005), "Causal Parameters, Structural Equations, Treatment Effects and Randomized Evaluation of Social Programs," *Econometrica*, 73, 669–738.
- Heckman, J. J., R. J. LaLonde, and J. A. Smith (1999), "The Economics and Econometrics of Active Labor Market Programs," in O. Ashenfelter and D. Card (eds.), *Handbook of Labor Economics*, Vol. 3A, Chapter 31, 1865–2097, New York: North-Holland.
- Hirano, K., G. W. Imbens, and G. Ridder (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189.
- Huber, M., M. Lechner, and C. Wunsch (2013), "The Performance of Estimators Based on the Propensity Score," *Journal of Econometrics*, 175, 1–21.
- Imbens, G. W. (2000), "The Role of the Propensity Score in Estimating Dose-Response Functions," *Biometrika*, 87, 706–710.
- Imbens, G. W. (2004), "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *Review of Economics and Statistics*, 86, 4–29.
- Imbens, G. W. (2010), "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)," *Journal of Economic Literature* 48, 399–423.
- Imbens, G. W., and J. D. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.
- Imbens, G. W., and J. M. Wooldridge (2009), "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86.
- Imbens, G. W., W. Newey, and G. Ridder (2007), "Mean-squared-error Calculations for Average Treatment Effects," IRP discussion paper.
- Klein, T. J. (2010), "Heterogeneous Treatment Effects: Instrumental Variables without Monotonicity?" *Journal of Econometrics*, 155(2), 99–116.
- Lechner, M. (2001), "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption," in M. Lechner and F. Pfeiffer (eds.), *Econometric Evaluation of Active Labour Market Policies*, 43–58, Heidelberg: Physica.
- Lechner, M. (2008), "Matching Estimation of Dynamic Treatment Models: Some Practical Issues," in D. Millimet, J. Smith, and E. Vytlacil (eds.), *Advances in Econometrics*, Volume 21, *Modelling and Evaluating Treatment Effects in Econometrics*, 289–333. Bingley: Emerald Group Publishing Limited.
- Lechner, M. (2009a), "Long-Run Labour Market and Health Effects of Individual Sports Activities," *Journal of Health Economics*, 28, 839–854.
- Lechner, M. (2009b), "Sequential Causal Models for the Evaluation of Labor Market Programs," *Journal of Business & Economic Statistics*, 27, 71–83.
- Lechner, M. (2011a), "The Estimation of Causal Effects by Difference-in-Difference Methods," *Foundations and Trends in Econometrics*, 4/3, 165–224.
- Lechner, M. (2011b), "The Relation of Different Concepts of Causality used in Time Series and Microeconomics," *Econometric Reviews*, 30, 109–127.
- Lechner, M., and R. Miquel (2010), "Identification of the Effects of Dynamic Treatments by Sequential Conditional Independence Assumptions," *Empirical Economics*, 39, 111–137.
- Lechner, M., and S. Wieseler (2013), "Does the Order and Timing of Active Labor Market Programs Matter?" *Oxford Bulletin of Economics and Statistics*, 72(2), 180–212.
- Lechner, M., and C. Wunsch (2013), "Sensitivity of Matching-Based Program Evaluations to the Availability of Control Variables," *Labour Economics*, 21, 111–121.

- Lechner, M., R. Miquel, and C. Wunsch (2011), “Long-Run Effects of Public Sector Sponsored Training in West Germany,” *Journal of the European Economic Association*, 9, 742–784.
- Mátyás, L., and P. Sevestre (2008), *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*, 3rd edn., Berlin-Heidelberg: Springer.
- Miquel, R. (2002), “Identification of Dynamic Treatments Effects by Instrumental Variables,” University of St. Gallen, Department of Economics, Discussion Paper, 2002-11.
- Miquel, R. (2003), “Identification of Effects of Dynamic Treatments with a Difference-in-Differences Approach.” University of St. Gallen, Department of Economics, Discussion paper, 2003-06.
- Robins, J. M. (1986), “A New Approach to Causal Inference in Mortality Studies with Sustained Exposure Periods—Application to Control of the Healthy Worker Survivor Effect,” *Mathematical Modeling*, 7, 1393–1512, with 1987 “Errata to: A New Approach to Causal Inference in Mortality Studies with Sustained Exposure Periods—Application to Control of the Healthy Worker Survivor Effect,” *Computers and Mathematics with Applications*, 14, 917–921; 1987 “Addendum to: A New Approach to Causal Inference in Mortality Studies with Sustained Exposure Periods—Application to Control of the Healthy Worker Survivor Effect,” *Computers and Mathematics with Applications*, 14, 923–945; and 1987 Errata to: “Addendum to ‘A New Approach to Causal Inference in Mortality Studies with Sustained Exposure Periods—Application to Control of the Healthy Worker Survivor Effect,’” *Computers and Mathematics with Applications*, 18, 477.
- Robins, J. M. (1989), “The Analysis of Randomized and Nonrandomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies,” in L. Sechrest, H. Freeman, and A. Mulley (eds.), *Health Service Research Methodology: A Focus on Aids*, 113–159, Washington, DC: Public Health Service, National Centre for Health Services Research.
- Robins, J. M. (1997), “Causal Inference from Complex Longitudinal Data: Latent Variable Modelling and Applications to Causality,” in M. Berkane (ed.), *Lecture Notes in Statistics* (120), 69–117, New York: Springer.
- Robins, J. M. (1999), “Association, Causation, and Marginal Structural Models,” *Synthese*, 121, 151–179.
- Robins, J. M., S. Greenland, and F. Hu, (1999), “Estimation of the Causal Effect of a Time-varying Exposure on the Marginal Mean of a Repeated Binary Outcome,” *Journal of the American Statistical Association*, 94, 687–700.
- Rosenbaum, P. R., and D. B. Rubin (1983), “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” *Biometrika*, 70, 41–55.
- Rubin, D. B. (1974), “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies,” *Journal of Educational Psychology*, 66, 688–701.
- Rubin, D. B. (1979), “Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies,” *Journal of the American Statistical Association*, 74, 318–328.
- Snow, J. (1855), *On the Mode of Communication of Cholera*, 2nd edn., London: John Churchill.
- Vytlačil, E. (2002), „Independence, Monotonicity and Latent Variable Models: An Equivalence Result,” *Econometrica*, 70, 331–341.

## CHAPTER 10

---

# NONPARAMETRIC PANEL DATA REGRESSION MODELS

---

YIGUO SUN, YU YVETTE ZHANG, AND QI LI

### 10.1 INTRODUCTION

---

THE increasing availability of panel data has nourished fast-growing development in panel data econometrics analysis. Textbooks and survey articles have been published to help readers get a quick jump-start into this exciting field. For parametric panel data analysis, the most popularly cited econometrics textbooks include Arellano (2003), Baltagi (2005), and Hsiao (2003). For non-/semi-parametric panel data analysis, recent survey articles include Arellano and Honore (2001), Ai and Li (2008), and Su and Ullah (2011). Arellano and Honore (2001) and Ai and Li (2008), reviewed parametric and semiparametric panel data censored, discrete choice, and sample selective models with fixed effects, while Su and Ullah (2011) focused on nonparametric panel data models, partially linear and varying coefficient panel data models and nonparametric panel data models with cross-sectional dependence and under nonseparability. This chapter joins the others to provide the readers with a selective survey on nonparametric panel data analysis in the framework of conditional mean or conditional quantile regressions. Unavoidably, our survey has some overlaps with Su and Ullah's (2011) survey on nonparametric panel data estimation.

Given a panel data set  $\{(X_{i,t}, Y_{i,t}) : i = 1, \dots, n; t = 1, \dots, T\}$ , we consider a nonparametric panel data model

$$Y_{i,t} = m(X_{i,t}) + u_{i,t}, \quad i = 1, 2, \dots, n, \quad t = 1, \dots, T, \quad (1.1)$$

where  $Y_{i,t}$  and  $X_{i,t}$  are real-valued random variables,  $m(\cdot)$  is an unknown smooth measurable function, and  $u_{i,t}$  is the error term. This chapter mainly focuses on estimation methods and test statistics developed for nonparametric panel data models, with the

exception that semiparametric models as well as nonparametric models are included in the presence of cross-sectional dependent errors in Section 10.2.3.

Note that model (1.1) does allow  $X_{i,t}$  to be a random vector of discrete and/or continuous components. However, we choose to consider the case that  $X_{i,t}$  is a continuously distributed scalar random variable out of two concerns. First, the estimation methods and test statistics included in this chapter can be naturally extended to the case that  $X_{i,t}$  is a random vector. However, this extra generality increases notation complexity in matrix representation which might hurt the readers' focus on grasping the essence of the methodologies explained in the chapter. Second, the curse-of-dimensionality becomes evident in finite samples if  $X_{i,t}$  contains more than two continuous variables, which makes semiparametric models strong competitors to pure nonparametric regression models. We refer the readers to Li and Racine's (2007) textbook for popularly studied semiparametric models for cross-sectional and time series data.

This rest of this chapter is organized as follows. Section 10.2 focuses on nonparametric estimation of conditional mean regression models in panel data setup. Section 10.3 is devoted to nonparametric conditional quantile regression models. Section 10.4 discusses nonseparable panel data models. Section 10.5 contains nonparametric tests on poolability and cross-sectional independence/uncorrelationess. The last section concludes.

## 10.2 CONDITIONAL MEAN REGRESSION MODELS

---

Consider a decomposition of the error term  $u_{i,t}$  into the following three cases:  
(a)  $u_{i,t} = \mu_i + \epsilon_{i,t}$  with  $\mu_i \sim i.i.d.(0, \sigma_\mu^2)$ ; (b)  $u_{i,t} = \lambda_t + \epsilon_{i,t}$  with  $\lambda_t \sim i.i.d.(0, \sigma_\lambda^2)$ ;  
(c)  $u_{i,t} = \mu_i + \delta_i^T g_t + \epsilon_{i,t}$ , where  $\epsilon_{i,t} \sim i.i.d.(0, \sigma_\epsilon^2)$ , and  $\delta_i$  and  $g_t$  are unobservable  $d \times 1$  vectors independent of the idiosyncratic errors  $\{\epsilon_{i,t}\}$ . Case (a) and case (b) assume that the error term contains a cross-sectional and time fixed effect, respectively. In case (c),  $g_t$  are unobserved time varying common factors and  $\delta_i$  are factor loadings. As case (b) can be studied in the same spirit as case (a), we choose not to consider case (b) in this survey.

This section contains three subsections, where Section 10.2.1 discusses random effects panel data models, Section 10.2.2 deals with fixed effects panel data models, and Section 10.2.3 allows for cross-sectional dependence.

### 10.2.1 Random Effects Panel Data Models

This subsection focuses on random-effects panel data models where  $E(u_{i,t}|X_{i,t}) \equiv 0$  for all  $i$  and  $t$  for large  $n$  and small  $T$ . And,  $\{(X_{i,t}, Y_{i,t})\}$  are assumed to be i.i.d. across index  $i$ .

with possible within-group correlation, so that  $V = E(uu^T|X) = \text{diag}(V_1, \dots, V_n)$  with  $V_i = E(u_i u_i^T|X_i)$ , where  $w_i = (w_{i,1}, \dots, w_{i,T})^T$  and  $w = (w_1^T, \dots, w_n^T)^T$  with  $w = u$  or  $X$ . In the presence of within-group correlation, the  $(nT) \times (nT)$  conditional covariance matrix  $V$  is not diagonal. We review the current literature on whether and how to take into consideration the within-group correlation, while estimating  $m(\cdot)$  nonparametrically.

Assuming that the unknown function  $m(\cdot)$  is continuously twice-differentiable around an interior point of interest,  $x$ , and applying Taylor expansion give

$$m(X_{i,t}) = m(x) + m'(x)(X_{i,t} - x) + r_{i,t},$$

where  $r_{i,t} = O((X_{i,t} - x)^2)$  is negligible if  $X_{i,t}$  is sufficiently close to  $x$ . Plugging this approximation into model (1.1) yields

$$Y_{i,t} \approx m(x) + m'(x)(X_{i,t} - x) + u_{i,t}. \quad (2.1)$$

Denoting  $w = (w_{1,1}, \dots, w_{1,T}, w_{2,1}, \dots, w_{2,T}, \dots, w_{n,1}, \dots, w_{n,T})^T$  for  $w = Y, X$  or  $u$ , and a 2 by 1 vector  $\beta(x) = [m(x), m'(x)]^T$ , we calculate the local linear weighted least square (LLWLS) estimator of  $\beta(x)$  as follows

$$\hat{\beta}(x) = \arg \min_{\beta(x)} [Y - Z(x)\beta(x)]^T W(x) [Y - Z(x)\beta(x)] \quad (2.2)$$

where the typical row vector of the  $(nT)$  by 2 matrix  $Z(x)$  is  $(1, X_{i,t} - x)$  and  $W(x)$  is a  $(nT)$  by  $(nT)$  weighting matrix depending on  $x$ . Taking the first order derivative with respect to  $\beta(x)$  gives a nonparametric *estimating equation*

$$Z(x)^T W(x) [Y - Z(x)\hat{\beta}(x)] = 0, \quad (2.3)$$

which yields

$$\hat{\beta}(x) = [Z(x)^T W(x) Z(x)]^{-1} Z(x)^T W(x) Y. \quad (2.4)$$

Taking  $W(x) \equiv K_h(x)$  yields the local linear least squares (LLS) estimator of  $\beta(x)$ , where the typical element of the  $(nT)$  by  $(nT)$  diagonal matrix  $K_h(x)$  is  $K((X_{i,t} - x)/h)$ , and  $K(\cdot)$  is a second-order kernel function and  $h$  is the bandwidth. Suppose that  $h \rightarrow 0$  as  $n \rightarrow \infty$  such that  $K((X_{i,t} - x)/h)$  assigns a positive weight to  $X_{i,t}$  only if  $|X_{i,t} - x| \leq ch$  for some positive constant  $c$ . The LLS estimator evidently does not take into account possible correlation between  $u_{i,t}$  and  $u_{i,s}$  for  $t \neq s$ . In order to account for the structure of the panel data covariance matrix, Ullah and Roy (1998) set  $W(x) \equiv V^{-1/2} K_h(x) V^{-1/2}$ , which is equivalent to applying a generalized least squares transformation through  $V^{-1/2}$ , then applying the local linear approach to the transformed data. If  $V$  were known,  $V^{-1/2}u$  becomes a sequence of i.i.d. errors. Lin and Carroll (2000) proposed two alternative estimators: one sets  $W(x) \equiv \sqrt{K_h(x)} V^{-1} \sqrt{K_h(x)}$ , which is equivalent to localizing points closer to  $x$  first then weighting the chosen data by  $V^{-1/2}$ ; the other is a hybrid estimator with

$W(x) \equiv V^{-1}K_h(x)$ . We denote Ullah and Roy's (1998) estimator by  $\hat{\beta}_{UR}(x)$ , and Lin and Carroll's (2000) estimators by  $\hat{\beta}_{LC}(x)$  and  $\hat{\beta}_{LCH}(x)$ , where  $\hat{\beta}_{LCH}(x)$  refers to the hybrid estimator. These three estimators are equivalent when the covariance matrix  $V$  is a diagonal matrix. Further, for a general matrix  $V$  and with *large n* and *small T*, Lin and Carroll (2000) showed that the asymptotic bias of  $\hat{\beta}_{LC}(x)$  does not depend on the density function of  $X_{i,t}$  and  $V$  for the local linear regression approach and that the same result holds for  $\hat{\beta}_{LCH}(x)$  with homogeneous Gaussian errors. Under *working independence* assumption (i.e., assuming  $V$  is a diagonal matrix or ignoring the within-group correlation in errors),  $\hat{\beta}(x)$  is the LLLS estimator and is shown to be asymptotically more efficient than the LLWLS estimators by Lin and Carroll (2000) under some regularity conditions. Hence, these results indicate that *ignoring the within-group correlation in  $\{u_{i,t}\}$  can be beneficial if one estimates  $m(\cdot)$  by the kernel-based local linear regression method.*

To facilitate further exposition on whether and how to explore the within-group correlation (i.e.,  $E(u_{i,t}u_{i,s}|X_{i,t}, X_{i,s}) \neq 0$  for some  $t \neq s$ ), we introduce two useful concepts below. Let us write the  $i$ th unit's covariance matrix  $V_i$  in a variance-correlation form,

$$V_i = A_i^{1/2} R A_i^{1/2}, \quad (2.5)$$

where  $A_i = \text{diag}(\sigma_{i,1}^2, \dots, \sigma_{i,T}^2)$  with  $\sigma_{i,t}^2 = \text{Var}(u_{i,t}|X_{i,t})$ , and  $R$  is a correlation matrix assumed to be *common* to all the cross-sectional units. This construction allows possible heteroskedastic  $u_i$  with the same correlation structure across units. A *working independence matrix* is a correlation matrix assuming no within-group correlation (i.e.,  $R = I_{nT}$ , the  $(nT)$  by  $(nT)$  identity matrix), while a *working correlation matrix* is a correlation matrix subjectively assumed to capture the within-group correlation structure of the error terms and may or may not equal the true correlation matrix (i.e.,  $R$  can be any correlation matrix).

Wang (2003) indicated that the reason that the LLWLS estimators perform no better than the LLLS estimator in the presence of the within-group correlation is that the kernel estimators asymptotically use at most one observation from each unit to calculate  $\hat{\beta}(x)$  with a sufficiently large  $n$  and finite  $T$ , so that the working independence assumption is reasonable with the local linear least squares estimation method. Assuming  $V_i \equiv \Sigma$  or cross-sectional homoskedasticity and localizing data first then weighting the localized data by  $V^{-1/2}$ , Welsh, Lin, and Carroll (2002) showed that, with a working independence matrix, the smoothing spline method is equivalent to the kernel method with a specific higher-order kernel. With a working correlation matrix, however, the estimators based on smoothing spline and penalized regression splines behave differently than the kernel-based LLWLS estimators. Specifically, Welsh, Lin, and Carroll (2002, p. 486) showed that the LLWLS estimator,  $\hat{\beta}_{LC}(x)$ , is asymptotically local and most efficient under working independence, while the smoothing spline and penalized regression spline estimators generally are not asymptotically local and are most efficient when the working covariance matrix equals the true within-group covariance matrix. When  $V$  is not a diagonal matrix, their theoretical results and simulation

results indicate that the smoothing spline and penalized regression spline methods perform better than the usual kernel estimation method. Lin and Carroll (2006) proposed a profile likelihood approach to estimate  $m(x)$  with  $W(x) \equiv V^{-1}K_h(x)$ , where all the observations from chosen unit  $i$  are used to calculate the nonparametric estimator.

The above-mentioned results are applied to nonparametric panel data models with a general non-diagonal conditional variance matrix,  $V$ , defined by (2.5). Ruckstuhl, Welsh, and Carroll (2000) studied the classical one-way component random-effects panel data model with  $u_{i,t} = \mu_i + \epsilon_{i,t}$ , where  $\mu_i \sim i.i.d.(0, \sigma_\mu^2)$ ,  $\epsilon_{i,t} \sim i.i.d.(0, \sigma_\epsilon^2)$  with  $\sigma_\mu^2 > 0$  and  $\sigma_\epsilon^2 > 0$ , and  $\{\mu_i\}$ ,  $\{\epsilon_{i,t}\}$  and  $\{X_{i,t}\}$  are mutually independent. Then, the one-way component random-effects model has  $V_i = E(u_i u_i^T | X_i) = \sigma_\mu^2 I_T \iota_T \iota_T^T + \sigma_\epsilon^2 I_T$  for all  $i$ , where  $\iota_T$  is a  $T \times 1$  vector of ones and  $I_T$  is the  $T \times T$  identity matrix. Given a known  $V$ , under some regularity conditions, Ruckstuhl, Welsh, and Carroll (2000, Th.1) showed that the asymptotic bias of the LLLS estimator equals  $h^2 m^{(2)}(x)/2$  and its asymptotic variance equals  $(\sigma_\epsilon^2 + \sigma_\mu^2) (\int K^2(u) du) / [nh \sum_{t=1}^T f_t(x)]$ , where  $f_t(x)$  is the probability density function of  $\{X_{i,t}\}_{i=1}^n$ , and  $h \rightarrow 0$  and  $nh \rightarrow \infty$  as  $n \rightarrow \infty$  with a finite  $T$ . They then proposed a two-step estimator, based on the observation that  $V^{-1/2}u$  is an i.i.d. sequence. Specifically, denoting  $Y_i^* = \tau V_i^{-1/2} Y_i + (I_{nT} - \tau V^{-1/2}) m(X)$  gives

$$Y^* = m(X) + \tau V^{-1/2} u, \quad (2.6)$$

where  $m(X) = [m(X_{1,1}), \dots, m(X_{1,T}), m(X_{2,1}), \dots, m(X_{2,T}), \dots, m(X_{n,1}), \dots, m(X_{n,T})]^T$ . Given  $\tau$  and  $V$ , the two-step estimator is constructed as follows: (i) calculate the LLLS estimator,  $\hat{m}(X)$ , from the original one-way component model (1.1) with a bandwidth  $h_0$  and construct  $\hat{Y}^* = \tau V^{-1/2} Y + (I_{nT} - \tau V^{-1/2}) \hat{m}(X)$ ; (ii) for an interior point  $x$ , estimate  $m(x)$  by the LLLS estimator with  $\hat{Y}^*$  as the working dependent variable with a bandwidth  $h$  and denote the two-step estimator by  $\tilde{m}(x)$ . The asymptotic variance of the two-step estimator,  $\tilde{m}(x)$ , is smaller than the LLLS estimator,  $\hat{m}(x)$ , if  $\tau^2 < \sigma_\epsilon^2 + \sigma_\mu^2$  as shown by Ruckstuhl, Welsh, and Carroll (2000, Th. 4) and Su and Ullah (2007, Th. 3.1), where the former set  $h = h_0$  while the latter allowed  $h \neq h_0$ . Actually, the asymptotic variance of the two-step estimator is the same whether  $h = h_0$  or not under some regularity conditions. However, allowing  $h \neq h_0$  in the two-step estimator, Su and Ullah (2007) showed that the two-step estimator has the same asymptotic bias term as the LLLS estimator, if  $h_0$  converges to zero faster than the optimal rate  $n^{-1/5}$ . Therefore, Su and Ullah (2007) showed that the two-step estimator can be asymptotically more efficient than the LLLS estimator if one chooses  $h_0 = o(n^{-1/5})$ ,  $h = cn^{-1/5}$  and  $\tau^2 < \sigma_\epsilon^2 + \sigma_\mu^2$ , where  $c$  can be chosen by the cross-validation method. Lastly, the unknown parameters  $\sigma_\epsilon^2$  and  $\sigma_\mu^2$  can be estimated by

$$\hat{\sigma}_\mu^2 = n^{-1} \sum_{i=1}^n \bar{u}_i^2 - T^{-1} \hat{\sigma}_\epsilon^2 \text{ and } \hat{\sigma}_\epsilon^2 = \frac{1}{n(T-1)} \sum_{i=1}^n \sum_{t=1}^T (\hat{u}_{i,t} - \bar{u}_i)^2, \quad (2.7)$$

respectively, where  $\widehat{u}_{i,t} = Y_{i,t} - \widehat{m}(X_{i,t})$  and  $\overline{\widehat{u}}_i = T^{-1} \sum_{t=1}^T \widehat{u}_{i,t}$ . As  $\widehat{\sigma}_\mu^2 = \sigma_\mu^2 + O_p(n^{-1/2})$ ,  $\widehat{\sigma}_\epsilon^2 = \sigma_\epsilon^2 + O_p(n^{-1/2})$ ,  $\widehat{m}(x) - m(x) = O_p(h^2) + O_p((nh)^{-1/2})$ , and  $n^{-1/2}$  converges to zero faster than  $O_p(h^2) + O_p((nh)^{-1/2})$ ,  $\sigma_\mu^2$  and  $\sigma_\epsilon^2$  can be considered known without affecting asymptotic results of the two-step estimator of  $m(x)$ . Similarly, (2.7) is also used by Henderson and Ullah (2005) in constructing feasible version of Ullah and Roy's (1998) and Lin and Carroll's (2000) estimators via a two-step procedure for large  $n$  and small  $T$ . In the first step, Henderson and Ullah (2005) calculated (2.7). In the second step, the feasible LLWLS estimators are calculated with  $V$  replaced by its consistent estimator  $\widehat{V} = I_n \otimes (\widehat{\sigma}_\mu^2 \iota_T \iota_T^T + \widehat{\sigma}_\epsilon^2 I_T)$ , where “ $\otimes$ ” stands for the Kronecker product.

So far we have seen that the existing literature has established that more efficient estimators than the LLLS estimator can be achieved by taking into account within-group correlation. However, other than the classical one-way error component case, more researches are required to further understand how to incorporate the within-group correlation for general cases when  $V$  is known as well as when  $V$  is unknown. Below, we include some existing works allowing an unknown conditional variance matrix  $V$  satisfying (2.5).

In a semiparametric varying coefficient partially linear regression framework for longitudinal data with large  $n$  and finite  $T$ , Fan, Huang, and Li (2007) proposed to estimate the unknown covariance matrices  $V_i$ 's semiparametrically. Specifically, they estimated  $\sigma_{i,t}^2 \equiv \sigma^2(X_{i,t}) = E(u_{i,t}^2 | X_{i,t})$  as in (2.5) by the local linear regression approach proposed by Fan and Yao (1998), where  $u_{i,t}$  is replaced by  $\widehat{u}_{i,t} = Y_{i,t} - \widehat{m}(X_{i,t})$  with  $\widehat{m}(\cdot)$  being the LLLS estimator given above. Second, assuming the working correlation matrix  $R$  is known up to a finite number of unknown parameters, i.e.,  $R = R(\theta)$ ; e.g.,  $\theta$  can be parameters appearing in a stationary ARMA(p,q) model of  $u_{i,t}$  with  $t = 1, \dots, T$ . Let  $\Gamma(\widehat{A}, \theta, x)$  be the estimated variance of smoothing spline or penalized regression spline estimator,  $\widehat{m}(x)$ , where  $\widehat{A}_i = \text{diag}(\widehat{\sigma}_{i,1}^2, \dots, \widehat{\sigma}_{i,T}^2)$ . Then, in the spirit of Fan, Huang, and Li (2007), we can estimate  $\theta$  by  $\widehat{\theta} = \arg \min_{\theta} \int \Gamma(\widehat{A}, \theta, x) w(x) dx$ , where  $w(x)$  is a weighting function. Let  $\widehat{V} = \text{diag}(\widehat{V}_1, \dots, \widehat{V}_n)$  and  $\widehat{V}_i = \widehat{A}_i^{1/2} R(\widehat{\theta}) \widehat{A}_i^{1/2}$ . If the parametric assumption on the working correlation matrix  $R$  is a good approximation of the true correlation structure, the spline estimator augmented with the estimated covariance  $\widehat{V}$  is expected to provide satisfactory results.

Alternatively, Qu and Li (2006) suggested to approximate  $R^{-1}$  by a linear combination of basis matrices as in Qu, Lindsay and Li (2000). As  $V_i^{-1} = A_i^{-1/2} R^{-1} A_i^{-1/2}$ , they approximated  $R^{-1} \approx a_0 I_T + a_1 M_1 + \dots + a_r M_r$ , where  $M_j$ 's are symmetric matrices for  $j = 1, \dots, r \ll T$ , and  $a_j$ 's are unknown constants. Approximating  $m(x)$  by series method first gives  $m(x) \approx \sum_{j=0}^p \delta_j P_j(x)$ , where  $P_j(x)$  are a set of basis functions of a functional space to which  $m(x)$  is assumed to belong; e.g., power functions, truncated power splines, B-splines, etc. Then, the nonparametric estimating equation can be expressed as a linear combination of the following estimating function

$$\bar{g}_n(\delta) \equiv \frac{1}{n} \sum_{i=1}^n g_i(\delta) = \begin{pmatrix} n^{-1} \sum_{i=1}^n P^T(X_i) A_i^{-1} \left( Y_i - \sum_{j=0}^p \delta_j P_j(X_i) \right) \\ n^{-1} \sum_{i=1}^n P^T(X_i) A_i^{-1/2} M_1 A_i^{-1/2} \left( Y_i - \sum_{j=0}^p \delta_j P_j(X_i) \right) \\ \vdots \\ n^{-1} \sum_{i=1}^n P^T(X_i) A_i^{-1/2} M_r A_i^{-1/2} \left( Y_i - \sum_{j=0}^p \delta_j P_j(X_i) \right) \end{pmatrix}$$

where  $P(X_i) = [P_1(X_i), \dots, P_{p+1}(X_i)]$  is a  $T$  by  $(p+1)$  matrix and  $P_j(X_i) = [P_j(X_{i,1}), \dots, P_j(X_{i,T})]^T$  is a  $T \times 1$  vector. The model is then estimated by the penalized generalized method of moments (PGMM) method

$$\hat{\delta} = \arg \min_{\delta} \bar{g}_n(\delta) \Omega^{-1} \bar{g}_n(\delta) + \lambda \delta^T \delta, \quad (2.8)$$

where  $\Omega = \text{Var}(g_i)$  and  $\lambda \geq 0$  is a regularization parameter that shrinks  $\delta$  towards zero. The purpose of introducing the penalty term in the objective function is to avoid over-parameterization when approximating  $m(x)$  by the series method. It is possible that one applies other regularization methods; e.g., Zou's (2006) adaptive LASSO penalty and Fan and Li's (2001) smoothly clipped absolute deviation (SCAD) penalty. Replacing the unknown matrices  $A_i$ 's by the nonparametric kernel estimators explained above and estimating the unknown matrix  $\Omega$  by  $\hat{\Omega} = n^{-1} \sum_{i=1}^n g_i(\tilde{\delta}) g_i^T(\tilde{\delta})$  with  $\tilde{\delta}$  being the estimator obtained under working independence, we obtain another estimator of  $m(\cdot)$ , the derivation of the limiting distribution of this new estimator is left as a future research topic. For the possible choice of the basis matrices,  $M_j$ 's, the readers are referred to Qu and Li (2006, p. 382).

The two methods mentioned above hold for finite  $T$  and sufficiently large  $n$ . If one allows  $T$  to increase as  $n$  increases, Bickel and Levina's (2008) covariance regularization method may be used to estimate the true covariance matrix,  $V$  under the condition that  $\ln(T)/n \rightarrow 0$  as  $n \rightarrow \infty$ . To the best of our knowledge, the nonparametric random-effects panel data models are usually studied with a finite  $T$  and more research needs to be done for constructing efficient nonparametric estimator of  $m(x)$  when both  $T$  and  $n$  are large.

All the papers discussed above either assume a known or pre-determined or pre-estimated within-group correlation/covariance structure before estimating  $m(x)$ . We conclude this section with a recent paper that estimates  $m(x)$  and the within-group correlation matrix simultaneously under cross-section homoskedasticity.

Yao and Li (2013) proposed a novel estimator that uses Cholesky decomposition and profile least squares techniques. Consider again model (1.1). Let  $u_i = (u_{i,1}, \dots, u_{i,T})^T$ ,  $i = 1, \dots, n$ . Suppose that  $\text{Cov}(u_i|X_i) = \Sigma$  for all  $i$ , assuming cross-section homoskedasticity. A Cholesky decomposition of the within-group covariance gives

$$\text{Cov}(\Phi u_i) = \Phi \Sigma \Phi^T = D,$$

where  $\Phi$  is a lower triangle matrix with 1's on the main diagonal and  $D = \text{diag}(d_1^2, \dots, d_T^2)$  is a diagonal matrix. Let  $\phi_{t,s}$  be the negative of the  $(t,s)$ -th element of  $\Phi$  and  $e_i = (e_{i,1}, \dots, e_{i,T})^T = \Phi u_i$ . One can rewrite  $u_i$  as follows:

$$\begin{aligned} u_{i,1} &= e_{i,1}, \\ u_{i,t} &= \phi_{t,1} u_{i,1} + \cdots + \phi_{t,t-1} u_{i,t-1} + e_{i,t}, \quad t = 2, \dots, T. \end{aligned}$$

Since  $D$  is a diagonal matrix,  $e_{i,t}$ 's are uncorrelated with  $\text{Var}(e_{i,t}) = d_t^2$ ,  $t = 1, \dots, T$ . If  $\{u_1, \dots, u_n\}$  were known, we would proceed with the following partially linear model with uncorrelated error term  $e_{i,t}$ :

$$\begin{aligned} Y_{i,1} &= m(X_{i,1}) + e_{i,1}, \\ Y_{i,t} &= m(X_{i,t}) + \phi_{t,1} u_{i,1} + \cdots + \phi_{t,t-1} u_{i,t-1} + e_{i,t}, \quad i = 1, \dots, n, \quad t = 2, \dots, T. \end{aligned} \tag{2.9}$$

However, in practice,  $u_{i,t}$ 's are not available but can be replaced by their consistent estimates. For instance, one can use  $\widehat{u}_{i,t} = Y_{i,t} - \widehat{m}(X_{i,t})$ , where  $\widehat{m}(\cdot)$  is the local linear estimator of  $m(\cdot)$  under the working independence assumption.

Replacing  $u_{i,t}$ 's in (2.9) with  $\widehat{u}_{i,t}$ 's yields

$$Y_{i,t} \approx m(X_{i,t}) + \phi_{t,1} \widehat{u}_{i,1} + \cdots + \phi_{t,t-1} \widehat{u}_{i,t-1} + e_{i,t}, \quad i = 1, \dots, n, \quad t = 2, \dots, T. \tag{2.10}$$

Let  $Z = (Z_{1,2}, \dots, Z_{1,T}, \dots, Z_{n,2}, \dots, Z_{n,T})^T$  for  $Z = X$  or  $Y$  or  $e$ ,  $\phi = (\phi_{2,1}, \dots, \phi_{T,T-1})^T$ , and  $\widehat{F}_{i,t} = (0_{(t-1)(t-2)/2}^T, \widehat{u}_{i,1}, \dots, \widehat{u}_{i,T-1}, 0_{(T-1)T/2-(t-1)t/2}^T)^T$ , where  $0_k$  is a  $k \times 1$  vector of 0's. Then we can rewrite model (2.10) in the following matrix form:

$$Y \approx m(X) + \widehat{F}\phi + e,$$

where  $\widehat{F} = (\widehat{F}_{1,2}, \dots, \widehat{F}_{1,T}, \dots, \widehat{F}_{n,T})^T$ . Define  $Y^* = Y - \widehat{F}\phi$ . Then

$$Y^* \approx m(X) + e. \tag{2.11}$$

Therefore if  $\Sigma$  and thus  $\Phi$  were known, we could use Cholesky decomposition to transform the original model to model (2.11), wherein the errors are uncorrelated. Yao and Li (2013) proposed a profile least squares estimator for model (2.11). Define

$$W_d(x) = \text{diag}\{K_h(X_{1,2} - x)/\widehat{d}_1^2, \dots, K_h(X_{1,T} - x)/\widehat{d}_T^2, \dots, K_h(X_{n,T} - x)/\widehat{d}_T^2\},$$

where  $K_h(t) = h^{-1}K(t/h)$ ,  $K(\cdot)$  is a kernel function and  $h$  is the bandwidth, and  $\widehat{d}_t^2$  is a consistent estimate of  $d_t^2$ . It follows that a weighted (by  $d_t^2$ 's) local linear estimator of  $m(\cdot)$  can be constructed as in (2.2) with  $W$  replaced by  $W_d$ , and the estimator is given by

$$\widehat{m}(x) = (1, 0) \left[ Z^T(x) W_d(x) Z(x) \right]^{-1} Z^T(x) W_d(x) Y^* \equiv S_h(x) Y^*$$

where

$$Z(x) = \begin{bmatrix} 1 & \dots & 1 & \dots & 1 & \dots & 1 \\ X_{1,2-x} & \dots & X_{1,T-x} & \dots & X_{n,2-x} & \dots & X_{n,T-x} \end{bmatrix}.$$

Thus  $\widehat{m}(X)$  can be represented by  $\widehat{m}(X) = S_h(X)Y^*$ , where  $S_h(X)$  is a  $n(T-1) \times n(T-1)$  smoothing matrix, depending only on  $X$  and the bandwidth  $h$ . Replacing  $m(X)$  in model (2.11) with  $\widehat{m}(X)$  then yields the following linear model

$$[I_{n(T-1)} - S_h(X)]Y \approx [I_{n(T-1)} - S_h(X)]\widehat{F}\phi + e,$$

where  $I_{n(T-1)}$  is the  $n(T-1) \times n(T-1)$  identity matrix. Denote  $M_h(X) = I_{n(T-1)} - S_h(X)$  and  $\widehat{G} = \text{diag}(\widehat{d}_2^2, \dots, \widehat{d}_T^2, \dots, \widehat{d}_2^2, \dots, \widehat{d}_T^2)$ . The profile least squares estimate for  $\phi$  is then given by

$$\widehat{\phi} = [\widehat{F}^T M_h^T(X) \widehat{G}^{-1} M_h(X) \widehat{F}]^{-1} \widehat{F}^T M_h^T(X) \widehat{G}^{-1} M_h(X) Y. \quad (2.12)$$

Next, letting  $\widehat{Y}^* = Y - \widehat{F}\widehat{\phi}$ , we have

$$\widehat{Y}^* \approx m(X) + e, \quad (2.13)$$

where the  $e_{i,t}$ 's are uncorrelated. We can then estimate (2.13) using the conventional local linear estimator

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \arg \min_{(\beta_0, \beta_1)} [\widehat{Y}^* - Z(x)\beta]^T W_d(x) [\widehat{Y}^* - Z(x)\beta]. \quad (2.14)$$

The local linear estimator of  $m(x)$  is given by  $\widehat{m}(x; \widehat{\phi}) = \widehat{\beta}_0$ .

Yao and Li (2013) established the asymptotic properties of their estimator. In particular, they showed that it is asymptotically efficient as if the true within-group covariance were known. Their Monte Carlo simulations demonstrated that this estimator outperforms the naive local linear estimator with working independence assumption and some existing two-step procedures.

### 10.2.2 Fixed Effects Panel Data Models

This subsection considers nonparametric estimation of fixed-effects panel data models, where  $X_{i,t}$  in model (1.1) is allowed to be correlated with the error term  $u_{i,t}$ ,  $i = 1, 2, \dots, n$ ,  $t = 1, \dots, T$ . Specifically, we consider case (a), where  $u_{i,t} = \mu_i + \epsilon_{i,t}$  and the unobserved individual effects,  $\mu_i \sim i.i.d. (0, \sigma_\mu^2)$  are allowed to be correlated with  $\{X_{i,t}\}_{t=1}^T$  in an unspecified way. We consider two cases: (i)  $\{\epsilon_{i,t}\}$  is independent of  $\{X_{i,t}\}$ , which excludes the case that  $X_{i,t}$  contains lagged dependent variable(s); (ii)  $X_{i,t}$  includes the lagged dependent variable.

We consider case (i) first. For large  $n$  and finite  $T$ , Henderson, Carroll and Li (2008) applied a first-differencing method and backfitting algorithm to estimate the unknown function,  $m(\cdot)$ . Taking the first difference removes the unobserved fixed effects  $\mu_i$ , which gives

$$Y_{i,t} - Y_{i,1} = m(X_{i,t}) - m(X_{i,1}) + v_{i,t} \quad (2.15)$$

where  $v_{i,t} = \epsilon_{i,t} - \epsilon_{i,1}$ ,  $V_i = E(v_i v_i^T | X_{i,1}, \dots, X_{i,T}) = \sigma_\epsilon^2 (I_{T-1} + \iota_{T-1} \iota_{T-1}^T)$ , and  $v_i = (v_{i,1}, \dots, v_{i,T})^T$ . Evidently, the first-differencing method cancels out any time-invariant component in  $m(\cdot)$ . As in model (1.1),  $E[m(X_{i,t})] = E(Y_{i,t})$  and  $m(\cdot)$  is identified if it does not contain time-invariant component other than the non-zero constant mean.

Applying Lin and Carroll's (2006) profile likelihood method, Henderson, Carroll, and Li (2008) constructed a local linear estimator with the backfitting algorithm. Due to the complexity of the backfitting algorithm, they conjectured the asymptotic bias and variance term of the proposed estimator without providing detailed proofs. As model (2.15) is a nonparametric additive panel data model, Mammen, Støve, and Tjøstheim's (2009) iterative smooth backfitting algorithm is applicable; however, they did not derive the limit distribution of their proposed estimator, either. Qian and Wang (2012) also considered estimating  $m(\cdot)$  from model (2.15) but used the marginal integration method of Linton and Nielsen (1995). They derived the consistency and asymptotic normality results of the proposed estimator for finite  $T$  and large  $n$  not only for cross-sectional fixed-effects models but also for panel data models with both unobserved cross-sectional and time fixed effects.

Alternatively, Su and Ullah (2006), Sun, Carroll, and Li, (2009) and Lin, Li, and Sun (2014) estimated model (1.1) directly, treating unobserved fixed effects  $\mu_i$  as parameters. These estimation methods can be considered as a nonparametric version of the least squares dummy variable method. Su and Ullah (2006) assumed  $\sum_{i=1}^n \mu_i = 0$  for a finite  $T$  and large  $n$ , and the same restriction is also imposed in, Li, Chen, and Gao (2011) and Chen, Gao, and Li (2013a, 2013b) for estimating semiparametric fixed effects panel data models. The use of “fixed effects” carries on the old tradition although the models with  $\sum_{i=1}^n \mu_i = 0$  exhibit cross-sectional dependence among the composite errors,  $u_{i,t}$ , as  $\mu_i$ 's have to be dependent to satisfy the restriction. Alternatively, Lin, Li, and Sun (2014) showed that this assumption can be dropped for large  $T = O(n^{1/4})$  and large  $n$ . Specifically, introducing  $T$  by  $n$  dummy variables,  $D$ , as in parametric fixed-effects panel data models, one considers the following objective function

$$\min_{(\mu, m(\cdot))} (Y - m(X) - D\mu)^T K_h(x) (Y - m(X) - D\mu).$$

Assuming  $m(\cdot)$  is known and imposing  $\sum_{i=1}^n \mu_i = 0$ , Su and Ullah (2006) obtained  $\widehat{\mu}_i(X_{i,t})$  when  $x = X_{i,t}$  and set  $\widehat{Y}_{i,t}^* = Y_{i,t} - \widehat{\mu}_i(X_{i,t})$  for all  $i$  and  $t$ . They then estimated  $m(x)$  by the local linear regression approach with the new dependent variable  $\widehat{Y}_{i,t}^*$ . Instead, Lin, Li, and Sun (2014) removed the impact of  $\mu_i$ 's asymptotically. Although their estimator can be obtained through the local polynomial approach, we present here a local constant estimator for simplicity:

$$\tilde{m}(x) = n^{-1} \sum_{i=1}^n \sum_{t=1}^T \omega_{i,t} Y_{i,t} \quad (2.16)$$

where  $\omega_{i,t} = \lambda_{i,t}/\sum_{t=1}^T \lambda_{i,t}$  and  $\lambda_{i,t} = K((X_{i,t} - x)/h)$ . By construction,  $\tilde{m}(x)$  is an average of the dependent variable with a within-group weight  $n^{-1}\omega_{i,t}$  attached to  $Y_{i,t}$  for a given  $i$  and  $t$ . Assuming that  $\{(X_{i,t}, \epsilon_{i,t})\}$  is strictly stationary with some mixing properties and other regularity conditions, if  $h \rightarrow 0$ ,  $Th \rightarrow \infty$  as  $T \rightarrow \infty$  and  $nTh^5 = O(1)$  as  $n \rightarrow \infty$  and  $T \rightarrow \infty$ , Lin, Li, and Sun (2014, Th. 1) showed that at an interior point  $x$ ,

$$\sqrt{nTh}[\tilde{m}(x) - m(x) - h^2B(x) - \bar{\mu}] \xrightarrow{d} N\left(0, \frac{\zeta_0\sigma_\epsilon^2}{f(x)}\right), \quad (2.17)$$

where  $B(x) = \kappa_2 \left[ 2m'(x)f'(x)/f(x) + m^{(2)}(x) \right]$ ,  $\kappa_2 = \int K(v)v^2 dv$ ,  $\zeta_0 = \int K(v)^2 dv$  and  $\bar{\mu} = n^{-1} \sum_{i=1}^n \mu_i$ , and  $f(x)$  is the common Lebesgue marginal p.d.f. of  $X_{i,t}$ . This result implies that  $\tilde{m}(x) = m(x) + h^2B(x) + \bar{\mu} + O_p((nTh)^{-1/2}) \xrightarrow{P} m(x)$  since  $\bar{\mu} \xrightarrow{P} 0$  and  $h \rightarrow 0$  as  $n \rightarrow \infty$ . In addition, the optimal bandwidth  $h_{opt} = O(nT)^{-1/5}$  with  $T = O(n^{1/4})$  allows  $T$  to grow at a slower speed than  $n$ .

When the sample size  $n$  is small and/or  $Var(\mu_i)$  is rather large compared to the variances of  $\epsilon_{it}$  and  $X_{it}$ , the existence of the fixed effects term  $\bar{\mu} = O_p(n^{-1/2})$  may affect the finite sample performance of the proposed estimator. However, Lin, Li and Sun (2014, Th. 2) found that  $\check{m}(x) \equiv \tilde{m}(x) - \bar{Y}$  is robust to the presence of  $\mu_i$ 's, where  $\bar{Y} = (nT)^{-1} \sum_{i=1}^n \sum_{t=1}^T Y_{i,t}$ . Since  $E(Y_{i,t}) = E[m(X_{i,t})] \equiv c_0$  with  $E(\mu_i) = E(\epsilon_{i,t}) = 0$  for all  $i$  and  $t$ , they obtained

$$\sqrt{nTh}[\check{m}(x) - (m(x) - c_0) - h^2B(x)] \xrightarrow{d} N\left(0, \frac{\zeta_0\sigma_\epsilon^2}{f(x)}\right), \quad (2.18)$$

so that  $\check{m}(x)$  is a consistent estimator of the demeaned curve  $m(x) - c_0$ .

Next we consider case (ii) where the lagged dependent variable  $Y_{i,t-1}$  is included in the explanatory variables. Again, the first-differencing method is conventionally used to completely remove the impact of the unobserved individual effects, which gives

$$\Delta Y_{i,t} = m(Y_{i,t-1}, X_{i,t}) - m(Y_{i,t-2}, X_{i,t-1}) + v_{i,t}, \quad (2.19)$$

where  $v_{i,t}$  is defined below eq. (2.15). Since the error terms under this framework are generally correlated with  $Y_{i,t-1}$ , traditional kernel methods based on the marginal integration or backfitting algorithm do not yield a consistent estimator of  $m(\cdot)$ . Lee (2010) considered the case where  $X_{i,t} \equiv Y_{i,t-1}$  and proposed a sieve estimation of the dynamic panel model via within-group transformation. He found that the series estimator is asymptotically biased and proposed a bias-corrected series estimator, which is shown to be asymptotically normal.

An alternative approach is employed by Su and Lu (2013). Their method is motivated by two observations: (i) the two additive terms of (2.19) share the same functional form; (ii)  $E[v_{i,t}|Y_{i,t-2}, X_{i,t-1}] = 0$  by assuming  $E(\epsilon_{i,t}|X_{i,t}, Y_{i,t-1}, X_{i,t-2}, Y_{i,t-2}, \dots) = 0$  holds for all  $i$  and  $t$ . Denote  $U_{i,t-2} = (Y_{i,t-2}, X_{i,t-1})^T$ . It follows that

$$E[\Delta Y_{i,t} - m(Y_{i,t-1}, X_{i,t}) + m(Y_{i,t-2}, X_{i,t-1}) | U_{i,t-2}] = 0. \quad (2.20)$$

Denote by  $f_{t-2}(u)$  the p.d.f. of  $U_{i,t-2}$  and  $f_{t-1|t-2}$  the conditional p.d.f. of  $U_{i,t-1}$  given  $U_{i,t-2}$ . Also define  $r(u) \equiv -E(\Delta Y_{i,t} | U_{i,t-2} = u)$  and  $f(v|u) \equiv f_{t-1|t-2}(v|u)$ . One can rewrite (2.20) as follows:

$$m(u) = r(u) + \int m(v)f(v|u)dv. \quad (2.21)$$

Thus the parameter of interest,  $m$ , is implicitly defined as a solution to a Fredholm integral equation of the second kind.

Let  $L_2(f)$  be an infinite dimensional Hilbert space. Under certain regularity conditions, we have

$$m = r + \mathcal{A}m, \quad (2.22)$$

where  $\mathcal{A} : L_2(f) \rightarrow L_2(f)$  is a bounded linear operator defined by

$$\mathcal{A}m(u) = \int m(v)f(v|u)dv.$$

Suppose  $\mathcal{A}$  is well behaved, the solution to the above Fredholm integral equation is given by  $m = (I - \mathcal{A})^{-1}r = \sum_{j=0}^{\infty} \mathcal{A}^j r$ , where  $I$  is the identity operator. One can then use a successive substitution approach to solve for  $m$  numerically. In this case, the sequence of approximations

$$m^{(l)} = r + \mathcal{A}m^{(l-1)}, \quad l = 1, 2, 3, \dots,$$

is close to the truth from any starting point  $m^0$ .

Let  $\hat{r}$  and  $\hat{\mathcal{A}}$  be nonparametric estimators of  $r$  and  $\mathcal{A}$ . The plug-in estimator  $\hat{m}$  is then given by the solution of

$$\hat{m} = \hat{r} + \hat{\mathcal{A}}\hat{m}.$$

Su and Lu (2013) proceeded with an iterative estimator of the parameter of interest  $m$ . The following steps were proposed:

1. Calculate an initial estimate of  $m$ , say,  $\hat{m}^{(0)}$ .
2. Calculate the estimate of  $r(u)$ , say,  $\hat{r}$ , by regressing  $-\Delta Y_{i,t}$  on  $U_{i,t-2}$  via the local polynomial approach.
3. Calculate the estimate of  $\mathcal{A}\hat{m}^{(0)}$ , say,  $\hat{\mathcal{A}}\hat{m}^{(0)}$ , by regressing  $\hat{m}^{(0)}(U_{i,t-1})$  on  $U_{i,t-2}$  via the local polynomial approach.
4. Calculate  $\hat{m}^{(1)} = \hat{r} + \hat{\mathcal{A}}\hat{m}^{(0)}$ .
5. Repeat steps 3-4 until convergence.

Below we provide a brief description of the estimators used in the steps outlined above. As for the initial estimate  $\hat{m}^{(0)}$ , Su and Lu (2013) used a sieve estimator. Let

$q_{i,t-1} \equiv q^L(U_{i,t-1})$  be a  $L$ -dimensional vector of sieve basis functions and  $\Delta q_{i,t-1} \equiv q_{i,t-1} - q_{i,t-2}$ . The estimator takes the form

$$\Delta Y_{i,t} = \beta_m^T \Delta q_{i,t-1} + r_{i,t},$$

where  $r_{i,t}$  incorporates the error term  $\nu_{i,t}$  in model(2.19) and the approximation errors due to the sieve approximation. Since  $r_{i,t}$  are generally correlated with  $\Delta q_{i,t-1}$ , they used  $Z_{i,t} = q_{i,t-2}$  as instruments, following Anderson and Hsiao (1981). Su and Lu (2013) then used the local polynomial approach to estimate  $r$  and  $\mathcal{A}$  in steps 3 and 4, respectively. Since  $m$  is identified only up to a location shift, they suggested recentering  $\widehat{m}^{(l)}(u)$  in each iteration to obtain

$$\widehat{m}^{(l)}(u) + \frac{1}{nT} \sum_{i=1}^n \sum_{t=2}^T [Y_{i,t} - \widehat{m}^{(l)}(U_{i,t-1})],$$

which is used in the next iteration.

Su and Lu (2013) established the consistency, convergence rate and asymptotic normality of the proposed estimator under suitable regularity conditions. They further proposed a specification test for linear dynamic panel models. The proposed test is based on the comparison of the restricted linear estimator and the unrestricted estimator. They considered the following smooth functional

$$\Gamma \equiv \int [m(u) - \gamma_0 - \beta_0^T u]^2 a(u) f(u) du,$$

where  $a(u)$  is a user-specified nonnegative weighting function defined on the compact support of  $u$ . A feasible test statistic is then constructed as

$$\Gamma_{nT} \equiv \frac{1}{nT} \sum_{i=1}^n \sum_{t=2}^T [\widehat{m}(U_{i,t-1}) - \widehat{\gamma} - \widehat{\beta}^T U_{i,t-1}]^2 a(U_{i,t-1}).$$

They showed that after appropriate centering and scaling,  $\Gamma_{nT}$  is asymptotically normally distributed under some suitable assumptions under the null hypothesis.

### 10.2.3 Panel Data Models with Cross-Sectional Dependence

In this subsection we discuss models with cross-sectional dependence. Consider a nonparametric panel data model

$$Y_{i,t} = m(X_{i,t}) + u_{i,t}, \quad (2.23)$$

$$u_{i,t} = \mu_i + \delta_i^T g_t + \epsilon_{i,t}, \quad i = 1, \dots, n, t = 1, \dots, T, \quad (2.24)$$

where  $g_t$  is a  $d$  by 1 vector of unobserved common factors,  $\delta_i \neq 0$  is a  $d$  by 1 vector of factor loadings, and  $\epsilon_{i,t}$  is i.i.d. and independent of  $\delta_j^T g_s$  for all  $i, j, s$ , and  $t$ . As  $E(u_{i,t} u_{j,t}) = \delta_i^T E(g_t g_t^T) \delta_j$  can be none-zero if  $\delta_i \neq 0$  and  $\delta_j \neq 0$  for  $i \neq j$ , the error term exhibits cross-sectional dependence in the presence of the unobserved term  $\delta_i^T g_t$ . This model is a special case studied by Su and Jin (2012) and Huang (2013), where both papers include an additional linear component in (2.23). Also, eq. (2.24) nests fixed-effects models and random-effects two-way error component models with  $u_{i,t} = \mu_i + \lambda_t + \epsilon_{i,t}$ . Both papers extend Pesaran's (2006) common correlated effects (CCE) estimation method in linear panel data models to non-/semi-parametric panel data models with large  $n$  and large  $T$ .

Closely following Pesaran's (2006) methodology, Su and Jin (2012) estimated  $m(\cdot)$  via sieve estimation method. The validity of this method crucially relies on the following assumption

$$X_{i,t} = \alpha_i + \gamma_i^T g_t + \nu_{i,t}, \quad (2.25)$$

where  $\{(v_{i,t}, \epsilon_{i,t})\}$  is a strictly stationary strong mixing sequence and is independent of another strictly stationary strong mixing sequence  $\{g_t\}$ , and  $\{\nu_{i,t}\}$  is also independent of  $\{\epsilon_{i,t}\}$ . Let  $\bar{w}_t = n^{-1} \sum_{i=1}^n w_{i,t}$  for any  $w$  ( $w$  can be  $Y$ ,  $X$ ,  $\mu$ ,  $\alpha$ , etc). Taking cross-sectional average to (2.23) and (2.25) gives

$$\begin{pmatrix} \bar{Y}_t \\ \bar{X}_t \end{pmatrix} = \begin{pmatrix} \bar{\mu} \\ \bar{\alpha} \end{pmatrix} + \begin{pmatrix} \bar{\delta} \\ \bar{\gamma} \end{pmatrix} g_t + \begin{pmatrix} \bar{m}_t + \bar{\epsilon}_t \\ \bar{\nu}_t \end{pmatrix}. \quad (2.26)$$

Letting  $\bar{\theta} = (\bar{\delta}, \bar{\gamma})^T$ , we have

$$g_t = (\bar{\theta} \bar{\theta}^T)^{-1} \bar{\theta}^T \left[ \begin{pmatrix} \bar{Y}_t \\ \bar{X}_t \end{pmatrix} - \begin{pmatrix} \bar{\mu} \\ \bar{\alpha} \end{pmatrix} - \begin{pmatrix} \bar{m}_t + \bar{\epsilon}_t \\ \bar{\nu}_t \end{pmatrix} \right]$$

where  $\text{rank}(\bar{\theta} \bar{\theta}^T) \leq d+1$  as  $n \rightarrow \infty$ . If  $\sup_t |\bar{w}_t| \xrightarrow{p} 0$  as  $n \rightarrow \infty$  for  $w = m, \epsilon$ , and  $\nu$ , then

$$g_t - (\bar{\theta} \bar{\theta}^T)^{-1} \bar{\theta}^T \left[ \begin{pmatrix} \bar{Y}_t \\ \bar{X}_t \end{pmatrix} - \begin{pmatrix} \bar{\mu} \\ \bar{\alpha} \end{pmatrix} \right] \xrightarrow{p} 0 \text{ as } n \rightarrow \infty,$$

which holds uniformly over all  $t$ . Thus  $g_t$  can be well approximated by  $(1, \bar{X}_t, \bar{Y}_t)^T$ . It follows that model (2.23)–(2.24) can be approximated by

$$Y_{i,t} \approx \pi_{i,0} + \pi_{i,1} \bar{Y}_t + \pi_{i,2} \bar{X}_t + m(X_{i,t}) + \epsilon_{i,t}, \quad (2.27)$$

which is a partially linear regression model. The identification condition is given by  $E[m(X_{i,t})] \equiv 0$ . Approximating  $m(X_{i,t})$  by a linear combination of basis functions, Su and Jin (2012) derived consistency and asymptotic normality results of the proposed estimator of  $m(x)$  when both  $n$  and  $T$  are sufficiently large and  $T/n^s \rightarrow 0$  for some  $s \in (0, 1)$ . When  $\delta_i = \gamma_i = 0$  for all  $i$ , model (2.23)–(2.24) becomes a nonparametric fixed-effects panel data considered by Lin, Li, and Sun (2014) in Section 10.2.3, and the argument based on (2.25) fails to be meaningful.

Huang (2013) followed the same methodology as in Pesaran (2006) in the sense that the unobserved time-varying factors  $g_t$  are to be instrumented with the cross-section averages,  $\bar{Y}_t$  and  $\bar{X}_t$ . Without imposing (2.25), Huang (2013) showed that  $\bar{\delta}\bar{\delta}^T$  having a full rank is sufficient to obtain a consistent estimator of  $m(\cdot)$ . For the fixed-effects panel data model, Huang (2013) used the within-group averages,  $\bar{Y}_i$  and  $\bar{X}_i$  as instrumental variables for  $\mu_i$  for all  $i$ , following similar argument as in using  $\bar{Y}_t$  and  $\bar{X}_t$  to predict  $g_t$ . In addition, assuming  $\delta'_i$ 's are random coefficients independent of  $(g_t, X_{i,t}, \epsilon_{i,t})$  for all  $i$  and  $t$ , Huang (2013) simplified model (2.27) to a model with common parameters across index  $i$ ,

$$Y_{i,t} \approx \pi_1 \bar{Y}_t + \pi_2 \bar{X}_t + \pi_3 \bar{Y}_i + \pi_4 \bar{X}_i + m(X_{i,t}) + \epsilon_{i,t}, \quad (2.28)$$

and estimated  $m(x)$  from the model above by local linear regression approach. Both (2.27) and (2.28) approximate the unobserved factors  $g_t$  by linear combination of  $\bar{Y}_t$  and  $\bar{X}_t$  globally in Su and Jin (2012) and locally in Huang (2013), respectively. Also, both papers assume that  $E[m(X_{i,t})] \equiv 0$  for identification purpose.

For a large  $T$  and finite  $n$ , Robinson (2012) considered the following model

$$Y_{i,t} = \alpha_i + \beta_t + u_{i,t}, \quad i = 1, \dots, n, t = 1, \dots, T \quad (2.29)$$

where  $\alpha_i$ 's and  $\beta_t$ 's are unknown individual and time effects,  $u_{i,t}$  are unobservable zero-mean random variables that exhibit cross-sectional dependence but cross-time independence. Under the assumption that  $\sum_{i=1}^n \alpha_i = 0$ , taking average across units yields

$$\bar{Y}_t = \beta_t + \bar{u}_t, \quad t = 1, \dots, T, \quad (2.30)$$

where  $\bar{w}_t = n^{-1} \sum_{i=1}^n w_{i,t}$  for  $w = Y$  or  $u$ . It is assumed that  $E(u_{i,t}) = 0$ ,  $E(u_{i,t}u_{j,t}) = \omega_{ij}$  and  $E(u_{i,t}u_{j,s}) = 0$  for all  $i, j, t \neq s$ . One can then estimate  $\beta_t$  by  $\bar{Y}_t$ . However, with small  $n$ , the estimated time trend at any given point of time may not be consistent.

Robinson (2012) proposed to estimate  $\beta_t$  as a smooth function of  $t$  using a kernel estimator, effectively borrowing information from neighboring time periods. Let  $\beta_t = \beta(t/T)$ ,  $t = 1, 2, \dots, T$ . Define

$$K_{t,\tau} = K\left(\frac{T\tau - t}{Th}\right) \text{ for } \tau \in (0, 1),$$

where  $K(\cdot)$  is a kernel function. The smooth time trend estimator is then given by

$$\tilde{\beta}(\tau) = \frac{\sum_{t=1}^T K_{t,\tau} \bar{Y}_t}{\sum_{t=1}^T K_{t,\tau}}.$$

Under some regularity conditions, he showed that the kernel estimator of  $\tilde{\beta}(\tau)$  for  $\tau \in (0, 1)$  from model (2.30) is consistent if  $h \rightarrow 0$  and  $Th \rightarrow \infty$  as  $T \rightarrow \infty$ . Let  $\Omega$  be the  $n \times n$  cross-sectional covariance matrix with the  $(i, j)$ -th element being  $\omega_{ij}$ . The mean squared error of  $\tilde{\beta}(\tau)$  is given by

$$\text{MSE}(\tilde{\beta}(\tau)) \sim \frac{\zeta_0}{Th} \frac{\iota_n' \Omega \iota_n}{n^2} + \frac{\kappa_2^2 h^4 [\beta''(\tau)]^2}{4}, \quad \text{as } T \rightarrow \infty, \quad (2.31)$$

where  $\iota_n$  is an  $n \times 1$  vector of 1's,  $\zeta_0 = \int K^2(u) du$ , and  $\kappa_2 = \int u^2 K(u) du$ .

The smooth time trend estimator,  $\tilde{\beta}(\tau)$ , has the standard convergence rate of kernel estimators. However, it only utilizes the average of each time period. Robinson (2012) then proceeded to show that  $\tilde{\beta}(\tau)$  is dominated by an improved estimator that takes advantage of cross-sectional data, which offers the possibility of variance reduction. Let  $w = (w_1, \dots, w_n)^T$  be a vector of weights with  $w^T \iota_n = 1$  and  $Y_t = (Y_{1,t}, \dots, Y_{n,t})^T$ . A weighted kernel estimator of the time trend is constructed as

$$\tilde{\beta}^{(w)}(\tau) = \frac{\sum_{t=1}^T K_{t,\tau} w^T Y_t}{\sum_{t=1}^T K_{t,\tau}}, \quad \tau \in (0, 1).$$

Robinson (2012) showed that the weight  $w$  affects the variance of  $\tilde{\beta}^{(w)}(\tau)$  but not its bias and that the corresponding mean squared error is given by

$$\text{MSE}(\tilde{\beta}^{(w)}(\tau)) \sim \frac{\zeta_0}{Th} w^T \Omega w + \frac{\kappa_2^2 h^4 [\beta''(\tau)]^2}{4}, \quad \text{as } T \rightarrow \infty. \quad (2.32)$$

Minimizing (2.32) with respect to  $w$  subject to  $w^T \iota_n = 1$  then yields the optimal weight

$$w^* = \frac{\Omega^{-1} \iota_n}{\iota_n^T \Omega \iota_n}.$$

It follows that

$$\text{MSE}(\tilde{\beta}^{(w^*)}(\tau)) \sim \frac{\zeta_0}{Th} \left( \iota_n^T \Omega^{-1} \iota_n \right)^{-1} + \frac{\kappa_2^2 h^4 [\beta''(\tau)]^2}{4}, \quad \text{as } T \rightarrow \infty. \quad (2.33)$$

Thus  $\tilde{\beta}^{(w^*)}$  generally dominates  $\tilde{\beta}$ . As expected, the cross-sectional variation of the panel data provides useful information to improve the estimation of time trend in terms of efficiency. Robinson (2012) further derived the optimal bandwidth of the proposed estimator and reported Monte Carlo simulations to confirm its improved performance. Lee and Robinson (2012) extended Robinson (2012) to allow for stochastic covariates, conditional heteroskedasticity, and intertemporal dependence.

Chen, Gao, and Li (2012a) considered a semiparametric partially linear panel model with cross-sectional dependence,

$$Y_{i,t} = X_{i,t}^T \beta + f_t + \mu_i + \epsilon_{i,t} \quad (2.34)$$

$$X_{i,t} = g_t + x_i + v_{i,t}, \quad (2.35)$$

where  $X_{i,t}$  is a  $d$ -dimensional covariates that are allowed to be nonstationary with a trending component and correlated with  $\mu_i$  and  $\epsilon_{i,t}$ , and  $f_t \equiv f(t/T)$  is an unknown

smooth measurable function of a time trend. In addition,  $\{\epsilon_{i,t}\}$  is a martingale process across index  $t$  and exhibits cross-section dependence across index  $i$ . For the purpose of identification, they assumed  $\sum_{i=1}^n \mu_i = 0$  and  $\sum_{i=1}^n x_i = 0$ . They proposed to estimate (2.34) via a semiparametric profile least squares method, wherein the individual effects are treated as unknown parameters and the time trend is approximated by local linear estimator.

Define the following quantities:

$$\begin{aligned} Y &= (Y_{1,1}, \dots, Y_{1,T}, Y_{2,1}, \dots, Y_{2,T}, \dots, Y_{n,1}, \dots, Y_{n,T})^T, \\ X &= (X_{1,1}, \dots, X_{1,T}, X_{2,1}, \dots, X_{2,T}, \dots, X_{n,1}, \dots, X_{n,T})^T, \\ \mu &= (\mu_2, \dots, \mu_n)^T, D = (-\iota_{n-1}, I_{n-1})^T \otimes \iota_T, \\ f &= \iota_n \otimes (f_1, \dots, f_T)^T, \\ \epsilon &= (\epsilon_{1,1}, \dots, \epsilon_{1,T}, \epsilon_{2,1}, \dots, \epsilon_{2,T}, \dots, \epsilon_{n,1}, \dots, \epsilon_{n,T})^T, \end{aligned}$$

where  $I_n$  is the  $n \times n$  identity matrix and  $\iota_T$  is a  $T \times 1$  vector of ones. Model (2.34) can then be rewritten as

$$Y = X\beta + f + D\mu + \epsilon.$$

Denote

$$z(\tau) = \begin{pmatrix} 1 & \frac{1-\tau T}{Th} \\ \vdots & \vdots \\ 1 & \frac{T-\tau T}{Th} \end{pmatrix}, \tau \in (0, 1),$$

and  $Z(\tau) = \iota_n \otimes z(\tau)$ . Then by Taylor expansion we have  $f(\tau) \approx Z(\tau)[f(\tau), hf'(\tau)]^T$ .

Let  $K(\cdot)$  be a kernel function and  $h$  be a bandwidth. Define  $w(\tau) = \text{diag}(K(\frac{1-\tau T}{Th}), \dots, K(\frac{T-\tau T}{Th}))$  and  $W(\tau) = \iota_n \otimes w(\tau)$ . A pooled local linear estimator is then defined by the minimization of the following objective function

$$[Y - X\beta - D\mu - Z(\tau)(a, b)^T]^T W(\tau) [Y - X\beta - D\mu - Z(\tau)(a, b)^T].$$

Chen, Gao, and Li (2012a) proceeded by first concentrating out the time trend. That is, for given  $\mu$  and  $\beta$ , they estimated  $f(\tau)$  and  $f'(\tau)$  by

$$\begin{aligned} (\widehat{f}_{\mu, \beta}(\tau), \widehat{f}'_{\mu, \beta}(\tau)) &= \arg \min_{a, b} \sum_{i=1}^n \sum_{t=1}^T \left[ Y_{i,t} - X_{i,t}^T \beta - \mu_i - a - b \left( \frac{t}{T} - \tau \right) \right]^2 \\ &\quad K \left( \frac{t - \tau T}{Th} \right). \end{aligned}$$

Define  $S(\tau) = [Z^T(\tau)W(\tau)Z(\tau)]^{-1}Z^T(\tau)W(\tau)$ , one can show that

$$\widehat{f}_{\mu,\beta} = (1, 0)S(\tau)(Y - X\beta - D\mu) = s(\tau)(Y - X\beta - D\mu), \quad (2.36)$$

where  $s(\tau) = (1, 0)S(\tau)$ . And, the profile least square estimator of  $\mu$  and  $\beta$  is then defined by

$$(\widehat{\mu}^T, \widehat{\beta}^T)^T = \arg \min_{\mu, \beta} \sum_{i=1}^n \sum_{t=1}^T \left[ Y_{i,t} - X_{i,t}^T \beta - \mu_i - \widehat{f}_{\mu,\beta} \left( \frac{t}{T} \right) \right]^2. \quad (2.37)$$

Further define  $R^* = (I_{nT} - S)R$  for  $R = Y, X$  or  $D$  and  $M^* = I_{nT} - D^{*T}(D^{*T}D^*)^{-1}D^{*T}$ .

Then, (2.37) is simplified to the usual least squares problem of modified data  $(Y^*, X^*, D^*)$ , which gives:

$$\begin{aligned} \widehat{\beta} &= (X^{*T} M^* X^{*T})^{-1} X^{*T} M^* Y^*, \\ \widehat{\mu} &= (D^{*T} D^*)^{-1} D^{*T} (Y^* - X^{*T} \widehat{\beta}). \end{aligned}$$

Finally plugging  $\widehat{\beta}$  and  $\widehat{\mu}$  into (2.36) yields the local linear estimator of the trend function,

$$\widehat{f}(\tau) = s(\tau)(Y - X\widehat{\beta} - D\widehat{\mu}).$$

Under fairly general conditions, Chen, Gao, and Li (2012a) showed that the parametric parameters are asymptotically normally distributed with a root- $nT$  convergence rate as the time series length  $T$  and cross-sectional size  $n$  both tend to infinity, while the nonparametrically estimated trend function converges at a root- $nTh$  rate. They also showed that if the cross-sectional averaging approach of Robinson (2012) is employed, the convergence rate of the estimator of the parametric component is root- $T$  while that of the nonparametric trend estimator remains the same. These observations were confirmed by their Monte Carlo simulations.

Lastly, we note that many new research works have taken place as we were writing this survey. For example, for large  $n$  and finite  $T$ , Freyberger (2012) considered nonparametric panel data models with interactive fixed effects in the form of  $Y_{i,t} = m_t(X_{i,t}, \delta_i^T g_t + \epsilon_{i,t})$ , where  $m_t(\cdot, \cdot)$  is an unknown, time-varying, smooth measurable function that is strictly increasing in the second argument, while for large  $n$  and large  $T$ , Su and Zhang (2013) focused on the sieve estimation and specification testing for nonparametric dynamic panel data models with interactive fixed effects. We thank the referee for providing with us these two references.

## 10.3 CONDITIONAL QUANTILE REGRESSION MODELS

---

Section 10.2 focuses on studying conditional mean regression models, which can be used to predict the average response of a dependent variable with respect to changes in covariates. However, with heteroskedastic data, mean regression models disguise heterogeneity and are not robust to tail reactions. Also, in the presence of strong data heteroskedasticity, one may find it attractive to estimate extreme causal relationship of an explanatory variable to a dependent variable, which can be suitably studied in quantile regression framework.

Since the seminal works of Koenker and Bassett (1978, 1982), quantile regression models have gradually appeared on the radar of both econometricians and applied economists in the presence of data heteroskedasticity where quantile estimations across different probability masses provide a better view of causal relationship across a conditional distribution than the average information revealed by conditional mean regression studies. For example, Yu, Lu, and Stander (2003), Buchinsky (1998), and a special issue of Empirical Economics (2001, Vol. 26) presented empirical applications of parametric quantile regressions to labor, microeconomics, macroeconomics, and finance. Koenker (2005) gave a full-length survey on parametric quantile regressions for estimation, tests and computational issues. Early works on nonparametric estimation include Bhattacharya and Gangopadhyay (1990), Chaudhuri (1991), Fan, Hu, and Truong (1994), Koenker, Ng, and Portnoy (1994), Yu and Jones (1998), Cai (2002), Lee (2003), and Sun (2005). Zheng (1998) extended the residual-based nonparametric test to test for parametric linear quantile regression (QR) models against nonparametric QR models. All these research works considered either cross-sectional or time series data. In this section, we review some recent development of quantile regression models for panel data.

Let  $F_{i,t}(y|x) = \Pr(Y_{i,t} \leq y|X_{i,t} = x)$  be the conditional distribution of  $Y_{i,t}$  given  $X_{i,t} = x$ . Let  $Q_p(x)$  be the 100pth conditional quantile function of  $Y_{i,t}$  given  $X_{i,t} = x$  for  $p \in (0, 1)$ . Define  $Q_p(x) = \inf\{y : F_{i,t}(y|x) \geq p\}$ . When  $F_{i,t}(y|x)$  is absolutely continuous in  $y$  for all  $x$ ,  $Q_p(x)$  is the unique solution to  $F_{i,t}(y|x) = p$ .

This section contains two subsections. Section 10.3.1 considers nonparametric pooled panel data quantile regression (QR) models when  $u_{i,t} = \sigma(X_{i,t})\epsilon_{i,t}$  and  $\epsilon_{i,t} \sim i.i.d.(0, 1)$  is independent of  $\{X_{i,t}\}$ . Section 10.3.2 considers fixed-effects panel data QR models with  $u_{i,t} = \mu_i + \sigma(X_{i,t})\epsilon_{i,t}$ , where  $\mu_i \sim i.i.d.(0, \sigma_\mu^2)$ , and  $\epsilon_{i,t} \sim i.i.d.(0, 1)$  is independent of  $\{X_{i,t}\}$ .

### 10.3.1 Pooled Panel Data Quantile Regression Models

Consider a nonparametric panel data mean regression model

$$Y_{i,t} = m(X_{i,t}) + \sigma(X_{i,t})\epsilon_{i,t}, \quad i = 1, \dots, n, t = 1, \dots, T,$$

where  $\epsilon_{i,t} \sim i.i.d. (0, 1)$ . Suppose that  $\epsilon_{i,t}$  has a common Lebesgue probability function  $f_\epsilon(z) > 0$  for all  $z \in R$ . Let  $\gamma_p$  be the  $100p$ th quantile of  $\epsilon_{i,t}$ . Then, we can define a  $100p$ th nonparametric panel data QR model by

$$Y_{i,t} = m_p(X_{i,t}) + \sigma(X_{i,t})v_{i,t} \quad (3.1)$$

where  $v_{i,t} = \epsilon_{i,t} - \gamma_p$  with  $\Pr(v_{i,t} \leq 0 | X_{i,t}) \equiv p \in (0, 1)$ , and  $m_p(x) \equiv m(x) + \gamma_p\sigma(x)$ . If  $\sigma(x) \equiv \sigma_0$  almost surely for all  $x \in R$ ,  $m_p(x)$  is parallel to each other for different values of  $p$ .

When  $m_p(\cdot)$  is known up to a finite number of unknown parameters, say  $m_p(x) \equiv m_p(x, \theta)$ , where  $m_p(\cdot)$  has a known functional form and  $\theta$  is a  $d \times 1$  unknown parameter vector with a finite integer  $d \geq 1$ , a consistent estimator of  $\theta$  can be obtained from the following optimization problem

$$\hat{\theta}_p = \arg \min_{\theta \in \Theta} \sum_{i=1}^n \sum_{t=1}^T \omega_{i,t} \rho_p(Y_{i,t} - m_p(X_{i,t}, \theta)), \quad (3.2)$$

where the first element of  $X_{i,t}$  is 1,  $\Theta \subset R^d$  is a compact subset of  $R^d$ ,  $\rho_p(u)$  denotes the “check” function  $\rho_p(u) = u[p - I(u < 0)]$ ,  $I(A)$  is the indicator function equal to 1 if event  $A$  occurs, and  $\omega_{i,t}$  is a weight attached to the  $(i, t)$ -th observation. The objective function (3.2) gives the following *estimating equation*

$$\sum_{i=1}^n \sum_{t=1}^T \omega_{i,t} \frac{\partial m_p(X_{i,t}, \theta)}{\partial \theta^T} [p - I(Y_{i,t} \leq m_p(X_{i,t}, \theta))] = 0. \quad (3.3)$$

For a linear regression case, eq. (3.3) becomes  $\sum_{i=1}^n \sum_{t=1}^T \omega_{i,t} X_{i,t} \xi_{i,t} = 0$ , where  $\xi_{i,t} = p - I(Y_{i,t} \leq m_p(X_{i,t}, \theta))$ . If one chooses  $\omega_{i,t} \equiv 1$ , one ignores possible within-group correlation and cross-sectional heteroskedasticity. Although the resulted estimator under working independence and homoskedasticity is consistent and asymptotically normally distributed if  $m_p(x, \theta)$  is correctly specified for  $p \in (0, 1)$ , with a stationary AR(1) error term, the bootstrap results given in Karlsson (2009) indicated that the quantile estimator with  $\omega_{i,t} \equiv 1$  exhibits larger bias than with  $\omega_{i,t} \equiv \hat{\sigma}_{ii}$ , the  $i$ th group’s sample standard deviation of estimated residuals. Assuming  $\Pr(v_{i,t} \leq 0 \text{ and } v_{i,s} \leq 0) = \delta$ , a constant for any  $t \neq s$ , Fu and Wang (2012) obtained an *exchangeable covariance matrix* for  $\xi_i = (\xi_{i,1}, \dots, \xi_{i,T})^T$ ; i.e.,

$$V_i = \text{Var}(\xi_i \xi_i^T | X_i) = p(1-p) \left[ (1-\gamma) I_T + \gamma \boldsymbol{\iota}_T \boldsymbol{\iota}_T^T \right] \quad (3.4)$$

where  $\gamma = \text{Corr}(\xi_{i,t}, \xi_{i,s}) = (\delta - p^2) / (p - p^2)$  for any  $t \neq s$ . Following Jung's (1996) quasi-likelihood approach, Fu and Wang (2012) then introduced a weighted estimating equation

$$M_p(\theta) = \sum_{i=1}^n X_i^T V_i^{-1} \xi_i.$$

Setting  $M_p(\theta) = 0$  gives a weighted estimator of  $\theta$  denoted by  $\widehat{\theta}_w$ , which would be consistent and asymptotically normally distributed if  $\gamma$  were known. The simulation results showed that  $\widehat{\theta}_w$  is more efficient than the working independence estimator when strong within-group correlation exists, although  $\widehat{\theta}_w$  has larger bias than the working independence estimator. Fu and Wang (2012) then split the weighted estimating equation  $M_p(\theta)$  into two components, the within-group and between-group estimating equations, and proposed a combined estimator that optimizes the combination of the two estimating functions in the way similar to the generalized method of moments approach. To save space, we omit the details here. As Fu and Wang (2012) considered the case with large  $n$  and finite  $T$ , our survey on estimating working correlation matrix in Section 10.2.1 can be relevant here; e.g., Qu, Lindsay, and Li (2000) and Fan, Huang, and Li (2007).

When  $p = 0.5$ , He, Fu, and Fung (2003) compared the estimator calculated under working independence, Jung's (1996) quasi-likelihood estimator and Huggins's (1993) robust estimator, and their Monte Carlo simulation results showed that Huggins' (1993) estimator under exchangeable covariance assumption performs significantly more efficient than the estimator under working independence. However, the exchangeable variance structure imposed by Fu and Wang (2012) is rather restrictive for panel data models. For seemingly unrelated quantile regression models with dependent cross-equation errors conditional on regressors, Jun and Pinkse (2009) used k-nearest neighbor methods to estimate the unknown optimal weight appearing in the estimating equation and showed that the resulted estimator is asymptotically efficient in the sense that the proposed estimator achieves corresponding semiparametric efficient bound.<sup>1</sup>

When  $m_p(x)$  is an unknown smooth measurable function, following the local linear regression approach of Fan, Hu, and Truong (1994), Sun (2006) estimated the unknown 100pth quantile curve  $m_p(x)$  at an interior point  $x$  as follows:

$$(\widehat{\theta}_0, \widehat{\theta}_1) = \arg \min_{\theta \in \Theta} \sum_{i=1}^n \sum_{t=1}^T \rho_p(Y_{i,t} - \theta_0 - \theta_1(X_{i,t} - x)) K\left(\frac{X_{i,t} - x}{h}\right) \quad (3.5)$$

where  $\rho_p(u)$  is “check” function defined above,  $K(\cdot)$  is a second-order kernel function, and  $h$  is the bandwidth parameter. Also,  $\widehat{m}_p(x) = \widehat{\theta}_0$  estimates  $m_p(x)$ , and  $\widehat{\theta}_1$  estimates  $m'_p(x)$ , the first-order derivative of  $m_p(x)$ . Under some regularity conditions,

Sun (2006, Lemma 1) showed that  $\widehat{m}_p(x) \xrightarrow{P} m_p(x)$  if  $h \rightarrow 0$  and  $nh \rightarrow \infty$  as  $n \rightarrow \infty$  for a finite  $T$ . She did not derive the asymptotic normality result for the estimator as

the primary interest of Sun (2006) is to conduct a nonparametric poolability test (see Section 10.4).

He, Zhu, and Fung (2002) and Wang, Zhu, and Zhou (2009) studied partially linear panel data QR models, where the latter also considered semiparametric varying coefficient models. Both papers assumed an unknown smooth function of time and took nonparametric panel data QR models as a special case. Applying series approximation method, both papers derived the consistency result of the nonparametric estimator for large  $n$  and finite  $T$ . Parallel to our review in Section 10.2.1, we conjecture that the kernel-based estimator is better calculated with working independence, while the sieve estimator benefits from a properly chosen working dependence matrix as a weighting matrix.

Further, combining information from different quantile regressions, a *composite quantile regression* (CQR) method is showed to be asymptotically more efficient than the quantile curve estimators for a given probability mass when  $\sigma(x) \equiv \sigma_0$ ; see, Koenker (1984), Zou and Yuan (2008), and Kai, Li, and Zou (2010).

### 10.3.2 Fixed Effects Panel Data Quantile Regression Models

We will show that fixed effects panel data quantile regression (QR) models naturally result from panel data mean regression models with unobserved individual factors,

$$Y_{i,t} = m(X_{i,t}) + \mu_i + \sigma(X_{i,t})\epsilon_{i,t}, \quad i = 1, \dots, n, t = 1, \dots, T, \quad (3.6)$$

where  $\mu_i$  is the unobserved factor from the  $i$ th unit and is independent of the error term  $\{\epsilon_{i,t}\}_{t=1}^T$ ,  $\epsilon_{i,t} \sim i.i.d. (0, 1)$  has a common probability function  $f_\epsilon(z) > 0$  for all  $z \in R$ , and  $\{\epsilon_{i,t}\}_{t=1}^T$  is independent of  $\{X_{i,t}\}_{t=1}^T$ . Let  $\gamma_p$  be the 100pth quantile of  $\epsilon_{i,t}$ . Then, we obtain a 100pth fixed-effects panel data QR model

$$Y_{i,t} = m_p(X_{i,t}) + \mu_i + \sigma(X_{i,t})v_{i,t}, \quad (3.7)$$

where  $v_{i,t} = \epsilon_{i,t} - \gamma_p$  and  $\Pr(v_{i,t} \leq 0 | X_{i,t}) \equiv p \in (0, 1)$ , and  $m_p(x) \equiv m(x) + \gamma_p\sigma(x)$ . For individual unit  $i$ , the 100pth regression quantile of  $Y_{i,t}$  given  $X_{i,t}$  and  $\mu_i$  is

$$m_{p,i}(X_{i,t}) \equiv m_p(X_{i,t}) + \mu_i, \quad (3.8)$$

where unit  $i$  shares the same individual unobservable factor  $\mu_i$ .

Naturally, one attempts to remove the unobserved fixed effects via first-differencing method. However, the first-differencing method does not work for panel data QR models with fixed effects because the quantile operator is nonlinear in nature and does not preserve additive property.

If one applies the first difference to model (3.7), he/she obtains

$$Y_{i,t} - Y_{i,t-1} = m_p(X_{i,t}) - m_p(X_{i,t-1}) + \sigma(X_{i,t})v_{i,t} - \sigma(X_{i,t-1})v_{i,t-1}, \quad (3.9)$$

which cancels out not only  $\mu_i$  but also time-invariant components in  $m_p(X_{i,t})$ . Letting  $\varepsilon_{i,t} = \sigma(X_{i,t})\nu_{i,t} - \sigma(X_{i,t-1})\nu_{i,t-1}$ , this model looks like a nonparametric additive panel data model. However, it is generally untrue that  $E(\varepsilon_{i,t}|X_{i,t}, X_{i,t-1}) = -[\sigma(X_{i,t}) - \sigma(X_{i,t-1})]\gamma_p$  equals a constant or  $\Pr(\varepsilon_{i,t} \leq 0|X_{i,t}, X_{i,t-1}) = p$ . In addition, it is not clear how to set the identification condition in model (3.9) so that  $m_p(x)$  is monotonic in  $p \in (0, 1)$ ; see, Rosen (2012) for identification issue with large n and fixed T.

The concern given above also holds for a special case like linear regression models. Assuming  $\sigma(X_{i,t}) = X_{i,t}^T\delta$  and letting  $\beta_p = \beta + \gamma_p\delta$  and  $\varepsilon_{i,t} = (X_{i,t}^T\delta)\nu_{i,t} - (X_{i,t-1}^T\delta)\nu_{i,t-1}$ , model (3.7) becomes

$$Y_{i,t} = X_{i,t}^T\beta_p + \mu_i + (X_{i,t}^T\delta)\nu_{i,t}, \quad (3.10)$$

and the first-differencing model becomes  $Y_{i,t} - Y_{i,t-1} = (X_{i,t} - X_{i,t-1})^T\beta_p + \varepsilon_{i,t}$ , where  $E(\varepsilon_{i,t}|X_{i,t}, X_{i,t-1}) = -(X_{i,t} - X_{i,t-1})^T\delta\gamma_p$ . Unless  $\delta\gamma_p = 0$ , one can not identify  $\beta_p$  if OLS estimator is applied to the first-differencing model. In addition, the 100pth quantile of  $\nu_{i,t}$  equal to 0 does not automatically ensure the 100pth conditional quantile of  $\varepsilon_{i,t}$  equal to 0 since the quantile operator is not a linear operator.

Assuming  $m(\cdot)$  and  $\sigma(\cdot)$  are linear functions, Koenker (2004) considered a linear fixed-effects panel data 100pth QR model

$$Y_{i,t} = X_{i,t}^T\beta_p + \mu_i + e_{p,i,t}, \quad (3.11)$$

where  $\Pr(e_{p,i,t} \leq 0|X_{i,t}) = p \in (0, 1)$ . He proposed a *penalized quantile regression fixed effects* (PQRFE) estimator

$$\begin{aligned} (\widehat{\beta}, \{\widehat{\mu}_i\}_{i=1}^n) = \arg \min_{\beta, \{\mu_i\}_{i=1}^n} & \sum_{k=1}^q w_k \sum_{i=1}^n \sum_{t=1}^T \rho_{p_k}(Y_{i,t} - X_{i,t}^T\beta_{p_k} - \mu_i) \\ & + \lambda \sum_{i=1}^n |\mu_i|, \end{aligned} \quad (3.12)$$

where  $\beta = \{\beta_{p_k}\}_{k=1}^q$ ,  $w_k$  is a subjectively-chosen weight assigned to the  $k$ th quantile and controls the relative influence of the quantiles at probability mass  $(p_1, \dots, p_k)$  on the estimation of the common individual effects,  $\mu_i$ 's. The  $l_1$  penalty function serves to shrink  $\mu_i$ 's toward zero and other penalty functions may also be used.  $\lambda \geq 0$  is called a *tuning parameter* or a *regularization parameter*. When  $\lambda \rightarrow 0$ ,  $\widehat{\mu}$ 's are not shrunk and we have a fixed effects QR model; when  $\lambda \rightarrow \infty$ ,  $\widehat{\mu}$  is close to zero and we have a pooled panel data QR model. Lamarche (2010) proposed to choose  $\lambda$  under the assumption that  $\mu_i$  is independent of  $\{X_{i,t}\}_{t=1}^T$  for all  $i$ . When setting  $\lambda = 0$ , both Koenker (2004) and Kato, Galvao, and Montes-Rojas (2012) showed consistency of  $\widehat{\beta}$  if  $n/T^s \rightarrow 0$  for some  $s > 0$  as  $T \rightarrow \infty$  and  $n \rightarrow \infty$ . However, they found that the asymptotic normality result of  $\widehat{\beta}$  requires a stronger condition on  $T$ ; i.e.,  $n^2(\log n)^3/T \rightarrow 0$  as  $n \rightarrow \infty$  and  $T \rightarrow \infty$ .

Galvao (2011) considered fixed effects dynamic panel data linear QR models,

$$Y_{i,t} = Y_{i,t-1}\delta_p + X_{i,t}^T\beta_p + \mu_i + e_{p,i,t}, \quad (3.13)$$

where  $X_{i,t}$ ,  $\mu_i$ , and  $e_{p,i,t}$  are generated under the same assumptions described in model (3.10). For fixed effects dynamic panel data linear mean regression models, the fixed-effects least squared estimator may be biased for a moderate  $T$  and its consistency depends critically on the assumptions on initial conditions; see Hsiao (2003), Arellano (2003), Phillips and Sul (2007) for examples. One way to construct a consistent estimator independent of initial conditions is to applying the instrumental variables approach; e.g., Anderson and Hsiao (1981, 1982) and Arellano and Bond (1991). Similar arguments also apply to the PQRFE estimator; see Monte Carlo evidence in Galvao (2011). Following the methodology in Chernozhukov and Hansen (2008) and Harding and Lamarche (2009), Galvao (2011) proposed an *instrumental variables quantile regression with fixed effects* (IVQRFE) estimator

$$\begin{aligned} \widehat{\delta} &= \arg \min_{\delta} \|\widehat{\gamma}(\delta)\|_A \\ &= (\widehat{\beta}(\delta), \{\widehat{\mu}_i(\delta)\}_{i=1}^n, \widehat{\gamma}(\delta)) \\ &= \arg \min_{(\beta, \{\mu_i\}_{i=1}^n, \gamma)} \sum_{k=1}^q w_k \sum_{i=1}^n \sum_{t=2}^T \rho_{p_k} \left( Y_{i,t} - Y_{i,t-1}\delta_{p_k} - X_{i,t}^T\beta_{p_k} - \mu_i - Z_{i,t}^T\gamma_{p_k} \right), \end{aligned}$$

where  $\alpha = \{\alpha_{p_k}\}_{k=1}^q$  for  $\alpha = \beta, \delta$  or  $\gamma$ ,  $Z_{i,t}$ 's are proper instrumental variables correlated with  $Y_{i,t-1}$  but uncorrelated with  $e_{p,i,t}$  for all  $i$  and  $t$ , and  $\|x\|_A = \sqrt{x^T A x}$  and  $A$  is a positive definite matrix. Through a grid search over  $\delta$ , one obtains  $\widehat{\delta}$  then defines  $\widehat{\beta} = \widehat{\beta}(\widehat{\delta})$ . If  $T \rightarrow \infty$  as  $n \rightarrow \infty$  and  $n^a/T \rightarrow 0$  for some  $a > 0$  and under some other conditions, Galvao (2011) showed consistency and asymptotic normality of the proposed estimator.

Galvao and Montes-Rojas (2010, p. 3479) commented that “*Unfortunately, in estimation of dynamic panel data models IVs become less informative when the autoregressive parameter increases toward one, and also when the variability of the FE increases.*” Hence, combining Koenker (2004) and Galvao (2011), Galvao and Montes-Rojas (2010) introduced a *penalized IVQRFE* (IVQRFE) estimator

$$\begin{aligned} \widehat{\delta}(\lambda) &= \arg \min_{\delta} \|\widehat{\gamma}(\delta, \lambda)\|_A \\ &= (\widehat{\beta}(\delta, \lambda), \{\widehat{\mu}_i(\delta, \lambda)\}_{i=1}^n, \widehat{\gamma}(\delta, \lambda)) \\ &= \arg \min_{(\beta, \{\mu_i\}_{i=1}^n, \gamma)} \sum_{k=1}^q w_k \sum_{i=1}^n \sum_{t=2}^T \rho_{p_k} \left( Y_{i,t} - Y_{i,t-1}\delta_{p_k} - X_{i,t}^T\beta_{p_k} - \mu_i - Z_{i,t}^T\gamma_{p_k} \right) + \lambda \sum_{i=1}^n |\mu_i| \end{aligned}$$

and the final parameter estimators are  $\widehat{\delta}(\lambda)$  and  $\widehat{\beta}(\lambda) = \widehat{\beta}(\widehat{\delta}(\lambda), \lambda)$ . With the same assumption on  $\lambda$ ,  $T$  and  $n$  and other conditions, Galvao and Montes-Rojas (2010)

showed consistency and delivered asymptotic normality result for the IVQRFE estimator. For the choice of  $\lambda$  and the comparative performance of the PQRFE, IVQRFE, and PIVQRFE estimators, interested readers are referred to their original papers for details.

Alternatively, Canay (2011) proposed a different estimator than Koenker (2004) for model (3.10), based on the following observation: when both  $n \rightarrow \infty$  and  $T \rightarrow \infty$ , the least-squared dummy variable approach can be used to estimate both  $\beta$  and  $\mu_i$ 's consistently from the fixed-effects panel data linear mean regression model. Let us call these estimates  $\tilde{\mu}_i$ 's. Canay (2011) then proposed to estimate  $\beta_p$  after subtracting  $\tilde{\mu}_i$  from  $Y_{i,t}$  and showed that the estimator of  $\beta_p$  is consistent and asymptotically normally distributed with a convergence rate of  $\text{root}(nT)$  if  $n/T^s \rightarrow 0$  for some  $s \in (1, \infty)$  as  $n \rightarrow \infty$  and  $T \rightarrow \infty$ .

To the best of our knowledge, we have not found published works estimating nonparametric fixed-effects panel data QR models at the point of our writing this survey. However, for both large  $T$  and  $n$ , we conjecture that Canay's (2011) methodology is applicable to nonparametric framework through two steps. In the first step, Henderson, Carroll, and Li (2008) or Lin, Li, and Sun (2014) method can be used to estimate the unobserved fixed effects  $\mu_i$ 's. Naming the estimator  $\hat{\mu}_i$  and we have  $Y_{i,t} - \hat{\mu}_i = m_p(X_{i,t}) + v_{i,t}$  with  $v_{i,t} = \sigma(X_{i,t})\epsilon_{i,t} + \mu_i - \hat{\mu}_i$ . Intuitively,  $\sup_i |\mu_i - \hat{\mu}_i| = o_p(T^{\xi-1/2})$  holds for some small  $\xi \in (0, 1/2)$ . In the second step, one can estimate  $m_p(x)$  from the pooled model  $Y_{i,t} - \hat{\mu}_i = m_p(X_{i,t}) + v_{i,t}$  as long as one can show that the impact of  $\mu_i - \hat{\mu}_i$  is asymptotically ignorable as both  $n$  and  $T$  grow.

### 10.3.3 Other Issues

Our survey does not include the literature of generalized QR regression models with nonlinearly transformed dependent variable. Intuitively we expect that the methodology discussed above can be extended to the generalized QR models through the so-called *equivariance to monotone transformation* property of quantile operator. Specifically, let  $w(\cdot)$  be a *nondecreasing* function along the real line, and for any random variable  $Y$ , let  $Q_p^w(x)$  and  $Q_p(x)$  be the respective 100pth conditional quantile of  $w(Y)$  and  $Y$  given  $X = x$ . The '*equivariance to monotone transformations*' property means that

$$Q_p^w(x) = w(Q_p(x)). \quad (3.14)$$

If  $w(\cdot)$  is a known monotonic function, then the above-mentioned methods can be used to consistently estimate panel data QR models with nonlinearly transformed dependent variables (e.g., Box-Cox transformed dependent variable; Powell (1986) for censored data, Koenker and Bilias (2001) for duration data). When it is unknown,  $w(\cdot)$  can be estimated consistently by nonparametric methods; see, e.g., Mu and Wei (2009).

A word of caution is given here: fitted quantile curves, by both parametric and nonparametric methods, can suffer *quantile crossing* problem; that is, a 100pth (nonparametric) quantile curve can cross over a 100qth (nonparametric) quantile curve for

some  $p < q$ . The quantile crossing may result from model misspecification, collinearity, or outliers. Neocleous and Portnoy (2008) explained how to correct this problem for linear quantile regression models, while He (1997), Dette and Volgushev (2008), Chernozhukov, Fernández-Val, and Galichoni (2010), and Bondell, Reich, and Wang (2010) suggested solutions for nonparametric quantile regression models for independent data.

## 10.4 NONSEPARABLE MODELS

All models reviewed in the previous sections assume that the individual effects and idiosyncratic error terms enter the models additively. In this section we review methods on nonseparable nonparametric panel data models that relax this assumption. Depending on the treatment of the error terms, there are two types of nonseparable models. The first type is a *partially separable* model

$$Y_{i,t} = m(X_{i,t}, \mu_i) + \epsilon_{i,t}, \quad i = 1, \dots, n, \quad t = 1, \dots, T,$$

where the individual effect  $\mu_i$  enters the conditional mean in an unknown form, while the idiosyncratic error term is assumed to be additive. The second type is a *fully nonseparable* model

$$Y_{i,t} = m(X_{i,t}, \mu_i, \epsilon_{i,t}), \quad i = 1, \dots, n, \quad t = 1, \dots, T,$$

where neither  $\mu_i$  nor  $\epsilon_{i,t}$  is separable from the conditional mean.

Su and Ullah (2011, Section 7) provided an in-depth review of nonseparable nonparametric models, focusing on two papers: Evdokimov (2009) and Altonji and Matzkin (2005). The former considered the partially nonseparable model. Assuming that  $\{X_{i,1}, \dots, X_{i,T}, \mu_i, \epsilon_i\}$  are i.i.d. sequence and that the conditional density of  $\epsilon_{i,t}$  given  $(X_{i,1}, \dots, X_{i,T}, \mu_i)$  equals that of  $\epsilon_{i,t}$  given  $X_{i,t}$  and other regularity conditions, Evdokimov (2009) proposed a kernel-based conditional deconvolution method to estimate the structural function  $m(x, \mu)$ . Altonji and Matzkin (2005) considered the fully nonseparable model, assuming a conditional density restriction

$$f(\mu, \epsilon | X', Z') = f(\mu, \epsilon | X'', Z'')$$

for specific values  $(X', Z')$  and  $(X'', Z'')$  and a conditional independence condition

$$f(\mu, \epsilon | X, Z) = f(\mu, \epsilon | Z).$$

Under these assumptions, Altonji and Matzkin (2005) proposed a nonparametric control function estimator to estimate the local average response function

$$\beta(x) = \int \frac{\partial}{\partial x} m(x, \mu, \epsilon) dF(\mu, \epsilon),$$

where  $F(\mu, \epsilon)$  is the joint distribution of  $\mu$  and  $\epsilon$ . Interested readers are referred to the original papers and Su and Ullah (2011) for discussions of relevant works.

In this section, we focus on a recent paper by Chernozhukov et al. (2013) that considered both average and quantile effects in fully nonseparable panel models. A key assumption of that paper is a time homogeneity condition which prescribes the stationarity of the conditional distribution of  $\epsilon_{i,t}$  given  $X_i$  and  $\mu_i$ , where  $X_i = (X_{i,1}, \dots, X_{i,T})^T$ . Formally, this condition is given by

$$\epsilon_{i,t} | (X_i, \mu_i) \stackrel{d}{=} \epsilon_{i,1} | (X_i, \mu_i), \text{ for all } t. \quad (4.1)$$

This condition, as suggested by the authors, “*it is like ‘time is randomly assigned’ or ‘time is an instrument’ with the distribution of factors other than  $x$  not varying over time, so that changes in  $x$  over time can help identify the effect of  $x$  on  $y$ .*”

The authors focused on two objects of interest, the average structural function (ASF) and the quantile structural function (QSF). The ASF is given by

$$g(x) = E[m(x, \mu_i, \epsilon_{i,t})] = \int m(x, \mu, \epsilon) dF(\mu, \epsilon).$$

The average treatment effect (ATE), as in the treatment effect literature, of changing  $x$  from  $x^b$  (before) to  $x^a$  (after) is then

$$\Delta = g(x^a) - g(x^b).$$

The authors showed that the ATE obtained this way is identical to that of the conditional mean model, which is given by  $\int [m(x^a, \mu) - m(x^b, \mu)] dF(\mu)$ .

The QSF,  $Q(p, x)$ , is the 100pth quantile of  $m(x, \mu_i, \epsilon_{i,t})$ , given by

$$Q(p, x) = G^{-1}(p, x), \text{ where } G(y, x) = E[I(m(x, \mu_i, \epsilon_{i,t}) \leq y)].$$

The 100pth quantile treatment effect (QTE) of changing  $x$  from  $x^b$  to  $x^a$  is given by

$$\Delta_p = Q(p, x^a) - Q(p, x^b).$$

The authors further assumed that the regressors are discrete with finite support. Let  $I(X_{i,t} = x)$  denote the indicator function that is equal to one when  $X_{i,t} = x$  and zero otherwise and  $T_i(x) = \sum_{t=1}^T I(X_{i,t} = x)$ . Also, let  $D_i = I(T_i(x^a) > 0)I(T_i(x^b) > 0)$  be the indicator that  $X_i$  includes both  $x^a$  and  $x^b$  for some period. The ATE is then defined as

$$\delta = E[m(x^a, \mu_i, \epsilon_{i,1}) - m(x^b, \mu_i, \epsilon_{i,1}) | D_i = 1].$$

A simple estimator of the conditional ATE  $\delta$  is

$$\hat{\delta} = \frac{\sum_{i=1}^n D_i [\bar{Y}_i(x^a) - \bar{Y}_i(x^b)]}{\sum_{i=1}^n D_i},$$

where

$$\bar{Y}_i(x) = \begin{cases} T_i(x)^{-1} \sum_{t=1}^T I(X_{i,t} = x) Y_{i,t}, & T_i(x) > 0; \\ 0, & T_i(x) = 0. \end{cases}$$

Consistency of this ATE estimator can be established based on

$$E \left\{ D_i \left[ \bar{Y}_i(x^a) - \bar{Y}_i(x^b) \right] \right\} = E \left\{ D_i \left[ m(x^a, \mu_i, \epsilon_{i,1}) - m(x^b, \mu_i, \epsilon_{i,1}) \right] \right\}.$$

A consistent estimator of the asymptotic variance of  $\sqrt{n}(\hat{\delta} - \delta)$  is given by  $n^{-1} \sum_{i=1}^n \hat{\psi}_i^2$  where  $\hat{\psi}_i = nD_i[\bar{Y}_i(x^a) - \bar{Y}_i(x^b) - \hat{\delta}] / \sum_{i=1}^n D_i$ . The authors also showed that the usual panel data within (linear fixed effects) estimator is not a consistent estimator of  $\delta$  because the within estimator constrains the slope coefficients to be the same for each  $i$  when the slope is actually varying with  $i$ .

We can identify and estimate the conditional QTE in a similar manner. Let  $G(y, x|D_i = 1) = \Pr(m(x, \mu_i, \epsilon_{i,1}) \leq y|D_i = 1)$  denote the cumulative distribution function of  $m(x, \mu_i, \epsilon_{i,1})$  conditional on  $D_i = 1$ . The QTE conditional on  $D_i = 1$  is

$$\delta_p = G^{-1}(p, x^a|D_i = 1) - G^{-1}(p, x^b|D_i = 1).$$

To calculate this estimator, we need an estimator of  $G(y, x|D_i = 1)$ . Let  $\Phi$  be the standard Gaussian kernel function and  $h$  be a bandwidth. A smoothed estimator of  $G(y, x|D_i = 1)$  is given by

$$\hat{G}(p, x|D_i = 1) = \frac{\sum_{i=1}^n D_i \bar{G}_i(y, x)}{\sum_{i=1}^n D_i},$$

where

$$\bar{G}_i(x) = \begin{cases} T_i(x)^{-1} \sum_{t=1}^T I(X_{i,t} = x) \Phi(\frac{y - Y_{i,t}}{h}), & T_i(x) > 0; \\ 0, & T_i(x) = 0. \end{cases}$$

An estimator of  $\delta_p$  is then given by

$$\hat{\delta}_p = \hat{G}^{-1}(p, x^a|D_i = 1) - \hat{G}^{-1}(p, x^b|D_i = 1).$$

A consistent estimator of the asymptotic variance of  $\sqrt{n}(\hat{\delta}_p - \delta_p)$  is given by  $n^{-1} \sum_{i=1}^n \hat{\psi}_{p,i}^2$  where

$$\hat{\psi}_{p,i} = \frac{nD_i}{\sum_{i=1}^n D_i} \left[ \frac{\bar{G}_i(\hat{q}^a, x^a) - p}{\hat{G}'(\hat{q}^a, x^a|D_i = 1)} - \frac{\bar{G}_i(\hat{q}^b, x^b) - p}{\hat{G}'(\hat{q}^b, x^b|D_i = 1)} \right]$$

and  $\hat{G}'(y, x|D_i = 1) = \partial \hat{G}(y, x|D_i = 1) / \partial y$ .

For  $x$ 's that do not appear in  $X_i$ , their effects can not be identified. The authors proposed a method to estimate their bounds nonparametrically. Suppose for all  $x$ ,  $B_l \leq m(x, \mu_i, \epsilon_{i,1}) \leq B_u$  for known constants  $B_l$  and  $B_u$ . Let  $\bar{P}(x) = \sum_{i=1}^n I(T_i(x) = 0)/n$  be

the sample frequency of  $x$  not occurring in any time period. Estimated lower and upper bounds for  $g(x)$  are

$$\begin{aligned}\widehat{g}_l(x) &= n^{-1} \sum_{i=1}^n \bar{Y}_i(x) + \bar{P}(x)B_l, \\ \widehat{g}_u(x) &= \widehat{g}_l(x) + \bar{P}(x)(B_u - B_l).\end{aligned}$$

Corresponding estimated lower and upper bounds for the ATE are  $\widehat{\Delta}_l = \widehat{g}_l(x^a) - \widehat{g}_u(x^b)$  and  $\widehat{\Delta}_u = \widehat{g}_u(x^a) - \widehat{g}_l(x^b)$ . The width of these estimated bounds is then given by

$$\widehat{\Delta}_u - \widehat{\Delta}_l = [\bar{P}(x^a) + \bar{P}(x^b)](B_u - B_l).$$

The bounds for the QSF can be obtained in a similar manner, based on the known lower bound of 0 and upper bound of 1 for the cumulative distribution function. For both types of bounds, the authors proposed consistent estimators of their asymptotic variance and established asymptotic normality.

The static model presented above is based on the time homogeneity condition (4.1). The authors also considered a dynamic panel model under the condition

$$\epsilon_{i,t} | (X_{i,t}, \dots, X_{i,1}, \mu_i) \stackrel{d}{=} \epsilon_{i,1} | (X_{i,1}, \mu_i), \quad \text{for all } t.$$

This is a “predetermined” version of time homogeneity under which the conditional distribution given only current and lagged regressors must be time invariant. The conditioning on  $X_{i,1}$  is a way of accounting for the initial conditions of dynamic models.

The authors also considered other extensions such as incorporation of time effects and semiparametric multinomial models. To save space, we do not discuss these extensions. Interested readers are referred to Chernozhukov et al.’s (2013) original paper for details.

A remarkable feature of the method proposed by Chernozhukov et al.’s (2013) is its simplicity. No optimizations or iterations are required to calculate most of the quantities of interest. However, the methods are only applicable to discrete regressors. Given their focus on treatment effect estimation, this restriction is rather innocuous. We conjecture that the methods can be used for models with continuous regressors if they are properly discretized.

## 10.5 NONPARAMETRIC TESTS

---

In nonparametric test literature, it is a common practice to test for functional form specification or conditional variance structure, although tests for random-effects models against fixed-effects models, for data poolability, for cross-sectional independence, and for within-group independence are only relevant to panel/longitudinal data.

Henderson, Carroll, and Li (2008) and Lin, Li, and Sun (2014) constructed nonparametric tests to test for a linear fixed-effects panel data mean regression model against a nonparametric fixed-effects panel data mean regression model, where the former is based on an  $L_2$ -distance in the spirit of Härdle and Mammen (1993) and the latter is essentially a residual-based test in the spirit of Zheng (1996). Henderson, Carroll, and Li (2008) relied on bootstrap methods to remove asymptotic centers and to obtain critical values. Lin, Li and Sun (2014) test statistic converges to a standard normal distribution under the linear fixed effects panel data model and is explosive under the alternative hypothesis. Another advantage of Lin, Li and Sun (2014) model specification test is that it works under both fixed-effects and random-effects model and it is not necessary to pre-test whether the model is a fixed-effects model. However, if the readers are interested in testing a nonparametric random-effects panel data model against a nonparametric fixed-effects panel data model, Sun, Carroll, and Li's (2009) residual-based test can be used. In a semiparametric varying coefficient panel data model setup, they showed that their proposed test is a consistent test and converges to a standard normal distribution under the random-effects null hypothesis.

For the rest of this section, our interests lie in nonparametric poolability test in Section 10.5.1 and nonparametric test for cross-sectional independence in Section 10.5.2.

### 10.5.1 Poolability Tests

Baltagi, Hidalgo, and Li (1996) tested for no structural changes across time for a panel data with large  $n$  and finite  $T$ . Specifically, they considered the following nonparametric panel data model

$$Y_{i,t} = m_t(X_{i,t}) + u_{i,t}, \quad i = 1, 2, \dots, n, t = 1, \dots, T, \quad (5.1)$$

where  $(X_{i,t}, Y_{i,t})$  are independent across index  $i$  with no restriction across index  $t$ ,  $m_t(\cdot)$  is an unknown smooth measurable function that may vary across time  $t$ ,  $E(u_{i,t}|X_{i,t}) = 0$ ,  $E(u_{i,t}^4|X_{i,t}) < \infty$ , and  $\sigma^p(x) = E(u_{i,t}^p|X_{i,t} = x)$  is continuous for  $p = 2$  and 4. The null and alternative hypotheses are

$$H_0 : \Pr\{m_t(X) = m_s(X)\} = 1 \text{ for all } t \neq s. \quad (5.2)$$

$$H_1 : \Pr\{m_t(X) \neq m_s(X)\} > 0 \text{ for some } t \neq s. \quad (5.3)$$

Define  $m(x) \equiv T^{-1} \sum_{t=1}^T m_t(x)$ . Under  $H_0$ ,  $m(x) \equiv m_t(x)$  almost surely for all  $t$ , while under  $H_1$ ,  $m(x) \neq m_t(x)$  with a positive probability over some interval for some  $t$ . Let  $e_{i,t} \equiv Y_{i,t} - m(X_{i,t})$ . Then,  $E[e_{i,t}E(e_{i,t}|X_{i,t})] = E[m_t(X_{i,t}) - m(X_{i,t})]^2 \geq 0$  so that the null hypothesis holds true if and only if  $E[e_{i,t}E(e_{i,t}|X_{i,t})] = 0$  for all  $t$ . Let  $\hat{m}(x)$  be the kernel estimator estimated from the pooled model  $Y_{i,t} = m(X_{i,t}) + e_{i,t}$ . The

test statistic is constructed by estimating the unconditional mean  $E(\cdot)$  by the sample average and replacing  $E(e_{i,t}|X_{i,t})$  by its kernel estimator and  $e_{i,t}$  by  $\hat{e}_{i,t} = Y_{i,t} - \hat{m}(X_{i,t})$ . To remove the random denominator due to the kernel estimation of  $E(e_{i,t}|X_{i,t})$ , the final test is based on  $E[e_{i,t}E(e_{i,t}|X_{i,t})f_t(X_{i,t})]$ , where  $f_t(x)$  is the Lebesgue probability density function of  $X_{i,t}$  for all  $i$  and a given  $t$ . The final test statistic, which is a residual-based test statistic in the spirit of Zheng (1996), is shown to be asymptotically pivotal and consistent and has a standard normal distribution under the null hypothesis.

In nonparametric panel data QR framework, Sun (2006) considered a poolability test for survey data sets comprising  $J$  subgroups and each subgroup has  $n_j$  members obeying a nonparametric QR model

$$Y_{i,j} = m_j(X_{i,j}) + u_{i,j}, \quad i = 1, \dots, n_j, j = 1, \dots, J, \quad (5.4)$$

where  $\Pr(u_{i,j} \leq 0|X_{i,j}) \equiv p \in (0, 1)$ , the total sample size  $n = \sum_{j=1}^J n_j$ , and  $n_j$  is large such that  $\lim_{n \rightarrow \infty} n_j/n = a_j \in (0, 1)$  for all  $j$  and  $J$  is a finite positive integer. The observations  $(X_{i,j}, Y_{i,j})$ ,  $i = 1, \dots, n_j$ ,  $j = 1, \dots, J$  are assumed to be i.i.d. across index  $i$  and independent across index  $j$ . A relevant empirical exercise can be testing median food Engle curves equality across different family types. Zheng (1998) studied the residual-based nonparametric test to test for a parametric QR model against a nonparametric one for cross-sectional data, while Jeong, Härdle, and Song (2012) studied the same test but for weakly dependent time series data. Following Zheng (1998), Sun (2006) constructed her test statistic based on  $\varepsilon_{i,t} = I(Y_{i,t} \leq m(X_{i,t})) - p$ , where  $I(A) = 1$  if event  $A$  occurs and zero otherwise, and  $m(x) \equiv J^{-1} \sum_{j=1}^J m_j(x)$ . Define  $e_{i,j} \equiv Y_{i,t} - m(X_{i,j})$ . Then,  $E[\varepsilon_{i,j}E(\varepsilon_{i,j}|X_{i,j})] = E[\Pr(e_{i,j} \leq 0|X_{i,j}) - p]^2 \geq 0$  so that the null hypothesis of equality holds true if and only if  $E[\varepsilon_{i,j}E(\varepsilon_{i,j}|X_{i,j})] = 0$  holds for all  $j$ . Under the same setup as in Sun (2006), Dette, Wagener, and Volgushev (2011) developed poolability tests based on an  $L_2$ -distance between non-crossing nonparametric quantile curve estimators of Dette and Volgushev (2008), which is measured by

$$M^2 \equiv \sum_{j_1=1}^J \sum_{j_2=1}^{j_1-1} \int [m_{j_1}(x) - m_{j_2}(x)]^2 w_{j_1, j_2}(x) dx \geq 0,$$

where  $w_{j_1, j_2}(\cdot) > 0$  are subjectively-chosen weight functions. The null hypothesis of equality holds if and only if  $M^2 = 0$ . Both Sun's (2006) and Dette, Wagener, and Volgushev's (2011) test statistics are shown to be consistent and asymptotically normal under the null hypothesis of data poolability across subgroups.

Jin and Su (2013) constructed a nonparametric poolability test for a panel data model with cross-sectional dependence that can be applied to test a null model satisfying (2.23), (2.24), and (2.25) from an alternative model  $Y_{i,t} = m_i(X_{i,t}) + u_{i,t}$  with (2.24) and (2.25), where  $\{\nu_{i,t}\}_{t=1}^T$  is a strictly stationary  $\alpha$ -mixing sequence and is independent of  $\{\varepsilon_{i,t}\}$ ,  $\{g_t\}$  is independent of  $\{(\varepsilon_{i,t}, \nu_{i,t})\}$ , and  $\{(\varepsilon_{i,t}, \nu_{i,t})\}$  is independently distributed across index  $i$ . The authors are interested in testing whether

$m_1(x) \equiv m_2(x) \equiv \dots \equiv m_n(x)$  or homogeneous relationship across different cross-sectional units for large  $n$  and large  $T$ . The measure used to differentiate the null hypothesis from the alternative hypothesis is

$$\Gamma_n = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=i+1}^n \int [m_i(x) - m_j(x)]^2 w(x) dx,$$

where  $w(x)$  is a known weight function. With a larger  $n$ ,  $\Gamma_n$  can not tell  $H_0$  from the case that only an ignorable number of units having different curves, so the null and alternative hypotheses are revised as

$$H_0 : \Delta_m = 0 \text{ vs. } H_1 : \Delta_m > 0,$$

where  $\Delta_m = \lim_{n \rightarrow \infty} \Gamma_n$ . Applying the sieve estimation method presented in Jin and Su (2013), they first estimated the unknown functions  $m_i(\cdot)$ ,  $i = 1, \dots, n$ , under the alternative hypothesis and then calculated the test statistic  $\widehat{\Gamma}_n$  with  $m_i(x)$  replaced by its estimator. The standardized version of the test statistic  $\widehat{\Gamma}_n$  is shown to be consistent and converge to a standard normal distribution under  $H_0$ .

### 10.5.2 Tests For Cross-Sectional Independence

Assuming the panel data  $\{(X_{i,t}, Y_{i,t})\}$  to be independent in index  $i$  is a common practice in empirical analysis. However, it is reasonable to believe that common economic status and other unobserved factors may induce cross-sectional dependence of  $Y_{i,t}$  even after controlling  $X_{i,t}$ . For example, one household's spending decision may not be totally independent of the others' spending decisions at a given time as they all live through the same economy, so that a panel data model with unobserved factors changing across units and time is more appropriate than a simple cross-sectional fixed-effects panel data model. Statistically, ignoring cross-sectional dependence structure may lead to inconsistent parametric and nonparametric estimators constructed from cross-sectional independent models. Sarafidis and Wansbeek (2012) provided an excellent overview on recent development for parametric panel data models with cross-section dependence, including the conceptual measurement of the degree of cross-sectional dependence, estimation methods under both strict and weak exogeneity and tests for cross-section dependence.

Chen, Gao, and Li (2012b) constructed a kernel-based test statistic to test for cross-sectional correlation and considered the following model

$$Y_{i,t} = m_i(X_{i,t}) + \sigma_i(X_{i,t}) \epsilon_{i,t}, \quad i = 1, \dots, n, t = 1, \dots, T, \quad (5.5)$$

where both  $m_i(\cdot)$  and  $\sigma_i(\cdot)$  are unknown smooth measurable functions and can be different for different unit  $i$ , and  $\{\epsilon_{i,t}\}$  is independent of  $\{X_{i,t}\}$  with  $E(\epsilon_{i,t}) = 0$  and

$E(\epsilon_{i,t}^2) = 1$ . The null and alternative hypotheses are

$$H_0 : E(\epsilon_{i,t}\epsilon_{j,t}) = 0 \text{ for all } t \geq 1 \text{ and all } i \neq j;$$

$$H_1 : E(\epsilon_{i,t}\epsilon_{j,t}) \neq 0 \text{ for some } t \geq 1 \text{ and some } i \neq j.$$

With sufficiently large  $T$ , they firstly calculated the LLLS estimator  $\hat{m}_i(x)$  using data only from unit  $i$  and defined  $\hat{u}_{i,t} = Y_{i,t} - \hat{m}_i(X_{i,t})$ . Following Pesaran (2004) for parametric linear panel data models, Chen, Gao, and Li (2012b) proposed a nonparametric cross-sectional uncorrelatedness test statistic

$$NCU = \sqrt{\frac{T}{n(n-1)}} \sum_{i=1}^n \sum_{j \neq i}^n \tilde{\rho}_{i,j} \quad (5.6)$$

where

$$\tilde{\rho}_{i,j} = \frac{T^{-1} \sum_{t=1}^T \bar{u}_{i,t} \bar{u}_{j,t}}{\sqrt{T^{-1} \sum_{t=1}^T \bar{u}_{i,t}^2} \sqrt{T^{-1} \sum_{t=1}^T \bar{u}_{j,t}^2}}$$

is the sample correlation between  $\bar{u}_{i,t}$  and  $\bar{u}_{j,t}$  with  $\bar{u}_{i,t} = \hat{u}_{it} \hat{f}_i(X_{i,t})$  and  $\hat{f}_i(X_{i,t})$  equal to the random denominator o the LLLS estimator  $\hat{m}_i(X_{i,t})$ . The test is applied to models with strictly exogenous explanatory variables and is asymptotically normally distributed under  $H_0$ .

However, for a sufficiently large  $n$ , the comment on Jin and Su's (2013) null and alternative hypotheses seems also applicable here: The test cannot distinguish an alternative model with an ignorable number of pairs of  $(\epsilon_{i,t}\epsilon_{j,t})$  being correlated from the null model as  $|\tilde{\rho}_{i,j}| \leq 1$  in (5.6).

With strictly exogenous variable,  $X_{i,t}$ , in model (5.5), Su and Zhang (2010) tested for pairwise cross-sectional independence by comparing the joint and marginal density functions of  $u_{i,t}$  and  $u_{j,t}$ . Assuming that  $\{u_{i,t}\}_{t=1}^T$  is a stationary time series for each unit  $i$  with a marginal density function  $f_i(\cdot)$ , and that  $\{(u_{i,t}, u_{j,t})\}_{t=1}^T$  has a joint probability density function  $f_{i,j}(\cdot, \cdot)$ . The null and alternative hypotheses are

$$H_0 : \Pr\{f_{i,j}(u_{i,t}, u_{j,t}) = f_i(u_{i,t})f_j(u_{j,t}) \text{ for all } i \neq j\} = 1$$

$$H_1 : \text{not } H_0.$$

The measurement used to differentiate the null against alternative hypothesis is an  $L_2$ -distance

$$\Gamma_n = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=i+1}^n \int [f_{i,j}(u, v) - f_i(u)f_j(v)]^2 du dv \geq 0$$

and the null hypothesis holds true if and only if  $\Gamma_n = 0$ . As Su and Zhang (2010) considered the case that both  $n$  and  $T$  are sufficiently large, for a given  $i$ , the unknown

curve  $m_i(\cdot)$  is consistently estimated via the local polynomial approach and the test is conducted over nonparametric residuals  $\widehat{u}_{i,t} = Y_{i,t} - \widehat{m}_i(X_{i,t})$  with  $f_i(\cdot)$  and  $f_{i,j}(\cdot, \cdot)$  replaced by their corresponding kernel density estimators.

With sufficiently large  $n$  and large  $T$ , for each given  $i$ ,  $X_{i,t}$  is assumed to be a strictly exogenous variable in both Chen, Gao, and Li (2012b) and Su and Zhang (2010), under which condition the presence of cross-sectional correlation or dependence only affects the estimation efficiency, not the consistency, of nonparametric estimation of the unknown curves. The strict exogeneity assumption, to some extent, limits the value of testing for cross-sectional dependence in practice. In addition, if one assumes that  $m_i(\cdot)$  are the same across units, our review in Section 10.2.1 may be extended to cross-sectional dependence case: the kernel estimator of the unknown curve is expected to be asymptotically more efficient with working independence across units and time than with working dependence across units and time, so that it is not a primary interest to test cross-sectional dependence when both  $n$  and  $T$  are large and the unknown curves are the same across time and units for strictly exogenous cases.

In linear fixed-effects panel data model framework, Sarafidis, Yamagata, and Robertson (2009) tested cross-sectional dependence for large  $n$  and finite  $T$ , using the generalized method of moments (GMM). It would be interesting to see tests for cross-sectional dependence for large  $n$  but small  $T$  in nonparametric panel data model framework.

## 10.6 CONCLUSION

---

The chapter selectively reviews some recent results on nonparametric panel data analysis, where the panel data are assumed to be stationary across time if  $T$  is large. For panel unit roots tests and panel cointegration literature, Banerjee (1999) and Phillips and Moon (2000) are two excellent survey papers in parametric regression framework. Also, our literature search on estimation and test of nonparametric panel data quantile regression models are not very fruitful, although there are many publications on estimation and tests of non-/semi-parametric panel data quantile regression models for independent and time series data; e.g., Lee (2003), Honda (2004), Sun (2005), Kim (2007), Cai and Xu (2008), and Cai and Xiao (2012), among which Cai and Xiao's (2012) semiparametric dynamic quantile regressions models with partially varying coefficients have the most general form. However, there is no parallel work extended to panel data models yet, to the best of our knowledge. Lastly, there is a large body of work on nonparametric panel data models using spline estimators. Interested readers are referred to Wu and Zhang's (2006) book on nonparametric methods for longitudinal data analysis for a general treatment of these methods.

---

## NOTES

---

- Zhao (2001) derived an asymptotically efficient estimator for linear median regression model in the presence of unknown heteroskedasticity, using nearest neighbour method. Oberhofer (1982), Weiss (1991), and Koenker and Park (1996) are among earlier contributions in estimating nonlinear regression quantiles. Mukherjee (2000) and Oberhofer and Haupt (2006) derived the limiting results of nonlinear quantile estimator for long range dependent and weakly dependent errors, respectively. Komunjer and Vuong (2010) derived efficient semiparametric estimator for dynamic QR models.

---

## REFERENCES

---

- Ai, C., and Li, Q., 2008. Semiparametric and nonparametric methods in panel data models. In *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice* (3rd edition). L. Mátyás and P. Sevestre (eds.), pp. 451–478. Springer, Berlin.
- Altonji, J., and Matzkin, R., 2005. Cross sectional and panel data estimator for nonseparable models with endogenous regressors. *Econometrica* 73, 1053–1102.
- Anderson, T.W., and Hsiao, C., 1981. Estimation of dynamic models with error components. *Journal of the American Statistical Association* 76, 598–606.
- Anderson, T.W., and Hsiao, C., 1982. Formulation and estimation of dynamic models using panel data. *Journal of Econometrics* 18, 47–82.
- Arellano, M., 2003. *Panel Data Econometrics*. Oxford University Press, New York.
- Arellano, M., and Bond, S., 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* 58, 277–279.
- Arellano, M., and Honore, B., 2001. Panel data models: Some recent developments. In *Handbook of Econometrics*, J. J. Heckman and E. Leamer (eds.), Vol. 5, Chap. 53, pp. 3229–3296. North-Holland, Amsterdam.
- Baltagi, B., 2005. *Econometrics Analysis of Panel Data* (2nd edition). Wiley, New York.
- Baltagi, B.H., Hidalgo, J., and Li, Q., 1996. A nonparametric test for poolability using panel data. *Journal of Econometrics* 75, 345–367.
- Banerjee, A., 1999. Panel data unit roots and cointegration: an overview. *Oxford Bulletin of Economics and Statistics* 61, 607–629.
- Bickel, P.J., and Levina, E., 2008. Covariance regularization by thresholding. *The Annals of Statistics* 36, 2577–2604.
- Bhattacharya, P.K., and Gangopadhyay, A.K., 1990. Kernel and nearest-neighbor estimation of a conditional Quantile. *The Annals of Statistics* 18, 1400–1415.
- Bondell, H.D., Reich, B.J., and Wang, H., 2010. Noncrossing quantile regression curve estimation. *Biometrika* 97, 825–838.
- Buchinsky, M., 1998. Recent advances in Quantile Regression Models: A Practical Guideline for Empirical Research. *The Journal of Human Resources* 33, 88–126.
- Cai, Z., and 2002. Regression quantiles for time series. *Econometric Theory* 18, 169–192.
- Cai, Z., and Xu, X., 2008. Nonparametric quantile estimations for dynamic smooth coefficient models. *Journal of the American Statistical Association* 103, 1596–1608.
- Cai, Z., and Xiao, Z., 2012. Semiparametric quantile regression estimation in dynamic models with partially varying coefficients. *Journal of Econometrics* 167, 413–425.

- Canay, I.A., 2011. A simple approach to quantile regression for panel data. *Econometrics Journal* 14, 368–386.
- Chaudhuri, P. 1991. Nonparametric estimate of regression quantiles and their local Bahadur representation. *The Annals of Statistics* 2, 760–777.
- Chen, J., Gao, J., and Li, D., 2012a. Semiparametric trending panel data models with cross-sectional dependence. *Journal of Econometrics*, 171, 71–85.
- Chen, J., Gao, J., and Li, D., 2012b. A new diagnostic test for cross-sectional uncorrelatedness in nonparametric panel data model. *Econometric Theory* 28, 1144–1168.
- Chen, J., Gao, J., and Li, D., 2013a. Estimation in single-index panel data models with heterogeneous link functions. *Econometric Reviews* 32, 928–955.
- Chen, J., Gao, J., and Li, D., 2013b. Estimation in partially linear single-index panel data models with fixed effects. *Journal of Business & Economics Statistics* 31, 315–330.
- Chernozhukov, V., and Hansen, C., 2008. Instrumental variable quantile regression: a robust inference approach. *Journal of Econometrics* 142, 379–398.
- Chernozhukov, V., Fernández-Val, I., and Galichoni, A., 2010. Quantile and probability curves without crossing. *Econometrica* 78, 1093–1125.
- Chernozhukov, V., Fernández-Val, I., Hahn, J., Newey, W., 2013. Average and quantile effects in nonseparable panel models. *Econometrica* 81, 535–580.
- Dette, H., and Volgushev, S., 2008. Non-crossing non-parametric estimates of quantile curves. *Journal of the Royal Statistical Society (Series B)* 70, 609–627.
- Dette, H., Wagener, J., and Volgushev, S., 2011. Comparing Conditional Quantile Curves. *Scandinavian Journal of Statistics* 38, 63–88.
- Evdokimov, K., 2009. Identification and estimation of a nonparametric panel data model with unobserved heterogeneity. Working paper.
- Fan, J., and Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association: Theory and Method* 96, 1348–1360.
- Fan, J., and Yao, Q., 1998. Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* 85, 645–660.
- Fu, L., and Wang, Y., 2012. Quantile regression for longitudinal data with a working correlation model. *Computational Statistics and Data Analysis* 56, 2526–2538.
- Fan, J., Hu, T., and Truong, Y.K., 1994. Robust nonparametric function estimation. *Scandinavian Journal of Statistics* 21, 433–446.
- Fan, J., Huang, T., and Li, R., 2007. Analysis of longitudinal data with semiparametric estimation of covariance function. *Journal of the American Statistical Association* 102, 632–641.
- Freyberger, J., 2012. Nonparametric Panel Data Models with Interactive Fixed Effects. Working paper, Northwestern University.
- Galvao, A.F., 2011. Quantile regression for dynamic panel data with fixed effects. *Journal of Econometrics* 164, 142–157.
- Galvao, A.F., and Montes-Rojas, G.V., 2010. Penalized quantile regression for dynamic panel data. *Journal of Statistical Planning and Inference* 140, 3476–3497.
- Harding, M., and Lamarche, C., 2009. A quantile regression approach for estimating panel data models using instrumental variables. *Economics Letters* 104, 133–135.
- Härdle, W., and Mammen, E., 1993. Comparing nonparametric versus parametric regression fits. *The Annals of Statistics* 21, 1926–1947.

- He, X., 1997. Quantile curves without crossing. *The American Statistician* 51, 186–192.
- He, X., Fu, B., and Fung, W.K., 2003. Median regression for longitudinal data. *Statistics in Medicine* 22, 3655–3669.
- He, X., Zhu, Z., and Fung, W., 2002. Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika* 89, 579–590.
- Henderson, D.J., and Ullah, A., 2005. A nonparametric random effects estimator. *Economics Letters* 88, 403–407.
- Henderson, D.J., Carroll, R.J., and Li, Q., 2008. Nonparametric estimation and testing of fixed effects panel data models. *Journal of Econometrics* 144, 257–275.
- Honda, T., 2004. Quantile regression in varying coefficient models. *Journal of Statistical Planning and Inferences* 121, 113–125.
- Hsiao, C., 2003. *Analysis of Panel Data* (2nd edition). Cambridge University Press, New York.
- Huang, X., 2013. Nonparametric estimation in large panel with cross-section dependence. *Econometric Reviews* 32, 754–777.
- Huggins, R.M., 1993. A robust approach to the analysis of repeated measures. *Biometrics* 49, 715–720.
- Jeong, K., Härdle, W.K., and Song, S., 2012. A consistent nonparametric test for causality in quantile. *Econometric Theory* 28, 861–887.
- Jin, S., and Su, L., 2013. A nonparametric poolability test for panel data models with cross section dependence. *Econometric Reviews* 32, 469–512.
- Jun, S.J., and Pinkse, J., 2009. Efficient semiparametric seemingly unrelated quantile regression estimation. *Econometric Theory* 25, 1392–1414.
- Jung, S., 1996. Quasi-likelihood for median regression models. *Journal of the American Statistical Association* 91, 251–257.
- Kai, B., Li, R., and Zou, H., 2010. Local composite quantile regression smoothing: an efficient and safe alternative to local polynomial regression. *Journal of the Royal Statistical Society (Series B)* 72, 49–69.
- Karlsson, A., 2009. Bootstrap methods for bias correction and confidence interval estimation for nonlinear quantile regression of longitudinal data. *Journal of Statistical Computation and Simulation* 79, 1205–1218.
- Kato, K., Galvao, A.F., and Montes-Rojas, G.V., 2012. Asymptotics for panel quantile regression models with individual effects. *Journal of Econometrics* 170, 76–91.
- Kim, M.O., 2007. Quantile regression with varying coefficients. *The Annals of Statistics* 35, 92–108.
- Koenker, R., 1984. A note on  $L_1$ -estimators for linear models. *Statistics and Probability Letters* 2, 323–325.
- Koenker, R., 2004. Quantile regression for longitudinal data. *Journal of Multivariate Analysis* 91, 74–89.
- Koenker, R., 2005. *Quantile Regression*. Cambridge University Press, New York.
- Koenker, R., and Bassett, G., 1978. Regression quantiles. *Econometrica* 46, 33–50.
- Koenker, R., and Bassett, G., 1982. Robust test for heteroskedasticity based on regression quantiles. *Econometrica* 50, 43–61.
- Koenker, R., and Bilias, Y., 2001. Quantile regression for duration data: a reappraisal of the Pennsylvania reemployment bonus experiments. *Empirical Economics* 26, 199–220.
- Koenker, R., and Park, B., 1996. An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics* 71, 265–283.

- Koenker, R., Ng, P., and Portnoy, S., 1994. Quantile smoothing spline. *Biometrika* 81, 673–680.
- Komunjer, I., and Vuong, Q., 2010. Efficient estimation in dynamic conditional quantile models. *Journal of Econometrics* 157, 272–285.
- Lamarche, C., 2010. Robust penalized quantile regression estimation for panel data. *Journal of Econometrics* 157, 396–408.
- Lee, J., and Robinson, P., 2012. Panel non-parametric common regression model with fixed effects. Working Paper.
- Lee, S. 2003. Efficient semiparametric estimation of a partially linear quantile regression. *Econometric Theory* 19, 1–31.
- Lee, Y., 2010. Nonparametric estimation of dynamic panel models with fixed effects. Working Paper.
- Li, D., Chen, J., and Gao, J., 2011. Nonparametric time-varying coefficient panel data models with fixed effects. *Econometrics Journal* 14, 387–408.
- Li, Q., and Racine, J., 2007. *Nonparametric Econometrics: Theory and Practice*. Princeton University, Princeton and Oxford.
- Lin, X., and Carroll, R.J., 2000. Nonparametric function estimation for clustered data when the predictor is measured without/with error. *Journal of the American Statistical Association* 95, 520–534.
- Lin, X., and Carroll, R.J., 2006. Semiparametric estimation in general repeated measures problems. *Journal of the Royal Statistical Society (Series B)* 68, 69–88.
- Lin, Z., Li, Q., and Sun, Y., 2014. A consistent nonparametric test of parametric regression functional form in fixed effects panel Data Models. *Journal of Econometrics* 178, 167–179.
- Linton, O.B., and Nielsen, J.P., 1995. A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* 82, 93–100.
- Mammen, E., Støve, B., and Tjøstheim, D., 2009. Nonparametric additive models for panels of time series. *Econometric Theory* 25, 442–481.
- Mu, Y., and Wei, Y., 2009. A dynamic quantile regression transformation model for longitudinal data. *Statistica Sinica* 19, 1137–1153.
- Mukherjee, A., 2000. Linearization of randomly weighted empiricals under long range dependence with applications to nonlinear regression quantiles. *Econometric Theory* 16, 301–323.
- Neocleous, T., and Portnoy, S., 2008. On monotonicity of regression quantile functions. *Statistics and Probability Letters* 78, 1226–1229.
- Oberhofer, W., 1982. The consistency of nonlinear regression minimizing the  $L_1$  norm. *The Annals of Statistics* 10, 316–319.
- Oberhofer, W., and Haupt, H., 2006. Nonlinear quantile regression under dependence and heterogeneity. University of Regensburg Discussion Papers in Economics 388.
- Pesaran, M.H., 2004. General diagnostic tests for cross section dependence in panels. Working Paper.
- Pesaran, M.H., 2006. Estimation and inference in large heterogeneous panels with multifactor error. *Econometrica* 74, 967–1012.
- Phillips, P.C.B., and Moon, H.R., 2000. Nonstationary panel data analysis: An overview of some recent developments. *Econometric Reviews* 19, 263 – 286.
- Phillips, P.C.B., and Sul, D., 2007. Bias in dynamic panel estimation with fixed effects, incidental trends and cross section dependence. *Journal of Econometrics* 137, 162–188.

- Powell, J.L., 1986. Censored regression quantiles. *Journal of Econometrics* 32, 143–155.
- Qian, J., and Wang, L., 2012. Estimating semiparametric panel data models by marginal integration. *Journal of Econometrics* 167, 483–493.
- Qu, A., and Li, R., 2006. Quadratic inference functions for varying-coefficient models with longitudinal data. *Biometrics* 62, 379–391.
- Qu, A., Lindsay, B.G., and Li, B., 2000. Improving generalized estimating equations using quadratic inference functions. *Biometrika* 87, 823–836.
- Rosen, A., 2012. Set identification via quantile restrictions in short panels. *Statistical Science* 6, 15–32.
- Robinson, P.M., 2012. Nonparametric trending regression with cross-sectional dependence. *Journal of Econometrics* 169, 4–14.
- Ruckstuhl, A.F., Welsh, A.H., and Carroll, R.J., 2000. Nonparametric function estimation of the relationship between two repeatedly measured variables. *Statistica Sinica* 10, 51–71.
- Sarafidis, V., and Wansbeek, T., 2012. Cross-sectional dependence panel data analysis. *Econometric Reviews* 31, 483–531.
- Sarafidis, V., Yamagata, T., and Robertson, D., 2009. A test of cross section dependence for a linear dynamic panel model with regressors. *Journal of Econometrics* 148, 149–161.
- Su, L., and Jin, S., 2012. Sieve estimation of panel data models with cross section dependence. *Journal of Econometrics* 169, 34–47.
- Su, L., and Lu, X., 2013. Nonparametric dynamic panel data models: kernel estimation and specification testing. *Journal of Econometrics* 176, 112–133.
- Su, L., and Ullah, A., 2006. Profile likelihood estimation of partially linear panel data models with fixed effects. *Economics Letters* 92, 75–81.
- Su, L., and Ullah, A., 2007. More efficient estimation of nonparametric panel data models with random effects. *Economics Letters* 96, 375–380.
- Su, L., and A. Ullah, 2011. Nonparametric and semiparametric panel econometric models: estimation and testing. In A. Ullah and D. E. A. Giles (eds.), *Handbook of Empirical Economics and Finance*, pp. 455–497. Taylor & Francis Group, New York.
- Su, L., and Zhang, Y., 2010. Testing cross-sectional dependence in nonparametric panel data models. Working paper.
- Su, L., and Zhang, Y., 2013. Sieve estimation of nonparametric dynamic panel data models with interactive fixed effects. Working paper, Singapore Management University.
- Sun, Y., 2005. Semiparametric efficient estimation of partially linear quantile regression models. *The Annals of Economics and Finance* 6, 105–127.
- Sun, Y., 2006. A consistent nonparametric equality test of conditional quantile functions. *Econometric Theory* 22, 614–632.
- Sun, Y., Carroll, R.J., and Li, D., 2009. Semiparametric estimation of fixed effects panel data varying coefficient models. *Advances in Econometrics* 25, 101–130.
- Ullah, A., and Roy, N., 1998. Nonparametric and semiparametric econometrics of panel data by: A. Ullah and D.E.A. Giles (eds.), *Handbook of Applied Economics Statistics*, vol. 1, pp. 579–604. Marcel Dekker, New York.
- Wang, H., Zhu, Z., and Zhou, J., 2009. Quantile regression in partially linear varying coefficient models. *The Annals of Statistics* 37, 3841–3866.
- Wang, N., 2003. Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika* 90, 43–52.
- Weiss, A.A., 1991. Estimating nonlinear dynamic models using least absolute error estimation. *Econometric Theory* 7, 46–68.

- Welsh, A., Lin, X., and Carroll, R.J., 2002. Marginal longitudinal nonparametric regression: locality and efficiency of spline and kernel methods. *Journal of the American Statistical Association* 97, 484–493.
- Wu, H., and Zhang, J., 2006. Nonparametric regression methods for longitudinal data analysis. Wiley-Interscience.
- Yao, W., and Li, R., 2013. New local estimation procedure for a non-parametric regression for longitudinal data. *Journal of the Royal Statistical Society (Series B)* 75, 123–138.
- Yu, K., and Jones, M.C., 1998. Local linear quantile regression. *Journal of the American Statistical Association* 93, 228–238.
- Yu, K., Lu, Z., and Stander, J., 2003. Quantile regression: applications and current research area. *The Statistician* 52, 331–350.
- Zhao, Q., 2001. Asymptotically efficient median regression in the presence of heteroskedasticity of unknown form. *Econometric Theory* 17, 765–784.
- Zheng, J., 1996. A consistent of functional form via nonparametric estimation techniques. *Journal of Econometrics* 75, 263–289.
- Zheng, J., 1998. A consistent nonparametric test of parametric regression models under conditional quantile restrictions. *Econometric Theory* 14, 123–138.
- Zou, H., 2006. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101, 1418–1429.
- Zou, H., and Yuan, M., 2008. Composite quantile regression and the oracle model selection theory. *The Annals of Statistics* 36, 1108–1126.

## CHAPTER 11

---

# MEASUREMENT ERROR IN PANEL DATA

---

ERIK MEIJER, LAURA SPIERDIJK, AND TOM WANSBEEK

### 11.1 INTRODUCTION

---

MEASUREMENT error and panel data are strange bedfellows. This holds at least in one specific sense: their popularity could not be more apart. While panel data analysis is an exponentially growing subfield of econometrics, both theoretically and empirically, the analysis of models allowing for measurement error has never become a major research area in econometrics. While empirical researchers often recognize that their data contain measurement error or that, out of necessity, they employ proxies for some of their variables, one can only guess about this marginal position of measurement error analysis in econometrics. It may be due to the identification problem that is present in the simplest regression model. In more complex models that do not suffer from identification problems, there may be insufficient awareness of the available methods. And when these methods are used,  $t$ -values usually go down.

Panel data models may not suffer from the identification problem inherent in the simplest model, and hence this is an area where one may expect above-average interest in handling measurement error. Although this is hard to quantify, it is certainly striking that the seminal publication in the area of measurement error in panel data, Griliches and Hausman (1986), has been cited hundreds of times thus becoming a classic, and has in fact been cited more than any other publication on measurement error, at least in econometrics.

The Griliches-Hausman paper basically offered three important insights, which will return in some form or another in the sequel. First, it argued that the within transformation exacerbates the bias due to measurement error. Second, when estimation is based on differencing the data in order to remove the individual effect, the bias

decreases when differences are taken of the data further apart in time. And third, exploiting zeros in the measurement error covariance matrix allows for consistent estimation.

Our review is structured as follows. In Section 11.2 we explore the basics of measurement error in the simplest possible panel data model with measurement error. We show that OLS is inconsistent, in very much the same way as with a single cross-section. We also show that we can exploit the panel character of the data to decrease the inconsistency. Correlated (or “fixed”) effects are next taken into account. There are various ways to eliminate them, all leading to estimators that are inconsistent in the presence of measurement error but in different ways. We present the ordering, and in particular consider the effect of taking short versus long differences. The inconsistency decreases when differences are taken far apart in time. We then discuss the random effects model, which is more similar to the cross-sectional case, and we conclude the section by a discussion of the identification of the model. It appears that restrictions on the various parameter matrices can be helpful in achieving identification and thus the construction of consistent estimators.

Following on this, we investigate in Section 11.3 how those restrictions can be put to work. We first do so in an ad hoc way for a simple case and show how to write this as an instrumental variables (IV) issue. We next analyze the presence of restrictions in a systematic way. Restrictions on the parameter matrices are not the only source of IVs. These can also be implied by the third moments of the data for those cases where they are nonzero, and from additional, exogenous regressors. In Section 11.4, we show the advantages of writing the model as a particular case of a structural equation model (SEM), for which many computer programs are available.

We next turn to the dynamic model, where the lagged dependent variable is among the regressors. This model has spawned a huge literature and applications mushroom. In Section 11.5 we discuss this model when there is measurement error and present a consistent estimator. Again, we show the benefits of formulating the model as an SEM.

Up till then, the measurement error was in the form of an error added to the true value of the regressor, and independent of it. In Section 11.6 we consider a departure from this, where the measurement error and the true value of the regressor are correlated. We pay some attention to the Berkson model for panel data. In Section 11.7 we consider measurement error that is multiplicative rather than additive. In Section 11.8, we discuss nonlinear models and in Section 11.9, we discuss how validation studies can be used to address measurement error.

Although this review covers many topics, it is also limited in various respects. For expository reasons, we stick to the simplest cases. Throughout, when discussing asymptotics, they are of the kind where the cross-sectional dimension goes to infinity while the time dimension stays fixed. Also, our coverage of non-linear panel data models with measurement error is limited, thus reflecting the very limited attention paid to the topic up till now.

## 11.2 BASIC RESULTS

---

In this Section we look at some basic issues at the nexus of panel data analysis and measurement error modeling. Overviews of panel data analysis are Arellano (2003), Hsiao (2003), Baltagi (2008), and Wooldridge (2010); Wansbeek and Meijer (2000) provide an overview of measurement error modeling in econometrics.

Panel data models have, by their nature, one dimension more than models for time series or cross-sections. Hence the notation is sometimes cumbersome. Since various interesting properties of the panel data model with measurement error can be illustrated by considering just a single regressor, most of the discussion here will concentrate on this case, saving one dimension in the notation. We begin by an even more simplified case, that is, we leave out individual effects. Throughout, time effects are considered to have been eliminated by transforming all variables into deviations from their means per time period (“wave”).

This leaves us with the following. Individuals are denoted by the subscript  $n = 1, \dots, N$  and waves by  $t = 1, \dots, T$ . The  $T$ -vector  $y_n$  depends on the  $T$ -vector  $\xi_n$ . Now,  $\xi_n$  is unobservable, and instead a proxy  $x_n$  is observed:

$$y_n = \xi_n \beta + \varepsilon_n \quad (1)$$

$$x_n = \xi_n + \nu_n, \quad (2)$$

with  $\varepsilon_n$  the  $T$ -vector of the errors in the equations and  $\nu_n$  the  $T$ -vector of the measurement errors. We assume  $\xi_n \sim (0, \Sigma_\xi)$ ,  $\varepsilon_n \sim (0, \Sigma_\varepsilon)$ , and  $\nu_n \sim (0, \Sigma_\nu)$ , which are mutually independent. We do not yet impose a structure on  $\Sigma_\varepsilon$ ,  $\Sigma_\nu$ , or  $\Sigma_\xi$ .

### 11.2.1 What Goes Wrong When Measurement Error is Neglected

We consider estimation of (1)–(2) by OLS of  $y_n$  on  $x_n$ , thus neglecting the presence of measurement error. We show how this affects the results. To assess OLS, we eliminate  $\xi_n$  leads to obtain the reduced form

$$\begin{aligned} y_n &= x_n \beta + u_n \\ u_n &\equiv \varepsilon_n - \nu_n \beta. \end{aligned}$$

As a result,  $E(x'_n u_n) = -\text{tr}(\Sigma_\nu)\beta \neq 0$ . Hence OLS is inconsistent. With  $E(x'_n y_n) = \text{tr}(\Sigma_\xi)\beta$  and  $\Sigma_x \equiv \Sigma_\xi + \Sigma_\nu$ , OLS estimation of the reduced form gives

$$\hat{\beta}_{\text{OLS}} = \frac{\sum_n x'_n y_n}{\sum_n x'_n x_n} \xrightarrow{p} \frac{\text{tr}(\Sigma_\xi)}{\text{tr}(\Sigma_x)} \beta. \quad (3)$$

Since  $0 < \text{tr}(\Sigma_\xi) \leq \text{tr}(\Sigma_x)$ , OLS suffers from an attenuation bias towards zero, much in the same way as in the case of a single cross-section. The factor  $\pi \equiv \text{tr}(\Sigma_\xi)/\text{tr}(\Sigma_x)$  is the so-called reliability of  $x$  as a proxy of  $\xi$ .

We now turn to  $\Sigma_\varepsilon$ . Let  $\hat{\varepsilon}_n \equiv y_n - x_n \hat{\beta}_{\text{OLS}}$ , the residual. An obvious estimator  $\hat{\Sigma}_\varepsilon$  of  $\Sigma_\varepsilon$  is given by  $\sum_n \hat{\varepsilon}_n \hat{\varepsilon}'_n / N$ . Then

$$\begin{aligned}\hat{\Sigma}_\varepsilon &= \frac{1}{N} \sum_n \left( y_n y'_n - \hat{\beta}_{\text{OLS}} x_n y'_n - \hat{\beta}_{\text{OLS}} y_n x'_n + \hat{\beta}_{\text{OLS}}^2 x_n x'_n \right) \\ &\xrightarrow{P} \beta^2 \Sigma_\xi + \Sigma_\varepsilon - 2\pi\beta^2 \Sigma_\xi + \pi^2 \beta^2 (\Sigma_\xi + \Sigma_\nu) \\ &= \Sigma_\varepsilon + \beta^2 [(1-\pi)^2 \Sigma_\xi + \pi^2 \Sigma_\nu].\end{aligned}\quad (4)$$

Clearly,  $\text{plim} \hat{\Sigma}_\varepsilon \geq \Sigma_\varepsilon$ , with equality when there is no measurement error, so when  $\Sigma_\nu = 0$  and hence  $\pi = 1$ . The inconsistency of  $\beta_{\text{OLS}}$  when the measurement error is neglected leads to too large residuals. Hence the variance of the errors in the equations is overestimated.

### 11.2.2 Reducing the Inconsistency

Panel data can help improve on the result in (3). Define  $\Sigma_{xy} \equiv E(x_n y'_n) = \beta \Sigma_\xi$  and  $\Sigma_y \equiv E(y_n y'_n) = \beta^2 \Sigma_\xi + \Sigma_\varepsilon$ . Because  $\Sigma_\xi$  must be positive semidefinite,  $\beta$  is positive if  $\Sigma_{xy}$  is positive semidefinite (but not zero), negative if  $\Sigma_{xy}$  is negative semidefinite (but not zero), and zero if  $\Sigma_{xy} = 0$ , so we can estimate the sign of  $\beta$  consistently. Furthermore,  $\Sigma_\nu = \Sigma_x - \Sigma_\xi$  is positive semidefinite, which implies that  $\Sigma_x - \Sigma_{xy}/\beta$  must be positive semidefinite. Assuming that  $\beta > 0$  (the analysis for  $\beta < 0$  is analogous), this implies that  $\beta I_T - \Sigma_x^{-1/2} \Sigma_{xy} \Sigma_x^{-1/2}$  must be positive definite, which in turn means that  $\beta$  must be larger than or equal to the largest eigenvalue of  $\Sigma_x^{-1/2} \Sigma_{xy} \Sigma_x^{-1/2}$ , or, equivalently, the largest eigenvalue of  $\Sigma_x^{-1} \Sigma_{xy}$ . Hence, this largest eigenvalue forms a lower bound for  $\beta$ . Conversely,  $\Sigma_\varepsilon = \Sigma_y - \beta^2 \Sigma_\xi = \Sigma_y - \beta \Sigma_{xy}$  must be positive semidefinite, an upper bound for  $\beta$  is the smallest eigenvalue of  $\Sigma_y \Sigma_{xy}^{-1}$ . Note that, for the model to be consistent with the observable covariance matrices, this lower bound must be lower than this upper bound. If this is the case, the bounds are sharp: a set of parameters exists that is consistent with the model and the observable covariance matrices and in which  $\beta$  attains the lower bound, and similarly for the upper bound.

To illustrate this, consider the (rather arbitrarily chosen) case with  $T = 5$ , and  $\xi$  and  $\nu$  are AR(1) with autocorrelation parameters 0.8 and 0, respectively, stationary variances of 0.8 and 0.2, respectively, and  $\beta = 1$ . Then  $\text{tr}(\Sigma_\xi)/\text{tr}(\Sigma_x) = 0.8$  whereas the largest eigenvalue of  $\Sigma_x^{-1} \Sigma_{xy}$  is 0.936. So the OLS estimator has a downward bias of 20%, whereas the least-attenuated estimator has a downward bias of only 6%. Unfortunately, the upper bound is typically less informative, because of the low  $R^2$ 's in most microeconomic regressions.

### 11.2.3 Fixed Effects

We now introduce individual effects in model (1)–(2). With  $\iota_T$  denoting a  $T$ -vector of ones, we do so in the traditional way by decomposing  $\varepsilon_n$  into a component that is constant over time,  $\iota_T \alpha_n$ , and a time-varying component  $e_n$ , so

$$\varepsilon_n = \iota_T \alpha_n + e_n.$$

We first discuss the fixed effects model, where  $\alpha_n$  is taken to be a fixed parameter. There are two cases where this can be relevant. First, one may view  $\alpha_n$  as a parameter of interest that must be estimated, rather than as a random variable. We believe that with the large  $N$ , small  $T$  panels that we are concerned with here, it will typically be more fruitful to view  $\alpha_n$  as an unobserved random variable. In the other case,  $\alpha_n$  is viewed as random in principle, but it correlates with  $\xi_n$  (or with at least one of the regressors in the more general setting with more than a single regressor), for example, due to information asymmetry between the agents under study (whose behavior is supposedly optimal) and the econometrician who studies their behavior. To account for this correlation we use the projection of  $\alpha_n$  on  $\xi_n$ ,

$$\alpha_n = \sigma'_{\xi \alpha} \Sigma_{\xi}^{-1} \xi_n + w_n,$$

where  $\sigma_{\xi \alpha}$  is the covariance between  $\xi_n$  and  $\alpha_n$ , and  $w_n$  by construction does not correlate with  $\xi_n$ . Substituting this in the model  $y_n = \xi_n \beta + \iota_T \alpha_n + e_n$  gives

$$y_n = \xi_n \beta + \iota_T \sigma'_{\xi \alpha} \Sigma_{\xi}^{-1} \xi_n + \iota_T w_n + e_n. \quad (5)$$

When the correlation between  $\alpha_n$  and  $\xi_n$  is nonzero but neglected, (3) becomes

$$\hat{\beta}_{OLS} \xrightarrow{P} \frac{\text{tr}(\Sigma_{\xi}) \beta + \iota'_T \sigma_{\xi \alpha}}{\text{tr}(\Sigma_x)} = \beta - \frac{\text{tr}(\Sigma_v) \beta - \iota'_T \sigma_{\xi \alpha}}{\text{tr}(\Sigma_x)}, \quad (6)$$

where we have maintained the assumption that  $e_n$  is uncorrelated with  $\alpha_n$ ,  $\xi_n$ , and  $v_n$ , and also assume that  $\alpha_n$  is uncorrelated with the measurement error  $v_n$ . (The latter can be relaxed for the estimators discussed below that eliminate  $\alpha_n$ .) So a positive correlation between  $\alpha_n$  and  $\xi_n$  induces a bias that goes against the bias due to measurement error if  $\beta$  is also positive, but there is of course no reason why the two might come close to canceling out, and if  $\xi_n$  and  $\alpha_n$  are negatively correlated (with still  $\beta > 0$ ), the two bias terms have the same sign.

Let  $C_T \equiv \frac{1}{T} \iota_T \iota'_T$  and  $A_T \equiv I_T - C_T$ . Premultiplication of (5) by  $A_T$  and, separately, by its complement  $C_T$  (or equivalently by its only nonzero eigenvector  $\frac{1}{T} \iota'_T$ ) gives, with tildes denoting within-transformed variables and bars denoting averages over time,

$$\tilde{y}_n = \tilde{\xi}_n \beta + \tilde{e}_n \quad (7a)$$

$$\bar{y}_n = \bar{\xi}'_n \theta + (\bar{w}_n + \bar{e}_n), \quad (7b)$$

with  $\theta \equiv \frac{1}{T}\iota_T\beta + \Sigma_\xi^{-1}\sigma_{\xi\alpha}$ , which is one-to-one in  $\sigma_{\xi\alpha}$ . From (7a) it appears that, with correlated effects and in the absence of measurement error, the optimal estimator of  $\beta$  is the OLS estimator in the model in within-transformed variables, which equals the OLS estimator in the model in the original variables plus  $\alpha_n$  as a fixed parameter (i.e., with  $N$  dummies for the cross-sectional units added as covariates), thus justifying the habit of calling a model with correlated effects a fixed effects model. Testing for correlated effects can be based on (7b), the regression of  $\bar{y}_n$  on the  $T$  regressors contained in  $\xi_n$ , cf. Arellano (1993).

#### 11.2.4 Eliminating the Fixed Effects

Transforming the data by  $A_T$  is one way to eliminate the additional bias term in (6) due to correlated effects. It is the optimal transformation when  $\xi_n$  is observed, that is, it is consistent and efficient. When there is measurement error, consistency no longer holds, and it is worthwhile to consider other transformations than  $A_T$  for their first-order properties. Let  $Q$  be a  $T \times T$  matrix satisfying  $Q\iota_T = 0$ . One choice is of course  $Q = A_T$ .

Other choices are such that  $Q = Q_\tau$  transforms the data into difference across  $\tau$  waves,  $1 \leq \tau \leq T - 1$ ; the choice  $\tau = 1$  corresponds with taking first differences and  $\tau = T - 1$  corresponds with the longest difference that one can take given the available data. So, by way of example, for  $T = 3$ ,

$$Q_1 = \begin{pmatrix} -1 & 0 \\ 1 & -1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}$$

$$Q_2 = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} (-1, 0, 1) = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix}.$$

In general, the relation between  $A_T$  and the  $Q_\tau$  is given by

$$A_T = \frac{1}{T} \sum_{\tau=1}^{T-1} Q_\tau, \quad (8)$$

which holds because  $Q_\tau$  has  $\tau$ th pseudo-diagonal equal to  $-1$  and all other pseudo-diagonals equal to 0;  $\sum_{\tau=1}^{T-1} Q_\tau$  has diagonal elements equal to  $T - 1$  since all rows have to add to zero. For any  $Q$  we have

$$\hat{\beta}_Q \equiv \frac{\sum_n x'_n Q y_n}{\sum_n x'_n Q x_n} \xrightarrow{p} \frac{\text{tr}(Q\Sigma_\xi)}{\text{tr}(Q\Sigma_x)} \beta = \left(1 - \frac{\text{tr}(Q\Sigma_v)}{\text{tr}(Q\Sigma_x)}\right) \beta.$$

The bias factor is  $\text{tr}(Q\Sigma_v)/\text{tr}(Q\Sigma_x)$ . The estimate of the regression coefficient based on panel data may quite well be more attenuated than the estimate based on a single cross-section. This is the case, intuitively speaking, when the regressor is more correlated over time than the measurement error, as will be the case in most cases met in practice. The transformation  $Q$  will eliminate a larger part of the variance of  $x$  than of the variance of  $v$ , and

$$\frac{\text{tr}(Q\Sigma_v)}{\text{tr}(Q\Sigma_x)} > \frac{\text{tr}(\Sigma_v)}{\text{tr}(\Sigma_x)}, \quad (9)$$

the latter being the bias factor without the transformation if the individual effect and the regressor do not correlate. This effect was a major insight of Griliches and Hausman (1986).

To illustrate this effect in some more detail, consider the case where  $\xi_n$  and  $v_n$  are both AR(1):

$$\Sigma_\xi = \sigma_\xi^2 \Psi_\xi \text{ with } (\Psi_\xi)_{ij} = \rho_\xi^{|i-j|},$$

and analogously for  $v$ . Let

$$\begin{aligned} \psi_\xi &\equiv \text{tr}(A_T \Psi_\xi) = T - 1 - 2 \frac{\rho_\xi}{1 - \rho_\xi} \phi_\xi \\ \phi_\xi &\equiv 1 - \frac{1}{T} \frac{1 - \rho_\xi^T}{1 - \rho_\xi}, \end{aligned} \quad (10)$$

and again analogously for  $v$ . For the within-estimator  $\hat{\beta}_W$  ( $Q = A_T$ ), the estimator  $\hat{\beta}_\tau$  based on the  $\tau$ th difference ( $Q = Q_\tau$ ) and, added for reference, the OLS estimator  $\hat{\beta}_{OLS}$  under the assumption of no correlation between  $\alpha_n$  and  $\xi_n$  ( $Q = I_T$ ), the relative biases of the various estimators satisfy

$$\begin{aligned} \frac{\hat{\beta}_W - \beta}{\beta} &\xrightarrow{p} - \frac{\sigma_v^2 \psi_v}{\sigma_\xi^2 \psi_\xi + \sigma_v^2 \psi_v} \\ \frac{\hat{\beta}_\tau - \beta}{\beta} &\xrightarrow{p} - \frac{\sigma_v^2 (1 - \rho_v^\tau)}{\sigma_\xi^2 (1 - \rho_\xi^\tau) + \sigma_v^2 (1 - \rho_v^\tau)}, \quad \tau = 1, \dots, T-1 \\ \frac{\hat{\beta}_{OLS} - \beta}{\beta} &\xrightarrow{p} - \frac{\sigma_v^2}{\sigma_\xi^2 + \sigma_v^2}. \end{aligned}$$

In the empirically likely case that  $\rho_\xi > \rho_v$ , in which the true value of the regressor is more correlated over time than the measurement error, these biases satisfy

$$0 < |\text{bias}(\hat{\beta}_{OLS})| < |\text{bias}(\hat{\beta}_{T-1})| < |\text{bias}(\hat{\beta}_W)| < |\text{bias}(\hat{\beta}_1)| < |\beta|.$$

So the within estimator has an intermediate position relative to the estimators based on the first and longest difference, thus reflecting their relationship as given in (8). Taking the longest possible differences is optimal from a bias point of view. From an MSE point

of view, the conclusion is less clear since the number of data points decreases linearly with  $\tau$ . Another implication of the above is that, with  $\rho_\xi > \rho_v$  still assumed,  $|\hat{\beta}_\tau|$  is an increasing function of  $\tau$ . The example is quite specific but the phenomenon will arise much more in general in the presence of measurement error and, as stated by Goolsbee (2000), provides a basis for testing for the presence of measurement error. Yet, the topic still has been hardly explored. One simple approach starts with the following. Under the model assumptions, the null-hypothesis of absence of measurement error translates into

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_{T-1},$$

the alternative being that the  $\beta$ 's are ordered from low to high. A test statistic could be the number of violations of this ordering. The distribution of the test statistic can be found through the parametric bootstrap.

The situation is illustrated in Figure 11.1. For the case of  $T = 5$ ,  $\beta = 1$ ,  $\sigma_\xi^2 = 2\sigma_v^2$ , and  $\rho_\xi = .8$ , it shows, as a function of the measurement error persistence parameter  $\rho_v$ , the bias of estimators for  $\beta$  based on first differences and longest (i.c., fourth) differences and the bias of the within estimator. It also shows the behavior of the OLS estimator and the between estimator when there would be no correlation between  $\alpha$  and  $\xi_n$ . The bias of the OLS estimator does not depend on  $\rho_v$  and is hence reflected by a horizontal line. In the empirically likely case  $\rho_\xi > \rho_v$ , the longest difference estimator performs hardly worse than OLS. Otherwise, the ordering is reversed. In the extreme case of  $\rho_v$  approaching unity, measurement error has become constant over time and will be simply removed by any kind of differencing. Hence, no bias remains.

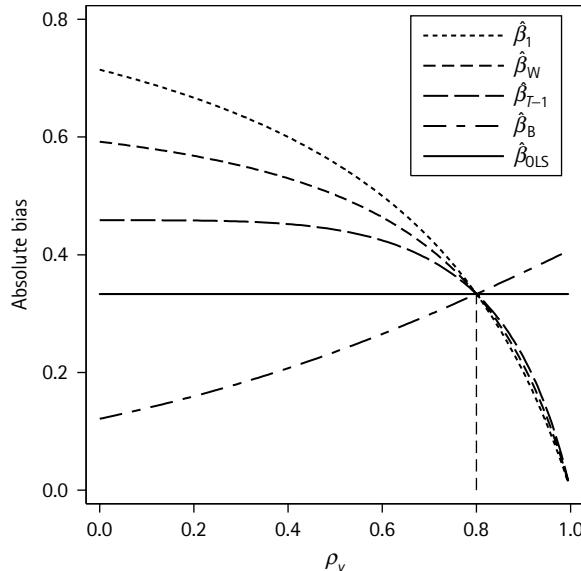


FIGURE 11.1 Bias of  $\hat{\beta}_1$ ,  $\hat{\beta}_{T-1}$ ,  $\hat{\beta}_W$ ,  $\hat{\beta}_{OLS}$ , and  $\hat{\beta}_B$ , for  $T = 5$ ,  $\beta = 1$ ,  $\sigma_\xi^2 = 2\sigma_v^2$ , and  $\rho_\xi = .8$ .

### 11.2.5 Random Effects

When  $\alpha_n$  is independent of  $\xi_n$ , the model is commonly called the random effects model. This opens up the scope for estimators that are based on any kind of linear transformation with some matrix  $Q$ . A first observation is that we can take  $Q = A_T = I_T - C_T$  in (9) and rearrange it as

$$\frac{\text{tr}(\Sigma_v)}{\text{tr}(\Sigma_x)} > \frac{\text{tr}(C_T \Sigma_v)}{\text{tr}(C_T \Sigma_x)}.$$

This implies that the between estimator, based on averaging the data over the waves to obtain a single cross-section,

$$\hat{\beta}_B = \frac{\sum_n x'_n C_T y_n}{\sum_n x'_n C_T x_n},$$

has less bias than the OLS estimator when  $v_n$  has less correlation over time than  $\xi_n$ . In the limit, the relative bias becomes

$$\frac{\hat{\beta}_B - \beta}{\beta} \xrightarrow{p} -\frac{\sigma_v^2(T - \psi_v)}{\sigma_\xi^2(T - \psi_\xi) + \sigma_v^2(T - \psi_v)}$$

when  $\xi_n$  and  $v_n$  are again taken to be AR(1) as above. This bias is represented in Figure 11.1 as the lower curve, left of  $\rho_\xi = .8$ , and to the right of that point the upper curve.

Usually, in the random effects model, it is assumed that  $e_n$  is i.i.d. across both time and cross-sectional units,  $e_n \sim (0, \sigma_e^2 I_T)$ . With  $\alpha_n \sim (0, \sigma_\alpha^2)$  and  $w_n \sim (0, \sigma_w^2)$ , this implies for the covariance structure of  $\varepsilon_n$ :

$$\Sigma_\varepsilon = \sigma_w^2(\lambda C_T + A_T),$$

with

$$\lambda \equiv \frac{T\sigma_\alpha^2 + \sigma_e^2}{\sigma_e^2}.$$

So  $\Sigma_\varepsilon^{-1}$  is proportional to  $C_T + \lambda A_T$ . In this model, the GLS estimator is the optimal estimator when there is no measurement error. As usual with GLS estimators, estimation takes place in two rounds. In the first round,  $\lambda$  is estimated consistently based on the OLS residuals, and the next round involves feasible GLS to obtain  $\hat{\beta}$ . With  $\hat{\varepsilon}_{Wn} \equiv y_n - x_n \hat{\beta}_W$  and  $\hat{\varepsilon}_{Bn}$  defined analogously, the obvious estimator of  $\lambda$  that is consistent in the absence of measurement error (cf. (7a)–(7b)) and is

$$\hat{\lambda} = \frac{\sum_n \hat{\varepsilon}'_{Bn} C_T \hat{\varepsilon}_{Bn}}{\sum_n \hat{\varepsilon}'_{Wn} A_T \hat{\varepsilon}_{Wn}/(T-1)}.$$

When there is measurement error, this is not consistent anymore. With  $\pi_B \equiv \text{tr}(C_T \Sigma_\xi) / \text{tr}(C_T \Sigma_x)$  the reliability of the between estimator and  $\pi_W$  defined analogously,

$$\hat{\lambda} \xrightarrow{P} \lambda^* \equiv \frac{\sigma_w^2 \lambda + \beta^2 \pi_B \text{tr}(C_T \Sigma_v)}{\sigma_w^2 + \beta^2 \pi_W \text{tr}(A_T \Sigma_v) / (T - 1)}.$$

For the assumed case of  $\xi_n$  more correlated over time than  $v_n$ ,  $\pi_B > \pi_W$ . When  $v_n$  is only weakly correlated over time,  $\text{tr}(C_T \Sigma_v)$  and  $\text{tr}(A_T \Sigma_v) / (T - 1)$  will not differ much. Yet this is insufficient to get an idea about  $\lambda$  being overestimated or underestimated when there is measurement error. Whatever  $\lambda^*$ , the random effects estimator satisfies

$$\begin{aligned} \hat{\beta}_{\text{RE}} &= \frac{\hat{\lambda} \sum_n x'_n A_T y_n + \sum_n x'_n C_T y_n}{\hat{\lambda} \sum_n x'_n A_T x_n + \sum_n x'_n C_T x_n} \\ &\xrightarrow{P} \left( 1 - \frac{\sigma_v^2 (T + (\lambda^* - 1)\psi_v)}{\sigma_\xi^2 (T + (\lambda^* - 1)\psi_\xi) + \sigma_v^2 (T + (\lambda^* - 1)\psi_v)} \right) \beta, \end{aligned}$$

again for the case where  $\xi_n$  and  $v_n$  are AR(1).

### 11.2.6 Identification

A major issue in measurement error models concerns its identification. For many practical purposes, identification is equivalent to the existence of consistent estimators for the model's parameters (e.g., Bekker, Merckens, and Wansbeek, 1994). A notable exception may occur when there are incidental parameters, that is, the number of parameters increases with sample size. As we mentioned above, in fixed- $T$  panel data with fixed effects, the latter may be either viewed as random variables (that may be correlated with other variables in the model) or as parameters, which would make them incidental parameters. The identification of  $\beta$  does not depend on this distinction, nor on the distinction between fixed and random effects, and therefore, we analyze it here for the random effects model.

A classical result due to Reiersøl (1950) says that the cross-sectional bivariate linear regression model with normally distributed errors in the equation and measurement errors is identified if and only if  $\xi_n$  is not normally distributed. This is a special case of our model (with  $T = 1$ ). Wansbeek and Meijer (2000, Chapter 4) extensively review identification issues in measurement error models.

A formal theory of identification of the measurement error model in the panel data setting in general does not seem to be available. (Hsiao and Taylor, 1991, discuss the special case with measurement errors that are i.i.d. across both individuals and time.) Hence we proceed somewhat intuitively and take the case of all random variables being normal as the most conservative case from an identification point of view. Since we take means zero throughout, all information from the data that can be brought to bear

on estimation is contained in the second moments of the data, these jointly forming a complete, sufficient statistic. In the absence of fixed effects, the second-order implications of the model are

$$\Sigma_y = \beta^2 \Sigma_\xi + \Sigma_\varepsilon \quad (11a)$$

$$\Sigma_{xy} = \beta \Sigma_\xi \quad (11b)$$

$$\Sigma_x = \Sigma_\xi + \Sigma_v. \quad (11c)$$

In this normality-based setting, the model is identified if this system can be solved uniquely for the parameters. When we start solving this system, we see from (11a) that we need the information contained in the second moment of  $y_n$  to identify  $\Sigma_\varepsilon$  since the latter only occurs there. Analogously, we need the information contained in the second moment of  $x_n$  to identify  $\Sigma_v$ . So the covariance of  $y_n$  and  $x_n$  is needed to identify both  $\beta$  and  $\Sigma_\xi$ . However, these occur as a product and cannot be disentangled.

We conclude that the basic panel data model with measurement error is not identified. For all  $T$ , the degree of underidentification is one. This includes the cross-sectional case,  $T = 1$ . So the presence of panel data by itself does not provide identification. More is needed in some way. This state of affairs sets measurement error problems in econometrics somewhat apart from, for example, physics and the medical sciences. In particular, in econometrics it is commonly assumed that the measurement error variance is unknown. For the case it is known, see, for example, Buonaccorsi, Demidenko, and Tosteson (2000) or Fan, Sutradhar, and Prabhakar Rao (2012).

As usual, the quest for consistent estimation begins with the search for instrumental variables (IVs). Sometimes they can be found outside the model (see Klette, 1999, for an example), but the panel data context offers scope for finding IVs from within the model. This is possible when the researcher is willing to impose restrictions on  $\Sigma_\varepsilon$  or  $\Sigma_v$ . IVs can also be derived when the model contains an exogenous variable in addition to  $\xi_n$ , or when  $\xi_n$  is not normally distributed. We now turn to an elaboration of these ideas.

### 11.3 CONSISTENT ESTIMATION

---

We first discuss the benefits of restrictions on  $\Sigma_\varepsilon$  or  $\Sigma_v$ , and then see how consistent estimation can be based on employing third moments. Also, the presence of an exogenous variable can be beneficial. Throughout this Section, we assume that individual effects are not correlated with the regressors. Meijer, Spierveldijk, and Wansbeek (2012) discuss the fixed effects case in considerable detail, which for the most part means that  $y_n$  is first transformed (e.g., within transformation or first differences) to remove the fixed effects, and  $x_n$  accordingly where they act as regressors. But  $x_n$  appears untransformed as part of the IVs. With these adaptations, the estimators discussed here can be

used. These estimate  $\beta$  consistently if  $T$  is sufficiently large but they do not generally identify all elements of the covariance matrices of the errors.

### 11.3.1 The Benefits of Restrictions

To appreciate the opportunities offered by restrictions on the parameter matrices, consider the simple yet not unrealistic case where the measurement errors are independently and identically distributed over time. Then  $\Sigma_\nu = \sigma_\nu^2 I_T$ , say. Provided that  $\Sigma_\xi$  is not diagonal, any estimator of the form

$$\hat{\beta}_{\hat{W}} \equiv \frac{\sum_n x'_n \hat{W} y_n}{\sum_n x'_n \hat{W} x_n} \xrightarrow{p} \frac{\text{tr}(W \Sigma_\xi)}{\text{tr}(W \Sigma_x)} \beta = \frac{\text{tr}(W \Sigma_\xi)}{\text{tr}(W \Sigma_\xi) + \text{tr}(W \Sigma_\nu)} \beta$$

is consistent if the diagonal of  $\hat{W}$  contains only zeros. So when  $x_n$  is thought to suffer from measurement error and the structure of that measurement error is thought to be of the simplest form, it is straightforward to adapt the OLS estimator of  $\beta$  to obtain consistency. One choice of  $W$  would be the matrix with ones on the pseudo-diagonal above the diagonal and zeros elsewhere, implying estimation by using the lagged  $x_n$  as an instrument.

As a start of a more formal approach, write (11c) in vectorized form,  $E(x_n \otimes x_n) = \text{vec } \Sigma_\xi + \sigma_\nu^2 \text{vec } I_T$ . Let  $F$  be a matrix of order  $T^2 \times T(T-1)/2$ , with elements 0 or 1, such that  $F'(x_n \otimes x_n)$  stacks the elements of  $x_n x'_n$  below its diagonal. For example, for  $T = 3$ ,

$$F' = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

Since  $F' \text{vec } I_T = 0$ ,  $E(F'(x_n \otimes x_n)) = F' \text{vec } \Sigma_\xi$  and, with  $R'_n \equiv F'(x_n \otimes I_T)$ ,  $E(R'_n u_n) = 0$ . With  $u_n = y_n - x_n \beta$ , this is a set of  $T(T-1)/2$  moment conditions that allow for consistent estimation of  $\beta$  by IV. In general, for  $R_n$  being a  $T \times g$  matrix satisfying  $E(R'_n u_n) = 0$  while  $E(R'_n x_n) \neq 0$ , the general form of an IV estimator is

$$\hat{\beta}_{\text{IV}} = \frac{x' R W R' y}{x' R W R' x}, \quad (12)$$

with the  $y_n$  and  $x_n$  collected in  $NT$ -vectors  $y$  and  $x$  and the  $R_n$  collected in the  $NT \times g$  matrix  $R$  with  $R' \equiv (R'_1, \dots, R'_N)$  and  $W$  a weight matrix, with  $W = (R'R)^{-1}$  for 2SLS and  $W = (R' \hat{\Omega} R)^{-1}$  for 3SLS (or efficient GMM). Here,  $\hat{\Omega} = I_N \otimes \sum_n \tilde{u}_n \tilde{u}'_n / N$ , in which  $\tilde{u}_n = y_n - x_n \tilde{\beta}$ , where  $\tilde{\beta}$  is an initial consistent (but not efficient) estimator, for example, the 2SLS estimator. Cameron and Trivedi (2005) provide an extensive summary of IV estimation in a panel data context. When consistent estimation by IV is possible, testing for the presence of measurement error is straightforward, by comparing the OLS and IV estimates, cf. Wansbeek and Meijer (2000, Section 6.2).

This example shows how restrictions on the parameter matrices provide a shortcut to consistent estimation. Also, it raises two issues. First, the IV estimator given above is not optimal. When we would apply our logic to the case where  $\Sigma_v$  is merely diagonal rather than scalar, the same estimator is still consistent, so the equality of measurement error variances over time has not been exploited. A second issue concerns  $\Sigma_\varepsilon$ . We may also be willing to restrict this matrix, for example, by adopting a random effects structure. Since  $\Sigma_\varepsilon$  occurs in an equation involving  $\beta^2$ , we may give  $\Sigma_\varepsilon$  the same treatment as  $\Sigma_v$ , but we might encounter an issue of non-linearity. We now discuss these issues.

### 11.3.2 Restrictions on $\Sigma_\varepsilon$ and $\Sigma_v$

Meijer, Spierdijk, and Wansbeek (2012), MSW12 here in after, consider restrictions on  $\Sigma_\varepsilon$ . Here we extend their approach and include restrictions on  $\Sigma_v$ , which have been studied extensively in the literature; see, for example, Biørn and Klette (1998), Biørn (2000), Wansbeek (2001), Shao, Xiao, and Xu (2011), Xiao et al. (2007), and Xiao, Shao, and Palta (2010a, b).

The restrictions we consider are linear and can hence be expressed as  $\text{vech } \Sigma_\varepsilon = C_\varepsilon \pi_\varepsilon$  and  $\text{vech } \Sigma_v = C_v \pi_v$ , with  $C_\varepsilon$  and  $C_v$  known and  $\pi_\varepsilon$  ( $r_\varepsilon \times 1$ ) and  $\pi_v$  ( $r_\pi \times 1$ ) unknown. For example, in the random-effects model with  $T = 5$ :

$$\pi_\varepsilon = \begin{pmatrix} \sigma_\varepsilon^2 \\ \sigma_\alpha^2 \end{pmatrix}$$

$$C'_\varepsilon = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix},$$

with  $\sigma_\varepsilon^2$  the variance of the idiosyncratic error and  $\sigma_\alpha^2$  the variance of the individual effect. We denote the elimination matrix by  $L_T$  (of order  $T^2 \times T(T+1)/2$ ) and the duplication matrix by  $D_T$  (of order  $T^2 \times T(T+1)/2$ ). The former, when multiplied with the vec of a matrix, stacks the lower triangular elements of a matrix into a vector, and the latter is defined by the property  $D_T \text{vech}(A) = \text{vec}(A)$  for any symmetric matrix  $A$ . So  $L'_T D_T = I_{T(T+1)/2}$ . See, for example, Magnus and Neudecker (1986) for these matrices and their properties. We define the symbol  $\bar{\otimes}$  by the relation  $w_1 \bar{\otimes} w_2 = L'_T(w_1 \otimes w_2)$  for any  $T$ -vectors  $w_1$  and  $w_2$ .

We can view (11) as moment conditions. With  $\sigma_\xi \equiv \text{vech } \Sigma_\xi$  they can be written in stacked form as  $E(h_n) = 0$ , with

$$h_n \equiv \begin{pmatrix} y_n \bar{\otimes} y_n - D_T \sigma_\xi \beta^2 - D_T C_\varepsilon \pi_\varepsilon \\ x_n \bar{\otimes} y_n - D_T \sigma_\xi \beta \\ x_n \bar{\otimes} x_n - D_T \sigma_\xi - D_T C_v \pi_v \end{pmatrix}. \quad (13)$$

Let  $C_\varepsilon^+ \equiv (C'_\varepsilon C_\varepsilon)^{-1} C'_\varepsilon$  and let  $C_\varepsilon^\perp$  be an orthogonal complement of  $C'_\varepsilon$ . Let  $C_\nu^+$  and  $C_\nu^\perp$  be analogous with respect to  $C_\nu$ . Then

$$\begin{aligned} & E \left( \begin{array}{cccc} C_\varepsilon^\perp & 0 & 0 & 0 \\ C_\varepsilon^+ & 0 & 0 & 0 \\ 0 & C_\nu^\perp & 0 & 0 \\ 0 & C_\nu^+ & 0 & 0 \\ 0 & 0 & L'_T & 0 \\ 0 & 0 & 0 & I_{T(T+1)/2} \end{array} \right) \left( \begin{array}{ccc} L'_T & -\beta L'_T & 0 \\ 0 & L'_T & -\beta L'_T \\ 0 & I_{T^2} - K_T & 0 \\ 0 & 0 & L'_T \end{array} \right) h_n \\ & = E \left( \begin{array}{c} C_\varepsilon^\perp (I_T \bar{\otimes} y_n) u_n \\ C_\varepsilon^+ (u_n \bar{\otimes} y_n) - \pi_\varepsilon \\ C_\nu^\perp (x_n \bar{\otimes} I_T) u_n \\ C_\nu^+ (x_n \bar{\otimes} u_n) + \pi_\nu \beta \\ x_n \bar{\otimes} y_n - y_n \bar{\otimes} x_n \\ x_n \bar{\otimes} x_n - \sigma_\xi - C_\nu \pi_\nu \end{array} \right) = 0, \end{aligned}$$

where  $K_T$  is the commutation matrix with property  $K_T(w_1 \otimes w_2) = w_2 \otimes w_1$  for any  $T$ -vectors  $w_1$  and  $w_2$  (see, e.g., Wansbeek, 1989). In this double transformation, no information from the moment conditions gets lost (although there are some elements that are identically zero in the fifth row), allowing for estimation that efficiently uses the second moments (11). The second, fourth, and sixth rows just-identify  $\pi_\varepsilon$ ,  $\pi_\nu$ , and  $\sigma_\xi$ , respectively. For estimating  $\beta$ , the first and third row are available, so we obtain

$$R'_{\varepsilon,n} = C_\varepsilon^\perp (I_T \bar{\otimes} y_n) \quad (14a)$$

$$R'_{\nu,n} = C_\nu^\perp (x_n \bar{\otimes} I_T) \quad (14b)$$

as the instrument matrices per individual  $n$  that, when collected over  $n$  in  $R$ , can be used in (12). Simulation results presented in MSW12 for the case of restrictions on  $\Sigma_\varepsilon$  show the approach works well.

Not all restrictions of potential interest are linear restrictions. A notable example of a non-linear restriction is the assumption that  $\varepsilon_n$  and/or  $\nu_n$  is AR(1). A computationally convenient way to obtain an asymptotically efficient estimator for this case is the following. First estimate the parameters of the model that only assumes that all elements of the  $i$ th pseudodiagonal are equal ( $i = 1, \dots, T-2$ ). This is a linear restriction that fits in the framework presented here. It is also much weaker than the AR(1) assumption. A multitude of consistent estimators of the autocorrelation parameter can be obtained from this, the easiest one being the estimate of the one-lag autocovariance (the value of all elements on the first pseudodiagonal) divided by the variance (the value of all elements on the diagonal). In the second step, an efficient estimator is obtained by writing out the non-linear moment conditions with the AR(1) restriction and computing the linearized GMM estimator, which is a single iteration of a Gauss-Newton algorithm. See Wansbeek and Meijer (2000), pp. 239-241, for the details. A variant of this method

can be used for any ARMA model for the errors, and any other non-linear restrictions that are nested within an identified model with (weaker) linear restrictions.

### 11.3.3 Using Third Moments

Following on a literature for the cross-sectional case starting with Geary (1942) and extended by Pal (1980), Van Montfort, Mooijaart, and de Leeuw (1987), Lewbel (1996, 1997), Dagenais and Dagenais (1997), and Erickson and Whited (2002), the third moments of the error-ridden regressor provide another source of moment conditions. The essential assumptions are that  $\xi_n$ ,  $\varepsilon_n$ , and  $v_n$  are not just uncorrelated but mutually independent, and that the third moments of  $\xi_n$  do not vanish. Then  $E(y_{nt}^2 x_{nt}) = \beta^2 E(\xi_{nt}^3)$  and  $E(y_{nt} x_{nt}^2) = \beta E(\xi_{nt}^3)$  and thus

$$\frac{\sum_{n,t} y_{nt}^2 x_{nt}}{\sum_{n,t} y_{nt} x_{nt}^2}$$

is a consistent estimator of  $\beta$ . Note that this estimator can also be used in cross-sectional data, and that it can be viewed as an IV estimator with instrument  $R_{nt} = y_{nt} x_{nt}$ . Unfortunately, this estimator generally does not have good statistical properties in small to moderate samples.

MSW12 generalize the theory to the panel setting, which increases the number of potential instruments. In particular, they show that

$$R'_{yx,n} \equiv G'_T(y_n \otimes I_T \otimes x_n),$$

with  $G_T$  the triplication matrix (Meijer, 2005), is a valid IV matrix when at least one of the  $T^3$  elements of  $E(\xi_n \otimes \xi_n \otimes \xi_n)$  is nonzero. Their simulations show that the resulting estimator does not have good statistical properties when the efficient GMM estimator (with the asymptotically optimal weight matrix, i.e., the standard two-step GMM) is used. However, the method has good first-order properties when the identity weight matrix is used, and only slightly worse when 2SLS is done. Also, good second-order properties are obtained through bootstrapping with recentered moments (Horowitz 2001, Section 3.7).

### 11.3.4 Using an Exogenous Regressor

As MSW12 note, the presence of a regressor  $z_n$ , say, in addition to  $\xi_n$  offers scope for identification and hence consistent estimation. When  $z_n$  is strongly exogenous,  $E(z_n \otimes u_n) = 0$ , and when it is weakly exogenous,  $E(z_n \bar{\otimes} u_n) = 0$ . These equations constitute  $T^2$  and  $T(T+1)/2$  moment conditions, respectively, whereas the model has just a single additional parameter, the coefficient of  $z_n$ , thus suggesting ample scope for consistent estimation.

This idea may seem like the egg of Columbus. However, there is a catch. Focusing on the case of strong exogeneity, consider the the Jacobian

$$J \equiv E\left(\frac{\partial(z_n \otimes u_n)}{\partial(\beta, \gamma)}\right) = -(\text{vec } \Sigma_{xz}, \text{vec } \Sigma_z),$$

where  $\Sigma_{xz} = E(x_n z'_n)$  and  $\Sigma_z = E(z_n z'_n)$ . For identification of  $\beta$  and  $\gamma$ ,  $J$  has to be of full column rank. When, for example,  $x_n = cz_n + w_n$  with  $E(z_n w'_n) = 0$ ,  $\Sigma_{xz} = c\Sigma_z$  and the model is not identified since  $\text{rank}(J)=1$ . This is an extreme case and in practice  $J$  will be of full column rank. But the asymptotic variance of  $\hat{\beta}$  and  $\hat{\gamma}$  depends on the degree of collinearity of the two columns of  $J$ , and in many cases of empirical relevance this degree will be too high for this approach to be reliable.

Yet the presence of  $z_n$  can be helpful as soon as the relation between  $x_n$  and  $z_n$  is more complex, in particular when it is heteroskedastic. Building on Lewbel (2012), MSW12 elaborate this. Let the projection of  $\xi_n$  on  $z_n$  be  $\xi_n = Kz_n + \omega_n$ , where  $K \equiv E(\xi_n z'_n)[E(z_n z'_n)]^{-1}$ . The crucial assumption is

$$E(z_n \otimes \omega_n \otimes \omega_n) \neq 0,$$

which basically says that the relation between  $z_n$  and  $\omega_n$  is heteroskedastic in nature. With  $\kappa \equiv \text{vec } K$  and

$$\begin{aligned} w_n &\equiv v_n + \omega_n = x_n - (z_n \otimes I_T)' \kappa \\ h_n &\equiv \begin{pmatrix} z_n \otimes w_n \\ z_n \otimes u_n \\ z_n \otimes w_n \otimes u_n \end{pmatrix}, \end{aligned}$$

$E(h_n) = 0$  and thus  $h_n$  is a valid set of moments. In addition to the two parameters  $\beta$  and  $\gamma$ , there are now  $T^2$  nuisance parameters  $\kappa$  in the model. The Jacobian now is

$$J_\dagger = - \begin{pmatrix} \Sigma_z \otimes I_T & 0 & 0 \\ 0 & \text{vec } \Sigma_{xz} & \text{vec } \Sigma_z \\ 0 & q_1 & q_2 \end{pmatrix},$$

with

$$\begin{aligned} q_1 &= (I_T \otimes I_T \otimes K)E(z_n \otimes \omega_n \otimes z_n) + E(z_n \otimes \omega_n \otimes \omega_n) \\ q_2 &= E(z_n \otimes \omega_n \otimes z_n). \end{aligned}$$

The upper-left block of  $J_\dagger$  concerns  $\kappa$  and the rows and columns of zeros surrounding it indicate that the identification and asymptotic variance of  $\beta$  and  $\gamma$  stand apart from  $\kappa$ . The lower-right blocks of  $J_\dagger$  show the effect of the heteroskedasticity assumption. It basically means that the two columns of  $J$  are each stretched by  $T^3$  elements that, by assumption, are not all zero. This should reduce the degree of collinearity between the columns, thus leading to better-behaved estimators. MSW12 show that the non-linearity in the moment condition  $E(h_n) = 0$  can be circumvented through a simple

two-step procedure, in which the first step constitutes of estimating  $K$  by OLS of  $x_n$  on  $z_n$  and computing the resulting residuals  $\hat{w}_n$ , and the second step constitutes of estimating  $\beta$  and  $\gamma$  by GMM using the second and third element of  $h_n$ , with  $w_n$  replaced by  $\hat{w}_n$ , which can be written as a linear IV estimator. They present simulations that show good performance of the method.

## 11.4 STRUCTURAL EQUATION MODELS

---

The linear panel data model with measurement error belongs to a class of models called structural equation models (SEMs). Structural equation modeling is a general framework to deal with latent variables, of which variables measured with error are a special case. When a measurement error model can be written as an SEM, estimation can be done by one of the various computer programs for SEMs. Dedicated SEM software packages are LISREL (<http://www.ssicentral.com>), EQS (<http://www.mvsoft.com>), Mplus (<http://www.statmodel.com>), and Mx (<http://www.vcu.edu/mx/>). OpenMx (<http://openmx.psyc.virginia.edu/>) is a reimplementation of the latter for usage with R. SEM modules of general statistical software are R's sem package (<http://cran.r-project.org/web/packages/sem/index.html>), SPSS Amos (<http://www.spss.com/amos/>), SAS CALIS (<http://support.sas.com/rnd/app/da/stat/procedures/calis.html>), and the SEM module of Stata (<http://www.stata.com/stata12/structural-equation-modeling/>).

### 11.4.1 Lisrel

There are several general model structures of linear SEMs, which all encompass the same class of models. We use the specification developed by Jöreskog and others, which is the oldest and best-known and is used in the LISREL program, and refer to it as the Lisrel model. Omitting for simplicity of exposition the various vectors of intercept parameters, the Lisrel model is

$$\eta_n = B\eta_n + \Gamma\xi_n + \zeta_n \quad (15a)$$

$$y_n = \Lambda_y\eta_n + \varepsilon_n \quad (15b)$$

$$x_n = \Lambda_x\xi_n + \delta_n, \quad (15c)$$

where  $\eta_n$  is a vector of endogenous latent variables,  $\xi_n$  is a vector of exogenous latent variables,  $y_n$  and  $x_n$  are vectors of observed variables,  $\zeta_n$ ,  $\varepsilon_n$ , and  $\delta_n$  are error terms,  $B$  and  $\Gamma$  are matrices of regression coefficients among the latent variables, and  $\Lambda_y$  and  $\Lambda_x$  are matrices of factor loadings, linking observed and latent variables. The first equation is a simultaneous equations regression model for the latent variables, and the second and third equations are factor analysis submodels, also jointly called the

measurement model, relating the latent variables  $\eta_n$  and  $\xi_n$  to their “indicators”  $y_n$  and  $x_n$ , respectively. The random vectors  $\xi_n$ ,  $\zeta_n$ ,  $\varepsilon_n$ , and  $\delta_n$  are assumed to be mutually uncorrelated with means zero and covariance matrices  $\Phi \equiv E(\xi_n \xi'_n)$ ,  $\Psi \equiv E(\zeta_n \zeta'_n)$ ,  $\Theta_\varepsilon \equiv E(\varepsilon_n \varepsilon'_n)$ , and  $\Theta_\delta \equiv E(\delta_n \delta'_n)$ , respectively.

By imposing restrictions on the various parameter matrices a wide variety of specific models can be generated. Already Jöreskog (1978) showed how panel data models can be written as SEMs. Clearly, SEMs are well-suited to estimate measurement error models and, more generally, models with latent variables. SEMs are widely used in the behavioral sciences, marketing, and other fields, but their potential has been underexploited in economics. Aasness, Biørn, and Skjerpen (1993), who formulated a demand system for panel data where total expenditures is the exogenous variable measured with error, is an early exception.

### 11.4.2 The Measurement Error Model

We now indicate how the linear panel data with measurement error fits into the Lisrel mold. To avoid confusion, we add the superscript “L” to Lisrel symbols. The model, with individual effects, is

$$\begin{aligned} y_n &= \xi_n \beta + \iota_T \alpha_n + e_n \\ x_n &= \xi_n + v_n, \end{aligned}$$

where  $\alpha_n$  does or does not correlate with  $\xi_n$ . Considering the latter case first, the Lisrel specification has, for the coefficients,  $B^L \equiv 0$  as there is no interaction between the endogenous variables over time,  $\Gamma^L = \beta I_T$  (the diagonal elements of  $\Gamma^L$  are specified to be equal and the off-diagonal elements are specified to be zero) since the regression coefficient is taken to be constant over time, and  $\Lambda_y^L = \Lambda_x^L = I_T$ . As to the variance matrices,  $\Phi^L = \Sigma_\xi$  (so is left unrestricted),  $\Psi^L$  structures the errors in the equations, so in the simplest random effects model it has diagonal elements equal and off-diagonal elements equal,  $\Theta_\varepsilon^L = 0$  since the dependent variable does not contain measurement error (or it is subsumed in  $e_n$ ), and  $\Theta_\delta^L (= \Sigma_v)$  is structured as deemed appropriate, with  $\Theta_\delta^L = \sigma_v^2 I_T$  being the simplest choice.

In the case of correlated effects, one way to adapt this is to consider the individual effect as a regressor and not as a component part of the errors in the equations. This allows us to accommodate  $\sigma_{\xi\alpha}$  in the Lisrel specification. So now  $\Gamma^L = (\beta I_T, \iota_T)$  and  $\Lambda_x^L \equiv (I_T, 0)$ , and

$$\Phi^L = \begin{pmatrix} \Sigma_\xi & \sigma_{\xi\alpha} \\ \sigma'_{\xi\alpha} & \sigma_\alpha^2 \end{pmatrix}.$$

The error covariance matrix  $\Theta_\varepsilon^L (= E(e_n e'_n))$  is structured as deemed appropriate, for example,  $\Theta_\varepsilon^L = \sigma_e^2 I_T$  in the simplest case.

This shows how the simple measurement error model can be expressed in Lisrel. The same holds for many more general models, whose estimation does not require dedicated methods since Lisrel takes care of the estimation. Mixture models, stratified and clustered samples, and full information estimation with missing data have also been studied in the literature and are features of some of the more advanced programs.

The actual way in which a particular model can be written as an SEM is sometimes quite complicated algebraically. But, fortunately, applied researchers typically do not have to make these translations explicitly, because the software allows for a more intuitive model specification, either through almost literally writing the equations or through graphical user interfaces.

On the other hand, writing a model as an SEM does not absolve the researcher from considering all other intricacies. Most importantly, the researcher still needs to ascertain that the model is identified. In some cases, the restrictions on the model are sufficient to identify  $\beta$  but not all auxiliary parameters like  $\Sigma_\varepsilon$ . This may lead to numerical problems when trying to estimate the model as an SEM. To avoid this, either additional arbitrary identification restrictions should be added, or the variables should be transformed (e.g., put in first differences) before estimation to eliminate the underidentified parameters if possible.

SEMs were originally developed for dealing with latent variables with multiple proxies. In the example above, and in the discussion so far, we have assumed that we have only one proxy per variable (per time point), and as a result the factor loadings matrices  $\Lambda_y^L$  and  $\Lambda_x^L$  were identity matrices. If there are multiple proxies, these factor loadings matrices are augmented with additional rows reflecting the relations between the additional proxies and the latent regressor of interest, and the error variance matrices are augmented accordingly. If it can be assumed that measurement errors of different proxies are uncorrelated, the additional proxies increase the scope for identification and generally increase precision of the estimators.

## 11.5 THE DYNAMIC MODEL

---

In this Section we consider measurement error when the model is dynamic, that is, the one-period lagged dependent variable is among the regressors. This case has received some attention in the literature, for example, Wansbeek and Kapteyn (1992), Dynan (2000), and Antman and McKenzie (2007). The most extensive treatment is given by Biørn (2014), who studies IV estimation (with the model in levels and the IVs in differences, and vice versa) of the model the dependent variable and the exogenous regressor are both subject to measurement error. Here we discuss the basic elements of the dynamic model with measurement error, following Meijer, Spijeldijk, and Wansbeek (2013; MSW13 hereinafter).

We focus on the essentials and restrict our attention to the simplest case, where the lagged dependent variable is the only regressor. Thus,

$$\eta_{nt} = \eta_{n,t-1}\gamma + \alpha_n + e_{nt} \quad (16a)$$

$$y_{nt} = \eta_{nt} + v_{nt}, \quad (16b)$$

for  $n = 1, \dots, N$  and  $t = 1, \dots, T$ . In this model,  $\eta_{nt}$  is an unobserved variable subject to an AR(1) process with individual effects. The error term consists of an idiosyncratic part  $e_{nt}$  and an individual effect  $\alpha_n$ . Instead of  $\eta_{nt}$  we observe  $y_{nt}$  according to the measurement equation (16b), where  $v_{nt}$  is the measurement error; our notation is such that  $y_{n0}$  is observed. As before, all variables have mean zero over  $n$ , possibly after demeaning per time period thus accounting for fixed time effects. We make the (highly) simplifying assumption that  $e_{nt}$  and  $v_{nt}$  are uncorrelated over time, and have constant variance.

### 11.5.1 Properties of the Model

Since the same variable appears at the left and right hand side, there is measurement error at the left and at the right. The reduced-form model

$$y_{nt} = y_{n,t-1}\gamma + \alpha_n + (e_{nt} + v_{nt} - v_{n,t-1}\gamma)$$

has the usual correlation between regressor and error term since they share the measurement error term, leading to inconsistency if untreated in estimation. But, as is overly well-known, even without measurement error, there is a problem in estimating this model since the model implies that the regressor is correlated with the individual effect, which is part of the error term.

This has led to a simple consistent estimation method, which we will call AH as it is due to Anderson and Hsiao (1982). To obtain the AH estimator, transform the model in first differences thus shedding the individual effect, and next use the preceding value of the sole variable in the model as an instrument,

$$\hat{\gamma}_{AH} = \frac{\frac{1}{N} \sum_n y_{n,t-2}(y_{nt} - y_{n,t-1})}{\frac{1}{N} \sum_n y_{n,t-2}(y_{n,t-1} - y_{n,t-2})}. \quad (17)$$

The consistency of this estimator needs a lack of correlation over time of the  $e_{nt}$ . First-differencing is an effective way to eliminate the individual effect but it does not eliminate measurement error. Hence the AH estimator, and the estimators that were built on it later on and gained huge popularity like the one due to Arellano and Bond (1991), are inconsistent in the presence of measurement error. MSW13 show that

$$\hat{\gamma}_{AH} \xrightarrow{P} \gamma_* = \frac{\gamma \sigma_e^2}{\sigma_e^2 + (1 + \gamma) \sigma_v^2} = \frac{\gamma}{1 + (1 + \gamma) \lambda},$$

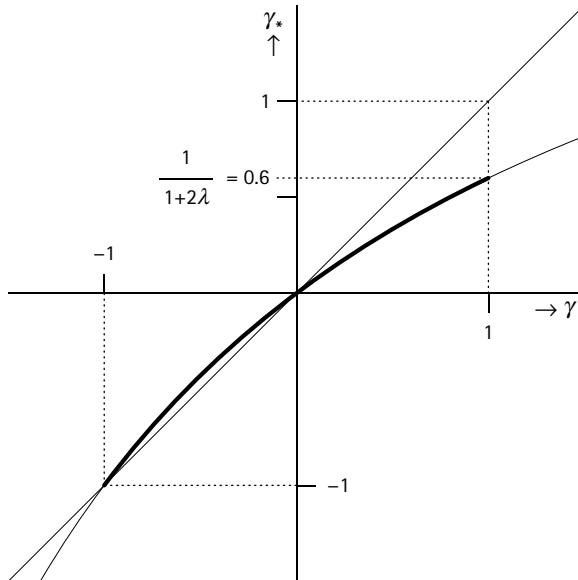


FIGURE 11.2 Probability limit  $\gamma_*$  of the Anderson-Hsiao estimator with  $\lambda = \frac{1}{3}$

where  $\lambda = \sigma_v^2/\sigma_e^2$ . Figure 11.2, adapted from MSW13, illustrates this result. Note that the attenuation bias is largest when the autocorrelation parameter is large, which may be the typical case in many applications. Fortunately, however, when, as we had assumed, the  $v_{nt}$  are uncorrelated over time, the AH estimator can be adapted in a simple way to restore its consistency. All that is needed is to replace in (17) the instrument  $y_{n,t-2}$  by  $y_{n,t-3}$ . But apart from consistency we may also be concerned with efficiency. We now turn to this issue.

### 11.5.2 Towards Optimal Estimation

Following MSW13, we now turn to a more systematic approach to estimating the dynamic model with measurement error. Define

$$\begin{aligned} y_n &\equiv (y_{n1}, \dots, y_{nT})' \\ y_{n,-1} &\equiv (y_{n0}, \dots, y_{n,T-1})' \\ y_{n,+} &\equiv (y_{n0}, \dots, y_{nT}). \end{aligned}$$

MSW13 consider IV estimation of  $\gamma$  with instrument  $A'y_{n,+}$ , for some  $(T+1) \times T$ -matrix  $A$ ,

$$\hat{\gamma}_A = \frac{\sum_n y'_{n,+} A y_n}{\sum_n y'_{n,+} A y_{n,-1}}.$$

There are three issues to be resolved. The first one concerns the existence of this estimator:  $A$  should be such that neither numerator nor denominator is identically equal to zero. The second one concerns consistency. Within the set of  $A$ 's that pass the existence test, the properties have to be found for  $A$  to make  $\hat{\gamma}_A$  consistent. The third issue concerns the choice of  $A$  such that  $\hat{\gamma}_A$  is not just consistent but has minimal asymptotic variance.

To describe the results, some notation is needed. Let  $C'_0 \equiv (I_T, 0_T)$  and let  $C'_1, \dots, C'_T$  be a series of matrices of order  $T \times (T + 1)$ , where  $C'_1 \equiv (0_T, I_T)$ , in  $C'_2$  the ones are moved one position to the right, and so on, ending with  $C'_T$ , which is zero, except for its  $(1, T + 1)$  element. Further,

$$\begin{aligned} C &\equiv (\text{vec } C_0, \dots, \text{vec } C_T) \\ a &\equiv \text{vec } A \\ F_\tau &\equiv (C'_\tau \otimes I_{T+1})D_{T+1}, \quad \tau = 0, 1 \\ s_y &\equiv \frac{1}{N} \sum_n (y_{n,+} \bar{\otimes} y_{n,+}), \end{aligned}$$

with  $D_{T+1}$  the duplication matrix of order  $(T + 1)^2 \times (T + 1)(T + 2)/2$ . MSW13 show that  $\hat{\gamma}$  can be rewritten as

$$\hat{\gamma}_A = \frac{a' F_1 s_y}{a' F_0 s_y},$$

hence the existence of  $\hat{\gamma}_A$  requires  $F'_0 a \neq 0$  and  $F'_1 a \neq 0$ .

As to consistency, the reduced form of the model is

$$\begin{aligned} y_n &= \gamma y_{n,-1} + v_n \\ v_n &\equiv \alpha_n \iota_T + e_n + \nu_n - \gamma \nu_{n,-1}, \end{aligned}$$

so consistency requires  $A$  to be such that

$$E(y'_{n,+} A v_n) = E(\alpha_n y'_{n,+} A \iota_T) + \text{tr}(E[(e_n + \nu_n - \gamma \nu_{n,-1}) y'_{n,+}] A) = 0.$$

In the “system GMM” case where  $E(\alpha_n y_{n,+}) = c \iota_{T+1}$  for some  $c$ , one requirement for consistency is  $\iota'_{T+1} A \iota_T = 0$  or  $\iota'_{T(T+1)} a = 0$ . When the correlation between individual effect and  $y$  is not structured (let alone constant over time), the wider set of conditions  $A \iota_T = 0$  or  $(\iota'_T \otimes I_T)a = 0$  is required. Elaboration of the other term readily shows that  $C' a = 0_{T+1}$  is also required.

In search of optimality, MSW13 write the restrictions on  $a$ , which are all linear, in the condensed form  $a = Qb$ , where  $Q$  is full column rank, and  $b$  can be chosen freely, subject to the numerator and denominator of  $\hat{\gamma}$  not becoming identically zero. Then

$$\hat{\gamma}_A = \frac{b' Q' F_1 s_y}{b' Q' F_0 s_y}$$

where  $b$  is to be determined such that the asymptotic variance of  $\hat{\gamma}_A$  is minimal. With

$$\Psi_y \equiv \text{plim}_{N \rightarrow \infty} \frac{1}{N} \sum_n [(y_{n,+} \bar{\otimes} y_{n,+} - s_y)(y_{n,+} \bar{\otimes} y_{n,+} - s_y)'].$$

Because of the singularity of  $Q' F_1 \Psi_y F_1' Q$ , the optimum is elusive but using the Moore-Penrose inverse yields

$$b = (Q' F_1 \Psi_y F_1' Q)^+ Q' F_0 \sigma_y,$$

leading to

$$\hat{\gamma} = \frac{s_y' F_0' Q (Q' F_1 \hat{\Psi}_y F_1' Q)^+ Q' F_1 s_y}{s_y' F_0' Q (Q' F_1 \hat{\Psi}_y F_1' Q)^+ Q' F_0 s_y}.$$

The denominator is a consistent estimator of the asymptotic variance of  $\hat{\gamma}_{\text{OPT}}$ . This estimator shows very performance in the empirical example studied by MSW13.

### 11.5.3 The Dynamic Model as a SEM

Like the static model, the dynamic model can be estimated after formulating it as an SEM. To see how, take  $T = 3$  for simplicity of notation. Then we can write the dynamic model (16) as

$$\begin{aligned} \begin{pmatrix} \eta_{n1} \\ \eta_{n2} \\ \eta_{n3} \end{pmatrix} &= \begin{pmatrix} 0 & 0 & 0 \\ \gamma & 0 & 0 \\ 0 & \gamma & 0 \end{pmatrix} \begin{pmatrix} \eta_{n1} \\ \eta_{n2} \\ \eta_{n3} \end{pmatrix} + \begin{pmatrix} \gamma & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \eta_{n0} \\ \alpha_n \\ e_n \end{pmatrix} + \begin{pmatrix} e_{n1} \\ e_{n2} \\ e_{n3} \end{pmatrix} \\ \begin{pmatrix} y_{n1} \\ y_{n2} \\ y_{n3} \end{pmatrix} &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \eta_{n1} \\ \eta_{n2} \\ \eta_{n3} \end{pmatrix} + \begin{pmatrix} v_{n1} \\ v_{n2} \\ v_{n3} \end{pmatrix} \\ y_{n0} &= (1, 0) \begin{pmatrix} \eta_{n0} \\ \alpha_n \end{pmatrix} + e_{n0}. \end{aligned}$$

From here, the specification as a Lisrel model follows readily. The “trick” is to add the random individual effect  $\alpha_n$  to vector of regressors. In the same way we accommodated fixed effects in the Lisrel specification for the static model. Notice that one of the parameters to be estimated is the covariance between the two latent regressors,  $\eta_{n0}$  and  $\alpha_n$ .

## 11.6 NONCLASSICAL MEASUREMENT ERROR

An assumption up till now has been the independence of  $\xi_n$  and  $v_n$ . This case is often called “classical” measurement error, to distinguish it from “nonclassical measurement

error,” which is the case where the two are not independent. (Sometimes, the notion is extended to models where the errors in the equation and the measurement errors are dependent.) The empirical relevance of nonclassical measurement error was stressed by Bound and Krueger (1991). They compared self-reported wage data for two years with the corresponding payroll tax data, assumed to be relatively accurate. They found that the measurement errors in the self-reported data are positively correlated over time and negatively correlated with the administrative data; on average, poorer people overstate their wage and richer people underestimate it. Because of the latter, they call this type of measurement error “mean-reverting.” For a broad overview, see Bound, Brown, and Mathiowetz (2001). Kim and Solon (2005) investigate various implications of mean reversion in a panel data setting.

### 11.6.1 Implications

We now consider the implications of nonclassical measurement error for our analysis. Then,  $\Sigma_{v\xi} \equiv E(v_n \xi'_n) \neq 0$ , and (11) becomes

$$\begin{aligned}\Sigma_y &= \beta^2 \Sigma_\xi + \Sigma_\epsilon \\ \Sigma_{xy} &= \beta (\Sigma_\xi + \Sigma_{v\xi}) \\ \Sigma_x &= \Sigma_\xi + \Sigma_v + \Sigma_{v\xi} + \Sigma'_{v\xi}.\end{aligned}$$

The identification problem becomes more severe as the elements of  $\Sigma_{v\xi}$  yield  $T^2$  additional parameters. Yet the availability of panel data can render the model identified to some extent, if the parameter space is sufficiently restricted. Before considering identification issues, we first look at the bias of the OLS estimator. Consider the simple case where  $\Sigma_v = \sigma_v^2 I_T$  and  $\Sigma_{v\xi} = \sigma_{v\xi} I_T$ . With  $\bar{\sigma}_\xi^2 \equiv \text{tr}(\Sigma_\xi)/T$ , (3) becomes

$$\hat{\beta}_{\text{OLS}} \xrightarrow{P} \left( 1 - \frac{\sigma_v^2 + \sigma_{v\xi}}{\bar{\sigma}_\xi^2 + \sigma_v^2 + 2\sigma_{v\xi}} \right) \beta \equiv (1 - \zeta_{v\xi}) \beta.$$

This is the same bias as in the case of a single cross-section. Since

$$\frac{\partial \zeta_{v\xi}}{\partial \sigma_{v\xi}} = c (\bar{\sigma}_\xi^2 - \sigma_v^2) \quad \text{with } c > 0,$$

the absolute bias increases with  $\sigma_{v\xi}$  if  $\bar{\sigma}_\xi^2 > \sigma_v^2$ . Then, in the empirically likely case that  $\sigma_{v\xi} < 0$ , the attenuation is less than in the case of “classical” measurement error, where  $\sigma_{v\xi} = 0$ . In the best scenario, the two biasing terms cancel out,  $\sigma_{v\xi} = -\sigma_v^2$ . Then, quoting Bound and Krueger (1991), “two wrongs make a right.”

We now turn to identification and consistent estimation. In this simple case, panel data allow for consistent estimation of  $\beta$  in the presence of nonclassical measurement error since the approach, as sketched in Section 11.3.1, with  $\hat{W}$  taken to be a matrix

with zeros on the diagonal still applies;  $\sigma_{v\xi}$  only appears in elements on the diagonals. So  $\beta$  is identified; the method is robust when the measurement error becomes non-classical. The off-diagonal elements of  $\Sigma_\xi$  are identified, too. The other parameters are not. To see this, consider

$$\begin{aligned}\Sigma_y &= \beta^2 \Sigma_\xi + \sigma_\varepsilon^2 I_T & = \beta^2 (\Sigma_\xi + \varphi I_T) + (\sigma_\varepsilon^2 - \varphi \beta^2) I_T \\ \Sigma_{xy} &= \beta (\Sigma_\xi + \sigma_{v\xi} I_T) & = \beta ((\Sigma_\xi + \varphi I_T) + (\sigma_{v\xi} - \varphi) I_T) \\ \Sigma_x &= \Sigma_\xi + (\sigma_v^2 + 2\sigma_{v\xi} \varphi) I_T = (\Sigma_\xi + \varphi I_T) + ((\sigma_v^2 + \varphi) + 2(\sigma_{v\xi} - \varphi)) I_T.\end{aligned}$$

The equations show that we can add any  $\varphi$  (within certain bounds) to the diagonal elements of  $\Sigma_\xi$  and adapt the other parameters accordingly without observational implications. Hence these parameters are not identified.

So far for the simplest case. We now consider the more general case. Assume that we are willing to restrict  $\Sigma_{v\xi}$  in the same way as we did for  $\Sigma_\varepsilon$  and  $\Sigma_v$  in Section 11.3.2:

$$E(v_n \otimes \xi_n) = H\lambda. \quad (18)$$

A general analysis of the state of identification of this case seems daunting. The same holds for the question whether the instruments derived in Section 11.3.2 still hold under (18), in other words, whether estimators based on the moment conditions of Section 11.3.2 are robust against nonclassical measurement error, like in the simplest case discussed above. Anyhow, to cover nonclassical measurement error, we can adapt  $h_n$  from (13) to take (18) into account:

$$h_n^\dagger \equiv \begin{pmatrix} y_n \otimes y_n - D_T \sigma_\xi \beta^2 - D_T C_\varepsilon \pi_\varepsilon \\ x_n \otimes y_n - D_T \sigma_\xi \beta - H\lambda \beta \\ x_n \otimes x_n - D_T \sigma_\xi - D_T C_v \pi_v - (I_{T^2} + K_T) H\lambda \end{pmatrix},$$

where  $K_T$  denotes the commutation matrix. Note that  $I_{T^2} + K_T = 2D_T D_T^+$ , where  $D_T^+ = (D'_T D_T)^{-1} D'_T$  (e.g., Wansbeek and Meijer, 2000, p. 363). We have  $E(h_n^\dagger) = 0$  and we can proceed as before, that is, transform the model to find the instrument matrices for this case, generalizing (14). These are obtained by replacing in (14a)  $C_\varepsilon^\perp$  by  $C_{\dagger\varepsilon}^\perp$ , say, which is an orthogonal complement of  $(C_\varepsilon, L'_T H)'$ , and by replacing in (14b)  $C_v^\perp$  by  $C_{\dagger v}^\perp$ , say, which is an orthogonal complement of  $(C_v, D_T^+ H)'$ .

A model with nonclassical measurement error can also be formulated as an SEM. Consider the example in Section 11.4 with  $\alpha_n$  correlated with  $\xi_n$ . In order to allow for correlation between regressor and measurement error, we consider the latter as part of the former. In the formulation of the Lisrel model (and again with superscript "L" to denote Lisrel symbols), we add  $v_n$  to  $\xi_n^L$  and set  $\delta_n^L = 0$ . This requires resetting  $\Lambda_x^L$  and  $\Gamma^L$ , which become  $\Lambda_x^L \equiv (I_T, 0_T, I_T)$  and  $\Gamma^L = (\beta I_T, \iota_T, 0)$ . Finally, taking

$$\Phi^L = \begin{pmatrix} \Sigma_\xi & \sigma_{\xi\alpha} & \Sigma'_{v\xi} \\ \sigma'_{\xi\alpha} & \sigma_\alpha^2 & 0 \\ \Sigma_{v\xi} & 0 & \Sigma_v \end{pmatrix},$$

accommodates the nonclassical measurement error. However, as mentioned before, if not all parameters are identified, arbitrary identification restrictions or data transformations have to be applied to render all parameters of the transformed model identified, before it can be estimated with a typical SEM program.

### 11.6.2 The Berkson Model

The Berkson model can be seen as a special case of a nonclassical measurement error model. In a sense, it is the opposite of classical measurement error. The key equation in the Berkson model is the relation between the true regressor and the observed one, which is

$$\xi_n = x_n + \nu_n,$$

where now  $\nu_n$  is uncorrelated with  $x_n$ . This can emerge in several ways. One way is if the observed regressor is a set of dummies corresponding with a discretized version of the continuous true regressor, see Wansbeek and Meijer (2000, pp. 29–30). Another way in which the Berkson model emerges is that the researcher has access to multiple proxies collected in the vector  $z_{nt}$ , say, estimates a measurement model, and computes  $x_{nt} = E(\xi_{nt} | z_{nt})$ . Meijer, Kapteyn, and Andreyeva (2011) constructed an index of general health in this way. Hyslop and Imbens (2001) discuss nonclassical measurement error and argue that the Berkson model may be appropriate because the respondents may have access to more information about the regressor of interest, but still imperfect, and construct their best guess according to a process that resembles a conditional expectation.

Interestingly, the Berkson model is notable because OLS is consistent. To our knowledge, it has not yet been studied for panel data. Consistency carries over from the cross-sectional context, and robust standard errors are obtained straightforwardly. However, whether the panel context allows for much more efficient estimation is an open question.

## 11.7 MULTIPLICATIVE MEASUREMENT ERROR

---

So far, we have dealt exclusively with linear models with additive measurement errors. These have been most extensively studied in the literature, especially in the context of panel data. However, non-linear models and non-additive models are important in economics, and measurement error is not limited to linear additive models. In this Section, we discuss multiplicative measurement error and in the next Section, we discuss non-linear models. Because of both the essentially unlimited ways in which non-additive and non-linear models may emerge and the limited amount of literature

on them, this will take the form of examples that illustrate how one could proceed when faced with nonstandard situations, rather than a comprehensive treatment.

### 11.7.1 Euler Equations

Many economic problems involve trade-offs between current and future costs and benefits. The classical example is the lifecycle model in which consumption and saving in each period is chosen so as to expected discounted lifetime utility. More elaborate examples involve job search models, retirement, health investment, and portfolio choice. Solving intertemporal optimization problems is computationally very demanding and is usually done by dynamic programming. See, for example, Adda and Cooper (2003), Christensen and Kiefer (2009), Rust (1994), or Stokey and Lucas (1989). The computational burden becomes even more demanding for parameter estimation, as the intertemporal optimization problem must be solved for each observation in each iteration.

Euler equations are first order conditions that potentially greatly reduce the burden. Hall (1978) introduced the Euler equation in the study of the lifecycle model, and Hansen and Singleton (1982) were the first who estimated the parameters with GMM based on the Euler equation.

Consider a simple lifecycle model, in which the individual maximizes lifetime expected utility  $V$  as a function of consumption  $C_t$  in each period. We assume (like most of the literature) that utility is time-separable, with period utility function  $U(C_t)$  and discount factor  $\beta$ , so that

$$V = E_1 \left[ \sum_{t=1}^T \beta^t U(C_t) \right], \quad (19)$$

where  $E_s$  denotes an expectation conditional on all information known to the agent at the beginning of period  $s$ , and  $T$  is length of life (or planning horizon), which for simplicity we assume known to the agent. An individual receives income  $Y_t$  in period  $t$  (which may be deterministic or stochastic) and is able to invest in an asset with stochastic return  $R_t$  between period  $t - 1$  and period  $t$ . Net wealth at the beginning of period  $t$  is  $A_t$ , so wealth evolves according to

$$A_{t+1} = (1 + R_{t+1})(A_t + Y_t - C_t). \quad (20)$$

Initial wealth  $A_1$  is given, and the intertemporal budget constraint is  $A_{T+1} \geq \bar{A}$ , with  $\bar{A}$  a constant known to the agent. Realistic models include liquidity constraints and a consumption floor, but we ignore these for this exposition. In this model, the Euler equation is

$$E_t \left[ \frac{U'(C_{t+1})}{U'(C_t)} (1 + R_{t+1}) \beta \right] = 1. \quad (21)$$

This is a remarkable simplification: it only depends on consumption at two consecutive time points, the discount factor, and the return on assets; it does not depend on previous values of consumption, on assets or income at any point in time, nor on the budget constraint parameter  $\bar{A}$ . The period utility function is often taken to be isoelastic (or constant relative risk aversion; CRRA),  $U(C) = C^{1-\gamma}/(1-\gamma)$ . The parameter  $\gamma$  is the coefficient of relative risk aversion (or the reciprocal of the elasticity of intertemporal substitution). With this choice, the Euler equation becomes

$$E_t \left[ \left( \frac{C_{t+1}}{C_t} \right)^{-\gamma} (1 + R_{t+1}) \beta \right] = 1, \quad (22)$$

from which we obtain moment conditions of the form (reintroducing the agent subscript  $n$ )

$$E \left\{ \left[ \left( \frac{C_{n,t+1}}{C_{nt}} \right)^{-\gamma} (1 + R_{t+1}) \beta - 1 \right] Z_{nt} \right\} = 0, \quad (23)$$

where  $Z_{nt}$  is any variable known at the beginning of period  $t$  or earlier. This is a highly non-linear moment condition, and early attempts first loglinearized (22), and estimated the resulting linear dynamic panel data model by IV (or more generally GMM), but this is not strictly correct, because the error terms depend on assets and income (in a generally intractable way). Carroll (2001) shows that the approximations are bad and the results of the loglinearized Euler equations do not estimate the parameters well.

Estimation based on the non-linear moment condition (23) is correct in principle, but the results have been less than satisfactory, especially in relatively short panels. Like other microeconomic variables such as earnings and wealth, observed consumption is typically subject to substantial measurement error, and the disappointing results from GMM estimation of the Euler equation have been attributed to this measurement error. This has led several authors to develop models that incorporate measurement errors explicitly and estimators that are consistent under these model specifications. Because of the form of the Euler equation (22), multiplicative measurement error leads to a more tractable solution than additive measurement error, and thus these studies assume multiplicative measurement error.

Alan, Attanasio, and Browning (2009), later on expanded by Alan and Browning (2010), adapt the model as follows (where we use notation more similar to the notation used in this chapter). Let  $\eta_{nt}$  denote true consumption, which satisfies the exact Euler equation (22), and let  $y_{nt}$  denote observed consumption. Measurement error  $v_{nt}$  is assumed to be independent of all other variables in the model, serially independent, and stationary, and  $y_{nt} = \eta_{nt} v_{nt}$ . Hence,

$$E_t \left[ \left( \frac{y_{n,t+1}}{y_{nt}} \right)^{-\gamma} (1 + R_{t+1}) \beta \right] = E \left[ \left( \frac{v_{n,t+1}}{v_{nt}} \right)^{-\gamma} \right] = \kappa_n, \quad (24)$$

say. They then derive a second Euler equation, which links consumption in period  $t+2$  with consumption in period  $t$ , and combine this with (24) to eliminate  $\kappa_n$ , resulting in the equation

$$\mathbb{E}_t \left\{ \left[ \left( \frac{y_{n,t+1}}{y_{nt}} \right)^{-\gamma} (1 + R_{t+1}) \beta \right] - \left[ \left( \frac{y_{n,t+2}}{y_{nt}} \right)^{-\gamma} (1 + R_{t+1})(1 + R_{t+2}) \beta^2 \right] \right\} = 0, \quad (25)$$

from which moment conditions and GMM estimators (which they call GMM-D) can be readily derived. They also introduce an alternative estimator, GMM-LN, which uses the additional assumption that  $\log v_{nt}$  is normally distributed with mean  $\mu_{v,n}$  and variance  $\sigma_v^2$ . Then  $\kappa_n = \exp(\gamma^2 \sigma_v^2)$ , and (24) and its two-period analog are combined with a set of instruments observed at the beginning of period  $t$  or earlier to estimate the parameters, which now also includes  $\sigma_v^2$  (but not the mean of  $\log v_{nt}$ ). This framework has been expanded by Alan and Browning (2010).

## 11.7.2 Disclosure Avoidance

In the context of Euler equations, the choice for multiplicative measurement error was largely a choice of convenience, because of its tractable implications. We now go back to the original model and consider the implications of multiplicative measurement error there. One reason why this may be relevant is in the context of disclosure minimization, where data at the individual level are perturbed in some way. Clearly, large values need more perturbation, and multiplication by white noise is preferable over addition.

Following Ronning and Schneeweiss (2011), we consider the case where both  $y_n$  and  $x_n$  have been “masked,” so they are observed with multiplicative measurement error. We then have

$$x_{nt} = \xi_{nt}(1 + v_{nt}) \quad (26a)$$

$$y_{nt} = \eta_{nt}(1 + w_{nt}), \quad (26b)$$

with

$$\begin{pmatrix} v_{nt} \\ w_{nt} \end{pmatrix} \sim \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_v^2 & \sigma_{vw} \\ \sigma_{vw} & \sigma_w^2 \end{pmatrix} \right).$$

As usual we are interested in the second-order implications. We use the notation  $\Delta_\xi$  for the matrix obtained from  $\Sigma_\xi$  by setting the off-diagonal elements equal to zero, and analogously for  $\Delta_\varepsilon$ . Then the second-order implications of the model are

$$\Sigma_y = \beta^2 (\Sigma_\xi + \sigma_w^2 \Delta_\xi) + \Sigma_\varepsilon + \sigma_w^2 \Delta_\varepsilon \quad (27a)$$

$$\Sigma_{xy} = \beta (\Sigma_\xi + \sigma_{vw} \Delta_\xi) \quad (27b)$$

$$\Sigma_x = \Sigma_\xi + \sigma_v^2 \Delta_\xi. \quad (27c)$$

So with this structure we find, for example, for the within estimator

$$\hat{\beta}_A \xrightarrow{p} \frac{\text{tr}(A_T \Sigma_{\xi}) + \frac{T-1}{T} \sigma_{vw} \text{tr}(\Sigma_{\xi})}{\text{tr}(A_T \Sigma_{\xi}) + \frac{T-1}{T} \sigma_v^2 \text{tr}(\Sigma_{\xi})} \beta.$$

The nature of the bias depends on the particular parameter values.

In the case where  $\sigma_v^2$ ,  $\sigma_{vw}$ , and  $\sigma_w^2$  are known since they are used in the data perturbation process, consistent estimators of the model parameters can be easily found by adapting the GMM approach in Section 11.3.2. If  $\sigma_v^2$ ,  $\sigma_{vw}$ , and  $\sigma_w^2$  are not known, not all model parameters are identified. From (27), it is clear that  $\beta$  and the off-diagonal elements of  $\Sigma_{\xi}$  and  $\Sigma_{\varepsilon}$  are identified, but the diagonal elements are not.

## 11.8 NONLINEAR MODELS

---

Many economic models are non-linear and thus non-linear models with measurement error are of great importance. These models have received a fair amount of attention in recent years. Chen, Hong, and Nekipelov (2011) provide an overview. However, this almost exclusively pertains to cross-sections, and extensions to panel data are mostly lacking. The Euler equations encountered in Section 11.7.1 are a notable exception.

### 11.8.1 Polynomial Models

Polynomial models are arguably the most natural entrance into non-linear models in many situations. They emerge as a higher order Taylor series approximation to a general non-linear relation. Occasionally, there may be stronger arguments for this functional form; for example, Banks, Blundell, and Lewbel (1997) showed that economic theory restricts certain demand systems to be quadratic in the logarithm of total expenditure. Polynomial models with measurement error have also been studied in the context of estimating Engel curves, cf. Hausman et al. (1991) and Hausman, Newey, and Powell (1995).

Although often non-linear models are more complicated than linear models, polynomial models sometimes have better identification properties than linear models in the presence of measurement error. The reason is that, even if the explanatory variable is normally distributed, they lead to nonnormality in the dependent variable, with a tractable moment structure.

Wansbeek and Meijer (2000, Section 11.2) discuss the quadratic regression model with measurement error, for the cross-sectional setting. The true value of the regressor

and the measurement error were assumed to be normally distributed, as that constituted the interesting case since normality is the most conservative assumption from an identification point of view in the linear case. Their treatment can be generalized to the panel setting without individual effects or with random effects in a straightforward way. Here we discuss the somewhat more complicated case of fixed effects. The model is

$$\begin{aligned} y_{nt} &= \beta \xi_{nt} + \gamma \tilde{\xi}_{nt}^2 + \alpha_n + e_{nt} \\ x_{nt} &= \xi_{nt} + v_{nt}, \end{aligned}$$

with  $\xi_{nt}$  and  $v_{nt}$  assumed to be normally distributed. The quadratic term is  $\tilde{\xi}_{nt}^2 \equiv \xi_{nt}^2 - v_t$ , with  $v_t$  the mean over  $n$  of the  $\xi_{nt}^2$ , ensuring that the regressors have mean zero per wave, as in the linear case. Since we consider large  $N$  asymptotics we may take  $v_t$  known.

Let  $\Delta$  denoting first differences, so  $\Delta y_{nt} = \beta \Delta \xi_{nt} + \gamma \Delta \tilde{\xi}_{nt}^2 + \Delta e_{nt}$ , and let

$$q_{stu} \equiv E \left[ \xi_{ns} \left( \Delta \xi_{nt} \Delta \tilde{\xi}_{nu}^2 + \Delta \tilde{\xi}_{nt}^2 \Delta \xi_{nu} \right) \right].$$

We then readily find, extensively exploiting the assumed normality (which, in the quadratic case is beneficial since various terms have zero expectation then),

$$\begin{aligned} E(x_{ns} \Delta y_{nt} \Delta y_{nu}) &= \beta \gamma q_{stu} \\ E(x_{ns} \Delta x_{nt} \Delta y_{nu} + x_{ns} \Delta y_{nt} \Delta x_{nu}) &= \gamma q_{stu}. \end{aligned}$$

Hence, the ratio of these two (typically, after averaging over  $s, t$ , and  $u$ ) gives a consistent estimator of  $\beta$ , and consistent estimators of the other parameters then follow.

If it cannot be assumed that  $\xi_{nt}$  and  $v_{nt}$  are normally distributed, more information (e.g. independence of measurement errors over time) may be needed to obtain consistent estimators.

### 11.8.2 Limited-Dependent Variables

Limited-dependent variables are ubiquitous in econometrics. For our purposes, the most convenient way to study them is as a linear model with a latent continuous variable  $y_{nt}^*$  as the dependent variable, with an observation function  $H_t(\cdot)$  (which may depend on auxiliary parameters such as thresholds) that maps the latent continuous variable to the observed dependent variable. We will consider a binary dependent variable here, but other types of limited-dependent variables (ordinal, truncated, censored) can be handled in the same way. For a binary dependent variable, the typical observation function is  $H_t(y_{nt}^*) = I(y_{nt}^* \geq 0)$ , where  $I(\cdot)$  is the indicator function

that returns 1 if its argument is true and 0 otherwise. If the intercept is restricted elsewhere in the model, we may have  $H_t(y_{nt}^*) = I(y_{nt}^* \geq \phi_t)$ , where  $\phi_t$  is an auxiliary parameter.

Kao and Schnell (1987a, 1987b) made some early contributions in this area. Their first paper proposes a bias-adjusted estimator for the panel logit model, with properties derived under small- $\sigma$  asymptotics. The second paper studies maximum likelihood estimation of the panel probit model with correlated effects, where identification is achieved by the i.i.d. assumption on the measurement errors.

If the model for  $y_n^*$  can be written as an identified linear SEM without fixed effects, then the limited-dependent model is identified up to a set of arbitrary normalizations, just like a probit model is identified up to an arbitrary location and scale normalization. (Limited-dependent variable models with fixed effects do not generally allow for consistent estimation of the parameters even in the absence of measurement error, with the exception of logit models.) Typically, one assumes that all latent variables are normally distributed, conditional on a set of exogenous covariates  $z_n$  that are not subject to measurement error. In the simplest case,  $z_n$  contains only the constant. The resulting model is called the LISCOMP model, after the eponymous software (superseded by *Mplus*), but variants of the model have been implemented in many SEM programs. For this model, we can write

$$\left( \begin{array}{c} y_n^* \\ x_n \end{array} \right) \Big| z_n \sim N(\Pi z_n, \Sigma),$$

for some reduced form parameter matrices  $\Pi$  and  $\Sigma$ . Univariate regression models for each dependent variable (probits) and each error-ridden explanatory variable (linear regressions in the case of most models discussed so far) on  $z_n$  give consistent estimators of  $\Pi$ , and  $\Sigma$  is identified from bivariate regressions. The model parameters of interest can then be estimated in a second step by minimum distance estimation:

$$\min_{\theta} (s - \sigma(\theta))' W (s - \sigma(\theta)),$$

where  $s$  is a vector that contains the elements of the reduced form estimates of  $\Pi$  and  $\Sigma$ ,  $\theta$  are the model parameters,  $\sigma(\theta)$  writes the probability limit of  $s$  as a function of the model parameters, and  $W$  is an estimate of the inverse of the asymptotic covariance matrix of  $s$ . Statistical inference on the model parameters can then be done using standard minimum distance theory (similar to GMM). See, for example, Wansbeek and Meijer (2000, Section 11.4) and the references therein for the details. It is also possible to estimate the model parameters in one step by using (quasi) maximum likelihood with numerical integration, for example, maximum simulated likelihood. This was implemented (in a cross-sectional model) by Meijer, Kapteyn, and Andreyeva (2011), which allowed them to deal with missing data and highly skewed binary dependent variables and covariates.

## 11.9 VALIDATION STUDIES AND OTHER FORMS OF EXTERNAL INFORMATION

---

As we saw in Section 11.2.6, the basic linear panel data model with measurement errors is not identified in the absence of additional information. We then proceeded to explore different types of additional information, starting with restrictions on the parameters. A special (and particularly powerful) restriction is obtained when we know the value of one or more parameters from external sources. For example, suppose we know the measurement error variance at time  $\tau$ , that is,  $\Sigma_{v,\tau\tau}$ . From (11b) and (11c), we see that

$$\beta = \frac{\Sigma_{xy,\tau\tau}}{\Sigma_{xx,\tau\tau} - \Sigma_{v,\tau\tau}}.$$

Because  $\Sigma_{xy}$  and  $\Sigma_{xx}$  are consistently estimated by their sample analogs and  $\Sigma_{v,\tau\tau}$  is known, it then immediately follows that  $\beta$  can be consistently estimated. Unfortunately, in econometrics, measurement error variances are generally unknown. This has led to some validation studies, which collected data on key variables of interest from administrative records and compared these with the reported values in the survey. There is a large literature on validation studies. See Bound, Brown, and Mathiowetz (2001) for an overview. Especially the validation study of the Panel Study of Income Dynamics has been influential and has been used to study measurement error in survey earnings reports.

Clearly, if we can assume that the administrative records do not contain any errors, then the validation study provides us with observations on  $\xi_{nt}$ , although generally only for a subset of individuals and time points. With observations on the dependent variable  $y_{nt}$  in the survey data, and assuming these are not subject to measurement error or the measurement error is independent of  $\xi_{nt}$ , we can estimate the model of interest directly and consistently without the need to use  $x_{nt}$ . This raises the question why one would want to use the survey measure  $x_{nt}$  at all. There may be several reasons: (1) the validation sample may be a relatively small subset of individuals; (2) the validation sample may be limited to only one, or a few, time points in a longer panel; (3) access to the validation sample may be highly restricted. In all these cases, the aim of the validation study is to learn something about the relation between  $x_{nt}$  and  $\xi_{nt}$  and condense this into a parsimonious model, and then superimpose this model on the model for the general survey data without recourse to the validation data. Note that this requires assumptions about the generalizability of the estimated relation to other individuals or other time points (or even other data sets) in order to obtain consistent estimators.

Assume that one is willing to make such assumptions. Then the measurement error variance can be estimated consistently by the sample variance of  $x_{nt} - \xi_{nt}$ . The estimate can then be plugged into the model of interest, and consistent estimators of the model parameters are straightforwardly obtained.

Validation studies also allow us to study nonclassical measurement error: from the validation study, we can estimate the whole conditional distribution of  $x_{nt}$  given  $\xi_{nt}$  and we can test whether the mean of this conditional distribution is equal to  $\xi_{nt}$ . We can study heteroskedasticity, non-linearity, and other aspects of this conditional distribution, and we can study heterogeneity in the conditional distribution (e.g., are individuals with lower cognitive ability more likely to have gross measurement errors).

Validation data can be used to consistently estimate models when there is measurement error (see, e.g., Chen, Hong, and Tamer, 2005), without making assumptions on, for example, the measurement error variance but only assuming that the relation between the observed and validated variables in the validated sample is the same as in the non-validated sample.

Conversely, we can also estimate the conditional distribution of  $\xi_{nt}$  given  $x_{nt}$ . This is often even more useful. In a very general sense, we are interested in characteristics of the conditional distribution of  $y_{nt}$  given  $\xi_{nt}$ . However, we observe the conditional distribution of  $y_{nt}$  given  $x_{nt}$ . This is a missing data problem, with  $\xi_{nt}$  the missing data. With the consistently estimated conditional distribution of  $\xi_{nt}$  given  $x_{nt}$  (and some additional assumptions about conditional independence), we can use a method for dealing with missing data and apply it to the problem at hand. A particularly straightforward way to do this would be to generate multiple imputations of  $\xi_{nt}$  from the estimated conditional distribution, estimate the model for  $y_{nt}$  given  $\xi_{nt}$  using these imputations, and combine the results using the standard “Rubin rules” (Rubin, 1987). This method is both flexible and powerful, and easy to apply. Because it only uses regression models for observed variables (the imputations act as observed variables), it is straightforward to incorporate non-linear models, heteroskedasticity, and other departures from the simple linear model. Brownstone and Valletta (1996) apply this method, except that they use it for a model with nonclassical measurement error in the dependent variable rather than the regressor.

Traditional validation methods require matching the survey information with administrative records. Because of confidentiality reasons, general reluctance among companies or government agencies to cooperate, monetary costs, and difficulties transforming extracts from administrative databases to an analytic file suitable for statistical analysis, validation studies are often small-scale. However, government agencies often routinely publish statistics, and sometimes they make de-identified extracts from their micro-data available for statistical analysis. Hu and Ridder (2012) show that this can also be used to identify measurement error models. A simple example of this follows from (11b): suppose that our sample is a representative sample of some population (say, individuals receiving Social Security benefits) and that a government agency publishes statistics or data from which we can derive or estimate  $\Sigma_{\xi,\tau\tau}$  (the variance of Social Security benefits amounts in year  $\tau$ ) for this population. Then a consistent estimator of  $\beta$  follows immediately from (11b). Hu and Ridder consider much more sophisticated variants of this, in which the marginal distribution of  $\xi_{nt}$  is estimated from one data source, and under some assumptions, this allows identification of the

joint distribution of  $y_{nt}$  and  $\xi_{nt}$  from survey data on  $(y_{nt}, x_{nt})$ , and thus a wide variety of non-linear and nonparametric regression models.

When the relationship between the error-free and error ridden regressors, or the marginal distribution of the error-free regressor, is estimated from one sample and the model of interest from another sample, the statistical uncertainty in the estimator of the former relation should be taken into account in the inference of the parameters of the model of interest. This is a data combination problem. Ridder and Moffitt (2007) discuss statistical inference using multiple data sources at length.

## REFERENCES

---

- Aasness, J., E. Biørn, and T. Skjerpen. "Engel functions, panel data, and latent variables," *Econometrica*, 61:1395–1422, 1993.
- Adda, J., and R. Cooper. *Dynamic economics: quantitative methods and applications*. MIT Press, Cambridge, MA, 2003.
- Alan, S., O. Attanasio, and M. Browning. "Estimating Euler equations with noisy data: two exact GMM estimators," *Journal of Applied Econometrics*, 24:309–324, 2009.
- Alan, S., and M. Browning. "Estimating intertemporal allocation parameters using synthetic residual estimation," *Review of Economic Studies*, 77:1231–1261, 2010.
- Anderson, T.W., and C. Hsiao. "Formulation and estimation of dynamic models using panel data," *Journal of Econometrics*, 18:47–82, 1982.
- Antman, F., and D. McKenzie. "Poverty traps and nonlinear income dynamics with measurement error and individual heterogeneity," *Journal of Development Studies*, 43:1057–1083, 2007.
- Arellano, M. "On testing of correlated effects with panel data," *Journal of Econometrics*, 59:87–97, 1993.
- Arellano, M. *Panel data econometrics*. Oxford University Press, Oxford, 2003.
- Arellano, M., and S. Bond. "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations," *Review of Economic Studies*, 58:277–297, 1991.
- Baltagi, B.H. *Econometric analysis of panel data*. Wiley, 4th edition, Chichester, 2008.
- Banks, J., R. Blundell, and A. Lewbel. "Quadratic Engel curves and consumer demand," *Review of Economics and Statistics*, 79:527–539, 1997.
- Bekker, P.A., A. Merckens, and T.J. Wansbeek. *Identification, equivalent models, and computer algebra*. Academic Press, Orlando, FL, 1994.
- Biørn, E. "Panel data with measurement errors: instrumental variables and GMM procedures combining levels and differences," *Econometric Reviews*, 19:391–424, 2000.
- Biørn, E. "Panel data dynamics with mis-measured variables: Modeling and GMM estimation," *Empirical Economics*, forthcoming, 2014.
- Biørn, E., and J.T. Klette. "Panel data with errors-in-variables: essential and redundant orthogonality conditions in GMM estimation," *Economics Letters*, 59:275–282, 1998.
- Bound, J., and A.B. Krueger. "The extent of measurement error in longitudinal earnings data: do two wrongs make a right?" *Journal of Labor Economics*, 9:1–24, 1991.

- Bound, J., C. Brown, and N. Mathiowetz. "Measurement error in survey data." In J.J. Heckman and E.E. Leamer, editors, *Handbook of econometrics* 5. North-Holland, Amsterdam, 2001, 3705–3843.
- Brownstone, D., and R.G. Valletta. "Modeling earnings measurement error: a multiple imputation approach," *Review of Economics and Statistics*, 78:705–717, 1996.
- Buonaccorsi, J., E. Demidenko, and T. Tosteson. "Estimation in longitudinal random effects models with measurement error," *Statistica Sinica*, 10:885–903, 2000.
- Cameron, C., and P.K. Trivedi. *Microeconometrics*. Cambridge University Press, Cambridge, UK, 2005.
- Carroll, C. "Death to the log-linearized consumption Euler equation! (And very poor health to the second-order approximation)" *Advances in Macroeconomics*, 1(1):Article 6, 2001.
- Chen, X., H. Hong, and D. Nekipelov. "Nonlinear models of measurement error," *Journal of Economic Literature*, 49:901–937, 2011.
- Chen, X., H. Hong, and E. Tamer. "Measurement error models with auxiliary data," *Review of Economic Studies*, 72:343–366, 2005.
- Christensen, B.J., and N.M. Kiefer. *Economic modeling and inference*. Princeton University Press, Princeton, NJ, 2009.
- Dagenais, M.G., and D.L. Dagenais. "Higher moment estimators for linear regression models with errors in the variables," *Journal of Econometrics*, 76:193–221, 1997.
- Dynan, K.E. "Habit formation in consumer preferences: evidence from panel data," *American Economic Review*, 90:391–406, 2000.
- Erickson, T., and T.M. Whited. "Two-step GMM estimation of the errors-in-variables model using high-order moments," *Econometric Theory*, 18:776–799, 2002.
- Fan, Z., B.C. Sutradhar, and R. Prabhakar Rao. "Bias corrected generalized method of moments and generalized quasi-likelihood inferences in linear models for panel data with measurement error," *Sankhyā B*, 74:126–148, 2012.
- Geary, R.C. "Inherent relations between random variables," *Proceedings of the Royal Irish Academy A*, 47:63–67, 1942.
- Goolsbee, A. "The importance of measurement error in the cost of capital," *National Tax Journal*, 53:215–228, 2000.
- Griliches, Z., and J.A. Hausman. "Errors in variables in panel data," *Journal of Econometrics*, 31:93–118, 1986.
- Hall, R.E. "Stochastic implications of the life cycle–permanent income hypothesis: theory and evidence," *Journal of Political Economy*, 86:971–987, 1978.
- Hansen, L.P., and K.J. Singleton. "Generalized instrumental variables estimation of nonlinear rational expectations models," *Econometrica*, 50:1269–1286, 1982.
- Hausman, J.A., W.K. Newey, and J.L. Powell. "Nonlinear errors in variables estimation of some Engel curves," *Journal of Econometrics*, 65:205–233, 1995.
- Newey, W.K., Hausman J.A., H. Ichimura, and J.L. Powell. "Identification and estimation for polynomial errors-in-variables models," *Journal of Econometrics*, 50:273–295, 1991.
- Horowitz, J.L. "The bootstrap." In J.J. Heckman and E.E. Leamer, editors, *Handbook of econometrics* 5. North-Holland, Amsterdam, 2001, 3159–3228.
- Hsiao, C. *Analysis of panel data*. Cambridge University Press, Cambridge, 2nd edition, 2003.
- Hsiao, C., and G. Taylor. "Some remarks on measurement errors and the identification of panel data models," *Statistica Neerlandica*, 45:187–194, 1991.

- Hu, Y., and G. Ridder. "Estimation of nonlinear models with mismeasured regressors using marginal information," *Journal of Applied Econometrics*, 27:347–385, 2012.
- Hyslop, D.R., and G.W. Imbens. "Bias from classical and other forms of measurement error," *Journal of Business & Economic Statistics*, 19:475–481, 2001.
- Jöreskog, K.G. "An econometric model for multivariate panel data," *Annales de l'INSEE*, 30-31:355–366, 1978.
- Kao, C., and J.F. Schnell. "Errors in variables in panel data with binary dependent variable," *Economics Letters*, 24:45–49, 1987a.
- Kao, C., and J.F. Schnell. "Errors in variables in a random effects probit model for panel data," *Economics Letters*, 24:339–342, 1987b.
- Kim, B., and G. Solon. "Implications of mean-reverting measurement error for longitudinal studies of wages and employment," *Review of Economics and Statistics*, 87:193–196, 2005.
- Klette, J.T. "Market power, scale economies and productivity: estimates from a panel of establishment data," *Journal of Industrial Economics*, 47:451–476, 1999.
- Lewbel, A. "Demand estimation with expenditure measurement errors on the left and right hand side," *Review of Economics and Statistics*, 78:718–725, 1996.
- Lewbel, A. "Constructing instruments for regressions with measurement error when no additional data are available, with an application to patents and R&D," *Econometrica*, 65:1201–1213, 1997.
- Lewbel, A. "Using heteroskedasticity to identify and estimate mismeasured and endogenous regressor models," *Journal of Business & Economic Statistics*, 30:67–80, 2012.
- Magnus, J.R., and H. Neudecker. "Symmetry, 0-1 matrices and Jacobians: a review," *Econometric Theory*, 2:157–190, 1986.
- Meijer, E. "Matrix algebra for higher order moments," *Linear Algebra and its Applications*, 410:112–134, 2005.
- Meijer, E., A. Kapteyn, and T. Andreyeva. "Internationally comparable health indices," *Health Economics*, 20:600–619, 2011.
- Meijer, E., L. Spierdijk, and T.J. Wansbeek. *Consistent estimation of linear panel data models with measurement error*. Working paper, University of Groningen, 2012.
- Meijer, E., L. Spierdijk, and T.J. Wansbeek. "Measurement error in the linear dynamic panel data model." In B.C. Sutradhar, editor, *Longitudinal data analysis subject to measurement error, missing values and/or outliers*. Springer, New York, 2013, 77–92.
- Pal, M. "Consistent moment estimators of regression coefficients in the presence of errors in variables," *Journal of Econometrics*, 14:349–364, 1980.
- Reiersøl, O. "Identifiability of a linear relation between variables which are subject to error," *Econometrica*, 18:375–389, 1950.
- Ridder, G., and R. Moffitt. "The econometrics of data combination." In J.J. Heckman and E.E. Leamer, editors, *Handbook of econometrics 6B*. Elsevier, Amsterdam, 2007, 5469–5547.
- Ronning, G., and H. Schneeweiss. "Panel regression with multiplicative measurement errors," *Economics Letters*, 110:136–139, 2011.
- Rubin, D.B. *Multiple imputation for nonresponse in surveys*. Wiley, New York, 1987.
- Rust, J. "Structural estimation of Markov decision processes." In R. F. Engle and D. L. McFadden, editors, *Handbook of econometrics 4*. North-Holland, Amsterdam, 3081–3143.
- Shao, J., Z. Xiao, and R. Xu. "Estimation with unbalanced panel data having covariate measurement error," *Journal of Statistical Planning and Inference*, 141:800–808, 2011.

- Stokey, N.L., and R.E. Lucas, Jr. (with E. C. Prescott). *Recursive methods in economic dynamics*. Harvard University Press, Cambridge, MA, 1989.
- van Montfort, K., A. Mooijaart, and J. de Leeuw. "Regression with errors in variables: Estimators based on third order moments," *Statistica Neerlandica*, 41:223–239, 1987.
- Wansbeek, T.J. "Permutation matrix - II," In S. Kotz and N.L. Johnson, editors, *Encyclopedia of statistical sciences, supplement volume*. Wiley, New York, 1989, 121–122.
- Wansbeek, T.J. "GMM estimation in panel data models with measurement error," *Journal of Econometrics*, 104:259–268, 2001.
- Wansbeek, T.J., and A. Kapteyn. "Simple estimators for dynamic panel data models with errors in the variables," In R. Bewley and Tran Van Hoa, editors, *Contributions to consumer demand and econometrics*. McMillan, London, 1992, 238–251.
- Wansbeek, T.J., and E. Meijer. *Measurement error and latent variables in econometrics*. North-Holland, Amsterdam, 2000.
- Wooldridge, J.M. *Econometric analysis of cross section and panel data*. The MIT Press, Cambridge, MA, 2010.
- Shao, J., Xiao, Z. and M. Palta. "Instrumental variable and GMM estimation for panel data with measurement error," *Statistica Sinica*, 20:1725–1747, 2010a.
- Shao, J., Xiao, Z. and M. Palta. "GMM in linear regression for longitudinal data with multiple covariates measured with error," *Journal of Applied Statistics*, 37:791–805, 2010b.
- Xiao, Z., J. Shao, R. Xu, and M. Palta. "Efficiency of GMM estimation in panel data models with measurement error," *Sankhyā*, 69:101–111, 2007.

## CHAPTER 12

---

# SPATIAL PANEL DATA MODELS

---

LUNG-FEI LEE AND JIHAI YU

### 12.1 INTRODUCTION

---

THE consideration of interactions among regions or agents has become increasingly important in various fields of economics. In public economics, a state government's cigarettes tax rate will be influenced by its neighboring states due to bootlegging. To keep a balanced budget, the state's cigarette tax rate cannot be too high relative to that of its neighbors. Also, public expenditure on education or health would also have spatial effects because of political concerns. In market integration, fluctuations of price level in one region will cause the comovement of prices across regions due to arbitrage, because high price areas would attract imports from low price areas. In growth theory, due to labor mobility and technological spillover, income levels of countries and states are spatially correlated. In labor economics, a student's performance in the class would be influenced by his/her peers. These are a few examples.

Recently there has been much interest in panel data models with spatial interactions, because they take into account the dynamic and spatial dependence and also control for unobservable heterogeneity. Some related recent surveys are Anselin, Le Gallo, and Jayet (2008), Elhorst (2010a), Baltagi (2011), and Lee and Yu (2010c, 2011a). With panel data available, we can not only have a larger sample size to increase the efficiency of estimates, but also investigate some issues that cannot be handled by cross-sectional data, such as heterogeneity and state dependence across time. For the static panel data with spatial interaction, Anselin (1988) provides a panel regression model with error components and spatial auto-regressive (SAR) disturbances. Baltagi, Song, and Koh (2003) consider specification tests for spatial correlation in a static spatial panel regression model using the Lagrange multiplier (LM) approach. With a different specification on error components and an SAR structure in the overall disturbance, Kapoor, Kelejian, and Prucha (2007) suggest a method of moments (MOM) estimation; and Fingleton (2008) adopts similar approach to estimate a spatial panel model

with a spatial lag (SL) in the regression and a spatial moving average (SMA) structure in the disturbance. Nesting the Anselin (1988) and Kapoor, Kelejian, and Prucha (2007) models, Baltagi, Egger, and Pfaffermayr (2007b) suggest an extended model without restrictions on implied SAR structures in the error component. As an alternative to the random effects specification, Lee and Yu (2010a) investigate the estimation of spatial panel models under fixed effects specification. The fixed effects model has the advantage of robustness in that fixed effects are allowed to depend on regressors in the model. It also provides a unified model framework because different random effects models in Anselin (1988), Kapoor, Kelejian, and Prucha (2007) and Baltagi, Egger, and Pfaffermayr (2007b) reduce to the same fixed effects model. Lee and Yu (2012a) investigate a more general model allowing for additional SMA structure and serial correlations, considering both fixed and random effects specifications. It is shown that the random effects estimators are a pooling of fixed effects (within) estimators and between equation estimators, similar to Maddala (1971), but in the spatial and serial correlation panel data context. The static panel data models can be applied to agricultural economics (Druska and Horrace 2004), transportation research (Frazier and Kockelman 2005; Parent and LeSage 2010), good demand (Baltagi and Li 2006), real estate economics (Holly, Pesaran, and Yamagata 2010), to name a few.

The static spatial panel models do not incorporate time lagged dependent variables in the regression equation. Instead, a spatial panel data model can have dynamic features. Anselin, Le Gallo, and Jayet (2008) and Anselin (2001) divide spatial dynamic models into four categories, namely, “pure space recursive” if only a spatial time lag is included; “time-space recursive” if an individual time lag and a spatial time lag are included; “time-space simultaneous” if an individual time lag and a contemporaneous SL term are specified; and “time-space dynamic” if all forms of lags are included. Korniotis (2010) studies a time-space recursive model with fixed effects, which is applied to the growth of consumption in each state in the United States to investigate habit formation. As a recursive model, the parameters, including the fixed effects, can be estimated by the ordinary least square (OLS) procedure. Elhorst (2005) estimates a dynamic panel data model with spatial errors by using an unconditional maximum likelihood method, and Murt (2006) investigates the model using a three-step generalized method of moments (GMM). Su and Yang (2007) derive the quasi-maximum likelihood (QML) estimation of the above model under both fixed and random effects specifications. For the general “time-space dynamic” model, we can term it the spatial dynamic panel data (SDPD) model to better link the terminology to the dynamic panel data literature (see, e.g., Hsiao 1986; Alvarez and Arellano 2003). For this general SDPD model, Yu, de Jong, and Lee (2008, 2012) and Yu and Lee (2010) study, respectively, the stable, spatial co-integration, and unit root models where the individual time lag, spatial time lag and contemporaneous SL are all included. For the dynamic models, they can be applied to the growth convergence of countries and regions (Baltagi, Bresson, and Pirotte 2007a; Ertur and Koch 2007; Yu and Lee 2012), regional markets (Keller and Shiue 2007), public economics (Wildasin 2003; Franzese 2007), and other fields.

This chapter is organized as follows. Section 12.2 introduces various spatial panel data models, both static and dynamic. Section 12.3 investigates estimation methods that have been applied to various spatial panel data models. Testing procedures to detect spatial effects are also reviewed. Section 12.4 lists some ongoing researches and interesting possible future research topics. Section 12.5 concludes.

## 12.2 MODEL SPECIFICATIONS

---

In this section, we discuss models which have been studied and used in theoretical and empirical studies. The models include both static and dynamic spatial panel data ones.

### 12.2.1 Static Panel Data Models

The spatial effect can be specified in the disturbances or the regression part, or both. For the spatial error component model, Anselin (1988) and Baltagi, Song, and Koh (2003) specify

$$Y_{nt} = X_{nt}\beta_0 + \mu_n + U_{nt} \text{ with } U_{nt} = \lambda_0 W_n U_{nt} + V_{nt}, \quad (1)$$

where  $Y_{nt} = (y_{1t}, y_{2t}, \dots, y_{nt})'$  and  $V_{nt} = (v_{1t}, v_{2t}, \dots, v_{nt})'$  are  $n \times 1$  (column) vectors,  $v_{it}$ 's are *i.i.d.*  $(0, \sigma_0^2)$  across  $i$  and  $t$ , and  $W_n$  is an  $n \times n$  spatial weights matrix, which is predetermined and generates the spatial dependence among cross-sectional units. Here,  $X_{nt}$  is an  $n \times k_x$  matrix of time-varying regressors,  $\mu_n$  is an  $n \times 1$  vector of individual random components, and the spatial correlation is in  $U_{nt}$ . Kapoor, Kelejian, and Prucha (2007) consider a different specification with

$$Y_{nt} = X_{nt}\beta_0 + U_{nt} \text{ with } U_{nt} = \lambda_0 W_n U_{nt} + \mu_n + V_{nt}, \quad (2)$$

where  $\mu_n$  is the vector of individual random components. These two models are different in the location of individual effects. Baltagi, Egger, and Pfaffermayr (2007b) formulate a general model which allows for spatial correlations in both individual and error components with different spatial parameters:

$$\begin{aligned} Y_{nt} &= X_{nt}\beta_0 + \mu_n + U_{nt}, \\ U_{nt} &= \lambda_{20} W_{n2} U_{nt} + V_{nt} \text{ and } \mu_n = \lambda_{30} W_{n3} \mu_n + c_{n0}. \end{aligned} \quad (3)$$

The SAR features in the individual effects  $\mu_n$  could be regarded as permanent spillover effects. The Anselin (1988) is a special case with  $\lambda_{30} = 0$ , and the Kapoor, Kelejian, and Prucha (2007) specification has  $W_{n2} = W_{n3}$  and  $\lambda_{20} = \lambda_{30}$ .

These panel models are different also in terms of the variance matrices of the overall disturbances. The variance matrices in Anselin (1988) and Baltagi, Song, and

Koh (2003) are more complicated, and their inverses are computationally demanding for a sample with a large  $n$ . For the model in Kapoor, Kelejian, and Prucha (2007), spatial correlations in both individual effects and error components have the same spatial effect parameter, which gives rise to a variance matrix having a special pattern so that its inverse can be easier to compute. Under fixed effects specification, these panel models are identical. Lee and Yu (2010a) consider a spatial panel data model with the SL term in regression equation under the fixed effects:

$$Y_{nt} = \lambda_{10} W_{n1} Y_{nt} + X_{nt} \beta_0 + \mu_n + U_{nt} \text{ with } U_{nt} = \lambda_{20} W_{n2} U_{nt} + V_{nt}. \quad (4)$$

We may also have time effects in spatial panel data models. In short panels, time effects will not cause an incidental parameter problem and they can be treated as regressors. However, for long panels, to avoid the incidental parameter problem, we may eliminate them before the estimation, or allow them in the regression with a random effects specification. Also, the exogenous variables  $X_{nt}$  can include spatial features such that the regressor function includes  $X_{nt}$  and  $W_n X_{nt}$ , where  $W_n X_{nt}$  can capture the so-called spatial Durbin regressors (LeSage and Pace 2009). While the spatial Durbin regressors may be of interest in empirical applications, in general, it does not introduce additional complication in theoretical analysis, but might cause near-multicollinearity in some situations if  $X_{nt}$  and  $W_n X_{nt}$  were highly correlated.

These models specify the spatial error components in the SAR form. An alternative specification is the SMA disturbances. Anselin, Le Gallo, and Jayet (2008) introduce a spatial panel data with random effects and SMA disturbances,  $Y_{nt} = X_{nt} \beta_0 + \mu_n + U_{nt}$  with  $U_{nt} = (I_n + \lambda_{20} W_{n2}) V_{nt}$ . The different economic implications of SAR vs. SMA disturbances in terms of global and local interactions are also discussed in Anselin, Le Gallo, and Jayet (2008). Fingleton (2008) specifies a different SMA process due to the location of individual effects, and additionally includes an SL for the dependent variable:

$$Y_{nt} = \lambda_{10} W_n Y_{nt} + X_{nt} \beta_0 + U_{nt}, \text{ with } U_{nt} = (I_n + \lambda_{20} W_{n2}) \xi_{nt} \text{ and } \xi_{nt} = c_{n0} + V_{nt}.$$

Fingleton (2008) proposes GMM estimation, where bootstrap method is suggested as a way of testing the significance of the moving average parameter.

There is also consideration of additional serial correlation as in Baltagi, Song, Jung, and Koh (2007c) and Elhorst (2008):

$$Y_{nt} = X_{nt} \beta_0 + \mu_n + U_{nt}, \text{ with } U_{nt} = \lambda_{20} W_{n2} U_{nt} + V_{nt} \text{ and } V_{nt} = \rho_0 V_{n,t-1} + e_{nt}.$$

By including an SL term in regression equation, Lee and Yu (2012a) provide a general model which allows for different situations:

$$Y_{nt} = \lambda_{10} W_{n1} Y_{nt} + z_n b_0 + X_{nt} \beta_0 + \mu_n + U_{nt}, \quad (5)$$

$$U_{nt} = \lambda_{20} W_{n2} U_{nt} + (I_n + \delta_{20} M_{n2}) V_{nt} \text{ with } V_{nt} = \rho_0 V_{n,t-1} + e_{nt} \text{ for } t = 2, \dots, T,$$

$$\mu_n = \lambda_{30} W_{n3} \mu_n + (I_n + \delta_{30} M_{n3}) c_{n0},$$

where  $z_n$  is an  $n \times k_z$  matrix that captures nonstochastic time invariant regressors including the constant intercept. The  $W_{nj}$  and  $M_{nj}$  are  $n \times n$  nonstochastic spatial weights matrices that generate spatial dependences,  $\mu_n$  is an  $n$ -dimensional vector of individual effects with spatial interactions,  $U_{nt}$  is the SAR-SMA error which is also serially correlated, and  $c_{n0}$  and  $e_{nt}$  are independent with *i.i.d.* elements. The  $U_{nt}$  and  $\mu_n$  are allowed to incorporate the SMA feature. The mixing of both SAR and SMA operators in an equation is meaningful only if they do not cancel each other under  $M_n = W_n$ . This is a generalized spatial panel model which incorporates spatial correlation, heterogeneity, and serial correlation in disturbances. It can nest various existing spatial panels as mentioned earlier. We can summarize these models in the following table.

We can consider the estimation under both fixed effects and random effects specifications on  $\mu_n$ . In Lee and Yu (2012a)'s model, the parameter subvector  $\theta_{10} = (\beta'_0, \lambda_{10}, \lambda_{20}, \delta_{20}, \rho_0, \sigma_{e0}^2)'$  can be estimated from both fixed and random effects models. The remaining parameters in  $\theta_{20} = (b'_0, \lambda_{30}, \delta_{30}, \sigma_{c0}^2)'$  can only be estimated under the random effects specification. The random effects specification of  $\mu_n$  in (3) can be assumed to be an SAR process as in Baltagi, Egger, and Pfaffermayr (2007b) and Lee and Yu (2012a). If the process of  $\mu_n$  in (3) is correctly specified, estimates of the parameters can be more efficient than those of the fixed effects specification, as they utilize the variation of elements of  $\mu_n$  across spatial units. On the other hand, the fixed effects specification is known to be robust against possible correlation of  $\mu_n$  with included regressors in the model. The fixed effects specification can also be robust against the spatial specification of  $\mu_n$ . With the fixed effects specification, various random effects panel models have the same representation.

**Table 12.1 Summary of different static spatial panel models**

	Spatial Disturbance		
	SAR	SMA	both SAR and SMA
Seriously uncorrelated ( $\rho = 0$ )			
without spatial lag ( $\lambda = 0$ )	Kapoor, Kelejian, and Prucha (2007) Anselin (1988) Baltagi, Egger and Pfaffermayr (2007b)	Anselin, Le Gallo, and Jayet (2008)	
with spatial lag ( $\lambda \neq 0$ )	Lee and Yu (2010a)	Fingleton (2008)	
Seriously correlated ( $\rho \neq 0$ )			
without spatial lag ( $\lambda = 0$ )	Baltagi et al. (2007c) Elhorst (2008)		
with spatial lag ( $\lambda \neq 0$ )			Lee and Yu (2012a)

Parent and LeSage (2012) apply the Markov Chain Monte Carlo method to a linear panel regression model where spatially and serially correlated disturbances are present. The product of the quasi-difference over time and the spatial transformation is called the space-time filter in Parent and LeSage (2012):

$$U_{nt} = \lambda_0 W_n U_{nt} + e_{nt} \text{ with } e_{nt} = \gamma_0 e_{n,t-1} + V_{nt}, \quad (6)$$

which is a separable space-time filter. The  $U_{nt}$  in (6) is an SAR process with its noise term  $e_{nt}$  being a vector auto-regressive process of order 1 (VAR(1)). In place of (6), we also have the representation that

$$U_{nt} = \gamma_0 U_{n,t-1} + e_{nt} \text{ with } e_{nt} = \lambda_0 W_n e_{nt} + V_{nt}, \quad (7)$$

where  $U_{nt}$  is a VAR(1) process with its noise vector  $e_{nt}$  being an SAR process. Define  $L_n$  as the time shift operator such that  $L_n e_{nt} = e_{n,t-1}$ . The  $I_n - \gamma_0 L_n$  is a time filter and  $I_n - \lambda_0 W_n$  is a spatial filter. With these filters, we have for (6)  $(I_n - \gamma_0 L_n)(I_n - \lambda_0 W_n)U_{nt} = V_{nt}$ , and for (7),  $(I_n - \lambda_0 W_n)(I_n - \gamma_0 L_n)U_{nt} = V_{nt}$ . The spatial and time filters are separated and apparently commutative.

A more general specification has

$$U_{nt} = \lambda_0 W_n U_{nt} + \gamma_0 U_{n,t-1} + \rho_0 W_n U_{n,t-1} + V_{nt}. \quad (8)$$

Parent and LeSage (2012)'s model is a special case of the general one as it has imposed the restriction  $\rho_0 = -\lambda_0 \gamma_0$ . For the general spatial and time dynamic disturbances  $U_{nt}$  in (8), it has the general spatial-time filter  $(I_n - \lambda_0 W_n) - (\gamma_0 I_n + \rho_0 W_n)L_n$ . When  $\rho_0 = -\gamma_0 \lambda_0$ , this spatial-time filter can be decomposed into a product of the spatial filter  $(I_n - \lambda_0 W_n)$  and the time filter  $(I_n - \gamma_0 L_n)$ . Parent and LeSage (2012) argue that a Bayesian estimation of the model (8) with a nonseparable space-time filter is too complicated to be computationally attractive and recommend the use of the separable space-time filter in practice. However, the separable space-time filter does not accommodate many important features of a general space-time filter which provides diffusion of spatial interactions over time, and in particular, rules out the possibility of spatial cointegration in Yu, de Jong, and Lee (2012).

The parameter space for the process in (8) is an interesting issue, which will be discussed in the next section under the spatial dynamic panel data setting.

### 12.2.2 Dynamic Panel Data Models

Dynamic panel data models consider not only heterogeneity but also state dependence that cannot be handled by cross-sectional data or static panel data models. From Anselin, Le Gallo, and Jayet (2008) and Anselin (2001), the most general case for a dynamic panel data with spatial effect is the “time-space dynamic” model, which is termed spatial dynamic panel data (SDPD) model in Yu, de Jong, and Lee (2008). This general model can be specified as

$$Y_{nt} = \lambda_0 W_n Y_{nt} + \gamma_0 Y_{n,t-1} + \rho_0 W_n Y_{n,t-1} + X_{nt} \beta_0 + c_{n0} + \alpha_{t0} l_n + V_{nt}, \quad (9)$$

where  $c_{n0}$  is  $n \times 1$  column vector of fixed effects and  $\alpha_{t0}$ 's are time effects. Compared to the static model, we have included the dynamic terms  $Y_{n,t-1}$  and  $W_n Y_{n,t-1}$ . Here,  $\gamma_0$  captures the pure dynamic effect and  $\rho_0$  captures the spatial-time effect which is also called a diffusion. Due to the presence of fixed individual and time effects,  $X_{nt}$  will not include any time invariant or individual invariant regressors.

To investigate the dynamics of this general model, we can investigate its reduced form. Define  $S_n(\lambda) = I_n - \lambda W_n$  and  $S_n \equiv S_n(\lambda_0) = I_n - \lambda_0 W_n$ . Then, presuming that  $S_n$  is invertible and denoting  $A_n = S_n^{-1}(\gamma_0 I_n + \rho_0 W_n)$ , (9) can be rewritten as

$$Y_{nt} = A_n Y_{n,t-1} + S_n^{-1} X_{nt} \beta_0 + S_n^{-1} c_{n0} + \alpha_{t0} S_n^{-1} l_n + S_n^{-1} V_{nt}. \quad (10)$$

We can study the eigenvalues of  $A_n$  by focusing on the practical case with  $W_n$  being row-normalized. Let  $\varpi_n = \text{diag}\{\varpi_{n1}, \varpi_{n2}, \dots, \varpi_{nn}\}$  be the  $n \times n$  diagonal eigenvalues matrix of  $W_n$  such that  $W_n = \Gamma_n \varpi_n \Gamma_n^{-1}$  where  $\Gamma_n$  is the corresponding eigenvector matrix. Because  $A_n = S_n^{-1}(\gamma_0 I_n + \rho_0 W_n)$ , the eigenvalues matrix of  $A_n$  is  $D_n = (I_n - \lambda_0 \varpi_n)^{-1}(\gamma_0 I_n + \rho_0 \varpi_n)$  such that  $A_n = \Gamma_n D_n \Gamma_n^{-1}$ . As  $W_n$  is row-normalized, all the eigenvalues are less than or equal to 1 in absolute value, where it has definitely some eigenvalues being 1. Denote  $m_n$  as the number of unit eigenvalues of  $W_n$  and let the first  $m_n$  eigenvalues of  $W_n$  be the unity. Hence,  $D_n$  can be decomposed into two parts, one corresponding to the unit eigenvalues of  $W_n$ , and the other corresponding to the eigenvalues of  $W_n$  which are smaller than 1. Define  $\mathbb{J}_n = \text{diag}\{l'_{m_n}, 0, \dots, 0\}$  with  $l_{m_n}$  being an  $m_n \times 1$  vector of ones and  $\tilde{D}_n = \text{diag}\{0, \dots, 0, d_{n,m_n+1}, \dots, d_{nn}\}$ , where  $|d_{ni}| < 1$ , for  $i = m_n + 1, \dots, n$ , are assumed. We have  $A_n^h = (\frac{\gamma_0 + \rho_0}{1 - \lambda_0})^h \Gamma_n \mathbb{J}_n \Gamma_n^{-1} + B_n^h$  where  $B_n = \Gamma_n \tilde{D}_n \Gamma_n^{-1}$  for any  $h = 1, 2, \dots$ .

Denote  $W_n^u = \Gamma_n \mathbb{J}_n \Gamma_n^{-1}$ . Then, as derived in Lee and Yu (2011b), for  $t \geq 0$ ,  $Y_{nt}$  can be decomposed into a sum of a possible stable part, a possible unstable or explosive part, and a time effect part:

$$Y_{nt} = Y_{nt}^u + Y_{nt}^s + Y_{nt}^\alpha, \quad (11)$$

with

$$\begin{aligned} Y_{nt}^s &= \sum_{h=0}^{\infty} B_n^h S_n^{-1} (c_{n0} + X_{n,t-h} \beta_0 + V_{n,t-h}), \\ Y_{nt}^u &= W_n^u \left\{ \left( \frac{\gamma_0 + \rho_0}{1 - \lambda_0} \right)^{t+1} Y_{n,-1} \right. \\ &\quad \left. + \frac{1}{(1 - \lambda_0)} \left[ \sum_{h=0}^t \left( \frac{\gamma_0 + \rho_0}{1 - \lambda_0} \right)^h (c_{n0} + X_{n,t-h} \beta_0 + V_{n,t-h}) \right] \right\}, \end{aligned}$$

$$Y_{nt}^\alpha = \frac{1}{(1-\lambda_0)} l_n \sum_{h=0}^t \alpha_{t-h,0} \left( \frac{\gamma_0 + \rho_0}{1-\lambda_0} \right)^h.$$

As the absolute values of the elements in  $\tilde{D}_n$  are less than 1,  $Y_{nt}^s$  is a stable component. The  $Y_{nt}^u$  can be an unstable component when  $\frac{\gamma_0 + \rho_0}{1-\lambda_0} = 1$ , which occurs when  $\gamma_0 + \rho_0 + \lambda_0 = 1$  and  $\lambda_0 \neq 1$ . If  $c_{n0}$  and/or the time mean of  $X_{nt}\beta_0$  are nonzero, the term  $\sum_{h=0}^t (c_{n0} + X_{n,t-h}\beta_0)$  will generate a time trend. The  $\sum_{h=0}^t V_{n,t-h}$  will generate a stochastic trend. These imply the instability of  $Y_{nt}^u$ . When  $\gamma_0 + \rho_0 + \lambda_0 > 1$ , it implies  $\frac{\gamma_0 + \rho_0}{1-\lambda_0} > 1$  and, hence,  $Y_{nt}^u$  can be explosive. The component  $Y_{nt}^\alpha$  captures the time effect due to the time dummies. The  $Y_{nt}^\alpha$  can be rather complicated as it depends on what exactly the time dummies represent. The  $Y_{nt}$  can be explosive when  $\alpha_{t0}$  represents some explosive functions of  $t$ , even when  $\frac{\gamma_0 + \rho_0}{1-\lambda_0}$  were smaller than 1.

The unit roots case has all eigenvalues of  $A_n$  being 1. It occurs when  $\gamma_0 + \rho_0 + \lambda_0 = 1$  and  $\gamma_0 = 1$ , because  $A_n = I_n$ . For this unit roots case, the unit eigenvalues of  $A_n$  are not linked to the eigenvalues of  $W_n$ . Because  $W_n^u$  is defined completely from  $W_n$ , the decomposition in (11) is not revealing for the unit roots case; but one simply has

$$Y_{nt} = Y_{n,t-1} + S_n^{-1}(X_{nt}\beta_0 + c_{n0} + \alpha_{t0}l_n + U_{nt}). \quad (12)$$

Hence, depending on the value of  $\frac{\gamma_0 + \rho_0}{1-\lambda_0}$ , we may broadly divide the process into four cases.<sup>1</sup>

- Stable case when  $\gamma_0 + \rho_0 + \lambda_0 < 1$  (and with some other proper restrictions on the three parameters).
- Spatial cointegration case when  $\gamma_0 + \rho_0 + \lambda_0 = 1$  but  $\gamma_0 < 1$ .
- Unit roots case when  $\gamma_0 + \rho_0 + \lambda_0 = 1$  and  $\gamma_0 = 1$ .
- Explosive case when  $\gamma_0 + \rho_0 + \lambda_0 > 1$ .

In the following, we discuss the parameter spaces for different SDPD models, which are also relevant for static spatial panel data models with a space-time filter in disturbances. For the possibility of stability, or in terms of stationarity in the time series notion, there are more restrictions on the parameter space of  $(\gamma_0, \rho_0, \lambda_0)$  in addition to  $\gamma_0 + \rho_0 + \lambda_0 < 1$ . Such a parameter region can be revealed from conditions such that all the eigenvalues  $d_{ni}$ 's of  $A_n$  are less than one in absolute value. The eigenvalues of  $A_n$  are  $d_{ni} = \frac{\gamma_0 + \rho_0 \varpi_{ni}}{1 - \lambda_0 \varpi_{ni}}$ , where  $\varpi_{ni}$ 's are eigenvalues of  $W_n$ . By regarding  $d$  as a function of  $\varpi$ , we have  $\frac{\partial}{\partial \varpi} \left( \frac{\gamma_0 + \rho_0 \varpi}{1 - \lambda_0 \varpi} \right) = \frac{\rho_0 + \lambda_0 \gamma_0}{(1 - \lambda_0 \varpi)^2}$ . Thus, we have three different situations:

- (i)  $\rho_0 + \lambda_0 \gamma_0 > 0$  if and only if  $d_{ni}$  has the same increasing order as  $\varpi_{ni}$ .
- (ii)  $\rho_0 + \lambda_0 \gamma_0 = 0$ , i.e., separable space-time filter, if and only if  $d_{ni}$  is a constant (under this case,  $d_{ni} = \gamma_0$ ).
- (iii)  $\rho_0 + \lambda_0 \gamma_0 < 0$  if and only if  $d_{ni}$  has the decreasing order of  $\varpi_{ni}$ .

### Stationarity

The stationarity (stable) case of  $A_n$  is the one with all eigenvalues of  $A_n$  less than 1 in absolute value, i.e.,  $|d_{ni}| < 1$  for all  $i = 1, \dots, n$ . Let  $d_{n,\min}$  and  $d_{n,\max}$  be, respectively, the smallest and largest eigenvalues of  $A_n$ ;  $\varpi_{n,\min}$  and  $\varpi_{n,\max}$  be the corresponding ones of  $W_n$ . By considering  $-1 < d_{n,\min}$  and  $d_{n,\max} < -1$  under different situations about the sign of  $\rho + \lambda\gamma$ , we can figure out the corresponding ranges of values of  $(\lambda, \gamma, \rho)$  for stationarity. Combining the basic restrictions that  $\frac{1}{\varpi_{n,\min}} < \lambda_0 < 1$  ( $= \varpi_{n,\max}$  with  $W_n$  being row-normalized) for the stable spatial filter  $I_n - \lambda_0 W_n$ , we have the following parameter space for stationarity:

$$\begin{aligned} R_s = \{(\lambda, \gamma, \rho) : \gamma + (\rho - \lambda)\varpi_{n,\min} > -1, \gamma + \lambda + \rho < 1, \gamma + \rho - \lambda > -1, \\ \gamma + (\lambda + \rho)\varpi_{n,\min} < 1\}. \end{aligned} \quad (13)$$

These four linear planes in (13) may interact with each other. Thus, the stationary region is convex with the vertices where some three planes intersect. There is a total of four vertices which are  $(1, -1, 1)$ ,  $(\frac{1}{\varpi_{n,\min}}, 1, -\frac{1}{\varpi_{n,\min}})$ ,  $(\frac{1}{\varpi_{n,\min}}, -1, \frac{1}{\varpi_{n,\min}})$  and  $(1, 1, -1)$ .

The above stationary regions provide also stationary regions for models with two effects. For the model without spatial time lag, i.e.,  $\rho_0 = 0$ , the range is the interior of the convex set determined by four linear edges with vertices of  $(\lambda, \gamma)$  being  $\{(1, 0), (0, 1), (\frac{1}{\varpi_{n,\min}}, 0), (0, -1)\}$ , which is illustrated in Elhorst (2008). For the model with both contemporaneous and time SLs but no time lag, i.e.,  $\gamma_0 = 0$ , the stationary region on the  $(\lambda, \rho)$  plane is the interior of the convex set with vertices of  $(\lambda, \rho)$  being  $\{(1, 0), (\frac{1}{2}(1 + \frac{1}{\varpi_{n,\min}}), \frac{1}{2}(1 - \frac{1}{\varpi_{n,\min}})), (\frac{1}{\varpi_{n,\min}}, 0), (\frac{1}{2}(1 + \frac{1}{\varpi_{n,\min}}), -\frac{1}{2}(1 - \frac{1}{\varpi_{n,\min}}))\}$ . For the model with time and spatial time lag but no contemporaneous SL, i.e.,  $\lambda_0 = 0$ , the stationary region is the interior of the convex set with the vertices  $\{(1, 0), (-\frac{1+\varpi_{n,\min}}{1-\varpi_{n,\min}}, \frac{2}{(1-\varpi_{n,\min})}), (-1, 0), (\frac{1+\varpi_{n,\min}}{1-\varpi_{n,\min}}, -\frac{2}{(1-\varpi_{n,\min})})\}$ . The following figures illustrate the parameter spaces implied by (13) and the three special cases with two effects (in the figures, we specify  $\varpi_{n,\min} = -0.5$  for an illustrative purpose).

Elhorst (2012) has provided the following characterization of the stationarity on parameters: (a)  $\gamma < 1 - (\lambda + \rho)$  if  $\lambda + \rho \geq 0$ ; (b)  $\gamma < 1 - (\lambda + \rho)\varpi_{n,\min}$  if  $\lambda + \rho < 0$ ; (c)  $-1 + (\lambda - \rho) < \gamma$  if  $\lambda - \rho \geq 0$ ; and (d)  $-1 + (\lambda - \rho)\varpi_{n,\min} < \gamma$  if  $\lambda - \rho < 0$ . His characterization seems different, but the overall region implied by his characterization is the same as we have described.

### Spatial cointegration

We define spatial cointegration for the situation with  $\lambda_0 + \gamma_0 + \rho_0 = 1$ , which has a revealing error correction model (ECM) representation. Denote  $\Delta Y_{nt} = Y_{nt} - Y_{n,t-1}$  as a difference in time. From (9) and (10), we have the ECM representation

$$\Delta Y_{nt} = S_n^{-1}[(\rho_0 + \lambda_{10})W_n - (1 - \gamma_0)I_n]Y_{n,t-1} + S_n^{-1}(X_{nt}\beta_0 + c_{n0} + \alpha_{t0}l_n + U_{nt}).$$

For the spatial cointegration case,

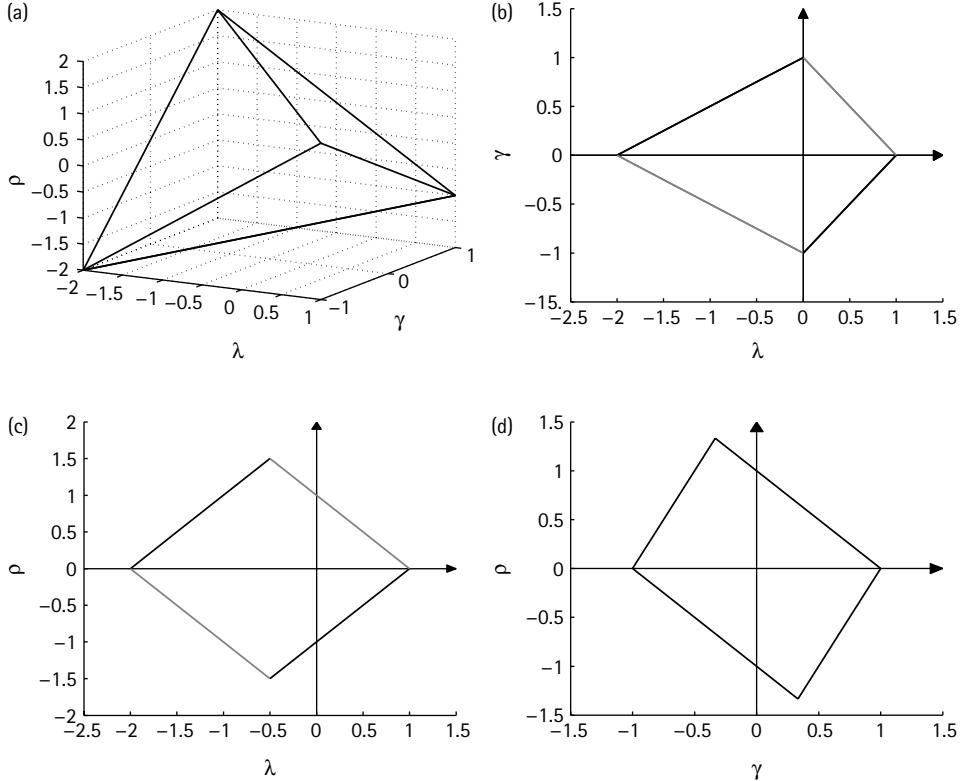
$$\Delta Y_{nt} = (1 - \gamma_0)S_n^{-1}(W_n - I_n)Y_{n,t-1} + S_n^{-1}(X_{nt}\beta_0 + c_{n0} + \alpha_{t0}l_n + U_{nt}).$$

Because  $W_n W_n^u = W_n^u$  and  $W_n l_n = l_n$ , the components  $Y_{nt}^u$  and  $Y_{nt}^\alpha$  have the identities  $W_n Y_{nt}^u = Y_{nt}^u$ ,  $W_n Y_{nt}^\alpha = Y_{nt}^\alpha$ . In this situation,  $(I_n - W_n)$  is a cointegrating matrix for  $Y_{nt}$  because  $(I_n - W_n)Y_{nt} = (I_n - W_n)Y_{nt}^s$ , which depends only on the stable component. Thus,  $Y_{nt}$  is spatially cointegrated. The cointegration rank of  $I_n - W_n = \Gamma_n(I_n - \varpi_n)\Gamma_n^{-1}$  is equal to  $n - m_n$ , which is the number of eigenvalues of  $W_n$  different from 1.

For our estimation theory, we require also that all the eigenvalues  $d_{ni}$  of  $A_n$  which correspond to eigenvalues  $\varpi_{ni}$  with  $|\varpi_{ni}| < 1$  to have the property  $|d_{ni}| < 1$ . This makes sure that for all eigenvalues of  $A_n$  except 1 are all less than 1 in absolute value. The corresponding parameter range turns out to be

$$\left\{ (\lambda, \gamma, \rho) : \frac{1}{\varpi_{n,\min}} < \lambda < 1, \gamma < 1, \rho < 1 + (1 - \lambda) \left( \frac{1 + \varpi_{n,\min}}{1 - \varpi_{n,\min}} \right) \right\} \quad (14)$$

for the spatial cointegration case. Because  $|\varpi_{n,\min}| < 1$ , this does not seem to have imposed much restriction.



### Explosive case

This is the case  $\lambda_0 + \gamma_0 + \rho_0 = 1 + a$ , where  $a > 0$  and  $|\gamma_0| < 1$  is assumed. The explosive feature refers to cases with mixed spatial and time effects together so that the total

value of spatial effects  $\lambda + \rho$  plays an role. To satisfy the condition for the eigenvalues  $|d_{ni}| < 1$  whenever  $|\varpi_{ni}| < 1$ , the corresponding parameter region will depend on the excess level  $a$  of explosiveness. The conditions for the explosive case will be satisfied by the parameters in the set

$$\left\{ (\lambda, \gamma, \rho) : \frac{1}{\varpi_{n,\min}} < \lambda < 1, |\gamma| < 1, \rho < \frac{a}{(1 - \varpi_{n,\min})} \right. \\ \left. + 1 + (1 - \lambda) \frac{(1 + \varpi_{n,\min})}{(1 - \varpi_{n,\min})}, \frac{a}{(1 - \varpi_{n,(2)})} < (\rho + \lambda) \right\}, \quad (15)$$

where  $\varpi_{n,(2)}$  refers to the largest eigenvalue less than one of  $W_n$ . Thus, the effective constraint is on the total spatial value of  $\rho + \lambda$  to be positive and not too small to generate this explosive phenomenon.

### *Unit root case*

When  $\lambda_0 + \gamma_0 + \rho_0 = 1$  and  $\gamma_0 = 1$ , the eigenvalues of  $A_n$  do not depend on the eigenvalues of  $W_n$ . Under this situation, the parameter range is

$$\left\{ (\lambda, \gamma, \rho) : \frac{1}{\varpi_{n,\min}} < \lambda < 1, \gamma = 1, \lambda + \rho = 0 \right\}. \quad (16)$$

Equation (10) expresses the model in terms of a space-time multiplier (Anselin, Le Gallo, and Jayet 2008), which specifies how the joint determination of the dependent variables is a function of both spatial and time lags of explanatory variables and disturbances of all spatial units. This equation is useful for calculating marginal effects of changes of exogenous variables on outcomes over time and across spatial units. Empirical researchers find the marginal effect calculations useful. They explore the economic implications of spatial multiplier and spatial spillover effects in addition to the coefficients in a spatial model (Pace and LeSage 2009). LeSage and Pace (2009) have introduced the concept of direct impact, total impact, and indirect impact, for spatial data. In a typical SAR model  $Y_n = \alpha_0 l_n + \lambda_0 W_n Y_n + \sum_{k=1}^{k_x} \beta_{k0} X_{nk} + \epsilon_n$ , for the impact of a regressor  $X_{nk}$  on  $Y_n$ , where  $W_n$  does not depend on  $X_n$ , the impact matrix is  $\frac{\partial Y_n}{\partial X_{nk}} = (I_n - \lambda_0 W_n)^{-1} \beta_{k0}$  for the  $k$ th regressor. The average direct impact, average total impact and average indirect impact are defined as, respectively,  $f_{k,direct}(\theta_0) \equiv \frac{1}{n} tr((I_n - \lambda_0 W_n)^{-1} \beta_{k0})$ ,  $f_{k,total}(\theta_0) \equiv \frac{1}{n} l_n' [(I_n - \lambda_0 W_n)^{-1} \beta_{k0}] l_n$ , and  $f_{k,indirect}(\theta_0) \equiv f_{k,total}(\theta_0) - f_{k,direct}(\theta_0)$ , with  $l_n$  being a  $n$ -dimensional column of ones. Debarsy, Ertur and LeSage (2012) extend such impact analyses to spatial dynamic panel models to investigate diffusion effects over time. Lee and Yu (2012b) further extend the impact analysis to the situation with time-varying spatial weights. Elhorst (2012) points out various restrictions on spatial dynamic panel models with marginal effects implied by specified models.

For an impact analysis for spatial dynamic panel data model, the change in exogenous variables will only influence the dependent variable of current period, but not the future ones. By changing the value of a regressor by the same amount across all

spatial units in some consecutive time periods, say, from the time period  $t_1$  to  $t_2$ , where  $t_1 \leq t_2 \leq t$ . Thus, we have  $\frac{\partial E(Y_{nt})}{\partial x} = \beta_0 \sum_{h=t-t_2}^{t-t_1} A_n^h S_n^{-1} l_n$ , where  $x$  is a regressor with its coefficient being  $\beta_0$ . Hence, by denoting  $\theta = (\lambda, \gamma, \rho, \beta')$ , the average total impact  $f_{t,tot}(t_0) \equiv \frac{1}{n} l_n' \frac{\partial E(Y_{nt})}{\partial x}$  is  $\sum_{h=t-t_2}^{t-t_1} \frac{1}{n} [\beta_0 l_n' A_n^h S_n^{-1} l_n]$  and similarly for other impacts. Debarsy, Ertur and LeSage (2012) study the case of  $t_2 = t$  so that  $f_{t,tot}(t_0)$  is written as  $\sum_{h=0}^{t-t_1} \frac{1}{n} [\beta_0 l_n' A_n^h S_n^{-1} l_n]$  and the object of interest is how a permanent change in  $X_{n,t_1}$  will affect the future horizons (accumulatively) until  $t$ .

### 12.2.3 Some Other Models with Cross-Sectional Dependence

Introducing factors into a model is a way to specify possible cross-sectional dependence in time series for macroeconomics. It also has wide applications in finance and real estate economics, e.g., Holly, Pesaran, and Yamagata (2010, 2011). Chudik, Pesaran and Tosetti (2011) introduce the concepts of time-specific weak and strong cross-section dependence in panel data models. In a panel setting, a factor model allows time effects to interact with spatial units with different intensity.

Pesaran and Tosetti (2011) investigate the following model with both common factors and spatial correlation:

$$y_{it} = \alpha'_i \mathbf{d}_t + \beta'_i \mathbf{x}_{it} + \gamma'_i \mathbf{f}_t + e_{it}, \quad (17)$$

where  $\mathbf{d}_t$  is a  $k \times 1$  vector of observed common effects,  $\mathbf{x}_{it}$  is the  $k_x \times 1$  vector of observed individual specific regressors on the  $i$ th cross-section unit at time  $t$ ,  $\mathbf{f}_t$  is an  $m$ -dimensional vector of unobservable common factors, and  $\gamma_i$  is the associated  $m \times 1$  vector of factor loadings. The common factors  $\mathbf{f}_t$  simultaneously affect all cross-section units, albeit with different intensity as measured by  $\gamma_i$ . The  $e_{it}$  follows some spatial process, either an SAR or SMA process.

To estimate the parameters of interest (the mean of  $\beta_i$ ), we can use the mean group estimator  $\hat{\beta}_{MG} = \frac{1}{n} \sum_{i=1}^n \hat{\beta}_i$ , where  $\hat{\beta}_i$  is the OLS estimate of  $y_i$  regressed on  $\mathbf{X}_i$  after the orthogonal projection of data by a  $T^*(k + k_{x+1})$  matrix  $\mathbf{M}_D$  where  $D$  consists of observed common factor and cross sectional average of dependent and independent variables, e.g.,  $\hat{\beta}_i = (\mathbf{X}'_t \mathbf{M}_D \mathbf{X}_i)^{-1} \mathbf{X}'_t \mathbf{M}_D y_i$  where  $\mathbf{X}_i$  is a  $T \times k_x$  vector of regressors for the  $i$ th unit,  $y_i$  is a  $T \times 1$  vector of the dependent variable for the  $i$ th unit, and  $\mathbf{M}_D = \mathbf{I}_T - D(D'D)^{-1}D'$ . Alternatively, the regression coefficient  $\beta$  can be obtained by a pooled estimate such as  $\hat{\beta}_P = (\sum_{i=1}^n \mathbf{X}'_t \mathbf{M}_D \mathbf{X}_i)^{-1} \sum_{i=1}^n \mathbf{X}'_t \mathbf{M}_D y_i$ . Given the spatial structure in  $e_{it}$ , Pesaran and Tosetti (2011) derive the asymptotic properties of  $\hat{\beta}_{MG}$  and  $\hat{\beta}_P$  under difference scenarios such as whether the unobserved common factor is present or not and whether the regression parameters are heterogeneous or not. Pesaran and Tosetti (2011) then consider the Common Correlated Effects (CCE) estimator advanced by Pesaran (2006), which continues to yield estimates of the slope coefficients that are consistent and asymptotically normal.

Holly, Pesaran, and Yamagata (2010) use (17) to analyze the changes in real house prices in the United States using state-level data. In their model, in addition to the common factor  $f_t$  in (17), they have specified a spatial process in  $e_{it}$ , which is shown to be significant after controlling for those unobserved common factors.

In (17), the number of factors is fixed. Chudik, Pesaran, and Tosetti (2011) introduce the concepts of time-specific weak and strong cross-section dependence, and extend the model by allowing additional weakly dependent common factors where its number can go to infinity. In their paper, both weakly dependent and strongly dependent common factors are defined accordingly, and the spatial process with row sum and column sum boundedness properties is regarded as weakly dependent. Their model is

$$y_{it} = \alpha'_i \mathbf{d}_t + \beta'_i \mathbf{x}_{it} + \gamma'_i f_t + \lambda'_i \mathbf{n}_t + e_{it}, \quad (18)$$

where  $\mathbf{n}_t$ , uncorrelated with the regressor  $\mathbf{x}_{it}$ , is the additional weakly dependent common factors with dimension of the factor loading coefficients  $\lambda_i$  going to infinity. They derive the asymptotic properties of the CCE estimators under such a setting, showing that the CCE method still yields consistent estimates of the mean of the slope coefficients and the asymptotic normal theory continues to be applicable.

Holly, Pesaran, and Yamagata (2011) provide a method for the analysis of the spatial and temporal diffusion of shocks in a dynamic system. They generalize VAR panel models with unobserved common factors in time series to incorporate spatial elements in the dynamic coefficient matrix. With coefficients being spatial unit specific, consistent estimation has emphasized on  $T$  tending to infinity while the number of spatial units can be moderate or large. They use changes in real house prices within the UK economy at the level of regions, and analyze the effect of shocks using generalized spatio-temporal impulse responses.

Chen and Conley (2002), earlier investigated a VAR model in a finite  $n$  setting. Unlike Holly, Pesaran, and Yamagata (2011), the dynamic coefficient matrix is from a semiparametric model where functions of agents' economic distances provide restrictions that enable estimation. A two-step sieve least-squares procedure is applied to estimate the model, where rates of convergence for the sieve estimators are provided, limiting distributions for the model's finite-dimensional parameters are derived, and a bootstrap method for inference is suggested.

The spatial panel data models discussed in Section 12.2 assume a time invariant spatial weights matrix. When the spatial weights matrix is constructed with economic/socioeconomic distances or demographic characteristics, it can be time varying. For example, Case, Hines, and Rosen (1993) on state spending have weights based on the difference in the percentage of the population that is black. Baicker (2005) constructs a weights matrix with the degree of population mobility between regions. So does Rincke (2010). Lee and Yu (2012b) investigate the QML estimation of SDPD models where spatial weights matrices can be time varying. They find that QML estimate is consistent and asymptotically normal. Monte Carlo results show that, when spatial weights matrices are substantially varying over time, a model misspecification of a time

invariant spatial weights matrix may cause substantial bias in estimation. Slowly time varying spatial weights matrices would be of less concern.

## 12.3 ESTIMATION AND INFERENCE

### 12.3.1 Static Models

For notational purposes, we define  $\tilde{Y}_{nt} = Y_{nt} - \bar{Y}_{nT}$  for  $t = 1, 2, \dots, T$  where  $\bar{Y}_{nT} = \frac{1}{T} \sum_{t=1}^T Y_{nt}$ .

#### 12.3.1.1 Static Model with Serially Uncorrelated Disturbances

For a fixed effects model, Lee and Yu (2010a) consider QMLE for the model (4), where both direct and transformation approaches are used to take care of the individual effects. In the direct approach, common parameters and individual effects are jointly estimated. The likelihood function with  $\mu_n$  concentrated out is

$$\begin{aligned} \ln L_{n,T}^d(\theta) = & -\frac{nT}{2} \ln(2\pi\sigma^2) + T[\ln|S_n(\lambda_1)| + \ln|R_n(\lambda_2)|] \\ & -\frac{1}{2\sigma^2} \sum_{t=1}^T \tilde{V}'_{nt}(\zeta) \tilde{V}_{nt}(\zeta), \end{aligned} \quad (19)$$

where  $\tilde{V}_{nt}(\zeta) = R_n(\lambda_2)[S_n(\lambda_1)\tilde{Y}_{nt} - \tilde{X}_{nt}\beta]$  and  $R_n(\lambda_2) = I_n - \lambda_2 W_{n2}$ . In the transformation approach, let  $[F_{T,T-1}, \frac{1}{\sqrt{T}}l_T]$  be the orthonormal eigenvector matrix of  $J_T = I_T - \frac{1}{T}l_T l'_T$ . By using the transformation  $F_{T,T-1}$  which transforms the  $T$  periods observations to a  $T-1$  periods observations, the individual effects are eliminated. In this regard, the use of  $F_{T,T-1}$  instead of  $J_T$  provides a proper way to evaluate the determinant of the proper Jacobian transformation and the degrees of freedom in the construction of the likelihood function. The log likelihood function is

$$\begin{aligned} \ln L_{n,T}(\theta) = & -\frac{n(T-1)}{2} \ln(2\pi\sigma^2) + (T-1)[\ln|S_n(\lambda_1)| + \ln|R_n(\lambda_2)|] \\ & -\frac{1}{2\sigma^2} \sum_{t=1}^T \tilde{V}'_{nt}(\zeta) \tilde{V}_{nt}(\zeta), \end{aligned} \quad (20)$$

where  $\tilde{V}_{nt}(\zeta) = R_n(\lambda_2)[S_n(\lambda_1)\tilde{Y}_{nt} - \tilde{X}_{nt}\beta]$ , in terms of the derivation from time mean variables.

For the direct ML approach, it will yield consistent estimates for the spatial and regression coefficients, except for the variance parameter  $\sigma_0^2$  of the disturbance when  $T$  is finite. Thus, the results are similar to Neyman and Scott (1948). The transformation

approach is the method of conditional likelihood, which is applicable when sufficient statistics can be found for the fixed effects. For similar spatial panel models but with time effects, with large  $T$ , the direct estimation approach would not provide consistent estimates even for the regression coefficients. In that case, it is desirable to eliminate them for estimation. For the model with both individual and time fixed effects, one may combine the transformations of deviation from time means and also deviation from cross-section means to eliminate those effects. The transformed equation can be regarded as a well-defined equation system when the spatial weights matrix is row-normalized. The resulting likelihood function can be interpreted as a partial likelihood (Cox 1975; Wong 1986).

By treating the individual effect as a random component in (2), Kapoor, Kelejian, and Prucha (2007) propose an MOM estimation with moment conditions in terms of  $(\lambda, \sigma_v^2, \sigma_1^2)$ , where  $\sigma_1^2 = \sigma_v^2 + T\sigma_\mu^2$ . The  $\beta$  can be consistently estimated by the OLS for its regression equation in (2). With MOM estimates of  $(\lambda, \sigma_v^2, \sigma_1^2)$  available from the least squares residuals, a feasible GLS estimate for  $\beta_0$  can then be implemented as  $\hat{\beta}_{GLS,n} = [\mathbf{X}'_{nT}(\hat{\Omega}_{nT}^{kkp})^{-1}\mathbf{X}_{nT}]^{-1}[\mathbf{X}'_{nT}(\hat{\Omega}_{nT}^{kkp})^{-1}\mathbf{Y}_{nT}]$ , where  $\hat{\Omega}_{nT}^{kkp}$  is an estimated variance matrix for the error components of the Kapoor, Kelejian, and Prucha (2007) model,  $\mathbf{Y}_{nT} = (Y'_{n1}, Y'_{n2}, \dots, Y'_{nT})'$  and other variables accordingly.

### 12.3.1.2 Static Model with Serially Correlated Disturbances

For the static spatial panel model with serially correlated disturbances, Parent and LeSage (2012) consider a Bayesian approach for estimation under the assumption that the disturbances are normally distributed. They apply the MCMC method to obtain random draws from posterior distributions of the parameters. Lee and Yu (2012a) consider a general spatial panel data model in (5) with serially and spatially correlated disturbances.

#### *Fixed individual effects*

With  $\mu_n$  being fixed parameters, after first difference and quasi difference, the estimation equation would be

$$\begin{aligned}\Delta Y_{nt,\rho_0} &= \lambda_{01} W_{n1} \Delta Y_{nt,\rho_0} + \Delta X_{nt,\rho} \beta_0 + \Delta U_{nt,\rho_0} \text{ with} \\ \Delta U_{nt,\rho_0} &= \lambda_{02} W_{n2} \Delta U_{nt,\rho_0} + (I_n + \delta_{02} M_{n2}) \Delta e_{nt}, t = 3, \dots, T,\end{aligned}\tag{21}$$

where  $\Delta Y_{nt,\rho_0} = \Delta Y_{nt} - \rho_0 \Delta Y_{n,t-1}$ . For the initial period,  $\Delta Y_{n2} = \lambda_{01} W_{n1} \Delta Y_{n2} + \Delta X_{n2} \beta_0 + \Delta U_{n2}$  with

$$\Delta U_{n2} = \lambda_{02} W_{n2} \Delta U_{n2} + (I_n + \delta_{02} M_{n2}) \Delta V_{n2},$$

$$\Delta V_{n2} = e_{n2} - (1 - \rho_0) V_{n1}.$$

As derived in Lee and Yu (2012a), with  $\theta_1 = (\beta', \lambda_1, \lambda_2, \delta_2, \rho, \sigma_e^2)'$ , the log likelihood of (21) is

$$\begin{aligned} \ln L_{w,nT}(\theta_1) &= -\frac{n(T-1)}{2} \ln(2\pi\sigma_e^2) + (T-1)(\ln|S_{n1}(\lambda_1)| + \ln|S_{n2}(\lambda_2)| - \ln|B_{n2}(\delta_2)|) \\ &\quad - \frac{n}{2} \ln|H_{T-1}(\rho)| - \frac{1}{2\sigma_e^2} \mathbf{V}'_{nT}(\theta_1) (\mathbb{J}_T(\rho) \otimes I_n) \mathbf{V}_{nT}(\theta_1), \end{aligned} \quad (22)$$

with  $\mathbf{V}_{nT}(\theta_1) = (I_T \otimes B_{n2}^{-1}(\delta_2) S_{n2}(\lambda_2)) [(I_T \otimes S_{n1}(\lambda_1)) \mathbf{Y}_{nT} - \mathbf{X}_{nT}\beta - \mathbf{c}_n]$ ,  $S_{nj} = I_n - \lambda_j W_{nj}$ , and  $B_{n2} = I_n - \delta_2 M_{n2}$ . Here,  $H_{T-1}(\rho)$  is the variance matrix of  $\mathbf{e}_{n,T-1}^d = (\Delta V'_{n2}, \Delta e'_{n3}, \dots, \Delta e'_{nT})'$  which is, under the assumption of a long past,

$$H_{T-1}(\rho) = \begin{pmatrix} \frac{2}{1+\rho} & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix},$$

and

$$\mathbb{J}_T(\rho) \equiv \mathbb{I}_T(\rho) - \frac{1-\rho}{T-(T-2)\rho} \mathbf{p}_T(\rho) \mathbf{p}_T'(\rho), \quad (23)$$

$$\text{with } \mathbb{I}_T(\rho) = \begin{pmatrix} 1 & -\rho & \cdots & 0 & 0 \\ -\rho & 1+\rho^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1+\rho^2 & -\rho \\ 0 & 0 & \cdots & -\rho & 1 \end{pmatrix} \text{ and } \mathbf{p}_T(\rho) = (1, 1-\rho, \dots, 1-\rho, 1)',$$

$1-\rho, 1)'$ , is the matrix that combines the quasi and first difference transformations of the data. This likelihood function is computationally tractable even with large  $n$  and  $T$ , because the determinants of  $S_n$ 's and  $B_n$  can be evaluated as in the cross-sectional SAR model and the analytical determinant and inverse of  $H_{T-1}$  are in Hsiao, Pesaran, and Tahmisioglu (2002).

### Random individual effects

The random effects specification can be seen from (5), which can be written for the  $nT$  observations as

$$\begin{aligned} \mathbf{Y}_{nT} &= l_T \otimes z_n b_0 + \lambda_{01} (I_T \otimes W_{n1}) \mathbf{Y}_{nT} + \mathbf{X}_{nT}\beta_0 \\ &\quad + l_T \otimes S_{n3}^{-1} B_{n3} c_{n0} + (I_T \otimes S_{n2}^{-1} B_{n2}) \mathbf{V}_{nT}. \end{aligned} \quad (24)$$

For the overall disturbance vector  $\xi_{nT} = l_T \otimes S_{n3}^{-1} B_{n3} c_{n0} + (I_T \otimes S_{n2}^{-1} B_{n2}) \mathbf{V}_{nT}$ , its variance matrix is

$$\Omega_{nT} = \sigma_{0c}^2 [l_T l'_T \otimes S_{n3}^{-1} B_{n3} B'_{n3} S_{n3}^{-1}] + \sigma_{0e}^2 [\Sigma_{T,\rho_0} \otimes S_{n2}^{-1} B_{n2} B'_{n2} S_{n2}^{-1}], \quad (25)$$

with  $\Sigma_{T,\rho_0} = \frac{1}{1-\rho_0^2} \begin{pmatrix} 1 & \rho_0 & \rho_0^2 & \cdots & \rho_0^{T-1} \\ \rho_0 & 1 & \rho_0 & \cdots & \rho_0^{T-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_0^{T-1} & \rho_0^{T-2} & \rho_0^{T-3} & \cdots & 1 \end{pmatrix}$  and  $E(\mathbf{V}_{nT}\mathbf{V}'_{nT}) = \sigma_{0e}^2 \Sigma_{T,\rho_0} \otimes I_n$ . By Lemma 2.2 in Magnus (1982),

$$|\Omega_{nT}| = |\Sigma_{T,\rho_0}|^n \cdot |\sigma_{0e}^2 S_{n2}^{-1} B_{n2} B'_{n2} S_{n2}'|^{T-1} \cdot |\mathcal{Z}_{n0}^{-1}|,$$

where  $\mathcal{Z}_{n0} = [d^2(1-\rho_0)^2 \sigma_{0c}^2 S_{n3}^{-1} B_{n3} B'_{n3} S_{n3}' + \sigma_{0e}^2 S_{n2}^{-1} B_{n2} B'_{n2} S_{n2}']^{-1}$  with  $d^2 = l_T^\rho l_T^\rho = \frac{1+\rho_0}{1-\rho_0} + (T-1)$ ,  $l_T^\rho = (\sqrt{\frac{1+\rho_0}{1-\rho_0}}, 1, \dots, 1)'$ , and

$$\begin{aligned} \Omega_{nT}^{-1} &= \frac{1}{d^2(1-\rho_0)^2} \Sigma_{T,\rho_0}^{-1} l_T l'_T \Sigma_{T,\rho_0}^{-1} \otimes \mathcal{Z}_{n0} + \left( \Sigma_{T,\rho_0}^{-1} - \frac{1}{d^2(1-\rho_0)^2} \Sigma_{T,\rho_0}^{-1} l_T l'_T \Sigma_{T,\rho_0}^{-1} \right) \\ &\otimes \left( \frac{1}{\sigma_{0e}^2} S_{n2}' B_{n2}'^{-1} B_{n2}^{-1} S_{n2} \right). \end{aligned}$$

The parameter subvector  $\theta_{01}$  can be estimated from both fixed and random effects models. The remaining parameters in  $\theta_2$  can only be estimated under the random effects specification.

The log likelihood function for (24) is

$$\begin{aligned} \ln L_{r,nT}(\theta) &= -\frac{nT}{2} \ln(2\pi) - \frac{1}{2} \ln |\Omega_{nT}(\theta)| + T \ln |S_{n1}(\lambda_1)| \\ &\quad - \frac{1}{2} \xi'_{nT}(\theta) \Omega_{nT}^{-1}(\theta) \xi_{nT}(\theta), \end{aligned} \tag{26}$$

where  $\xi_{nT}(\theta) = (I_T \otimes S_{n1}(\lambda_1)) \mathbf{Y}_{nT} - \mathbf{X}_{nT} \beta - l_T \otimes z_n b$ . Here, the computational burden comes mainly from the evaluation of matrix  $\mathcal{Z}_{n0}$ , which involves an inverse of a sum of square matrices of dimension  $n$ . However, the additional serial correlation in  $V_{nt}$  does not generate more computational complexity.

It is known for panel regression models that, if the random components model is properly specified, the random effects MLE will be more efficient relative to the fixed effects one (Maddala 1971). This can be revealed by relating the likelihood functions of the two settings in terms of the between and within equations and the pooling of these estimates.

### *The between equation*

Similar to linear panel data models, the likelihood of the random effects model can be written as a product of the within likelihood and the between likelihood. As derived in Lee and Yu (2012a), the within equation under the serial correlation is

$$S_{n1} \vec{Y}_{nT} = z_n b_0 + \vec{X}_{nT} \beta_0 + \mu_n + S_{n2}^{-1} B_{n2} \vec{V}_{nT}, \tag{27}$$

where  $\vec{V}_{nT} = [T - (T-2)\rho_0]^{-1} [V_{n1} + (1-\rho_0) \sum_{t=2}^{T-1} V_{nt} + V_{nT}]$ , and similarly for  $\vec{Y}_{nT}$  and  $\vec{X}_{nT}$ . When there is no serial correlation so that  $\rho_0 = 0$ ,  $\vec{Y}_{nT} = \bar{Y}_{nT}$ ,

$\vec{X}_{nT} = \bar{X}_{nT}$  and  $\vec{V}_{nT} = \bar{V}_{nT}$  are time averages, and the between equation becomes  $S_{n1}\bar{Y}_{nT} = z_n b_0 + \bar{X}_{nT}\beta_0 + \mu_n + S_{n2}^{-1}B_{n2}\bar{V}_{nT}$  in the familiar form. As the within equation does not involve  $\mu_n$ , identification of the spatial structure of  $\mu_n$  will solely depend on the between equation. This between equation highlights the main distinction of the random components model and the within equation.

The variance matrix of disturbances in the between equation (27) is

$$\Omega_{n1} = \sigma_{0c}^2 S_{n3}^{-1} B_{n3} B'_{n3} S'^{-1}_{n3} + \sigma_1^2 S_{n2}^{-1} B_{n2} B'_{n2} S'^{-1}_{n2},$$

and the log likelihood is

$$\ln L_{b,n}(\theta) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |\Omega_{n1}(\theta)| + \ln |S_{n1}(\lambda_1)| - \frac{1}{2} \xi'_n(\theta) \Omega_{n1}^{-1}(\theta) \xi_n(\theta), \quad (28)$$

where  $\xi_n(\theta) = S_{n1}(\lambda_1) \bar{Y}_{nT}(\rho) - \bar{X}_{nT}(\rho) \beta - z_n b$ . For parameters' identification,  $b_0$ ,  $\beta_0$ ,  $\lambda_{01}$  and  $\rho_0$  can be identified from the main equation in (27); but the variance matrix  $\Omega_{n1}$  will be the sole source for the identification of  $\lambda_{02}$ ,  $\lambda_{03}$ ,  $\delta_{02}$ ,  $\delta_{03}$ ,  $\sigma_{0e}^2$  and  $\sigma_{0c}^2$ .

### *Pooling of estimates*

The random effects model has all the parameter vector in  $\theta_0 = (\theta'_{01}, \theta'_{02})'$ , but that of the fixed effects model has only the subvector  $\theta_{01}$ . The excluded parameters in  $\theta_{02}$  would appear in the between equation. In order to compare the efficiency of estimates of the two models, one simple approach is to use the concentrated likelihood function  $L_{r,nT}^c(\theta_1)$  of the random effects model with  $\theta_2$  concentrated out, and compare it with  $L_{w,nT}^c(\theta_1)$  of the within equation. Similarly, one can have the concentrated likelihood  $L_{b,n}^c(\theta_1)$  of the between equation. Those concentrated likelihood functions of the random effects model and the between equation have the same common  $\theta_1$  as that of the within equation. The random effects estimate of  $\theta_{01}$  can be interpreted as an asymptotically weighted average of the within and between estimates, closely analogous to Maddala (1971) for the panel regression model:

$$\sqrt{n}(\hat{\theta}_{r1} - \theta_{01}) = A_{nT,1}\sqrt{n}(\hat{\theta}_{w1} - \theta_{01}) + A_{nT,2}\sqrt{n}(\hat{\theta}_{b1} - \theta_{01}) + o_p(1), \quad (29)$$

where  $A_{nT,1} = (\frac{1}{n} \frac{\partial^2 \ln L_{r,nT}^c(\theta_1)}{\partial \theta_1 \partial \theta'_1})^{-1} \frac{1}{n} \frac{\partial^2 \ln L_{w,nT}^c(\theta_1)}{\partial \theta_1 \partial \theta'_1}$  and  $A_{nT,2} = (\frac{1}{n} \frac{\partial^2 \ln L_{r,nT}^c(\theta_1)}{\partial \theta_1 \partial \theta'_1})^{-1} \frac{1}{n} \frac{\partial^2 \ln L_{b,n}^c(\theta_1)}{\partial \theta_1 \partial \theta'_1}$  are weights. The weighting above is valid even though those likelihood functions are quasi ones.

### *The Hausman specification test*

*The Hausman test under normality* The likelihood decomposition provides a useful device for a Hausman-type test of the random effects specification against the fixed effects specification where the individual effects could be correlated with exogenous regressors. Under the null hypothesis that  $c_{n0}$  is uncorrelated with exogenous regressors, the MLE  $\hat{\theta}_{r1}$  of the random effects model is consistent and asymptotically

efficient (assuming the likelihood function is correctly specified, in this case, with normal disturbances). However, under the alternative,  $\hat{\theta}_{r1}$  is inconsistent; while the within estimator  $\hat{\theta}_{w1}$  is consistent under both the null and alternative hypotheses. The Hausman-type statistic is  $n(\hat{\theta}_{r1} - \hat{\theta}_{w1})'\hat{\Omega}_n^+(\hat{\theta}_{r1} - \hat{\theta}_{w1})$ , where  $\hat{\Omega}_n$  is a consistent estimate of the limiting variance matrix of  $\sqrt{n}(\hat{\theta}_{r1} - \hat{\theta}_{w1})$  under the null hypothesis, and  $\hat{\Omega}_n^+$  is its generalized inverse. This test statistic will be asymptotically  $\chi^2$  distributed, and its degrees of freedom is the rank of the limiting matrix of  $\Omega_n$ . Asymptotically, this test is also equivalent to test the difference of the within and between estimates, because (29) implies  $\sqrt{n}(\hat{\theta}_{r1} - \hat{\theta}_{w1}) = A_{nT,2}\sqrt{n}(\hat{\theta}_{b1} - \hat{\theta}_{w1}) + o_p(1)$ .

Here, the rank of  $\Omega_n$  needs special attention. Suppose that  $B^{-1}$  is the limiting variance matrix of  $\sqrt{n}(\hat{\theta}_{w1} - \theta_{01})$ , and, as  $\hat{\theta}_{r1}$  is asymptotically efficient relative to  $\hat{\theta}_{w1}$ , the limiting variance matrix of  $\sqrt{n}(\hat{\theta}_{r1} - \theta_{01})$  can be written as  $(B + C)^{-1}$  for some nonnegative definite matrix  $C$ . The  $C$  is a semi definite matrix with rank  $m$  and its nonzero eigenvalues  $\lambda_1, \dots, \lambda_m$ , but not necessarily of full rank. The Hausman test statistic can be computed as  $n(\hat{\theta}_{r1} - \hat{\theta}_{w1})'Q\text{diag}\{\frac{1+\lambda_1}{\lambda_1}, \dots, \frac{1+\lambda_m}{\lambda_m}, 0, \dots, 0\}Q'(\hat{\theta}_{r1} - \hat{\theta}_{w1})$ , where  $Q$  is the eigenvector matrix of  $C$  in the metric of  $B$ . In general, if  $\theta_{01}$  can be identified and estimated from the between equation,  $C$  would have full rank  $k_{\theta_1}$ . As shown in Lee and Yu (2012a), the  $C$  in the Hausman test for Kapoor, Kelejian, and Prucha (2007) is not of full rank, while that for the Anselin (1988) model is.

*The Hausman test under non-normality* When the disturbances are not normally distributed, the random effects estimate is no longer efficient; hence, the variance of the difference of  $\hat{\theta}_{r1}$  and  $\hat{\theta}_{w1}$  will not necessarily equal the difference of the variances of  $\hat{\theta}_{r1}$  and  $\hat{\theta}_{w1}$ . We can compute the variance of  $\sqrt{n}(\hat{\theta}_{r1} - \hat{\theta}_{w1})$  explicitly and set up a corresponding Wald-type robust test. As derived in Lee and Yu (2012a), the Hausman test statistic can be calculated as  $(\hat{\theta}_{r1} - \hat{\theta}_{w1})'\hat{\Omega}_{rw,nT}^+(\hat{\theta}_{r1} - \hat{\theta}_{w1})$ , where  $\hat{\Omega}_{rw,nT}^+$  needs to take into account the third and fourth moments of disturbances and also the covariance matrix of the scores of the within and between equations.

Instead of the ML approach, if the main equation is estimated by the two-stage least squares (2SLS) method, Hausman test statistics based on the coefficients of the main equation can be constructed as in Mutl and Pfaffermayr (2011). Because the likelihood approach takes into account both the main regression equation, as well as the implied correlation of disturbances of the reduced form equation, the Hausman test statistic based on the likelihood function will be relatively more powerful than that based on 2SLS estimates. Debarsy (2012) extends the Mundlak approach to the spatial Durbin panel data model to determine the adequacy of the random effects specification. A likelihood ratio (LR) test is proposed that assesses the significance of the correlation between regressors and individual effects. His Monte Carlo simulations study show that in some cases, the LR test has better power than the Hausman test.

## 12.3.2 Dynamic Models

### 12.3.2.1 Dynamic Panel Data Models with SLs

Yu, de Jong, and Lee (2008, 2012) and Yu and Lee (2010) apply QML estimation to different SDPD models. Elhorst (2010b) uses Monte Carlo studies to evaluate performances of QML and GMM estimators. Including high order SLs in a spatial dynamic panel model, Lee and Yu (2014) investigate GMM estimation with linear and quadratic moments under the setting that  $T$  is small relative to  $n$ .

For the estimation for the SDPD model (9) when both  $n$  and  $T$  tend to infinity, the rates of convergence of QMLEs are  $\sqrt{nT}$  for the stable case, as shown in Yu, de Jong, and Lee (2008). For the spatial cointegration case, Yu, de Jong, and Lee (2012) show that the QMLEs for such a model are  $\sqrt{nT}$  consistent and asymptotically normal, but, the presence of the unstable components will make the estimators' asymptotic variance matrix singular. Consequently, a linear combination of the spatial and dynamic effects estimates can converge at a higher rate. For the unit roots case, from Yu and Lee (2010), the QMLEs of  $\gamma_0$  is  $\sqrt{nT^3}$  consistent and other estimates are  $\sqrt{nT}$  consistent; however, the estimate of the sum  $\rho_0 + \lambda_0$  is  $\sqrt{nT^3}$  consistent. For the explosive case, as seen from Lee and Yu (2011b), one may rely on a data transformation to estimate the model. In the following, we first focus on models without the time effects. We then discuss the case where the time effects are included but will be eliminated by some data transformations.

For notational purposes, we define  $\tilde{Y}_{n,t-1}^{(-1)} = Y_{n,t-1} - \bar{Y}_{nT,-1}$  for  $t = 1, 2, \dots, T$  where  $\bar{Y}_{nT,-1} = \frac{1}{T} \sum_{t=1}^T Y_{n,t-1}$ .

#### Stable Case

Denote  $\theta = (\delta', \lambda, \sigma^2)'$ , where  $\delta = (\gamma, \rho, \beta')'$ . By denoting  $Z_{nt} = (Y_{n,t-1}, W_n Y_{n,t-1}, X_{nt})$ , the likelihood function of the SDPD model (9) is

$$\ln L_{n,T}(\theta, \mathbf{c}_n) = -\frac{nT}{2} \ln 2\pi - \frac{nT}{2} \ln \sigma^2 + T \ln |S_n(\lambda)| - \frac{1}{2\sigma^2} \sum_{t=1}^T V'_{nt}(\theta) V_{nt}(\theta), \quad (30)$$

where  $V_{nt}(\theta) = S_n(\lambda)Y_{nt} - Z_{nt}\delta - c_n$ . The QMLEs  $\hat{\theta}_{nT}$  and  $\hat{c}_{nT}$  are the extremum estimators derived from the maximization of (30), and  $\hat{c}_{nT}$  can be consistently estimated when  $T$  goes to infinity. Using the first order condition for  $\mathbf{c}_n$ , the concentrated likelihood is

$$\ln L_{n,T}(\theta) = -\frac{nT}{2} \ln 2\pi - \frac{nT}{2} \ln \sigma^2 + T \ln |S_n(\lambda)| - \frac{1}{2\sigma^2} \sum_{t=1}^T \tilde{V}'_{nt}(\theta) \tilde{V}_{nt}(\theta), \quad (31)$$

where  $\tilde{V}_{nt}(\theta) = S_n(\lambda)\tilde{Y}_{nt} - \tilde{Z}_{nt}\delta$  with  $\tilde{Z}_{nt} = (\tilde{Y}_{n,t-1}^{(-1)}, W_n \tilde{Y}_{n,t-1}^{(-1)}, \tilde{X}_{nt})$ . The QMLE  $\hat{\theta}_{nT}$  maximizes the concentrated likelihood function (31). As is shown in Yu, de Jong, and

Lee (2008), we have

$$\begin{aligned} & \sqrt{nT} (\hat{\theta}_{nT} - \theta_0) + \sqrt{\frac{n}{T}} \varphi_{\theta_0, nT} + O_p \left( \max \left( \sqrt{\frac{n}{T^3}}, \sqrt{\frac{1}{T}} \right) \right) \\ & \xrightarrow{d} N(0, \lim_{T \rightarrow \infty} \Sigma_{\theta_0, nT}^{-1} (\Sigma_{\theta_0, nT} + \Omega_{\theta_0, nT}) \Sigma_{\theta_0, nT}^{-1}), \end{aligned} \quad (32)$$

where  $\theta_0$  is the true parameter vector,  $\varphi_{\theta_0, nT}$  is the leading bias term of order  $O(1)$ ,  $\Sigma_{\theta_0, nT}$  is the information matrix, and  $\Omega_{\theta_0, nT}$  captures the non-normality feature of the disturbances. Hence, for distribution of the common parameters, when  $T$  is large relative to  $n$ , the estimators are  $\sqrt{nT}$  consistent and asymptotically normal, with the limiting distribution centered around 0; when  $n$  is asymptotically proportional to  $T$ , the estimators are  $\sqrt{nT}$  consistent and asymptotically normal, but the limiting distribution is not centered around 0; and when  $n$  is large relative to  $T$ , the estimators are  $T$  consistent, but have a degenerate limiting distribution.

### *Spatial cointegration case*

The log likelihood function of the spatial cointegration model is the same as that of the stable case. However, the properties of the estimators are not the same. As shown in Yu, de Jong, and Lee (2012), one can still derive the expansion in (32) with a  $\varphi_{\theta_0, nT}$ , which has a stable part and a nonstable part. The distinctive feature of the spatial cointegration case is that  $\lim_{T \rightarrow \infty} \Sigma_{\theta_0, nT}^{-1}$  exists but is singular. This indicates that some linear combinations may have higher rates of convergence. Indeed, we have

$$\begin{aligned} & \sqrt{nT^3} (\hat{\lambda}_{nT} + \hat{\gamma}_{nT} + \hat{\rho}_{nT} - 1) + \sqrt{\frac{n}{T}} b_{\theta_0, nT} + O_p \left( \max \left( \sqrt{\frac{n}{T^3}}, \sqrt{\frac{1}{T}} \right) \right) \\ & \xrightarrow{d} N(0, \lim_{T \rightarrow \infty} \sigma_{1, nT}^2), \end{aligned}$$

where  $\sigma_{1, nT}^2$  is a positive scalar variance and  $b_{\theta_0, nT}$  is  $O(1)$ .

The spatial cointegration model is related to the cointegration literature. Here, the unit roots are generated from the mixed time and spatial dimensions. The cointegration matrix is  $(I_n - W_n)$ , and its rank is the number of eigenvalues of  $W_n$  being less than 1 in absolute value. Compared to conventional cointegration in time series literature, the cointegrating space is completely known and is determined by the spatial weights matrix; while in the conventional time series, it is the main object of inference. Also, in the conventional cointegration, the dimension of VAR is fixed and relatively small while the spatial dimension in the SDPD model is large.

### *Transformation approach with $J_n$ : the case with time dummies*

When we have time effects included in the SDPD model, as is derived in Lee and Yu (2010b), the direct estimation method by treating time dummies as regression coefficients will yield a bias of order  $O(\max(1/n, 1/T))$  for the common parameters. In

order to avoid the bias of the order  $O(1/n)$  due to the presence of many time dummies, we may use a data transformation approach to eliminate the time dummies. The transformed equation can be treated as a partial likelihood, and the resulting ML estimator may have the same asymptotic efficiency as the direct QML estimator. This transformation procedure is particularly useful when  $n/T \rightarrow 0$ , where the estimates of the transformed approach will have a faster rate of convergence than that of the direct estimates. Also, when  $n/T \rightarrow 0$ , the estimates under the direct approach have a degenerate limit distribution, but the estimates under the transformation approach are properly centered and asymptotically normal.

With the transformation  $F_{n,n-1}$  from the eigenvector matrix of  $J_n = I_n - \frac{1}{n}l_n l_n'$ , by denoting  $Y_{nt}^* = F'_{n,n-1} Y_{nt}$ , we have

$$Y_{nt}^* = \lambda_0 W_n^* Y_{nt}^* + \gamma_0 Y_{n,t-1}^* + \rho_0 W_n^* Y_{n,t-1}^* + X_{nt}^* \beta_0 + c_{n0}^* + V_{nt}^*, \quad (33)$$

where  $W_n^* = F'_{n,n-1} W_n F_{n,n-1}$  holds because  $W_n$  is row-normalized. The  $V_{nt}^*$  is an  $(n-1)$  dimensional disturbance vector with zero mean and variance matrix  $\sigma_0^2 I_{n-1}$ . Equation (33) is in the format of a typical SDPD model, where the number of observations is  $T(n-1)$ , reduced from the original sample observations by one for each period. Equation (33) is useful because a likelihood function for  $Y_{nt}^*$  can be constructed. Such a likelihood function is a partial likelihood. If  $V_{nt}$  is normally distributed  $N(0, \sigma_0^2 I_n)$ , the transformed  $V_{nt}^*$  will be  $N(0, \sigma_0^2 I_{n-1})$ . Thus, the log likelihood function of (33) can be written as

$$\begin{aligned} \ln L_{n,T}(\theta, \mathbf{c}_n) = & -\frac{(n-1)T}{2} \ln 2\pi - \frac{(n-1)T}{2} \ln \sigma^2 - T \ln(1-\lambda) \\ & + T \ln |I_n - \lambda W_n| - \frac{1}{2\sigma^2} \sum_{t=1}^T V'_{nt}(\theta) J_n V_{nt}(\theta). \end{aligned} \quad (34)$$

As is shown in Lee and Yu (2010b), the QMLE from the above maximization is free of  $O(1/n)$  bias.

### *Explosive case*

When some eigenvalues of  $A_n$  are greater than 1, an ML algorithm might be numerically unstable. Furthermore, asymptotic properties of the QML estimates of such a case are unknown. However, the explosive feature of the model can be handled by the data transformation  $I_n - W_n$ . The transformation  $I_n - W_n$  can eliminate not only time dummies but also the unstable component. By using the eigenvector matrix  $F_n$  from  $\Sigma_n = (I_n - W_n)(I_n - W_n)'$ , denoting  $W_n^* = \Lambda_n^{-1/2} F'_n W_n F_n \Lambda_n^{1/2}$  which is an  $n^* \times n^*$  matrix where  $n^* = n - m_n$ , we have

$$Y_{nt}^* = \lambda_0 W_n^* Y_{nt}^* + \gamma_0 Y_{n,t-1}^* + \rho_0 W_n^* Y_{n,t-1}^* + X_{nt}^* \beta_0 + c_{n0}^* + V_{nt}^*, \quad (35)$$

where  $Y_{nt}^* = \Lambda_n^{-1/2} F'_n (I_n - W_n) Y_{nt}$  and other variables are defined accordingly. Note that this transformed  $Y_{nt}^*$  is an  $n^*$ -dimensional vector. The eigenvalues of  $W_n^*$  are

exactly those eigenvalues of  $W_n$  less than 1 in absolute value. It follows that the eigenvalues of  $A_n^* = (I_{n^*} - \lambda_0 W_n^*)^{-1}(\gamma_0 I_{n^*} + \rho_0 W_n^*)$  can be all less than 1 in absolute values even when  $\gamma_0 + \rho_0 + \lambda_0 = 1$  with  $|\lambda_0| < 1$  and  $|\gamma_0| < 1$ . For the explosive case with  $\gamma_0 + \rho_0 + \lambda_0 > 1$ , all the eigenvalues of  $A_n^*$  can be less than 1 only if  $\frac{\rho_0 + \lambda_0}{1 - \gamma_0} < \frac{1}{\varpi_{n,(2)}}$ , where  $\varpi_{n,(2)}$  is the maximum eigenvalue of  $W_n$  less than 1. Hence, the transformed model (35) is a stable one as long as  $\gamma_0 + \rho_0 + \lambda_0$  is not too much larger than 1 and the parameter space is in (15).

The concentrated log likelihood of (35) is

$$\begin{aligned}\ln L_{n,T}(\theta) = & -\frac{n^*T}{2} \ln 2\pi - \frac{n^*T}{2} \ln \sigma^2 - (n - n^*)T \ln(1 - \lambda) + T \ln |I_n - \lambda W_n| \\ & - \frac{1}{2\sigma^2} \sum_{t=1}^T \tilde{V}'_{nt}(\theta)(I_n - W_n)' \Sigma_n^+ (I_n - W_n) \tilde{V}_{nt}(\theta),\end{aligned}\quad (36)$$

where  $\tilde{V}_{nt}(\theta) = S_n(\lambda) \tilde{Y}_{nt} - \tilde{Z}_{nt} \delta$ . From Lee and Yu (2011b), we have similar results to those of the stable model, where the bias term and the variance term would involve only the stable component left after the  $I_n - W_n$  transformation.

Therefore, we can use the spatial difference operator,  $I_n - W_n$ , which may eliminate not only the time effects but also the possible unstable or explosive components that are generated from the spatial cointegration, or explosive roots. This implies that the spatial difference transformation can be applied to DGPs with stability, spatial cointegration, or explosive roots, and the resulting estimate can have robustness to the various stochastic nature of the spatial processes. The transformation  $I_n - W_n$  provides a unified estimation procedure for the estimation of SDPD models. Elhorst (2012) has pointed out a practical use of this transformation in an empirical study.

### Bias correction

For each case above, we have bias due to initial condition and incidental parameter problems in dynamic panel data models with individual effects. This bias can be eliminated by using the bias corrected estimator. For a time-space recursive model with fixed effects, Korniotis (2010) has considered a bias adjusted within estimator, which generalizes Hahn and Kuersteiner (2002). For the SDPD models, we can similarly use  $\hat{\theta}_{nT}^1 = \hat{\theta}_{nT} - \frac{\hat{B}_{nT}}{T}$ , where  $\hat{B}_{nT} = \left[ \left( E \left( \frac{1}{nT} \frac{\partial^2 \ln L_{n,T}(\theta)}{\partial \theta \partial \theta'} \right) \right)^{-1} \varphi(\theta) \right]_{\theta=\hat{\theta}_{nT}}$ , which has to be evaluated differently under different SDPD models. When  $T$  grows faster than  $n^{1/3}$ , the correction will eliminate the bias of order  $O(T^{-1})$  and yield a properly centered confidence interval.

#### 12.3.2.2 Dynamic Panel Data Models with Spatial Errors

Elhorst (2005), Su and Yang (2007), and Yu and Lee (2010) consider the estimation of a dynamic panel data model with spatial disturbances but no SL,

$$\begin{aligned} Y_{nt} &= \gamma_0 Y_{n,t-1} + X_{nt}\beta_0 + z_n\eta_0 + U_{nt}, \quad t = 1, \dots, T, \\ U_{nt} &= \mu_n + \varepsilon_{nt}, \text{ and } \varepsilon_{nt} = \lambda_0 W_n \varepsilon_{nt} + V_{nt}. \end{aligned} \tag{37}$$

When  $T$  is moderate or large, this model with  $|\gamma_0| < 1$  can be estimated by the methods for SDPD models. The case  $\gamma_0 = 1$  is special in the sense that the model is a pure unit root case in the time dimension with spatial disturbances.

Elhorst (2005) and Su and Yang (2007) have focused on estimating the short panel case, i.e.,  $n$  is large but  $T$  is fixed. Elhorst (2005) uses the first difference to eliminate the fixed individual effects in  $\mu_n$ , and Su and Yang (2007) derive the asymptotic properties of QMLEs using both the random and fixed effects specifications. When  $T$  is fixed and we have the dynamic feature, the specification of the initial observation  $Y_{n0}$  is important. When  $Y_{n0}$  is assumed to be exogenous, the likelihood function can be easily obtained, either for the random effects specification, or for the fixed effects specification where the first difference is made to eliminate the individual effects. When  $Y_{n0}$  is assumed to be endogenous,  $Y_{n0}$  will need to be generated from a stationary process; or, in the presence of time varying exogenous variables, its distribution needs approximation. With the corresponding likelihood, QMLE can be obtained. We shall provide some more details on these approaches in later parts.

### Pure unit root case

In Yu and Lee (2010) for the SDPD model, when  $\gamma_0 = 1$  and  $\rho_0 + \lambda_0 = 0$ , we have  $A_n = I_n$  in (10) and hence the unit root case. In this case, the eigenvalues of  $A_n$  have no relation with the eigenvalues of  $W_n$  because all of them are equal to 1. This model includes the unit root panel model with SAR disturbances in (37) as a special case. The likelihood of the unit root SDPD model without imposing the constraints  $\gamma_0 = 1$  and  $\rho_0 + \lambda_0 = 0$  is similar to the stable case in (31), but the asymptotic distributions of the estimates are different.

For the unit root SDPD model, as shown in Yu and Lee (2010), the estimate of the pure dynamic coefficient  $\gamma_0$  is  $\sqrt{nT^3}$  consistent and the estimates of all the other parameters are  $\sqrt{nT}$  consistent; and they are asymptotically normal. Also, the sum of the contemporaneous spatial effect estimate of  $\lambda_0$  and the dynamic spatial effect estimate of  $\rho_0$ , i.e.,  $\lambda_0 + \rho_0$ , will converge at  $\sqrt{nT^3}$  rate. The rates of convergence of the estimates can be compared with those of the spatial cointegration case in Yu, de Jong, and Lee (2012). For the latter, all the estimates of parameters including  $\gamma_0$  are  $\sqrt{nT}$  consistent; only the sum of the pure dynamic and spatial effects estimates, i.e.,  $\gamma_0 + \lambda_0 + \rho_0$ , is convergent at the faster  $\sqrt{nT^3}$  rate. Also, there are differences in the bias orders of estimates. For the spatial cointegration case, the biases of all the estimates have the order  $O(1/T)$ . But for the unit root SDPD model, the bias of the estimate of  $\gamma_0$  is of the smaller order  $O(1/T^2)$ , while the biases for all the other estimates have the same  $O(1/T)$  order. These differences are due to different asymptotic behaviors of the two models, even though both models involve unit eigenvalues in  $A_n$ . The unit eigenvalues of the unit root SDPD model are not linked to the eigenvalues of the spatial

**Table 12.2 Summary of different SDPD models**

	Stable	Spatial cointegration	Unit root
Parameters	$\gamma_0 + \lambda_0 + \rho_0 < 1$	$\gamma_0 + \lambda_0 + \rho_0 = 1$ and $\gamma_0 < 1$	$\gamma_0 + \lambda_0 + \rho_0 = 1$ and $\gamma_0 = 1$
Eigenvalues of $A_n$	all less than 1 in absolute value	some unit roots	all unit roots
Dynamics	stable	cointegration cointegrating matrix $I_n - W_n$	no cointegration
Consistency rates	$\sqrt{nT}$ for $\hat{\gamma}_{nT}, \hat{\lambda}_{nT}, \hat{\rho}_{nT}$	$\sqrt{nT}$ for $\hat{\gamma}_{nT}, \hat{\lambda}_{nT}, \hat{\rho}_{nT}$ , $\sqrt{nT^3}$ for $\hat{\gamma}_{nT} + \hat{\lambda}_{nT} + \hat{\rho}_{nT}$	$\sqrt{nT}$ for $\hat{\lambda}_{nT}, \hat{\rho}_{nT}$ , $\sqrt{nT^3}$ for $\hat{\gamma}_{nT}$ and $\hat{\lambda}_{nT} + \hat{\rho}_{nT}$
Information matrix	nonsingular	singular	nonsingular after rescaling
Bias magnitude	$O(\frac{1}{T})$ for $\hat{\gamma}_{nT}, \hat{\lambda}_{nT}, \hat{\rho}_{nT}$	$O(\frac{1}{T})$ for $\hat{\gamma}_{nT}, \hat{\lambda}_{nT}, \hat{\rho}_{nT}$	$O(\frac{1}{T^2})$ for $\hat{\gamma}_{nT}$ $O(\frac{1}{T})$ for $\hat{\lambda}_{nT}, \hat{\rho}_{nT}$

Note: The parameter spaces for different cases are in (13), (14) and (16).

weights matrix. On the contrary, for the spatial cointegration model, the unit eigenvalues correspond exactly to the unit eigenvalues of the spatial weights matrix via a well defined relation. For the unit roots SDPD model, the outcomes of different spatial units do not show co-movements. For the spatial cointegration model, the outcomes of different spatial units can be cointegrated with a reduced rank, where the rank is the number of eigenvalues of  $W_n$  different from 1. Table 12.2 summarizes briefly the differences between the unit root SDPD model and the stable and spatial cointegration cases:

#### *Random effects specification with a fixed T*

For the stable SDPD model with fixed effects, the asymptotic properties of the QML estimation in Yu, de Jong, and Lee (2008) are developed under  $T \rightarrow \infty$  where  $T$  cannot be too small relative to  $n$ . When  $T$  is small, the bias will cause the QMLE estimators to be inconsistent. To handle the initial period for the QML estimation, Elhorst (2005, 2010b), Su and Yang (2007), and Parent and LeSage (2012) approximate the initial observation by some process.

For (37) under the random effects specification, as shown in Su and Yang (2007), the variance matrix of the disturbances is  $\sigma_v^2 \Omega_{nT} = \sigma_v^2 [\phi_\mu (l_T l'_T \otimes I_n) + I_T \otimes (S'_n S_n)^{-1}]$  where  $\phi_\mu = \frac{\sigma_\mu^2}{\sigma_v^2}$ . There are two cases under this specification.

Case I:  $Y_{n0}$  is exogenous. This will be the special situation where the process starts at period 1 coincided with sampling periods. Let  $\theta = (\beta', \eta', \gamma)', \delta = (\lambda, \phi_\mu)', \varsigma = (\theta', \sigma_\nu^2, \delta')'$ . The log likelihood is

$$\ln L(\varsigma) = -\frac{nT}{2} \log(2\pi) - \frac{nT}{2} \log(\sigma_\nu^2) - \frac{1}{2} \log |\Omega_{nT}| - \frac{1}{2\sigma_\nu^2} \mathbf{u}'_{nT}(\theta) \Omega_{nT}^{-1} \mathbf{u}_{nT}(\theta),$$

where  $\mathbf{u}_{nT}(\theta) = \mathbf{Y}_{nT} - \gamma \mathbf{Y}_{nT,-1} - \mathbf{X}_{nT}\beta - l_T \otimes z_n \eta$ . The problem of this approach is that if the process starts earlier than the sampling period at  $t = 1$ , the likelihood will be a misspecified conditional likelihood function and the ML estimates would not be consistent because  $T$  is fixed and finite.

Case II:  $Y_{n0}$  is endogenous. (37) implies that  $Y_{n0} = \tilde{Y}_{n0} + \zeta_{n0}$  where  $\tilde{Y}_{n0} = \sum_{j=0}^{\infty} \gamma_0^j X_{n,t-j} \beta_0 + \frac{z_n \eta_0}{1-\gamma_0}$  is an exogenous part of  $Y_{n0}$  and  $\zeta_{n0} = \frac{\mu_n}{1-\gamma_0} + \sum_{j=0}^{\infty} \gamma_0^j S_n^{-1} V_{n,t-j}$  is the endogenous part. The difficulty to handle  $\tilde{Y}_{n0}$  is due to the missing observations  $X_{nt}$  in  $\tilde{Y}_{n0}$  for  $t < 0$ . Under this situation, Su and Yang (2007) suggest the use of the Bhargava and Sargan (1983) approximation, where the initial value is specified as  $Y_{n0} = \mathcal{X}_{nT}\pi + \epsilon_n$  with  $\mathcal{X}_{nT} = [l_n, \mathbb{X}_{n,T+1}, z_n]$ ,  $\mathbb{X}_{n,T+1} = [X_{n0}, \dots, X_{nT}]$  and  $\pi = (\pi_0, \pi'_1, \pi'_2)'$ , or,  $\mathcal{X}_{nT} = [l_n, \bar{\mathbb{X}}_{n,T+1}, z_n]$  and  $\bar{\mathbb{X}}_{n,T+1} = \frac{1}{T} \sum_{t=0}^T X_{nt}$ . The disturbances of the initial period are specified as  $\epsilon_n = \zeta_n + \zeta_{n0}$  where  $\zeta_n$  is  $(0, \sigma_\zeta^2 I_n)$ . The motivation is that  $\mathcal{X}_{nT}\pi + \zeta_n$  approximates  $\tilde{Y}_{n0}$ . Hence, the disturbances vector would be  $\mathbf{u}_{n,T+1}^* = (\epsilon'_n, \mathbf{u}'_{nT})'$ . Its variance matrix is

$$\sigma_\nu^2 \Omega_{n,T+1}^* = \begin{pmatrix} \sigma_\zeta^2 I_n + \frac{\sigma_\mu^2}{(1-\gamma_0)^2} I_n + \frac{\sigma_\nu^2}{1-\gamma_0^2} (S'_n S_n)^{-1} & \frac{\sigma_\mu^2}{1-\gamma_0} l'_T \otimes I_n \\ \frac{\sigma_\mu^2}{1-\gamma_0} l_T \otimes I_n & \sigma_\nu^2 \Omega_{nT} \end{pmatrix}.$$

Let  $\theta = (\beta', \eta', \pi')', \delta = (\gamma, \lambda, \phi_\mu, \sigma_\zeta^2)'$  and  $\varsigma = (\theta', \sigma_\nu^2, \delta')'$ . The log likelihood is

$$\begin{aligned} \ln L(\varsigma) = & -\frac{n(T+1)}{2} \log(2\pi) - \frac{n(T+1)}{2} \log(\sigma_\nu^2) - \frac{1}{2} \log |\Omega_{n,T+1}^*| \\ & - \frac{1}{2\sigma_\nu^2} \mathbf{u}_{n,T+1}^{**}(\theta) \Omega_{n,T+1}^{*-1} \mathbf{u}_{n,T+1}^*(\theta). \end{aligned}$$

### *Fixed effects specification with a fixed T*

As is discussed in Elhorst (2005) and Su and Yang (2007), the model may also be first differenced to eliminate the individual effects. Thus, we have

$$\Delta Y_{nt} = \gamma_0 \Delta Y_{n,t-1} + \Delta X_{nt} \beta_0 + S_n^{-1} \Delta V_{nt},$$

for  $t = 2, \dots, T$ , and the difference of the first two periods is specified to be  $\Delta Y_{n1} = \Delta \mathcal{X}_{nT}\pi + e_n$ , where  $\Delta \mathcal{X}_{nT} = [l_n, X_{n1} - X_{n0}, \dots, X_{nT} - X_{n,T-1}]$  or  $\Delta \mathcal{X}_{nT} = [l_n, \frac{1}{T} \sum_{t=1}^T (X_{nt} - X_{n,t-1})]$ . Here,  $e_n$  is specified as  $(\xi_{n1} - E(\xi_{n1} | \Delta \mathcal{X}_{nT})) + \sum_{j=0}^m (\gamma_0^j S_n^{-1} \Delta V_{n,1-j})$ , where  $\xi_{n1} - E(\xi_{n1} | \Delta \mathcal{X}_{nT})$  is assumed to be  $(0, \sigma_e^2 I_n)$ . With this

specification, we have  $E(e_n|\Delta\mathcal{X}_{nT}) = 0$  and  $E(e_n e_n') = \sigma_e^2 I_n + \sigma_v^2 c_m (S_n' S_n)^{-1}$ , where  $\sigma_e^2$  and  $c_m$  are parameters or  $c_m$  is a function of parameters to be estimated. The variance matrix of the disturbances vector  $\Delta\mathbf{u}_{nT} = (e_n', \Delta u_{n2}', \dots, \Delta u_{nT}')'$  is  $\sigma_v^2 \Omega_{nT} = \sigma_v^2 (I_T \otimes S_n^{-1}) H_E (I_T \otimes S_n'^{-1})$ , where

$$H_E = \begin{pmatrix} E_n & -I_n & 0 & \cdots & 0 \\ -I_n & 2I_n & -I_n & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & -I_n \\ 0 & \cdots & 0 & -I_n & 2I_n \end{pmatrix},$$

and  $E_n = \frac{\sigma_e^2}{\sigma_v^2} (I_n + c_m (S_n' S_n)^{-1})$ . The log likelihood is

$$\ln L(\zeta) = -\frac{nT}{2} \log(2\pi) - \frac{nT}{2} \log(\sigma_v^2) - \frac{1}{2} \log |\Omega_{nT}| - \frac{1}{2\sigma_v^2} \Delta\mathbf{u}_{nT}'(\theta) \Omega_{nT}^{-1} \Delta\mathbf{u}_{nT}(\theta),$$

$$\text{where } \Delta\mathbf{u}_{nT}(\theta) = \begin{pmatrix} \Delta Y_{n1} - \Delta\mathcal{X}_{nT}\pi \\ \Delta Y_{n2} - \rho\Delta Y_{n1} - \Delta X_{n2}\beta \\ \vdots \\ \Delta Y_{nT} - \rho\Delta Y_{nT} - \Delta X_{nT}\beta \end{pmatrix}.$$

Su and Yang (2007) show that the ML estimates under both random and fixed effects specifications are consistent and asymptotically normally distributed, under the assumption that the specification of  $\Delta Y_{n1}$  is correct. In principle, one could show that the estimates would not be consistent for a short panel if the initial specification were misspecified. However, Elhorst (2005) and Su and Yang (2007) have provided some Monte Carlo results to demonstrate that their proposed approximation could be valuable even with initial period misspecification.

### 12.3.2.3 GMM Estimation

The likelihood approach would yield efficient estimates under a correct specification of distributions. If consistency of an estimate is the main concern, simpler instrumental variable (IV) or 2SLS estimation is feasible. For a panel with a finite periods  $T$ , Mtl (2006) suggests a feasible generalized 2SLS approach for the estimation of dynamic panel data model with fixed effects and SAR disturbances after first-difference of data. His feasible 2SLS is based on three steps, which extends the three steps feasible GLS approach in Kapoor, Kelejian, and Prucha (2007) for the panel regression model with random component and SAR disturbances to the estimation of dynamic panel model. Instead of QML estimation, by using the IV in dynamic panel data literature, Lee and Yu (2014) consider the best GMM estimation of the SDPD model to cover the scenario that  $T$  is large but relatively small to  $n$ .

The GMM estimation has also the advantage of being computationally simpler and hence can be applied to the estimation of spatial panels with high order SLs. The inclusion of high order SLs can allow spatial dependence with different sources of interactions such as geographical contiguity and economic interaction. Lee and Yu (2014) consider an SDPD model with both individual and time effects

$$\begin{aligned} Y_{nt} = & \sum_{j=1}^p \lambda_{j0} W_{nj} Y_{nt} + \gamma_0 Y_{n,t-1} + \sum_{j=1}^p \rho_{j0} W_{nj} Y_{n,t-1} \\ & + X_{nt} \beta_0 + c_{n0} + \alpha_{t0} l_n + V_{nt}. \end{aligned} \quad (38)$$

The  $W_{nj}$ 's are  $n \times n$  spatial weights matrices for  $j = 1, \dots, p$ , which are nonstochastic and generate the dependence of  $y_{it}$ 's across spatial units. If  $p \geq 2$ , (38) has a high order SAR structure. The  $W_{nj}$ 's may or may not be row-normalized. Let  $[F_{T,T-1}, \frac{1}{\sqrt{T}}l_T]$  be the orthonormal matrix of the eigenvectors of  $J_T$ , where  $F_{T,T-1}$  takes the form of forward orthogonal difference (FOD) transformation (also known as the Helmert transformation). By  $[Y_{n1}^*, Y_{n2}^*, \dots, Y_{n,T-1}^*] = [Y_{n1}, Y_{n2}, \dots, Y_{nT}]F_{T,T-1}$  and  $[Y_{n0}^{(*,-1)}, Y_{n1}^{(*,-1)}, \dots, Y_{n,T-2}^{(*,-1)}] = [Y_{n0}, Y_{n1}, \dots, Y_{n,T-1}]F_{T,T-1}$  to eliminate individual effects and the  $J_n$  transformation to eliminate the time effects, the estimation equation is

$$\begin{aligned} J_n Y_{nt}^* = & \sum_{j=1}^p \lambda_{j0} J_n W_{nj} Y_{nt}^* + \gamma_0 J_n Y_{n,t-1}^{(*,-1)} + \sum_{j=1}^p \rho_{j0} J_n W_{nj} Y_{n,t-1}^{(*,-1)} \\ & + J_n X_{nt}^* \beta_0 + J_n V_{nt}^*, \quad t = 1, \dots, T-1, \end{aligned} \quad (39)$$

because  $J_n l_n = 0$ . Estimation of (39) by the ML method has two issues. First, when  $W_{nj}$  is not row-normalized, (39) would not have a well-defined SAR structure for  $J_n Y_{nt}^*$ . Second, in the presence of time lags variables, the predetermined variable  $J_n Y_{n,t-1}^{(*,-1)}$  in (39) are correlated with the disturbances after the data transformation by  $F_{T,T-1}$ . For these reasons, we see that a likelihood function could not be formed directly from (39).

Lee and Yu (2014) propose GMM estimation of (39) taking advantage of the forwarding feature of the Helmert transformation, which does not require an SAR form for  $J_n Y_{nt}^*$  and it can be free of asymptotic bias. As argued in Lee and Yu (2014), compared to QML estimation, the GMM estimation has the following merits for the SDPD model: (1) GMM has a computational advantage over MLE, because GMM does not need to compute the determinant of the Jacobian matrix in the likelihood function for a spatial model, which is especially inconvenient for MLE when  $n$  is large or the model has high order SLs; (2) some GMM methods can be applied to a short SDPD model and is free of asymptotic bias, while ML estimation of the SDPD model requires a large  $T$  and a bias correction procedure is needed to eliminate the asymptotic bias. For a finite  $T$  case, an initial specification for the first time period observations would also be needed in order to formulate a likelihood function. (3) for the case that  $T$  tends to infinity (but can be more slowly than  $n$ ), with carefully designed moment

conditions, the GMM estimate can be more efficient than the QML estimate when the true distribution of the disturbances are not normal and has a nonzero degree of excess kurtosis; (4) GMM is also applicable for the SDPD model with time effects and non-row-normalized spatial weights matrices.

Here, for a vector (or matrix)  $b_n$  with  $n$  rows, we denote  $\mathcal{W}_n^s b_n$  with  $s$  being an integer as a matrix consisting of vectors (or matrices)  $W_{n1}^{s_1} W_{n2}^{s_2} \cdots W_{np}^{s_p} b_n$  where  $s_1, s_2, \dots, s_p$  are nonnegative integers such that  $s_1 + s_2 + \dots + s_p = s$ . For example,

$$\begin{aligned}\mathcal{W}_n^2 b_n = [W_{n1}^2 b_n, \dots, W_{n1} W_{np} b_n, W_{n2} W_{n1} b_n, \dots, W_{n2} W_{np} b_n, \dots, \\ W_{np} W_{n1} b_n, \dots, W_{np}^2 b_n].\end{aligned}$$

For the estimation equation (39), we note that  $Y_{n,t-1}^{(*,-1)}$  is correlated with  $V_{nt}^*$ . For this reason, IVs are needed for  $Y_{n,t-1}^{(*,-1)}$  and  $W_{nk} Y_{n,t-1}^{(*,-1)}$  for each  $t$  (and also for  $W_{nj} Y_{nt}^*$ ). Therefore, for the estimation of (39), we need to find IVs for the regressors

$$J_n[\mathcal{W}_n Y_{nt}^*, Y_{n,t-1}^{(*,-1)}, \mathcal{W}_n Y_{n,t-1}^{(*,-1)}]. \quad (40)$$

In addition to all strictly exogenous variables  $X_{ns}$  for  $s = 1, \dots, T - 1$ , the time lag variables  $Y_{n0}, \dots, Y_{n,t-1}$  can also be used to construct IVs for  $Y_{n,t-1}^{(*,-1)}$  as in the literature of dynamic panel data models (Alvarez and Arellano 2003, etc). Correspondingly, we may use  $W_{nk} X_{ns}$  for  $s = 1, \dots, T - 1$  and  $W_{nk} Y_{ns}$  for  $s = 0, \dots, t - 1$  as IVs for  $W_{nk} Y_{n,t-1}^{(*,-1)}$ .

For the linear moments, an IV matrix for (40) can take the form  $J_n Q_{nt}$  where  $Q_{nt}$  has a fixed column dimension. For example,  $Q_{nt}$  could be  $[Y_{n,t-1}, \mathcal{W}_n Y_{n,t-1}, \mathcal{W}_n^2 Y_{n,t-1}, X_{nt}^*, \mathcal{W}_n X_{nt}^*, \mathcal{W}_n^2 X_{nt}^*]$ . We can also use the many moment approach. By denoting  $h_{nt} = (Y_{n0}, \dots, Y_{n,t-1}, X_{n1}, \dots, X_{nT}, l_n)$ , we can use the IV matrix  $J_n H_{nt}$  with  $H_{nt} = (h_{nt}, \mathcal{W}_n h_{nt}, \dots, \mathcal{W}_n^{p_n} h_{nt})$ , where  $p_n$  is the order of spatial power series expansion.

In addition to the linear moments, due to the spatial correlation in the DGP, quadratic moments can capture correlations and may increase the efficiency of estimates. The use of quadratic moments is motivated by the likelihood function of the SAR model under normality disturbances (Lee 2007), as well as the Moran test statistic (Moran 1950). The vector  $P_n V_{nt}^*$  can be uncorrelated with  $J_n V_{nt}^*$  in (39) for an  $n \times n$  nonstochastic matrix  $P_n$  satisfying the property  $\text{tr}(P_n J_n) = 0$ , while it may correlate with  $J_n \mathcal{W}_n Y_{nt}^*$  in (40).

For the best IVs of (40), the ideal IV matrix is its conditional mean, which is not directly available but can be approximated. Lee and Yu (2014) derive this conditional mean and propose a linear combination of predetermined and exogenous variables to approximate that, which is valid when  $T$  is large. Also, under the many moment approach, they show that many IVs will approximate this conditional mean, but with some bias related to the ratio of the number of IVs relative to the total sample size  $nT$ , which is essentially the ratio of  $T$  over  $n$ . For the quadratic moments, the best ones are also derived in Lee and Yu (2014).

### 12.3.3 Testing Spatial Effect

Testing hypotheses is an important component in econometric analyses of an estimated model in practice. With a well-specified general model, after estimation, one may use Wald's test procedures for the relevant hypotheses. For spatial models, one of the important hypotheses is to test the presence of spatial effects. Because the restricted model without spatial correlation would be computationally simpler, the LM test is in general preferred in place of the Wald procedure. For testing spatial correlation in the disturbances of a regression equation, Moran's test is a well-known and is an LM test. Currently, most of the testing procedures for spatial correlation in the literature focus on cross-sectional spatial models, but there are a few for static or dynamic spatial panels.

The Moran I test is well known in testing spatial correlation in a regression equation with possible spatial correlated disturbances. The test statistic given by Moran (1950) and Cliff and Ord (1973) is  $I = \frac{e_n' W_n e_n}{e_n' e_n}$ , where  $e_n$  is the OLS residual and  $W_n$  is a spatial weights matrix. The Moran I test is an LM statistic with an unscaled denominator. The LM interpretation of Moran I test of spatial correlation is due to Burridge (1980), where the standardized version  $LM = \frac{n}{\sqrt{S_{n0}}} \frac{e_n' W_n e_n}{e_n' e_n}$  is proposed with  $S_{n0} = \text{tr}(W_n' W_n + W_n^2)$ . Burridge (1980) has also pointed out the LM test for a moving average alternative, i.e.,  $U_n = \epsilon_n + \rho W_n \epsilon_n$ , is exactly the same as that given above.

Baltagi, Song, Jung, and Koh (2007c) develop the LM test for serial correlation and spatial autocorrelation in a random effects panel data model. Their model is a special case of (5) with the specification

$$\begin{aligned} Y_{nt} &= X_{nt}\beta_0 + \mu_n + U_{nt}, \\ U_{nt} &= \lambda_{20} W_{n2} U_{nt} + V_{nt} \text{ and } V_{nt} = \rho_0 V_{n,t-1} + e_{nt}. \end{aligned} \tag{41}$$

They derive several LM tests for this panel data regression model including a joint test for serial correlation, spatial autocorrelation, and random effects. This paper also derives conditional LM and LR tests that do not ignore the correlations of marginal hypotheses and contrast them with their marginal LM and LR counterparts. Thus, it extends the LM test in Baltagi, Song, and Koh (2003), where the serial correlation is not considered.

For the model (3) with different spatial effects in individual effects and disturbances, Baltagi, Egger, and Pfaffermayr (2008) develop a pretest estimator based on a sequence of LM tests. Specifically, the following hypotheses are considered: (1)  $H_0^A : \lambda_{30} = \lambda_{02} = 0$  vs.  $H_1^A$ : at least one of  $\lambda_{30}$  or  $\lambda_{20} \neq 0$ ; (2)  $H_0^B : \lambda_{30} = \lambda_{02}$  vs.  $H_1^B : \lambda_{30} \neq \lambda_{20}$ ; and (3)  $H_0^C : \lambda_{30} = 0$  vs.  $H_1^C : \lambda_{30} \neq 0$ . The  $H_0^A$  is to test whether there is any spatial effect in the disturbances. If  $H_0^A$  is not rejected, the pretest estimator would revert to the classical random effects model. If  $H_0^A$  is rejected, we can further test whether the two spatial effects in the error components are the same or not. If  $H_0^B$  is not rejected, it would become the Kapoor, Kelejian, and Prucha (2007) model. If  $H_0^B$  is rejected, then we will

test  $H_0^C$  to see whether there is the spatial effect in the individual effects are zero or not. If  $H_0^C$  is not rejected, then we will have the Anselin (1988) model. If  $H_0^C$  is rejected, we will then have the general model in Baltagi, Egger, and Pfaffermayr (2007b) where both  $\lambda_{03}$  and  $\lambda_{02}$  are present and might not be equal to each other.

Under the panel data setting, Baltagi and Yang (2013) have developed a standardized LM test, which corrects both the mean and variance of the existing LM tests under more relaxed assumptions on the error distributions. It is shown that these LM tests are not only robust against distributional misspecification, but are also quite robust against changes in the spatial layout.

Instead of focusing on the error components, Debarsy and Ertur (2010) consider tests for spatial autocorrelation in the fixed effects panel data model in (4). Several LM test statistics as well as their LR counterparts are developed to discriminate between endogenous SL effect  $\lambda_{10}$  versus spatially autocorrelated errors effect  $\lambda_{20}$ . They have developed the following joint, marginal and conditional tests: (1)  $H_0^a : \lambda_{10} = \lambda_{20} = 0$ ; (2)  $H_0^b : \lambda_{10} = 0$  assuming  $\lambda_{20} = 0$ ; (3)  $H_0^c : \lambda_{20} = 0$  assuming  $\lambda_{10} = 0$ ; (4)  $H_0^d : \lambda_{10} = 0$  conditional on  $\lambda_{20} \neq 0$ ; and (5)  $H_0^e : \lambda_{20} = 0$  conditional on  $\lambda_{10} \neq 0$ .

There are some general tests for cross-sectional dependence in panel data setting, which might detect not only spatial correlations but also unobserved common factors. Details can be found in Pesaran (2004), Pesaran Ullah, and Yamagata (2008), Sarafidis, Yamagata, and Robertson (2009), and Chudik, Pesaran, and Tosetti (2011) among others. Also, Baltagi, Bresson, and Pirotte (2007a) study the performance of panel unit root tests when spatial effects are present that account for cross-section correlation. Monte Carlo simulations show that there can be considerable size distortions in panel unit root tests when the true specification exhibits spatial error correlation.

## 12.4 OTHER TOPICS

Lee and Yu (2012b) investigate the QML estimation of SDPD models where spatial weights matrices can be time varying. This is especially relevant when weights matrices are constructed from economic/socioeconomic distances or demographic characteristics. However, when using economic distances to construct the spatial weights matrix, we might have a separate issue of endogeneity as these economic distances could be endogenous. This calls for the estimation of spatial models with *endogenous spatial weights matrix*, which is an interesting topic for future research.

For a spatial panel data set, some spatial units or variables might be missing. For missing observations, researchers suggest different methods of imputation. For spatial panels, Wang and Lee (2013) investigate the case of static spatial panels with randomly missing dependent variables. A nonlinear least squares (NLS) method is suggested and a GMM estimation is developed for the model. A 2SLS estimation with imputation is proposed as well. They analytically compare these estimation methods and find that

the generalized NLS, best generalized 2SLS with imputation, and best GMM estimators have identical asymptotic variances. The robustness of these estimation methods against unknown heteroskedastic variances is also stressed. EM method is feasible but will not be consistent in the presence of unknown heteroskedasticity.

Forecasting has its applications in spatial panel data models. For a linear regression model  $y_t = x_t \beta_0 + \epsilon_t$ ,  $t = 1, \dots, T$ , with the covariance matrix  $\Omega$  for  $\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_T)'$ . Goldberger (1962) showed that the best linear unbiased predictor (BLUP) for  $y_{T+\tau}$  is

$$\hat{y}_{T+\tau} = x_{T+\tau} \hat{\beta}_{GLS} + \omega \Omega^{-1} \hat{\epsilon}_{GLS}, \quad (42)$$

where  $\hat{\beta}_{GLS}$  and  $\hat{\epsilon}_{GLS}$  are GLS estimator and residuals, and  $\omega$  is the row vector of covariances of  $\epsilon_{T+\tau}$  with  $\epsilon$ . The  $x_{T+\tau} \hat{\beta}_{GLS}$  estimates the expected value of  $y_{T+\tau}$  given  $x_{T+\tau}$ , and the  $\omega \Omega^{-1} \hat{\epsilon}_{GLS}$  utilizes a priori knowledge of the interdependence of disturbances to predict future disturbances. Fingleton (2009) considers prediction in a spatial panel data with Kapoor, Kelejian, and Prucha (2007)'s specification, with an additional SL regressor. Baltagi, Bresson, and Pirotte (2012) investigate the BLUP for a linear regression panel data with spatial error components. For the fixed effects model, the second component  $\omega \Omega^{-1}$  in (42) does not show up because it is a static process, and the  $x_{T+\tau} \hat{\beta}_{GLS} + \hat{c}_i$  which takes into account the fixed effect, is used to replace  $x_{T+\tau} \hat{\beta}_{GLS}$  for the first component in (42). For the random effects model with a finite (short) time  $T$ , one does not estimate the individual effect  $c_i$ , but explores the second component  $\omega \Omega^{-1}$  as it captures the correlation for future and current disturbances due to the individual random effects. Baltagi and Li (2004, 2006) investigate a panel model with spatial correlation and apply it respectively, to investigate the cigarettes demand equation in the United States and a liquor demand equation. Instead of the error components model, Baltagi, Fingleton, and Pirotte (2011) investigate forecasting under a dynamic panel data model with a SL term in the regression equation.

Under Baltagi, Bresson, and Pirotte (2012), due to the static feature of the model, the spatial effect is irrelevant in the formula of BLUP, so that the prediction does not involve the spatial weights matrix  $W_n$  which is the same as for the panel regression model with classical error components but without spatial effect. While the predictor formulae are the same, the MLEs would be different with or without spatial effect, which results in different estimates of parameters and estimated residuals and hence different forecasts with a sample. With a dynamic feature with diffusion over time, it is possible that the spatial effect (hence  $W_n$ ) will play a role in the prediction formula in general. Under the panel regression model with the nonseparable space-time filter in disturbances, Lee and Yu (2013) show that the spatial weights matrix  $W_n$  would be present in the second component  $\omega \Omega^{-1}$  in (42). However, under separable space-time filter case, the  $W_n$  will not play a role in  $\omega \Omega^{-1}$ , even though a dynamic feature is still present in the disturbance.

The research on spatial econometrics is expanding, both in cross-sectional and panel data models. There are some ongoing research on cross-sectional spatial econometrics, which is also promising for spatial panel data models. Here we mention only a few.

Bester, Conley, and Hansen (2011) presents an inference approach for dependent data in time series, spatial, and panel data applications. They construct the  $t$  and Wald statistics using a cluster covariance matrix estimator, which is consistent under *heteroskedasticity and autocorrelation*. Their approach is similar to Conley (1999), where a class of nonparametric, positive semi-definite covariance matrix estimators is developed that allows for general forms of dependence characterized by economic distance. Kelejian and Prucha (2007) suggests a nonparametric heteroskedasticity and autocorrelation consistent (HAC) estimator of the variance–covariance (VC) matrix for a vector of sample moments within a spatial context, which can be applied to a typical SAR model. Kim and Sun (2011) consider spatial heteroskedasticity and auto-correlation consistent (spatial HAC) estimation of covariance matrices of parameter estimators. They generalize the spatial HAC estimator introduced by Kelejian and Prucha (2007) to apply to linear and nonlinear spatial models with moment conditions. Based on the asymptotic truncated MSE criterion, they derive the optimal bandwidth parameter and suggest its data dependent estimation procedure using a parametric plug-in method. Vogelsan (2012) develops an asymptotic theory for test statistics in linear panel models that are robust to heteroskedasticity, autocorrelation and/or spatial correlation. Two classes of standard errors are analyzed, which are based on nonparametric heteroskedasticity autocorrelation covariance matrix estimators. The first class is based on averages of HAC estimators across individuals in the cross-section, and the second class is based on the HAC of cross-section averages.

Allers and Elhorst (2011) use a cross-sectional *simultaneous equations model* (SEM) to investigate fiscal policy interaction due to budget constraint and cost differences between jurisdictions. Jeanty, Partridge, and Irwin (2010) investigate local interactions between housing prices and population migration with their simultaneous and spatially interdependent relationship. In the panel data setting, Baltagi and Bresson (2011) investigate ML estimation and LM tests for panel seemingly unrelated regressions (SUR) with SL and spatial errors. Their proposed model is used to study hedonic housing prices in Paris. Baltagi and Pirotte (2011) investigate various estimators using panel data SUR with spatial error correlation. The true data-generating process is assumed to be SUR with spatial errors of the auto-regressive or moving average type. Moreover, the remainder term of the spatial process is assumed to follow an error component structure. Both ML and GMM estimations are used. Baltagi and Deng (2011) derive the EC3SLS estimator for a simultaneous system of spatial auto-regressive equations with random effects.

Robinson (2010) develops efficient *semiparametric and parametric estimates* for an SAR model, where series nonparametric estimates of the score function are employed for adaptive estimation of parameters of interest. These estimates can be as efficient as the ones based on a correct form; in particular, they are more efficient than QMLE at non-Gaussian distributions. In Robinson (2011), a nonparametric regression with spatial data is considered in a setting where the conditional mean of a dependent variable,

given explanatory ones, is a nonparametric function, while the conditional covariance reflects spatial correlation. Sufficient conditions are established for consistency and asymptotic normality of the kernel regression estimates. Su and Jin (2010) propose profile QML estimation of spatial auto-regressive models that are partially linear. Su (2012) proposes semiparametric GMM estimation of semiparametric SAR models under weak moment conditions. In comparison with the QML-based semiparametric estimator of Su and Jin (2010), Su (2012) allow for both heteroskedasticity and spatial dependence in the error terms. In the panel data setting, Robinson (2012) shows that a simple smoothed nonparametric trend estimate is dominated by an estimate which exploits availability of cross-sectional data. Asymptotically optimal bandwidth choices are justified for both estimates, where feasible optimal bandwidths, and feasible optimal trend estimates, are asymptotically justified.

## 12.5 CONCLUSION

---

This chapter briefly describes some recent developments in spatial panel data models. We first investigate different model specifications in the current literature, for both static and dynamic panel data models with spatial interactions. These model specifications differ in terms of the location of spatial effect (i.e., whether it is in the regression equation or disturbances; whether it is an SAR or SMA process; whether the individual effects have implicit spatial correlation). We then focus on the estimation for various static and dynamic models and briefly review testing procedures. For the static model, we can have estimation methods designed for fixed effects or random effects specifications. The random effects estimator is a pooling of fixed effects estimator and between equation estimator. For the dynamic models, the asymptotic properties of estimators are reviewed for stable and nonstable models. We finally conclude with some further issues and future research.

## NOTES

---

1. These cases would yield different rate of convergence for some estimates. See Section 12.3 for details.

## REFERENCES

---

- Allers, M. and J. Elhorst. 2011. A simultaneous equations model of fiscal policy interactions. *Journal of Regional Science* 51, 271–291.
- Alvarez, J. and M. Arellano. 2003. The time series and cross-section asymptotics of dynamic panel data estimators. *Econometrica* 71, 1121–1159.

- Anselin, L. 1988. *Spatial Econometrics: Methods and Models*. The Netherlands: Kluwer Academic: Berlin.
- Anselin, L. 2001. Spatial econometrics, in Badi H. Baltagi (ed.), *A Companion to Theoretical Econometrics*, 310–330, Blackwell Publishers Lte: Oxford.
- Anselin, L., J. Le Gallo, and H. Jayet. 2008. Spatial panel econometrics. chapter 19, in L., Matyas and P. Sevestre (eds.). *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*, pp. 625–660, Berlin: Springer.
- Arellano, M., 1993. On the testing of correlated effects with panel data. *Journal of Econometrics* 59, 87–97.
- Baicker, K., 2005. The spillover effects of state spending. *Journal of Public Economics* 89, 529–544.
- Baltagi, B. H. 2011. Spatial panels, in A. Ullah, and D. E. A. Giles (eds.), *Handbook on Empirical Economics and Finance*, pp. 397–434. New York: Taylor & Francis Group.
- Baltagi, B. and G. Bresson, 2011. Maximum likelihood estimation and Lagrange multiplier tests for panel seemingly unrelated regressions with spatial lag and spatial errors: An application to hedonic housing prices in Paris. *Journal of Urban Economics* 69, 24–42.
- Baltagi, B. and Y. Deng, 2011. EC3SLS estimator for a simultaneous system of spatial autoregressive equations with random effects. Manuscript.
- Baltagi, B. H. and D. Li. 2004. Prediction in the panel data model with spatial correlation, Chapter 13. in L. Anselin, R. J. G. M. Florax, and S. J. Rey (eds.). *Advances in Spatial Econometrics: Methodology, Tools and Applications*. pp. 283–295. Berlin: Springer.
- Baltagi, B. H. and D. Li. 2006. Prediction in the panel data model with spatial correlation: The case of liquor. *Spatial Economic Analysis* 1, 175–185.
- Baltagi, B. H. and A. Pirotte. 2011. Seemingly unrelated regressions with spatial error components. *Empirical Economics* 40, 5–49.
- Baltagi, B. H. and Z. Yang. 2013. Standardized LM tests for spatial error dependence in linear or panel regressions. *The Econometrics Journal*, Royal Economic Society. vol. 16(1), pages 103–134, 02.
- Baltagi, B. H., G. Bresson, and A. Pirotte. 2007a. Panel unit root tests and spatial dependence. *Journal of Applied Econometrics* 22, 339–360.
- Baltagi, B. H., G. Bresson, and A. Pirotte. 2012. Forecasting with spatial panel data. *Computational Statistics and Data Analysis* 56, 3381–3397.
- Baltagi, B. H., P. Egger, and M. Pfaffermayr. 2007b. A generalized spatial panel data model with random effects. Working paper.
- Baltagi, B. H., P. Egger, and M. Pfaffermayr. 2008. A Monte Carlo study for pure and pretest estimators of a panel data model with spatially autocorrelated disturbances. *Annales d'Economie et de Statistique* 87/88: 11–38.
- Baltagi, B. H., B. Fingleton, and A. Pirotte. 2011. Estimating and forecasting with a dynamic spatial panel data model. Manuscript.
- Baltagi, B. H., S. H. Song, and W. Koh., 2003. Testing panel data regression models with spatial error correlation. *Journal of Econometrics* 117, 123–150.
- Baltagi, B. H., S. H. Song, B. C. Jung, and W. Koh. 2007c. Testing for serial correlation, spatial autocorrelation and random effects using panel data. *Journal of Econometrics* 140, 5–51.
- Bester, C. A., T. G. Conley, and C. B. Hansen. 2011. Inference with dependent data using cluster covariance estimators. *Journal of Econometrics*, 165, 137–151.

- Bhargava, A. and J. D. Sargan. 1983. Estimating dynamic random effects models from panel data covering short time periods. *Econometrica* 51, 1635–1659.
- Burridge, P. 1980. On the Cliff-Ord test for spatial correlation. *Journal of the Royal Statistical Society, Series B*, 42, 1, 107–108.
- Case, A., J. R. Hines, and H. S. Rosen. 1993. Budget spillovers and fiscal policy interdependence: Evidence from the states. *Journal of Public Economics* 52, 285–307.
- Chen, X. and T. Conley, 2002. A new semiparametric spatial model for panel time series. *Journal of Econometrics* 105, 59–83.
- Chudik, A., M. H. Pesaran, and E. Tosetti. 2011. Weak and strong cross-section dependence and estimation of large panels. *The Econometrics Journal* 14, C45–C90.
- Cliff, A. D. and J. K. Ord. 1973. *Spatial Autocorrelation*. London: Pion Ltd.
- Conley, T. G. 1999. GMM estimation with cross-sectional dependence. *Journal of Econometrics* 92, 1–45.
- Cox, D.R. 1975. Partial likelihood. *Biometrika* 62, 269–276.
- Debarsy, N. 2012. The Mundlak approach in the spatial Durbin panel data model. *Spatial Economic Analysis* 7, 109–132.
- Debarsy, N., and C. Ertur. 2010. Testing for spatial autocorrelation in a fixed effects panel data model. *Regional Science and Urban Economics* 40, 453–470.
- Debarsy, N., C. Ertur, and J. LeSage. 2012. Interpreting dynamic space-time panel data models. *Statistical Methodology* 9, 158–171.
- Druska, V. and W. C. Horrace. 2004. Generalized moments estimation for spatial panel data: Indonesian rice farming. *American Journal of Agricultural Economics* 86, 185–198.
- Elhorst, J. 2005. Unconditional maximum likelihood estimation of linear and log-linear dynamic models for spatial panels. *Geographical Analysis* 37, 85–106.
- Elhorst, J. 2008. Serial and spatial error correlation. *Economics Letters* 100, 422–424.
- Elhorst, J. 2010a. Spatial panel data models., chapter C.2, in M. M. Fischer and A. Getis (eds.), *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*. London: pp. 377–407. Springer Heidelberg Dordrecht.
- Elhorst, J. 2010b. Dynamic panels with endogenous interaction effects when  $T$  is small. *Regional Science and Urban Economics* 40, 272–282.
- Elhorst, J. 2012. Dynamic spatial panels: Models, methods, and inferences. *Journal of Geographical Systems* 14, 5–28.
- Ertur C. and W. Koch., 2007. Growth, technological Interdependence and spatial externalities: Theory and evidence. *Journal of Applied Econometrics* 22, 1033–1062.
- Fingleton, B. 2008. A generalized method of moments estimators for a spatial panel model with an endogenous spatial lag and spatial moving average errors. *Spatial Economic Analysis* 3, 27–44.
- Fingleton, B. 2009. Prediction using panel data regression with spatial random effects. *International Regional Science Review* 32, 195–220.
- Franzese, R. J. 2007. Spatial econometric models of cross-sectional interdependence in political science panel and time-series-cross-section data. *Political Analysis* 15, 140–164.
- Frazier, C. and K.M. Kockelman, 2005. Spatial econometric models for panel data: Incorporating spatial and temporal data. *Transportation Research Record: Journal of the Transportation Research Board* 1902/2005, 80–90.

- Goldberger, A. S. 1962. Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association* 57, 369–375.
- Hahn, J., and G. Kuersteiner,. 2002. Asymptotically unbiased inference for a dynamic panel model with fixed effects when both  $n$  and  $T$  are large. *Econometrica* 70, 1639–1657.
- Holly, S., M. H. Pesaran, and T. Yamagata. 2010. A spatio-temporal model of house prices in the USA. *Journal of Econometrics* 158, 160–173.
- Holly, S., M. H. Pesaran, and T. Yamagata. 2011. The spatial and temporal diffusion of house prices in the UK. *Journal of Urban Economics* 69, 2–23.
- Hsiao, C. 1986. *Analysis of Panel Data*. Cambridge: Cambridge University Press.
- Hsiao, C., M. H. Pesaran, and A. K. Tahmisioglu. 2002. Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods. *Journal of Econometrics* 109, 107–150.
- Jeanty, P., M. Partridge, and E. Irwin. 2010. Estimation of a spatial simultaneous equation model of population migration and housing price dynamics. *Regional Science and Urban Economics* 40, 343–352.
- Kapoor, N.M., H. H. Kelejian, and I.R. Prucha. 2007. Panel data models with spatially correlated error components. *Journal of Econometrics* 140, 97–130.
- Kelejian, H. H. and I.R. Prucha, 2007. HAC estimation in a spatial framework. *Journal of Econometrics* 140, 131–154.
- Keller, W. and C. H. Shiue, 2007. The origin of spatial interaction. *Journal of Econometrics* 140, 304–332.
- Kim, M. S. and Y. Sun. 2011. Spatial heteroskedasticity and autocorrelation consistent estimation of covariance matrix. *Journal of Econometrics* 160, 349–371.
- Korniotis, G.M. 2010. Estimating panel models with internal and external habit formation. *Journal of Business and Economic Statistics* 28, 145–158.
- Lee, L. F. 2007. GMM and 2SLS estimation of mixed regressive, spatial autoregressive models. *Journal of Econometrics* 137, 489–514.
- Lee, L. F. and J. Yu., 2010a. Estimation of spatial autoregressive panel data models with fixed effects. *Journal of Econometrics* 154, 165–185.
- Lee, L. F. and J. Yu., 2010b. A spatial dynamic panel data model with both time and individual fixed effects. *Econometric Theory* 26, 564–597.
- Lee, L. F. and J. Yu. 2010c. Some recent developments in spatial panel data models. *Regional Science and Urban Economics* 40, 255–271.
- Lee, L. F. and J. Yu., 2011a. Estimation of spatial panels. *Foundations and Trends in Econometrics*, NOW publisher.
- Lee, L. F. and J. Yu. 2011b. A Unified Transformation Approach for the Estimation of Spatial Dynamic Panel Data Models: Stability, Spatial Cointegration and Explosive Roots, in: A. Ullah and D.E.A. Giles (eds.), *Handbook on Empirical Economics and Finance*, pp. 397–434. New York: Taylor & Francis Group.
- Lee, L. F. and J. Yu. 2012a. Spatial panels: Random components vs. fixed effects. *International Economic Review* 53, 1369–1412.
- Lee, L. F. and J. Yu. 2012b. QML estimation of spatial dynamic panel data models with time varying spatial weights matrices. *Spatial Economic Analysis* 7, 31–74.
- Lee, L. F. and J. Yu. 2013. Estimation of panel regression models with separable and non-separable spatial-time filters. Manuscript.

- Lee, L.F. and J. Yu, 2014. Efficient GMM estimation of spatial dynamic panel data models with fixed effects. *Journal of Econometrics* 180, 174–197.
- LeSage, J. and K. Pace. 2009. *Introduction to Spatial Econometrics*. CRC Press/Taylor & Francis Group: Boca Raton.
- Maddala, G.S. 1971. The use of variance components models in pooling cross subsection and time series data. *Econometrica* 39, 341–358.
- Magnus, J.R. 1982. Multivariate error components analysis of linear and nonlinear regression models by maximum likelihood. *Journal of Econometrics* 19, 239–285.
- Moran, P. A. P. 1950. Notes on continuous stochastic phenomena. *Biometrika* 37, 17–33.
- Mutl, J. 2006. Dynamic panel data models with spatially correlated disturbances. PhD thesis, University of Maryland, College Park.
- Mutl, J. and M. Pfaffermayr. 2011. The Hausman test in a Cliff and Ord panel model. *Econometrics Journal* 14, 48–76.
- Neyman, J. and E.L. Scott. 1948. Consistent estimates based on partially consistent observations. *Econometrica* 16, 1–32.
- Parent, O. and J. LeSage., 2010. A spatial dynamic panel model with random effects applied to commuting times. *Transportation Research Part B: Methodological* 44, 633–645.
- Parent, O.,and J. LeSage. 2012. Spatial dynamic panel data models with random effects. *Regional Science and Urban Economics* 42, 727–738.
- Pesaran, M. H. 2004. General diagnostic test for cross-section dependence in panels. Working Paper, University of Cambridge & USC.
- Pesaran, M. H. 2006. Estimation and inference in large heterogenous panels with multifactor error structure. *Econometrica* 74, 967–1012.
- Pesaran, M. H. and E. Tosetti. 2011. Large panels with common factors and spatial correlation. *Journal of Econometrics* 161, 182–202.
- Pesaran, M. H. A. Ullah., and T. Yamagata. 2008. Bias-adjusted LM test of error cross-section independence. *Econometrics Journal* 11, 105–127.
- Rincke, J. 2010. A commuting-based refinement of the contiguity matrix for spatial models, and an application to local police expenditures. *Regional Science and Urban Economics* 40, 324–330.
- Robinson, P. M. 2010. Efficient estimation of the semiparametric spatial autoregressive model. *Journal of Econometrics* 157, 6–17.
- Robinson, P. M. 2011. Asymptotic theory for nonparametric regression with spatial data. *Journal of Econometrics* 165, 5–19.
- Robinson, P. M. 2012. Nonparametric trending regression with cross-sectional dependence. *Journal of Econometrics* 169, 4–14.
- Ruud, P.A. 2000. *An Introduction to Econometric Theory*. New York: Oxford University Press.
- Sarafidis, V. T. Yamagata., and D. Robertson. 2009. A test of cross-section dependence for a linear dynamic panel model with regressors. *Journal of Econometrics* 148, 149–161.
- Su, L. 2012. Semiparametric GMM estimation of spatial autoregressive models. *Journal of Econometrics* 167, 543–560.
- Su, L. and S. Jin. 2010. Profile quasi-maximum likelihood estimation of partially linear spatial autoregressive models. *Journal of Econometrics* 157, 18–33.
- Su, L. and Z. Yang. 2007. QML estimation of dynamic panel data models with spatial errors. Manuscript, Singapore Management University.
- Vogelsan., T. J. 2012. Heteroskedasticity., autocorrelation, and spatial correlation robust inference in linear panel models with fixed-effects. *Journal of Econometrics* 166, 303–319.

- Wildasin., D. 2003. Fiscal competition in space and time. *Journal of Public Economics* 87, 2571–2588.
- Wang, W. and L.F. Lee, 2013. Estimation of spatial panel data models with randomly missing data in the dependent variable. *Regional Science and Urban Economics* 43, 521–538.
- Wong., W.H. 1986. Theory of partial likelihood. *The Annals of Statistics* 14, 88–123.
- Yu., J. and L. F. Lee. 2010. Estimation of unit root spatial dynamic data models. *Econometric Theory* 26, 1332–1362.
- Yu., J. and L. F. Lee. 2012. Convergence: a spatial dynamic panel data approach. *Global Journal of Economics* 1, 125006 (36 pages).
- Yu., J. R. de Jong., and L.F. Lee. 2008. Quasi-maximum likelihood estimators for spatial dynamic panel data with fixed effects when both  $n$  and  $T$  are large. *Journal of Econometrics* 146, 118–134.
- Y, J., R. de Jong., and L.F. Lee. 2012. Estimation for spatial dynamic panel data with fixed effects: The case of spatial cointegration. *Journal of Econometrics* 167, 16–37.

## CHAPTER 13

---

# RANDOM COEFFICIENTS MODELS IN PANELS

---

CHENG HSIAO

### 13.1 INTRODUCTION

---

PANEL data focus on individual outcomes. Factors affecting individual outcomes are numerous. If all essential factors ( $x, w, z, \dots$ ) affecting individual outcomes are used as conditioning variables, then presumably one can assume that the conditional distribution of  $y, f(y_{it} | x_{it}, w_{it}, z_{it}, \dots), i = 1, \dots, N$ , and  $t = 1, \dots, T$ , are independently identically distributed across  $i$  and over  $t$ . However, a model is not a mirror. It is a simplification of reality. To capture the essential relations between  $y$  and  $x$ , it may not be feasible to also include  $z, w, \dots$  as conditioning variables because of shortages of degree of freedom or multicollinearity, etc. Moreover  $z, w$  may simply not be observable.

Standard random or fixed effects models (referred to as variable intercept models in Hsiao 2003, chapters 3 and 4) for the analysis of panel data assume the heterogeneity across individuals and/or over time that are not captured by the  $K$  conditioning variables,  $x$ , do not interact with the variation of them. A more general framework would be to let the coefficients that also vary with  $i$  and  $t$ , say, in the case of linear model, we can let

$$y_{it} = x_{it}\beta_{\tilde{i}t} + u_{it}, \quad i = 1, \dots, N, \\ t = 1, \dots, T. \quad (1.1)$$

However, (1.1) is not estimable if no structure is imposed on  $\beta_{\tilde{i}t}$ . A parsimonious specification that takes account unobserved heterogeneity is to let  $\beta_{\tilde{i}t}$  be a random variable with common mean  $\bar{\beta}$ .

For ease of exposition we shall assume that the  $K \times 1$  parameters,  $\beta_{\tilde{i}t} = \beta_{\tilde{i}}$  is time invariant with mean

$$E\beta_{\tilde{i}} = \bar{\beta}, \quad (1.2)$$

and is independently distributed over  $i$  with covariance matrix

$$E(\tilde{\beta}_i - \bar{\beta})(\tilde{\beta}_i - \bar{\beta})' = \Delta. \quad (1.3)$$

In Section 13.2 we consider linear static models under this framework. Section 13.3 considers dynamic models. Section 13.4 provides an example demonstrating the importance of taking account of parameter heterogeneity in empirical analysis. Section 13.5 considers a test for heterogeneity and whether we should treat unobserved differences as fixed or random. Concluding remarks are in section 13.6.

## 13.2 LINEAR STATIC MODEL

---

Let  $\tilde{\beta}_i = \bar{\beta} + \alpha_i$ . We assume that  $\alpha_i$  and the  $K \times 1$  explanatory variables  $\tilde{x}_{it}$  are independent,

$$E\alpha_i \tilde{x}'_{it} = 0. \quad (2.1)$$

For ease of exposition, we assume that  $u_{it}$  is independently distributed across  $i$  and over  $t$  with mean 0 and variance  $\sigma_i^2$ .

### 13.2.1 Sampling Approach

Stacking all NT observations, we have

$$\begin{aligned} \tilde{y} &= \tilde{X}\tilde{\beta} + \tilde{u} \\ &= X\bar{\beta} + \tilde{X}\alpha + \tilde{u}, \end{aligned} \quad (2.2)$$

where

$$\begin{aligned} \underset{NT \times 1}{\tilde{y}} &= (\tilde{y}'_1, \dots, \tilde{y}'_N)' , \quad \tilde{y}'_i = (y_{i1}, \dots, y_{iT}), \\ \underset{NT \times K}{X} &= \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}, \quad \underset{NT \times NK}{\tilde{X}} = \begin{bmatrix} X_1 & & & 0 \\ & X_2 & & \\ & & \ddots & \\ 0 & & & X_N \end{bmatrix} = \text{diag}(X_1, \dots, X_N), \end{aligned}$$

$\tilde{u} = (\tilde{u}'_1, \dots, \tilde{u}'_N)'$ ,  $\tilde{u}'_i = (u_{i1}, \dots, u_{iT})'$ ,  $\tilde{\beta}' = (\tilde{\beta}'_1, \dots, \tilde{\beta}'_N)'$ , and  $\alpha = (\alpha'_1, \dots, \alpha'_N)'$ . The covariance matrix for the composite disturbance term  $\tilde{X}\alpha + \tilde{u}$  is block-diagonal, with the  $i$ th diagonal block given by

$$\Phi_i = X_i \Delta X_i' + \sigma_i^2 I_T. \quad (2.3)$$

Under the assumption (2.1), the least squares estimator of  $\bar{\beta}$  is consistent. However, it is inefficient. The best linear unbiased estimator of  $\bar{\beta}$  of (2.2) is the generalized least squares (GLS) estimator

$$\begin{aligned}\hat{\beta}_{GLS} &= \left( \sum_{i=1}^N X_i' \Phi_i^{-1} X_i \right)^{-1} \left( \sum_{i=1}^N X_i' \Phi_i^{-1} \gamma_i \right) \\ &= \sum_{i=1}^N W_i \hat{\beta}_i,\end{aligned}\tag{2.4}$$

where

$$W_i = \left\{ \sum_{i=1}^N [\Delta + \sigma_i^2 (X_i' X_i)^{-1}]^{-1} \right\}^{-1} [\Delta + \sigma_i^2 (X_i' X_i)^{-1}]^{-1},\tag{2.5}$$

and

$$\hat{\beta}_i = (X_i' X_i)^{-1} X_i' \gamma_i.\tag{2.6}$$

The last expression of (2.4) shows that the GLS estimator is a matrix-weighted average of the least-squares estimator for each cross-sectional unit, with the weights inversely proportional to their covariance matrices. It also shows that the GLS estimator requires only a matrix inversion of order  $K$ , and so it is not much more complicated to compute than the simple least-squares estimator.

The covariance matrix for the GLS estimator is

$$\begin{aligned}\text{Var}(\hat{\beta}_{GLS}) &= \left( \sum_{i=1}^N X_i' \Phi_i^{-1} X_i \right)^{-1} \\ &= \left\{ \sum_{i=1}^N [\Delta + \sigma_i^2 (X_i' X_i)^{-1}]^{-1} \right\}^{-1}.\end{aligned}\tag{2.7}$$

In general,  $\Delta$  and  $\sigma_i^2$  are unknown. Swamy (1970) proposed using the least-squares estimators  $\hat{\beta}_i = (X_i' X_i)^{-1} X_i' \gamma_i$  and their residuals  $\hat{u}_i = \gamma_i - X_i \hat{\beta}_i$  to obtain unbiased estimators of  $\sigma_i^2$  and  $\Delta$ ,

$$\begin{aligned}\hat{\sigma}_i^2 &= \frac{\hat{u}_i' \hat{u}_i}{T - K} \\ &= \frac{1}{T - K} \gamma_i' [I - X_i (X_i' X_i)^{-1} X_i'] \gamma_i,\end{aligned}\tag{2.8}$$

$$\begin{aligned}\hat{\Delta} &= \frac{1}{N-1} \sum_{i=1}^N \left( \hat{\beta}_i - N^{-1} \sum_{i=1}^N \hat{\beta}_i \right) \\ &\quad \left( \hat{\beta}_i - N^{-1} \sum_{i=1}^N \hat{\beta}_i \right)' - \frac{1}{N} \sum_{i=1}^N \hat{\sigma}_i^2 (X_i' X_i)^{-1}. \end{aligned} \quad (2.9)$$

Again, just as in the error-component model, the estimator (2.9) is not necessarily nonnegative definite. In this situation, Swamy (also see Judge et al. 1980) has suggested replacing (2.9) by

$$\hat{\Delta} = \frac{1}{N-1} \sum_{i=1}^N \left( \hat{\beta}_i - N^{-1} \sum_{i=1}^N \hat{\beta}_i \right) \left( \hat{\beta}_i - N^{-1} \sum_{i=1}^N \hat{\beta}_i \right)'. \quad (2.10)$$

This estimator, although not unbiased, is nonnegative definite and is consistent when both  $N$  and  $T$  tend to infinity.

Swamy (1970) proved that substituting  $\hat{\sigma}_i^2$  and  $\hat{\Delta}$  for  $\sigma_i^2$  and  $\Delta$  in (2.4) yields an asymptotically normal and efficient estimator of  $\bar{\beta}$ . The speed of convergence of the GLS estimator is  $N^{1/2}$ . This can be seen by noting that the inverse of the covariance matrix for the GLS estimator ((2.7)) is

$$\begin{aligned}\text{Var}(\hat{\beta}_{GLS})^{-1} &= N\Delta^{-1} - \Delta^{-1} \left[ \sum_{i=1}^N \left( \Delta^{-1} + \frac{1}{\sigma_i^2} X_i' X_i \right)^{-1} \right] \Delta^{-1} \\ &= O(N) - O(N/T).\end{aligned} \quad (2.11)$$

### 13.2.2 Bayesian Approach

Under the assumptions (1.1)–(1.3) and (2.1), there is a natural Bayesian interpretation:

- A1. Conditional on  $X_i, \beta_i, y_i$  is independently normally distributed across  $i$  with mean  $X_i \beta_i$  and covariance matrix  $C_i$ ,

$$P(y_i | X_i, \beta_i) \sim N(X_i \beta_i, C_i). \quad (2.12)$$

- A2. The prior of  $\beta_i$  is normally distributed with mean  $\bar{\beta}$  and covariance matrix  $\Delta$

$$P(\beta_i) \sim N(\bar{\beta}, \Delta). \quad (2.13)$$

- A3. There is no information on  $\bar{\beta}$ ,

$$P(\bar{\beta}) \propto \text{constant}. \quad (2.14)$$

Under A1 – A3,<sup>1</sup>

- (i) The posterior distribution of  $\tilde{\beta}$  given  $\tilde{\gamma}, X, C_i$ , and  $\Delta$  is

$$P(\tilde{\beta} | \tilde{\gamma}, X) \sim N(\tilde{\beta}^*, D_1), \quad (2.15)$$

where

$$\tilde{\beta}^* = D_1 \left\{ \sum_{i=1}^N X'_i (X_i \Delta X_i + C_i)^{-1} \tilde{\gamma}_i \right\}, \quad (2.16)$$

$$D_1 = \left\{ \sum_{i=1}^N X'_i (X_i \Delta X_i + C_i)^{-1} X_i \right\}^{-1} \quad (2.17)$$

- (ii) The posterior distribution of  $\tilde{\beta}$  conditional on  $\tilde{\gamma}, \tilde{X}, \tilde{\beta}, \Delta, C_i$ , is  $N(\tilde{\beta}^*, D_2)$ , where

$$\tilde{\beta}^* = D_2 \left\{ \tilde{X}' \Omega^{-1} \tilde{\gamma} + (I_N \otimes \Delta^{-1}) A \tilde{\beta} \right\}, \quad (2.18)$$

$$D_2 = \left\{ \tilde{X}' \Omega^{-1} \tilde{X} + (I_N \otimes \Delta^{-1}) \right\}^{-1}, \quad (2.19)$$

where

$$\Omega = \begin{pmatrix} C_1 & \tilde{0} & \cdot & \tilde{0} \\ \tilde{0} & C_2 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \tilde{0} \\ \tilde{0} & \tilde{0} & \tilde{0} & C_N \end{pmatrix}, \quad {}_{NK \times K}^A = \begin{pmatrix} I_K \\ \cdot \\ \cdot \\ \cdot \\ I_K \end{pmatrix}.$$

- (iii) The (unconditional) posterior distribution of  $\tilde{\beta}$  given  $\tilde{\gamma}, \tilde{X}, \Delta, \Omega$ , and (2.14) is  $N(\tilde{\beta}^*, D_3)$ , where

$$D_3 = \{ \tilde{X}' \Omega^{-1} \tilde{X} + (I_N \otimes \Delta^{-1}) - (I_N \otimes \Delta^{-1}) A [A' (I_N \otimes \Delta^{-1}) A]^{-1} A' (I_N \otimes \Delta^{-1}) \}^{-1} \quad (2.20)$$

$$\tilde{\beta}^* = D_3 \left\{ \tilde{X}' \Omega^{-1} \tilde{\gamma} \right\} \quad (2.21)$$

Given a quadratic loss function, the Bayes point estimator is the posterior mean. The Bayes estimator of  $\tilde{\beta}$  is the GLS estimator (2.16). The Bayes estimator of  $\tilde{\beta}$ , is  $\tilde{\beta}^*$  (2.21), which is equal to

$$\tilde{\beta}^* = D_2 \left\{ \tilde{X}' \Omega^{-1} \tilde{X} \hat{\beta} + (I_N \otimes \Delta^{-1}) A \tilde{\beta}^* \right\}, \quad (2.22)$$

where

$$\hat{\beta} = (\tilde{X}'\Omega^{-1}\tilde{X})^{-1}(\tilde{X}'\Omega^{-1}\tilde{y}). \quad (2.23)$$

In other words, the Bayes point estimate of  $\tilde{\beta}_i^*$  is the weighted average of the GLS estimator of  $\tilde{\beta}_i$  using the  $i$ th individual's time series data,  $\hat{\beta}_i = (X_i' C_i^{-1} X_i)^{-1} (X_i' C_i^{-1} \tilde{y}_i)$  and the GLS estimator of  $\bar{\beta}$ , ((2.4)), with the weights proportional to the inverse of respective estimates. The difference between the sampling approach estimates of  $\tilde{\beta}_i$  and Bayes approach is that the former only uses data of the  $i$ th individual to estimate  $\tilde{\beta}_i$  while the latter uses data for all cross-sectional units because the Bayes estimates impose the additional constraints that  $E\tilde{\beta}_i = \bar{\beta}$  ((2.13)).

In practice,  $\Delta$  and  $C_i$  are rarely known. One way to obtain the Bayes estimate of  $\bar{\beta}$  or  $\beta$  is to substitute unknown  $\Delta$  and  $C_i$  by the sampling estimates of  $\Delta$  and  $C_i$  (e.g., (2.8), (2.9), or (2.10)) into (2.16) or (2.22). Alternatively, a Bayes mode estimator of  $\Delta$  and  $C_i$  can be used. For instance, Lindley and Smith (1972), Smith (1973) suggest using

$$\Delta^* = \left\{ R + (N - 1)\hat{\Delta} \right\} / (N + \rho - K - 2), \quad (2.24)$$

where  $R$  and  $\rho$  are prior parameters, assuming that  $\Delta^{-1}$  has a Wishart distribution with  $\rho$  degrees of freedom and matrix  $R$ , where one may let  $R = \hat{\Delta}$  ((2.9) or (2.10)) and  $\rho = 2$  as in Hsiao, Pesaran, and Tahmisioglu (1999). We shall call the estimator based on estimated  $\Delta$  and  $C_i$  the empirical Bayes estimator.

Alternatively, we can assume a prior for  $\Delta$  and  $C_i$ . For instance, in the case of  $C_i = \sigma_i^2 I_T$ , Lindley and Smith (1972) assume that the prior distribution of  $\sigma_i^2$  and  $\Delta$  are independent and are distributed as

$$P(\Delta^{-1}, \sigma_1^2, \dots, \sigma_N^2) = W(\Delta^{-1} | (\rho R)^{-1}, \rho) \prod_{i=1}^N \sigma_i^{-1}, \quad (2.25)$$

where  $W$  represents the Wishart distribution with scale matrix  $(\rho R)$  and degrees of freedom  $\rho$  (e.g., Anderson 1985). Incorporating this prior into the model under (1.1)–(1.3) and (2.1), we can obtain the marginal posterior densities of the parameters of interest by integrating out  $\sigma_i^2$  and  $\Delta$  from the joint posterior density. However, the required integrations do not yield closed form solutions. Hsiao, Pesaran, and Tahmisioglu (1999) have suggested using Gibbs sampler to calculate marginal densities.

### 13.3 DYNAMIC MODELS

Because of inertia in human behavior and institutional and technological rigidities, behavioral models are often specified as a dynamic model in the form

$$\begin{aligned} y_{it} &= \gamma_i y_{i,t-1} + \beta_i' \underline{x}_{it} + u_{it}, \quad |\gamma_i| < 1, \\ i &= 1, \dots, N, \\ t &= 1, \dots, T, \end{aligned} \tag{3.1}$$

where  $\underline{x}_{it}$  is a  $K \times 1$  vector of exogenous variables, and the error term  $u_{it}$  is assumed to be independently, identically distributed (i.i.d.) over  $t$  with mean zero and variance  $\sigma_{u_i}^2$  and is independently distributed across  $i$ . The coefficients  $\underline{\theta}_i = (\gamma_i, \beta_i')'$  is assumed to be independently distributed across  $i$  with mean  $\bar{\underline{\theta}} = (\bar{\gamma}, \bar{\beta})'$  and covariance matrix  $\Delta$ . Let

$$\underline{\theta}_i = \bar{\underline{\theta}} + \underline{\alpha}_i. \tag{3.2}$$

where  $\underline{\alpha}_i = (\alpha_{i1}, \alpha_{i2}')$ , is now a  $(K+1) \times 1$  time invariant random vector with

$$E\underline{\alpha}_i = \underline{0}, E\underline{\alpha}_i \underline{\alpha}_j' = \Delta \text{ if } i = j \text{ and } \underline{0} \text{ otherwise,} \tag{3.3}$$

and

$$E\underline{\alpha}_i \underline{x}_{it}' = \underline{0}. \tag{3.4}$$

Stacking the  $T$  time series observations of the  $i$ th individuals in matrix form yields

$$\begin{matrix} \underline{y}_i \\ T \times 1 \end{matrix} = Q_i \underline{\theta}_i + \underline{u}_i, \quad i = 1, \dots, N. \tag{3.5}$$

where  $\underline{y}_i = (y_{i1}, \dots, y_{iT})'$ ,  $Q_i = (\underline{y}_{i,-1}', X_i)$ ,  $\underline{y}_{i,-1} = (y_{i0}, \dots, y_{iT-1})'$ ,  $X_i = (\underline{x}_{i1}, \dots, \underline{x}_{iT})'$ ,  $\underline{u}_i = (u_{i1}, \dots, u_{iT})'$ , and for ease of exposition, we assume that  $y_{i0}$  are observable.

We note that because  $y_{i,t-1}$  depends on  $\underline{\alpha}_i$ ,  $EQ_i \underline{\alpha}_i' \neq \underline{0}$ , i.e., the independence between the explanatory variables,  $Q_i$ , and  $\underline{\alpha}_i$ , is violated. Substituting  $\underline{\theta}_i = \bar{\underline{\theta}} + \underline{\alpha}_i$  into (3.5) yields

$$\underline{y}_i = Q_i \bar{\underline{\theta}} + \underline{y}_i, \quad i = 1, \dots, N, \tag{3.6}$$

where

$$\underline{y}_i = Q_i \underline{\alpha}_i + \underline{u}_i. \tag{3.7}$$

Since

$$y_{i,t-1} = \sum_{j=0}^{\infty} (\bar{\gamma} + \alpha_{i1})^j \underline{x}_{i,t-j-1}' (\bar{\beta} + \underline{\alpha}_{i2}) + \sum_{j=0}^{\infty} (\bar{\gamma} + \alpha_{i1})^j u_{i,t-j-1}, \tag{3.8}$$

it follows that  $E(\underline{y}_i | Q_i) \neq \underline{0}$ . Therefore, contrary to the static case, the least squares estimator of the common mean,  $\bar{\underline{\theta}}$  is inconsistent.

Pesaran and Smith (1995) have noted that as  $T \rightarrow \infty$ , the least squares regression of  $\underline{\hat{\beta}}_j$  on  $Q_i$  yields a consistent estimator of  $\underline{\hat{\beta}}_i$ . They suggest a mean group estimator of  $\underline{\hat{\beta}}$  by taking the average of  $\underline{\hat{\beta}}_i$  across  $i$ ,

$$\hat{\underline{\hat{\beta}}} = \frac{1}{N} \sum_{i=1}^N \underline{\hat{\beta}}_i. \quad (3.9)$$

The mean group estimator (3.9) is consistent and asymptotically normally distributed so long as  $\sqrt{N}/T \rightarrow 0$  as both  $N$  and  $T \rightarrow \infty$  (Hsiao, Pesaran, and Tahmisioglu (1999)).

However, panels with large  $T$  are typically the exception in economics. Nevertheless, under the assumption that  $y_{i0}$  are fixed and known and  $\alpha_i$  and  $u_{it}$  are independently normally distributed, we can implement the Bayes estimator of  $\underline{\hat{\beta}}$  conditional on  $\sigma_i^2$  and  $\Delta$  using the formula (2.16) just as in the static case discussed in section 2. The Bayes estimator conditional on  $\Delta$  and  $\sigma_i^2$  is equal to

$$\hat{\underline{\hat{\beta}}}_B = \left\{ \sum_{i=1}^N [\sigma_i^2 (Q'_i Q_i)^{-1} + \Delta]^{-1} \right\}^{-1} \sum_{i=1}^N [\sigma_i^2 (Q'_i Q_i)^{-1} + \Delta]^{-1} \hat{\underline{\beta}}_i, \quad (3.10)$$

which is a weighted average of the least squares estimator of individual units with the weights being inversely proportional to individual variances. When  $T \rightarrow \infty, N \rightarrow \infty$  and  $\sqrt{N}/T^{3/2} \rightarrow 0$ , the Bayes estimator is asymptotically equivalent to the mean group estimator (3.9).

Hsiao, Pesaran, and Tahmisioglu (1999) have conducted Monte Carlo experiments to study the finite sample properties of (3.10), referred as infeasible Bayes estimator, the Bayes estimator obtained through the Gibbs sampler, referred as hierarchical Bayes estimator, the empirical Bayes estimator, the group mean estimator (3.9), the bias corrected group mean estimator (Pesaran and Zhao 1999), obtained by directly correcting the finite  $T$  bias of the least squares estimator,  $\underline{\hat{\beta}}_i$ , using the formula of Kiviet (1993), Kiviet and Phillips (1995), then taking the average, and the pooled least squares estimator. For  $N = 50$ , the infeasible Bayes estimator performs very well. It has small bias even for  $T = 5$ . For  $T = 5$ , its bias falls within the range of 3 to 17%. For  $T = 20$ , the bias is at most about 2%. The hierarchical Bayes estimator also performs well followed by the empirical Bayes estimator when  $T$  is small, but it improves quickly as  $T$  increases. The empirical Bayes estimator gives very good results even for  $T = 5$  in some cases but the bias also appears to be quite high in certain other cases. As  $T$  gets larger its bias decreases considerably. The mean group and the bias corrected mean group estimator both have large bias when  $T$  is small, with the bias-corrected mean group estimator performs slightly better. However, the performance of both improve as  $T$  increase and both are still much better than the least squares estimator. The least squares estimator yields significant bias and its bias persists as  $T$  increases.

The Bayes estimator is derived under the assumption that the initial observations,  $y_{i0}$ , are fixed constants. As discussed in Anderson and Hsiao (1981, 1982), this

assumption is clearly unjustifiable for a panel with finite  $T$ . However, contrary to the sampling approach where the correct modeling of initial observations is quite important, Bayesian approach appears to perform fairly well in the estimation of the mean coefficients for dynamic random coefficients models even the initial observations are treated as fixed constants. The Monte Carlo study also cautions against the practice of justifying the use of certain estimators based on their asymptotic properties here. Both the mean group and the corrected mean group estimators perform poorly in panels with very small  $T$ . The hierarchical Bayes estimator appears preferable to the other consistent estimators unless the time dimension of the panel is sufficiently large.

### 13.4 AGGREGATE VERSUS DISAGGREGATE ANALYSIS

Table 13.1 provides Hsiao, Shen, and Fujiki (2005) Japanese aggregate time series analysis of the relations between (real) money demand, (real) GDP and (five year) bond rate. The estimated relations between real money demand and real GDP are unstable and sensitive to the time period covered. Depending on the sample period covered,

**Table 13.1 Least Squares Estimation of Aggregate Money Demand Function**

Dependent variable	Sample period	Variable	Parameter estimate	Standard error
M2	1980.IV-2000.IV	Intercept	1.30462	0.28975
		Real GDP	-0.15425	0.04538
		RM2(-1)	1.07022	0.02790
		Bond rate	-0.00186	0.00069
	1992.I-2000.IV	Intercept	-0.16272	0.85081
		Real GDP	0.00847	0.06772
		RM2(-1)	1.00295	0.02248
		Bond rate	-0.00250	0.00140
M1	1980.IV-2000.IV	Intercept	0.46907	0.21852
		Real GDP	-0.01857	0.01700
		RM(-1)	0.98964	0.01249
		Bond rate	-0.00566	0.00135
	1992.I-2000.IV	Intercept	-0.68783	2.10228
		Real GDP	0.08414	0.14898
		RM1(-1)	0.96038	0.019999
		Bond rate	-0.01005	0.00283

Source: Hsiao, Shen, and Fujiki (2005, table 5).

the estimated relations are either of wrong sign or statistically insignificant. The estimated long-run income elasticities are 75.23 for M1 and 11.04 for M2, respectively, an “incredible” magnitude.

Alternatively, Hsiao, Shen, and Fujiki (2005) use Japanese prefecture data and a random coefficient framework to estimate the mean relation between (real) money demand and (real) GDP and (five year) bond rate for the 40 Japanese prefectures. Table 13.2 shows that estimated short-run income elasticity for M1 and M2 is 0.88 and 0.47, respectively. The long-run income elasticity is 2.56 for M1 and 1.01 for M2. These results appear to be consistent with economic theory and the broadly observed facts about Japan. The average growth rate for M2 in the 1980s is about 9.34%. The inflation rate is 1.98%. The real M2 growth rate is 7.36%. The real growth rate of GDP during this period is 4.13%. Taking account the impact of Five-year bond rate fell from 9.332% at 1980.I to 5.767 at 1989.IV, the results are indeed very close to the estimated long-run income elasticities based on disaggregate data analysis.

A model is a simplification of the real world. The purpose is to capture the essential factors that affect the outcomes. One of the tool for reducing the real world detail is through “suitable” aggregation. However, for aggregation not to distort the fundamental behavioral relations among economic agents, certain “homogeneity” conditions must hold between the micro units. Many economists have shown that if micro units are heterogeneous, aggregation can lead to very different relations among macro variables from those of the micro relations (e.g., Lewbel 1992, 1994; Pesaran 2003;

**Table 13.2 Random Coefficient Estimates of Japan Prefecture Money Demand Equation**

	M1		M2	
	Coefficient	Standard Error	Coefficient	Standard error
lagged money	0.656	0.034	0.533	0.069
Income	0.881	0.114	0.473	0.064
Bond Rate	-0.0476	0.006	-0.009	0.003
Constant	-2.125	0.038	0.043	0.239
	Variance-covariance matrix of $M1(\gamma_i, \beta_i')$			
	0.015			
	-0.001	0.177		
	0.001	-0.059	0.0005	
	-0.024	-0.588	-0.023	2.017
	Variance-covariance matrix of $M2$ equation $(\gamma_i, \beta_i')$			
	0.068			
	-0.031	0.062		
	0.002	0.0003	0.0014	
	-0.13	-0.107	-0.009	0.8385

Source: Hsiao, Shen, and Fujiki (2005, table 1).

Stoker 1993; Theil 1954; Trivedi 1985). Hsiao, Shen, and Fujiki (2005) have shown that for the simple dynamic equation,

$$y_{it} = \gamma_i y_{i,t-1} + \underline{x}'_{it} \beta_i + \alpha_i + u_{it}, \quad |\gamma_i| < 1, \quad i = 1, \dots, N, \quad (4.1)$$

where the error  $u_{it}$  is covariance stationary, implies a long-run relation between the aggregate variables,  $y_t = \sum_{i=1}^N y_{it}$  and  $\underline{x}_t = \sum_{i=1}^N \underline{x}_{it}$ ,

$$y_t - \underline{x}'_t \underline{b} - c = v_t,$$

where  $v_t$  is stationary if and only if either of the following conditions hold:

- (i)  $\frac{1}{1-\gamma_i} \beta_i = \frac{1}{1-\gamma_j} \beta_j$  for all  $i$  and  $j$ ; or
- (ii) if  $\frac{1}{1-\gamma_i} \beta_i \neq \frac{1}{1-\gamma_j} \beta_j$ , then  $\underline{x}'_t = (\underline{x}'_{1t}, \dots, \underline{x}'_{Nt})$  must lie on the null space of  $D$  for all  $t$ , where  $D' = \left( \frac{1}{1-\gamma_1} \beta'_1 - b', \dots, \frac{1}{1-\gamma_N} \beta'_N - b' \right)$ .

Panel data provide information on micro units. They can be used to check if either of these two suitable aggregation conditions hold. When micro-relations are “heterogeneous,” one way to get around the aggregation issue is to estimate each micro-relation separately, then aggregate. However, there may not have enough time series observations to obtain reliable estimates of the micro relations. Moreover, policy makers are interested in the average relations, not the individual relations. A random coefficient framework is a reasonable compromise that can take into account individual heterogeneity while still allowing the estimation of average relation.

## 13.5 TEST OF HETEROGENEITY

### 13.5.1 Heterogeneity Test

When the regressors are strictly exogenous (e.g., the model under (1.1)–(1.3) and (2.1)), a simple  $F$ -test can be used to test for homogeneity versus heterogeneity.

When the regressors contain lagged dependent variables, one way to test for homogeneity is to compare the difference between the mean group estimator (3.9) and the pooled least squares estimator

$$\tilde{\hat{\theta}} = \left( \sum_{i=1}^N Q'_i Q_i \right)^{-1} \left( \sum_{i=1}^N Q'_i y_{\sim i} \right). \quad (5.1)$$

Under the null hypothesis of homogeneity, both  $\hat{\theta}$  and  $\tilde{\hat{\theta}}$  converge to  $\bar{\hat{\theta}}$ . However,  $\hat{\theta}$  is inefficient while  $\tilde{\hat{\theta}}$  is efficient. Under the alternative of heterogeneity,  $\hat{\theta}$  remains consistent but  $\tilde{\hat{\theta}}$  is inconsistent. Therefore a Hausman-type test statistic (Hausman 1978)

can be used to test for homogeneity against heterogeneity (e.g., Hsiao and Pesaran 2008).

The Hausman procedure of comparing the difference between the group mean estimator (3.9) and the pooled least squares estimator (5.1) cannot be applied if no exogenous variables appear in (3.1) because both converge to the same common value  $\bar{\gamma}$ . In this case, following the idea of Swamy (1970), Pesaran and Yamagata (2008) suggest testing homogeneity versus heterogeneity by testing the dispersion of individual estimates from a suitable pooled estimator when both  $N$  and  $T$  are large,

$$S = \sum_{i=1}^N \frac{1}{\hat{\sigma}_i^2} (\hat{\theta}_i - \tilde{\theta}^*)' Q_i' Q_i (\hat{\theta}_i - \tilde{\theta}^*), \quad (5.2)$$

where

$$\hat{\sigma}_i^2 = T^{-1} (\underline{y}_i - Q_i \tilde{\theta})' (\underline{y}_i - Q_i \tilde{\theta}), \quad (5.3)$$

$$\tilde{\theta}^* = \left( \sum_{i=1}^N \frac{1}{\hat{\sigma}_i^2} Q_i' Q_i \right)^{-1} \left( \sum_{i=1}^N \frac{1}{\hat{\sigma}_i^2} Q_i' \underline{y}_i \right). \quad (5.4)$$

### 13.5.2 Fixed or Random Specification

The above testing procedures can only tell whether heterogeneity exists. It cannot tell whether we should treat the difference as fixed or random because the various tests of homogeneity versus heterogeneity essentially exploit the implications of certain formulations in a specific framework. They are indirect in nature. The distribution of a test statistic is derived under a specific null, while the alternative is composite. The rejection of a null does not automatically imply the acceptance of a specific alternative. Therefore, it would appear more appropriate to treat the fixed coefficients, random coefficients or various forms of mixed fixed and random coefficients models as different models and use either a model selection criterion or model predictive density approach to select an appropriate specification (Hsiao and Sun 2000). For instance, well-known model selection criterion such as Akaike's (1973) information criterion or Schwarz's (1978) Bayesian information criterion that selects the model  $H_j$  among  $j = 1, \dots, J$  different specification if it yields the smallest value of

$$-2 \log f(\underline{y} | H_j) + 2m_j, \quad j = 1, \dots, J, \quad (5.5)$$

or

$$-2 \log f(\underline{y} | H_j) + m_j \log NT, \quad j = 1, \dots, J, \quad (5.6)$$

can be used, where  $\log f(\underline{y} | H_j)$  and  $m_j$  denote the log likelihood values of  $\underline{y}$  and the number of unknown parameters of model  $H_j$ . Alternatively, Hsiao and Sun (2000),

following Min and Zellner (1993), suggest selecting the model that yields the highest predictive density. In this framework, time series observations are divided into two periods, 1 to  $T_1$ , denoted by  $\tilde{y}^1$ , and  $T_1 + 1$  to  $T$ , denoted by  $\tilde{y}^2$ . The first  $T_1$  observations are used to obtain the probability distribution of the parameters associated with model  $H_j$ , say  $\tilde{\theta}^j$ ,  $P(\tilde{\theta}^j | \tilde{y}^1)$ . The predictive density is then evaluated as

$$\int f(\tilde{y}^2 | \tilde{\theta}^j) p(\tilde{\theta}^j | \tilde{y}^1) d\tilde{\theta}^j, \quad (5.7)$$

where  $f(\tilde{y}^2 | \tilde{\theta}^j)$  is the density of  $\tilde{y}^2$  conditional on  $\tilde{\theta}^j$ . Given the sensitivity of Bayesian approach to the choice of prior, the advantage of using (5.7) is that the choice of a model does not have to depend on the prior. One can use the non-informative (or diffuse) prior to derive  $P(\tilde{\theta}^j | \tilde{y}^1)$ . It is also consistent with the theme that “a severe test for an economic theory, the only test and the ultimate test is its ability to predict” (Klein 1988, p. 21 also see Friedman 1953).

When  $\tilde{y}^2$  only contains a limited number of observations, the choice of a model in terms of predictive density may become heavily sample dependent. If too many observations are put in  $\tilde{y}^2$ , then a lot of sample information is not utilized to estimate unknown parameters. One compromise is to modify (5.7) by recursively updating the estimates,

$$\begin{aligned} & \int f(\tilde{y}_T | \tilde{\theta}^j, \tilde{y}^{T-1}) P(\tilde{\theta}^j | \tilde{y}^{T-1}) d\tilde{\theta}^j \cdot \int f(\tilde{y}_{T-1} | \tilde{\theta}^j, \tilde{y}^{T-2}) P(\tilde{\theta}^j | \tilde{y}^{T-2}) \\ & d\tilde{\theta}^j \dots \int f(\tilde{y}_{T_1+1} | \tilde{\theta}^j, \tilde{y}^1) P(\tilde{\theta}^j | \tilde{y}^1) d\tilde{\theta}^j, \end{aligned} \quad (5.8)$$

where  $P(\tilde{\theta}^j | \tilde{y}^T)$  denotes the posterior distribution of  $\tilde{\theta}$  given observations from 1 to  $T$ . While the formula may look formidable, it turns out that the Bayes updating formula is fairly straightforward to compute. For instance, consider the model (3.1). Let  $\tilde{\theta} = (\tilde{y}, \beta)$  and  $\tilde{\theta}_t$  and  $V_t$  denote the posterior mean and variance of  $\tilde{\theta}$  based on the first  $t$ -observations, then

$$\tilde{\theta}_t = V_{t-1}(Q'_t \Omega^{-1} \tilde{y}_t + V_{t-1}^{-1} \tilde{\theta}_{t-1}), \quad (5.9)$$

$$V_t = (Q'_t \Omega^{-1} Q_t + V_{t-1}^{-1})^{-1}, \quad t = T_1 + 1, \dots, T, \quad (5.10)$$

and

$$\begin{aligned} P(\tilde{y}_{t+1} | \tilde{y}^t) &= \int P(\tilde{y}_{t+1} | \theta, \tilde{y}^t) P(\theta | \tilde{y}^t) d\theta \\ &\sim N(Q_{t+1} \tilde{\theta}_t, \Omega + Q_{t+1} V_t Q'_{t+1}), \end{aligned} \quad (5.11)$$

where  $\tilde{y}'_t = (y_{1t}, y_{2t}, \dots, y_{Nt})$ ,  $Q_t = (\tilde{x}'_t, \tilde{w}'_t)$ ,  $\tilde{x}_t = (\tilde{x}_{1t}, \dots, \tilde{x}_{Nt})$ ,  $\tilde{w}_t = (\tilde{w}_{1t}, \dots, \tilde{w}_{Nt})$ ,  $\Omega = E \tilde{u}_t \tilde{u}'_t$ , and  $\tilde{u}'_t = (u_{1t}, \dots, u_{Nt})$  (Hsiao, Appelbe, and Dineen 1993).

Hsiao and Sun (2000) have conducted limited Monte Carlo studies to evaluate the performance of these model selection criteria in selecting the random, fixed, and mixed

random-fixed coefficients specification. They all appear to have a very high percentage in selecting the correct specification.

## 13.6 CONCLUDING REMARKS

---

The conventional variable intercepts models (e.g., Hsiao 2003, chapters 3 and 4) to capture the unobserved heterogeneity across individuals over time do not allow the interaction between the variation in the coefficients and exogenous variables. There are many examples that variables not included in the specification could also impact the marginal influences of the included explanatory variables. For example, in the empirical growth model the per capita output growth rates are assumed to depend on two sets of variables. One set of variables consist of initial per capita output, savings, and population growth rates, variables that are suggested by the Solow growth model. The second set of variables consists of control variables that correspond to whatever additional determinants of growth a researcher wishes to examine (e.g., Durlauf 2001 Durlauf and Quah 1999). However, there is nothing in growth theory which would lead one to think that the marginal effects of a change in high school enrollment percentages on the per capita growth of the United States should be the same as the effect on a country in sub-Saharan Africa. Had all these factors been taken into account in the specification, a common slope coefficients model may seem reasonable. However, these variables could be unavailable or could be difficult to be observed with precision. Moreover, a model is not a mirror, it is a simplification of the real world to capture the relationships among the essential variables. As a matter of fact, any parsimonious regression will necessarily leave out many factors that would from the perspective of economic theory be likely to affect the parameters of the included variables (e.g., Canova 1999, Durlauf and Johnson 1995). In these situations, a random coefficient specification would appear to be more capable of capturing the unobserved heterogeneity than a model with only individual-specific and/or time-specific effects (variable-intercept models) while allowing a researcher to draw the implications of mean relations between the outcome and included explanatory variables.

In this chapter we selectively surveyed some formulations for linear panel data models and their implications. Parameter heterogeneity in nonlinear panel data models poses fundamentally new issues and deserves to be studied seriously.

## ACKNOWLEDGMENTS

---

This work is partially supported by the NSF of China grant #71131008.

---

## NOTES

---

1. For detail, see Hsiao (2003) or Hsiao, Appelbe, and Dineen (1993).

---

## REFERENCES

---

- Akaike, H. (1973). "Information Theory and an Extension of the Maximum Likelihood Principle," in *Proc. 2nd. International Symposium Information Theory*, ed. B.N. Petrov and F. Csaki, 267–281. Budapest: Akademiai Kiado.
- Anderson, T.W. (1985). *An Introduction to Multivariate Analysis*, 2nd edition. New York: Wiley.
- 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.
- Canova, F. (1999). "Testing for Convergence Clubs in Income Per Capita: A Predictive Density Approach," Mimeo, Universitat Pompeu Fabra.
- Durlauf, S.N. (2001). "Manifesto for Growth Econometrics," *Journal of Econometrics*, 100, 65–69.
- Durlauf, S.N., and P. Johnson (1995). "Multiple Regimes and Cross-Country Growth Behavior," *Journal of Applied Econometrics*, 10, 365–384.
- Durlauf, S.N., and D. Quah (1999). "The New Empirics of Economic Growth," in *Handbook of Macroeconomics*, ed. J. Taylor and M. Woodford. Amsterdam: North-Holland.
- Friedman, M. (1953). *Essays in Positive Economics*. Chicago: University of Chicago Press.
- Hausman, J.A. (1978), "Specification Tests in Econometrics," *Econometrica*, 46, 1251–1371.
- Hsiao, C. (2003). *Analysis of Panel Data*, 2nd edition. Econometric Society monograph no. 34. Cambridge: Cambridge University Press.
- Hsiao, C., and M.H. Pesaran (2008). "Random Coefficients Models," in *The Econometrics of Panel Data*, 3rd ed. edited by L. Matayas and P. Sevestre, 187–216. Berlin: Springer.
- Hsiao, C., and B.H. Sun (2000). "To Pool or Not to Pool Panel Data," in *Panel Data Econometrics: Future Directions, Papers in Honor of Professor Pietro Balestra*, ed. J. Krishnakumar and E. Ronchetti, 181–198. Amsterdam: North Holland.
- Hsiao, C., T.W. Appelbe, and C.R. Dineen (1993). "A General Framework for Panel Data Analysis—with an Application to Canadian Customer Dialed Long Distance Service," *Journal of Econometrics*, 59, 63–86.
- Hsiao, C., M.H. Pesaran and A.K. Tahmisioglu (1999). "Bayes Estimation of Short-Run Coefficients in Dynamic Panel Data Models," in *Analysis of Panels and Limited Dependent Variables Models*, ed. C. Hsiao, L.F. Lee, K. Lahiri, and M.H. Pesaran, 268–296. Cambridge: Cambridge University Press.
- Hsiao, C., Y. Shen, and H. Fujiki (2005). "Aggregate vs. Disaggregate Data Analysis—A Paradox in the Estimation of Money Demand Function of Japan under the Low Interest Rate Policy," *Journal of Applied Econometrics*, 20, 579–601.
- Kiviet, J.F. (1995). "On Bias, Inconsistency, and Efficiency in Various Estimators of Dynamic Panel Data Models," *Journal of Econometrics*, 68, 53–78.
- Kiviet, J.F., and G.D.A. Phillips (1993). "Alternative Bias Approximation with Lagged Dependent Variables," *Econometric Theory*, 9, 62–80.

- Klein, L.R., (1988). "The Statistical Approach to Economics," *Journal of Econometrics*, 37, 7–26.
- Lewbel, A. (1992). "Aggregation with Log-Linear Models," *Review of Economic Studies*, 59, 635–642.
- Lewbel, A. (1994). "Aggregation and Simple Dynamics," *American Economic Review*, 84, 905–918.
- Lindley, D.V., and A.F.M. Smith (1972). "Bayes Estimates for the Linear Model," and Discussion, *Journal of the Royal Statistical Society, Series B*, 34, 1–41.
- 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.
- Pesaran, M.H. (2003). "On Aggregation of Linear Dynamic Models: An Application to Life-Cycle Consumption Models under Habit Formation," *Economic Modeling*, 20, 427–435.
- Pesaran, M.H., and R. Smith (1995). "Estimation of Long-Run Relationships from Dynamic Heterogeneous Panels," *Journal of Econometrics*, 68, 79–114.
- Pesaran, M.H., and T. Yamagata (2008). "Testing Slope Homogeneity in Large Panels," *Journal of Econometrics*, 142, 50–93.
- Pesaran, M.H., and Z. Zhao (1999). "Bias Reduction in Estimating Long-Run Relationships from Dynamic Heterogeneous Panels," in *Analysis of Panels and Limited Dependent Variables*, ed. C. Hsiao, K. Lahiri, L.F. Lee, and M.H. Pesaran, 297–322. Cambridge: Cambridge University Press.
- Schwarz, G. (1978). "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–464.
- Smith, A.F.M. (1973). "A General Bayesian Linear Model," *Journal of the Royal Statistical Society, B*, 35, 67–75.
- Stoker, T.M. (1993). "Empirical Approaches to the Problem of Aggregation over Individuals," *Journal of Economic Literature*, 31, 1827–1874.
- Swamy, P.A.V.B., (1970). "Efficient Inference in a Random Coefficient Regression Model," *Econometrica*, 38: 311–323.
- Theil, H. (1954). *Linear Aggregation of Economic Relations*. Amsterdam: North-Holland.
- Trivedi, P.K. (1985). "Distributed Lags, Aggregation and Compounding: Some Econometric Implications," *Review of Economic Studies*, 52, 19–35.

## CHAPTER 14

---

# ROBUST PANEL DATA METHODS AND INFLUENTIAL OBSERVATIONS

---

BADI H. BALTAGI AND GEORGES BRESSON

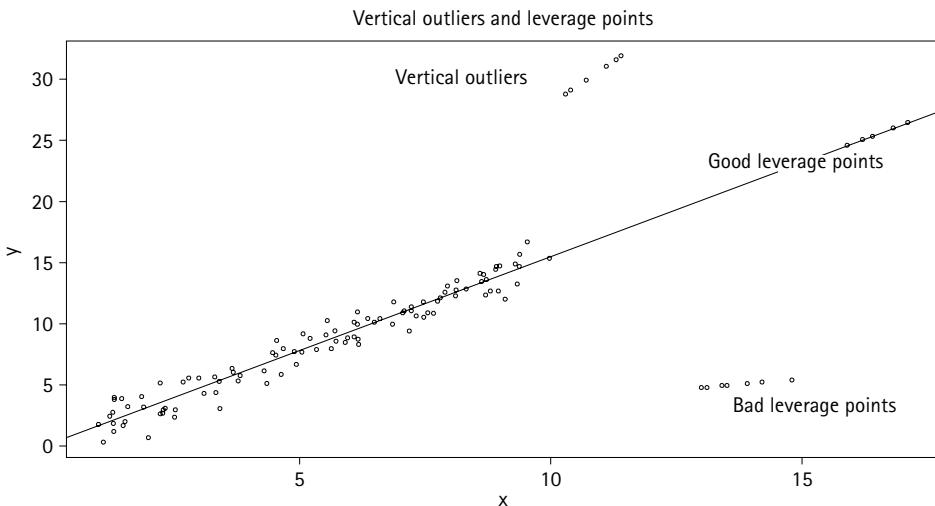
*“Normality is a myth; there never was and never will be a normal distribution.”*

Geary (1947, p. 241)

### 14.1 INTRODUCTION

---

WHAT is meant by an influential observation? The definition given by Belsley, Kuh, and Welsch (1980 p. 11) seems appropriate: “An influential observation is one which, either individually or together with several other observations, has demonstrably larger impact on the calculated values of various estimates (coefficients, standard errors, t-values, etc) than the case for most of the other observations.” Huber (1981) has shown that only 3% of *atypical* values in a set of observations is sufficient to reveal the weakness of classical consistent results. Panel data has a large number of observations but if 3% is enough to spoil the soup, this large number of “good points” is far from drowning a few “bad points.” But, what is meant by good or bad points? The robust statistics literature generally splits influential observations into three types: vertical outliers, good leverage points, and bad leverage points (see Rousseeuw and Leroy 2003). *Vertical outliers* are observations that have outlying values for the corresponding error term (the  $y$ -dimension) but are not outlying in the design space (the  $X$ -dimension). *Good leverage points* are observations that are outlying in the design space but are located close to the regression line. While, *bad leverage points* are observations located far away from the regression line. Least squares estimation is highly contaminated by these *bad leverage points* (see Dehon, Gassner, and Verardi 2009). To summarize, an observation is influential if the regression estimates change considerably when it is removed. Figure 14.1 shows a hypothetical cloud of points where the bulk of the data is located far from the vertical outliers, the good leverage points and the bad leverage points to illustrate their influence graphically.



**FIGURE 14.1** An example of vertical outliers and leverage points.

Arbitrary procedures for deleting observations, which are more frequent than it ought to be in empirical work, should be avoided. Even after one cleans up a data set, it is unlikely to better fit a model. With many observations, as is typical with panel data, it may be difficult to rely on graphical procedures, and it becomes necessary to find statistical methods to detect outliers and bad leverage points and to robustify the estimates. This is the object of robust statistics.<sup>1</sup> Although there is a huge literature in robust statistics spanning 60 years, there are but a few papers that we know of in the context of panel data econometrics. It seems surprising that, despite their relevance, studies of robust methods using panel data are rather limited. In order to appreciate their strength or weakness, we must first review some basic concepts and methods in robust statistics and in the panel data field in particular. This is the main object of Section 14.2, where we present robust estimators of linear static panel data models, see Bramati and Croux (2007). In Section 14.3, we present robust instrumental variables (IV) estimators, see Wagenvoort and Waldmann (2002), robust Hausman and Taylor (1981) estimation method, see Baltagi and Bresson (2012), as well as dynamic panel data GMM methods, see Lucas, Van Dijk, and Kloek (1994, 2007). Section 14.4 describes some robust nonlinear panel data methods. Section 14.5 focuses on the detection of influential observations and outliers in the context of panel data. While Section 14.6 concludes.

## 14.2 ROBUST ESTIMATORS OF LINEAR STATIC PANEL DATA MODELS

---

Outliers can be classified into two main categories: gross errors which come from recording errors, and outliers which are due to the approximate nature of the statistical

or econometric model (see Krasker, Kuh, and Welsch 1983; Hampel 1986; Lucas 1996, to mention a few). There is a large literature on influence measures and on the rules for rejection of outliers (see Chatterjee and Hadi 1986 and Barnett and Lewis 1994; to mention a few). But, as shown by Hampel (1985, 2002), some of these rules cannot even reject one distant outlier out of 20 observations. Hampel (1973, p. 88) states that “5 – 10% wrong values in a data set seem to be the rule rather than the exception.” Moreover, we must keep in mind that identifying multiple influential observations becomes much harder due to two effects called masking and swamping effects (Hadi and Simonoff 1993). The *masking effect* occurs when an outlying subset goes unnoticed because of the presence of another. While, the *swamping effect* refers to good observations wrongly identified as outliers because of the existence of a subset of influential observations far from the data cloud.

After Tukey's (1960) inspiring paper, Huber (1964, 1965), Hampel (1968), Donoho and Huber (1983), among others, developed some key concepts for a robust theory of statistics. We need to briefly review some of these concepts which are applied in the robust panel data literature. These concepts include  $M$ -estimators, the influence function, and the breakdown down point (BDP). These are explained in the next subsection.

### 14.2.1 Breakdown Point, $M$ , and *LTS* Estimators

Let our panel data sample be composed of  $NT$  observations  $\Omega = \{y_{it}, X_{it}\}$  where  $i = 1, \dots, N$  and  $t = 1, \dots, T$ ,  $y_{it}$  is an observation of the dependent variable and  $X_{it}$  is a  $(1 \times K)$  vector of explanatory variables. Let  $\theta(\Omega)$  be our estimator using the sample  $\Omega$ . Let  $\tilde{\Omega} = \{\tilde{y}_{it}, \tilde{X}_{it}\}$  be a contaminated set of  $NT$  observations, where any  $m$  of the original points of  $\Omega$  are replaced by arbitrary values, and  $\tilde{\theta}(\tilde{\Omega})$  is the corresponding estimator using the corrupted sample  $\tilde{\Omega}$ . If  $\omega(m, \theta, \Omega)$  is the supremum of  $\|\tilde{\theta}(\tilde{\Omega}) - \theta(\Omega)\|$ , then the BDP of  $\theta$  at  $\Omega$  is defined as:

$$\begin{aligned}\varepsilon_{NT}^*(m, \theta, \Omega) &= \min \left\{ \frac{m}{NT}; \omega(m, \theta, \Omega) \text{ is infinite} \right\} \\ &= \min \left\{ \frac{m}{NT}; \sup_{\tilde{\Omega}} \|\tilde{\theta}(\tilde{\Omega}) - \theta(\Omega)\| = \infty \right\}. \end{aligned} \quad (1)$$

BDP is the smallest proportion of observations replaced by outliers which can cause the estimator  $\theta$  to take on values arbitrarily far from  $\theta(\Omega)$  (see Bramati and Croux 2007 and Croux and Verardi 2008). The least squares estimator, for example, has a BDP of  $(1/NT)$  because just one leverage point can cause it to break down. As the number of observations increases, the BDP tends to 0 and the LS estimator has a BDP of 0%. The highest BDP we can hope for is 50%. If more than half of the data is contaminated, one cannot discriminate between good and bad points. The median tolerates slightly less than 50% outliers (its asymptotic BDP is  $1/2$ ) and the  $\alpha$ -trimmed mean

has approximately and asymptotically a  $BDP = \alpha(0 < \alpha < 1/2)$ .<sup>2</sup> Generally, robust regression techniques are compared in terms of some properties like relative efficiency and *equivariance* of an estimator.<sup>3</sup> Some of the earlier robust regression methods (like  $M$ -estimators) did not satisfy one or more of these equivariance properties. We study these methods because later more popular robust methods—which perform well in terms of efficiency and equivariance properties—are extensions of these earlier robust methods.

Huber (1964) generalized the median regression to a wider class of estimators, called  $M$ -estimators (or maximum likelihood type estimators), by considering other functions besides the absolute value of the residuals. This increases Gaussian efficiency while keeping robustness with respect to vertical outliers. Applying this to the case of a fixed effects panel data regression disturbances:  $r_{it}(\alpha, \beta) \equiv r_{it} = y_{it} - X_{it}\beta - \alpha_i$ , the  $M$ -estimator is defined as

$$\widehat{\theta}_M \left( \equiv (\alpha, \beta')' \right) = \arg \min_{\alpha, \beta} \sum_{i=1}^N \sum_{t=1}^T \rho \left( \frac{r_{it}}{\widehat{\sigma}} \right), \quad (2)$$

where  $\rho(\cdot)$  is a continuous, symmetric objective function. To guarantee *scale equivariance* (i.e., independence with respect to the measurement units of the dependent variable), residuals are standardized by a measure of dispersion  $\widehat{\sigma}$ . This can be implemented as an iterative weighted least squares. A popular estimator for  $\widehat{\sigma}$  is the ‘re-scaled mean absolute deviation’ (MAD):

$$\widehat{\sigma} = 1.4826 \times \text{med}(|r_{it} - \text{med}(r_{it})|). \quad (3)$$

It is highly resistant to outlying observations with BDP of 50% as it is based on the median rather than the mean. The estimator  $\widehat{\sigma}$  rescales MAD by the factor 1.4826 so that when the sample is large and  $r_{it}$  are distributed as  $N(0, \sigma^2)$ ,  $\widehat{\sigma}$  estimates the standard deviation.

The  $M$ -estimator of  $\theta$  based on the function  $\rho(\cdot)$  is the vector  $\widehat{\theta}_M$  of size  $(K \times 1)$  which is the solution of the following system:

$$\min_{\theta} \sum_{i=1}^N \sum_{t=1}^T \psi \left( \frac{r_{it}}{\sigma} \right) \frac{d(r_{it})}{d\theta_j} = 0, j = 1, \dots, K \quad (4)$$

$\psi(u) = \rho'(u)$  is called the *influence function*. If we define a weight function  $W_r(u) = \psi(u)/u$ , then  $\widehat{\theta}_M$  is a weighted least squares estimator:

$$\widehat{\theta}_M = \arg \min_{\theta} \sum_{i=1}^N \sum_{t=1}^T W_r(r_{it}) r_{it}^2 \quad (5)$$

and the first order condition which defines the class of  $M$ -estimators is given by:

$$\sum_{i=1}^N \sum_{t=1}^T X'_{it} r_{it} W_r(r_{it}) = 0. \quad (6)$$

The loss function  $\rho(\cdot)$  is a symmetric, positive-definite function with a unique minimum at zero. The robust  $M$ -estimator should have a bounded influence function  $\psi(r_{it}/\sigma)$ , and the robust estimator should be unique. This requires that the function  $\rho(\cdot)$  be convex in  $\theta$ . The literature on robust statistics proposed several specifications for the  $\rho$ -function. Choosing a weight function  $W_r$  to apply to the scaled residuals corresponds directly to choosing a probability distribution function for the errors. So, the choice of the loss function  $\rho(\cdot)$  is crucial to having good robustness properties and high Gaussian efficiency. Several suggestions for loss functions, and associated influence functions and weight functions, have been made in the literature (see Andersen 2008; Rousseeuw and Leroy 2003, for instance). The three commonly used  $M$ -estimators are given in Table 14.1.

$M$ -estimators are called *monotone* if the loss function  $\rho$  is convex over the entire domain and are called *redescending* if the influence function  $\psi$  is bounded. The Huber  $M$ -estimator corresponds to a probability distribution for the errors which is normal in the center but it is like a double exponential distribution in the tails. The associated weight function gives those observations, whose scaled residuals are within the central bound, weight one, whilst scaled residuals outside the bound have smaller weights. In contrast, the redescending Hampel  $M$ -estimator and the Tukey  $M$ -estimator are defined such as  $\psi(u) = 0$  and  $W_r(u) = 0$  if  $|u| > c$ , where  $c$  is the tuning constant. Redescending  $M$ -estimators have high breakdown points (close to 0.5) and their influence function can be chosen to redescend smoothly to 0 as in the Tukey biweight function (see Figure 14.2).

The Tukey bisquare weight function (or biweight function) is a common choice. In this case, the high breakdown point  $M$ -estimator is defined as:

$$\hat{\theta}_M = (X' W_r X)^{-1} X' W_r y, \quad (7)$$

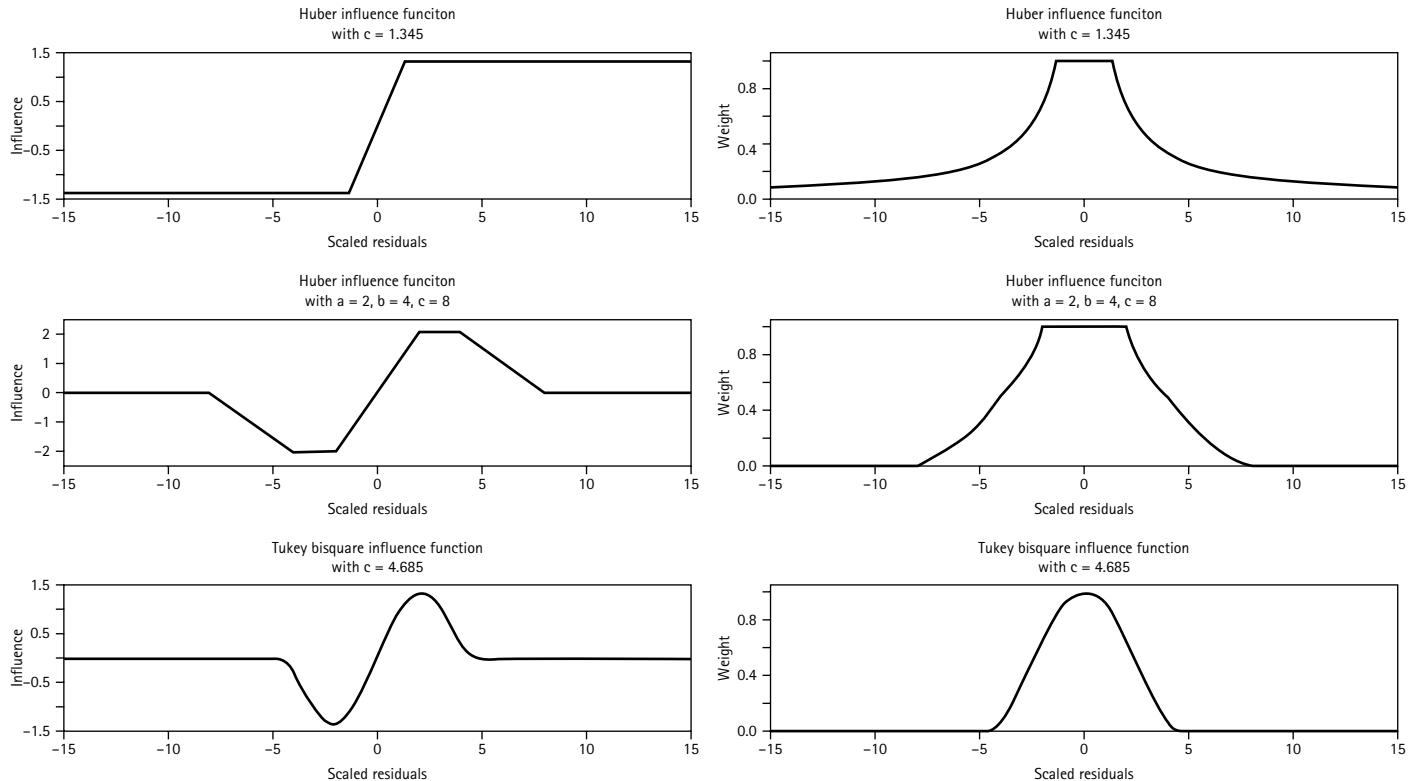
where  $y$  is the  $(NT \times 1)$  vector denoting the dependent variable, and  $X$  is the  $(NT \times K)$  matrix of the explanatory variables.  $W_r$  is an  $(NT \times NT)$  matrix with diagonal elements given by:

$$W_r(r_{it}) = \begin{cases} \left[ 1 - \left( \frac{r_{it}}{c\sigma} \right)^2 \right]^2 & \text{if } \left| \frac{r_{it}}{\sigma} \right| \leq c \\ 0 & \text{if } \left| \frac{r_{it}}{\sigma} \right| > c \end{cases}. \quad (8)$$

For the tuning constant  $c = 2.937$  (or  $c = 1.547$ ), the corresponding  $M$ -estimator resists contamination up to 25% (or up to 50%) of the outliers. In other words, it is said to have a breakdown point of 25% (or 50%). Unfortunately, this  $M$ -estimator suffers from some deficiencies. It does not consider leverage effects. One bad leverage point can cause the estimator to entirely break down and thus the  $M$ -estimator has an overall

**Table 14.1** The three commonly used *M*-estimators

	Loss function $\rho(u)$	Influence function $\psi(u)$	Weight function $W_r(u)$
Huber	$\begin{cases} \frac{u^2}{2} & \text{if }  u  < c, c > 0 \\ c u  - \frac{c^2}{2} & \text{if }  u  \geq c \end{cases}$	$\begin{cases} u & \text{if }  u  < c \\ c \operatorname{sign} u & \text{if }  u  \geq c \end{cases}$	$\begin{cases} 1 & \text{if }  u  < c \\ \frac{c}{ u } & \text{if }  u  \geq c \end{cases}$
Hampel	$\begin{cases} \frac{u^2}{2} & \text{if }  u  < a, a > 0 \\ a u  - \frac{a^2}{2} & \text{if } a \leq  u  < b, b > 0 \\ a\frac{c u  - \frac{a^2}{2}}{\frac{c-b}{a} - \frac{7a^2}{6}} - \frac{7a^2}{6} & \text{if } b \leq  u  \leq c, c > 0 \\ a(b + c - a) & \text{otherwise} \end{cases}$	$\begin{cases} u & \text{if }  u  < a \\ a \operatorname{sign} u & \text{if } a \leq  u  < b \\ a\frac{c \operatorname{sign} u - u}{c-b} & \text{if } b \leq  u  \leq c \\ 0 & \text{otherwise} \end{cases}$	$\begin{cases} 1 & \text{if }  u  < a \\ \frac{a}{ u } & \text{if } a \leq  u  < b \\ a\frac{c/ u -1}{c-b} & \text{if } b \leq  u  \leq c \\ 0 & \text{otherwise} \end{cases}$
Tukey bisquare	$\begin{cases} \frac{u^2}{2} - \frac{u^4}{2c^2} + \frac{u^6}{6c^4} & \text{if }  u  \leq c, c > 0 \\ \frac{c^2}{6} & \text{if }  u  > c \end{cases}$	$\begin{cases} u \left[ 1 - \left( \frac{u}{c} \right)^2 \right]^2 & \text{if }  u  \leq c \\ 0 & \text{if }  u  > c \end{cases}$	$\begin{cases} \left[ 1 - \left( \frac{u}{c} \right)^2 \right]^2 & \text{if }  u  \leq c \\ 0 & \text{if }  u  > c \end{cases}$



**FIGURE 14.2** Influence functions and weight functions of the three commonly used  $M$ -estimators.

BDP of 0%. In other words, if an  $M$ -estimator is able to identify isolated outliers, it is inappropriate in case of *masking effects*, i.e., in case there are clusters of outliers in which one outlier can mask the presence of another. Furthermore, the initial values for the iterative reweighted least squares algorithm are monotone  $M$ -estimators that are not robust to bad leverage points and may cause the algorithm to converge to a local instead of a global minimum (see Croux and Verardi 2008).

In contrast, other early robust estimators such as the Least Trimmed Squares (LTS) are inefficient but reach a BDP of 50% (see Andersen 2008; Rousseeuw 1984). The LTS estimator, which minimizes the sum of the smallest  $H$  ( $1 \leq H \leq NT$ ) squared residuals, is given by:

$$\hat{\beta}_{LTS} = \arg \min_{\beta} \sum_{h=1}^H \left[ (\tilde{y}_h - \tilde{X}_h \beta)^2 \right]_{h:NT}, \quad (9)$$

where the centered data are defined as

$$\tilde{y}_i = (\tilde{y}_{i1}, \dots, \tilde{y}_{iT})', \tilde{y}_{it} = y_{it} - \text{med}_t(y_{it}) \quad (10)$$

$$\tilde{X}_i = (\tilde{x}_{1,i}, \dots, \tilde{x}_{k,i}, \dots, \tilde{x}_{K,i}), \tilde{x}_{k,i} = (\tilde{x}_{k,i1}, \dots, \tilde{x}_{k,iT})', \tilde{x}_{k,it} = x_{k,it} - \text{med}_t(x_{k,it}) \quad (11)$$

and where

$$\left[ (\tilde{y}_i - \tilde{X}_i \beta)^2 \right]_{1:NT} \leq \left[ (\tilde{y}_i - \tilde{X}_i \beta)^2 \right]_{2:NT} \leq \dots \leq \left[ (\tilde{y}_i - \tilde{X}_i \beta)^2 \right]_{NT:NT} \quad (12)$$

are the ordered squared regression residuals. The common choice for the truncation value,  $H = 0.75 \times NT$ , leads to a BDP of 25%. Bramati and Croux (2007) show that  $\hat{\beta}_{LTS}$  is only *scale equivariant*.<sup>4</sup> The nonlinearity of the centering transformation by the median makes it impossible to achieve regression and affine equivariances. To improve the lack of efficiency of the LTS, Bramati and Croux (2007) use reweighted least squares using Tukey biweights, so the BDP of the initial LTS estimator is preserved.<sup>5</sup>

### 14.2.2 $S$ , $MS$ , and $WMS$ -estimators

Rousseeuw and Yohai (1987) proposed minimizing a measure of dispersion of the residuals that is less sensitive to extreme values. They call this class of estimators the  $S$ -estimators. In order to increase robustness, they suggest finding the smallest robust scale of the residuals. This robust dispersion, that is called  $\hat{\sigma}_S$ , satisfies

$$\frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \rho \left( \frac{r_{it}(\alpha, \beta)}{\hat{\sigma}_S} \right) = b, \quad (13)$$

where  $b = E[\rho(Q)]$  with  $Q \sim N(0,1)$ . The value of  $\theta$  that minimizes  $\widehat{\sigma}_S$  is an S-estimator defined as:

$$\widehat{\theta}_M^S = \arg \min_{\theta} \widehat{\sigma}_S(r_{11}(\theta), \dots, r_{NT}(\theta)) \quad (14)$$

with the corresponding  $\widehat{\sigma}_S$  being the robust estimator of scale.

Rousseeuw and Yohai (1987) computed the asymptotic efficiency of the S-estimator of a Gaussian model for different values of the breakdown point (see Table 14.2).

Unfortunately, this S-estimator has a Gaussian efficiency of only 28.7%. If the tuning constant ( $c$ ) of the Tukey biweight loss function  $\rho(\cdot)$  is high, for instance  $c = 5.182$ , the Gaussian efficiency climbs to 96.6% but the breakdown point drops to 10%. Monotone  $M$ -estimators are robust to outliers in the response variable, but are not resistant to outliers in the explanatory variables (leverage points). In contrast, redescending  $M$ -estimators are resistant to bad leverage points but are difficult to implement from a computational point of view. S-estimation, which finds a hyperplane that minimizes a robust estimate of the scale of the residuals, is highly resistant to leverage points, and is robust to outliers in the response. However, this method can be inefficient. MM-estimation tries to capture both the robustness and resistance of S-estimation, while at the same time gaining the efficiency of  $M$ -estimation. The method proceeds in three steps: (a) using an initial loss function, we get an initial  $M$ -estimator, (b) we obtain an  $M$ -estimate of the scale of the residuals, (c) the estimated scale is then held constant whilst an  $M$ -estimate of the parameters is located with a new loss function.

**Table 14.2 Breakdown point and asymptotic efficiency of the S-estimator of a Gaussian model (from Rousseeuw and Yohai 1987, Table 3, p. 268)**

breakdown point $\varepsilon^*$ (in %)	as.eff (in %)	tuning constant	
		c	b
50	28.7	1.547	0.1995
45	37.0	1.756	0.2312
40	46.2	1.988	0.2634
35	56.0	2.251	0.2957
30	66.1	2.560	0.3278
25	75.9	2.937	0.3593
20	84.7	3.420	0.3899
15	91.7	4.096	0.4194
10	96.6	5.182	0.4475

Note: Where  $as.eff = (\int \psi' d\Phi)^2 / (\int \psi^2 d\Phi)$  where  $\Phi$  is the c.d.f of  $N(0,1)$  and  $b = E[\rho(Q)]$  with  $Q \sim N(0,1)$ .

*MM*-estimators are robust and efficient but to our knowledge no panel data version has been developed.

Yohai (1987) introduced *M*-estimators that combine high-breakdown point and high efficiency. These estimators are redescending *M*-estimators, but where the scale is fixed at  $\widehat{\sigma}_S$ . The preliminary *S*-estimator guarantees a high breakdown point and the final *M*-estimate allows a high Gaussian efficiency. Following the proposition of Rousseeuw and Yohai (1987), the tuning constant can be set to  $c = 1.547$  for the *S*-estimator to guarantee a 50% breakdown point, and it can be set to  $c = 5.182$  for the second step *M*-estimator to guarantee 96% efficiency of the final estimator.<sup>6</sup>

Generally, the *S* and *M*-estimators use the algorithm of Salibian-Barrera and Yohai (2006) (see also Maronna and Yohai 2006). The algorithm starts by randomly picking  $p$  subsets of  $K$  observations where  $K$  is the number of regression parameters to estimate. For each of the  $p$ -subsets, residuals are computed and a scale estimate  $\widehat{\sigma}_S$  is obtained. An approximation for the final scale estimate  $\widehat{\sigma}_S$  is then given by the value that leads to the smallest scale over all  $p$ -subsets.<sup>7</sup> Maronna and Yohai (2006) introduce the *MS*-estimator that alternates an *S*-estimator and an *M*-estimator, until convergence. This estimator has been adapted for the fixed effects panel data case by Bramati and Croux (2007). They call this estimator the *WMS*-estimator (or Within *MS*-estimator).

Suppose that  $\beta$  is known, then the fixed effects vector  $\alpha = (\alpha_1, \dots, \alpha_N)'$  is obtained as an *M*-estimator:

$$\widehat{\alpha}(\beta) = \arg \min_{\alpha} \sum_{i=1}^N \sum_{t=1}^T \rho_M(r_{it}(\alpha, \beta)) \text{ or } \widehat{\alpha}_i(\beta) = \arg \min_{\alpha_i} \sum_{t=1}^T \rho_M(r_{it}(\alpha_i, \beta)), \forall i. \quad (15)$$

Bramati and Croux (2007) use  $\rho_M(r_{it}(\alpha, \beta)) = |r_{it}(\alpha, \beta)|$ , the absolute value of the loss function. This yields to a simple expression for the estimated fixed effects:

$$\widehat{\alpha}_i(\beta) = \text{med}_t(y_{it} - X_{it}\beta), i = 1, \dots, N. \quad (16)$$

So, the Within *MS*-estimator (or *WMS*-estimator), proposed by Bramati and Croux (2007), for a linear panel data model is defined as:

$$\widehat{\beta}_{WMS} = \arg \min_{\beta} \widehat{\sigma}_S(r_{11}(\widehat{\alpha}(\beta), \beta), \dots, r_{NT}(\widehat{\alpha}(\beta), \beta)), \quad (17)$$

where

$$r_{it}(\widehat{\alpha}(\beta), \beta) = y_{it} - X_{it}\beta - \text{med}_t(y_{it} - X_{it}\beta). \quad (18)$$

The *WMS*-estimator is the robust counterpart of the Least Squares Dummy Variables representation of the Within estimator. Given an initial estimate  $\beta_0$ , Bramati and Croux (2007) use an iterative algorithm based upon the generation of random subsamples suggested by Maronna and Yohai (2006) to compute the robust scale estimate of the residuals. They suggest iterating a fixed number of times ( $\max m = 20$ ) and to choose the  $\widehat{\beta}_{WMS}^{(m)}$  *WMS*-estimator which produces the minimum value of the objective

function in Eq. (17). The variance-covariance matrix of the WMS-estimate  $\hat{\beta}_{WMS}$  is given by:

$$\text{Var}(\hat{\beta}_{WMS}) = \hat{\sigma}_S^2 (X'D_1X)^{-1} X'D_2X (X'D_1X)^{-1}, \quad (19)$$

where  $D_1$  and  $D_2$  are diagonal matrices with diagonal elements given by:

$$D_{1,it} = \frac{d}{du_{it}} [u_{it} W_r(u_{it})] \text{ and } D_{2,it} = [u_{it} W_r(u_{it})]^2 \text{ with } u_{it} = \frac{r_{it}}{c.\hat{\sigma}_S}. \quad (20)$$

### 14.2.3 GM and WGM-estimators

In  $M$ -estimation, the goal is to choose the regression coefficients that minimize some function of the residuals.  $M$ -estimates were primarily used for identifying and down-weighting the effects of outliers in the  $y$ -direction. Krasker and Welsch (1982) warn against the possibility of arbitrarily large influence of an observation because of the multiplier  $X_{it}$  in the robust normal equations (Eq. (6)), which occurs when high leverage observations are present in the data. High breakdown point generalized- $M$  estimators (GM-estimators) (or bounded influence (BI) estimators, also termed Mallows's estimators (see Mallows 1975)) differ from  $M$ -estimators in that they try to account for both an observation's residual and its leverage value. The robust normal equations of Eq. (6) are modified by the use of  $W_x(X_{it})$  weights, which are only based on independent variables for a given observation and are chosen to vary indirectly with leverage. The first order condition which defines the class of generalized- $M$  estimators (GM-estimators) is given by:

$$\sum_{i=1}^N \sum_{t=1}^T X'_{it} W_x(X_{it}) r_{it} W_r(r_{it}) = 0. \quad (21)$$

$W_x(\cdot)$  and  $W_r(\cdot)$  are downweight leverage points and vertical outliers. The location weights  $W_x(\cdot)$  are based on robust distances  $RD_{it}$ . They are indirectly proportional to the values of covariates:

$$W_x(X_{it}) = \min \left( 1, \frac{\sqrt{\chi_{K,0.975}^2}}{RD_{it}} \right), \quad (22)$$

where  $\chi_{K,0.975}^2$  is the upper 97.5% quantile of a chi-squared distribution with  $K$  degrees of freedom. The distances  $RD_{it}$  are computed with the minimum volume ellipsoid (MVE) estimates of location  $\mu_x$  and scale  $V_x$  of  $X_{it}$ :

$$RD_{it} = \sqrt{(X_{it} - \mu_x)' V_x^{-1} (X_{it} - \mu_x)}. \quad (23)$$

Eq.(23) is a robust version of the Mahalanobis distance (or Rao's distance).<sup>8</sup> Following the work of Hinloopen and Wagenvoort (1997) and Wagenvoort and Waldmann (2002), Bramati and Croux (2007) propose the use of this class of generalized  $M$ -estimators in the panel data case. They call it Within GM-estimator (or WGM-estimator). In this case, the high breakdown point WGM-estimator is defined as:

$$\widehat{\theta}_{WGM} = (X' W_x W_r X)^{-1} X' W_x W_r y, \quad (24)$$

where  $W_x$  is the  $(NT \times NT)$  matrix with diagonal elements given by  $W_x(X_{it})$ . The WGM-estimator is also a weighted least squares estimator for median-transformed data. As for the LTS proposed by Bramati and Croux (see Section 14.2.1), it is only scale equivariant since the nonlinearity of the centering transformation by the median makes it impossible to achieve regression and affine equivariances. Moreover, its asymptotic distribution has not been derived. However, the Monte Carlo simulations show good properties of the WMS and WGM-estimators. In fact, when no outlier is present, the efficiency of the two robust estimators is very close to that of the Within estimator. The Within estimator performs badly when there are vertical outliers and even worse in the presence of bad leverage points. The WMS and the WGM-estimators yield good results and similar outcomes. When bad leverage points are present, the WMS-estimator gives slightly better results. Bramati and Croux (2007) conclude that they cannot clearly distinguish between the performance of the two robust estimators. Both estimators yield a large gain in mean square error with respect to the Within estimator in the presence of outliers. Bramati and Croux (2007) applied these methods to the macro application by Giavazzi, Jappelli, and Pagano (2000) on the response of the private sector to fiscal policy. Although the World Saving Data Base contains yearly national income and fiscal variables for a group of 150 industrial and developing countries from 1960 to 1995, the authors cleaned the data, selecting 101 developing countries out of 108 and restricting the sample period considered to 1970–1994. Bramati and Croux (2007) show that their robust estimates based on the complete data have the same signs but different magnitudes from the ones using the cleaned data set by the authors. They argue that instead of the subjective preliminary cleaning of the data, which has the risk of not detecting all outliers, and assigning zero or one weights to a country, a robust approach allows for a more careful, data-driven weighting of the observations.

Verardi and Wagner (2011) also use the WMS-estimator to estimate the exporter productivity premium (i.e., the productivity difference between exporting and non-exporting firms) in Germany. The panel data set is composed of 33,508 manufacturing firms over the period 1995–2006. About 3% of the enterprises are identified as outliers, which accounts for a little over 12% percent of the observations. Dropping these outliers leads to a smaller estimate for the exporter productivity premium which drops from 13% percent to 1% percent when robust fixed effects are used.

#### 14.2.4 First Difference, Pairwise Differences, *IRLS*, *REWLS* and *RLTS* Estimators

The WMS and the WGM-estimators proposed by Bramati and Croux (2007) are *not equivariant* with respect to various transformations of the data. Moreover, they must estimate the fixed effects leading to the bias if  $T$  is too small ( $T < 4$ ). Their Monte Carlo study used  $N = 100$  and  $T = 4$ , and 20. With  $T < 4$ , biases may come from the non linearity of the procedure. Recently, Aquaro and Čížek (2013) proposed a new robust technique which is equivariant with the usual data transformations. It is also consistent for small  $T = 3$ . This procedure only requires a unimodal distribution on the error distribution. Aquaro and Čížek (2013) use two different data transformations: *first difference*  $\Delta y_{it} = y_{it} - y_{it-1}$  and, following Honoré and Powell (2005), *pairwise differences*  $\Delta^s y_{it} = y_{it} - y_{it-s}$  with  $s = 1, \dots, t-1$ .<sup>9</sup> Then, they apply the standard robust LTS estimators of linear panel data models to the transformed data. Since both the difference and pairwise differences transformations are linear transformations, the LTS on such transformed data keeps regression and affine equivariances. This is in contrast to the LST applied to the median-transformed data proposed by Bramati and Croux (2007). Aquaro and Čížek (2013) establish the equivariance, robust and asymptotic properties of their estimators. Their estimators have asymptotic normal distributions as well as a BDP converging to 25%.

Following Bramati and Croux (2007), Aquaro and Čížek (2013) use the LTS estimator:

$$\hat{\beta}_{LTS,\mathcal{F}} = \arg \min_{\beta} \sum_{h=1}^H [r_{h,\mathcal{F}}^2(\beta)]_{h:NT} = \arg \min_{\beta} \sum_{h=1}^H \left[ (\gamma_{h,\mathcal{F}} - X_{h,\mathcal{F}}\beta)^2 \right]_{h:NT}, \quad (25)$$

where  $r_{h,\mathcal{F}}^2(\beta)$  is the  $h$ th smallest order statistic of the squared residuals and the  $(i,t)$ th residual is given by  $r_{it,\mathcal{F}}(\beta) = y_{it,\mathcal{F}} - X_{it,\mathcal{F}}\beta$ . The  $\mathcal{F}$ -transformed dependent variable  $y_{it,\mathcal{F}}$  is defined either as the median  $y_{it,\mathcal{F}} = \text{median}_t(y_{it})$ , the first difference  $y_{it,\mathcal{F}} = \Delta y_{it}$  or the pairwise differences  $y_{it,\mathcal{F}} = \Delta^s y_{it}$ . To improve the efficiency of the LTS estimator, efficient one-step methods have been developed. Aquaro and Čížek (2013) apply iterated reweighted least squares (*IRLS*) (see Rousseeuw and Leroy 2003) to the linear panel data model with fixed effects. This estimator removes observations having large absolute residuals according to some initial LTS estimator. Once  $\hat{\beta}_{LTS,\mathcal{F}}^{(0)}$  and  $\hat{\sigma}_{LTS,\mathcal{F}}^{(0)}$  of the initial LST are obtained, the residuals are given by:  $r_{h,\mathcal{F}}(\hat{\beta}_{LTS,\mathcal{F}}^{(0)}) = y_{it,\mathcal{F}} - X_{it,\mathcal{F}}\hat{\beta}_{LTS,\mathcal{F}}^{(0)}$  and the weights are defined as:

$$\hat{w}_{it}(\hat{\beta}_{LTS,\mathcal{F}}^{(0)}, \hat{\sigma}_{LTS,\mathcal{F}}^{(0)}, v) = I\left(\left|\frac{r_{h,\mathcal{F}}(\hat{\beta}_{LTS,\mathcal{F}}^{(0)})}{\hat{\sigma}_{LTS,\mathcal{F}}^{(0)}}\right| < v\right) \quad (26)$$

for some constant  $v > 0$  (usually,  $v = 2.5$ ). The *IRLS* estimator is given by

$$\widehat{\beta}_{IRLS,\mathcal{F}} = \arg \min_{\beta} \sum_{i=1}^N \sum_{t=1}^{T(\mathcal{F})} \widehat{w}_{it} \left( \widehat{\beta}_{LTS,\mathcal{F}}^{(0)}, \widehat{\sigma}_{LTS,\mathcal{F}}^{(0)}, \nu \right) \cdot r_{it,\mathcal{F}}^2(\beta) \quad (27)$$

with  $T(\mathcal{F}) = T, (T-1), T(T-1)/2$  for the median, the first difference and the pairwise differences transformation, respectively. To achieve the efficiency for normally distributed data, Aquaro and Čížek (2013) apply the robust and efficient weighted least squares (REWLS) (see Gervini and Yohai 2002) to the linear panel data with fixed effects. This estimator is a data-adaptive version of the IRLS estimator. Consider the distribution  $F^+$  which is associated with the absolute standardized residuals  $|r_{h,\mathcal{F}}(\beta_{LTS,\mathcal{F}}^{(0)})/\sigma_{LTS,\mathcal{F}}^{(0)}|$  and  $F_0^+$  which is associated with the distribution of the absolute standardized residuals of the standard linear fixed effect model.  $F^+$  (and similarly for  $F_0^+$ ) is estimated with the empirical distribution functions  $\widehat{F}^+$  using the estimated absolute standardized residuals  $|r_{h,\mathcal{F}}(\widehat{\beta}_{LTS,\mathcal{F}}^{(0)})/\widehat{\sigma}_{LTS,\mathcal{F}}^{(0)}|$ . The maximum difference  $\widehat{d}_{NT}$  between the tails of the two empirical distributions functions is obtained from

$$\widehat{d}_{NT} = \sup_{\nu \geq \eta} \{ [\widehat{F}_0^+(\nu) - \widehat{F}^+(\nu)] \cdot I(\widehat{F}_0^+(\nu) - \widehat{F}^+(\nu)) \geq 0 \}, \quad (28)$$

where  $\eta$  is a large quantile of  $F_0^+$ . Aquaro and Čížek (2013) use  $\eta = 2.5$  for Gaussian errors with  $F_0^+ = N(0, 1)$ . The robust and efficient weighted least squares (REWLS) estimator is then defined as:

$$\widehat{\beta}_{REWLS,\mathcal{F}} = \arg \min_{\beta} \sum_{i=1}^N \sum_{t=1}^{T(\mathcal{F})} \widehat{w}_{it} \left( \widehat{\beta}_{LTS,\mathcal{F}}^{(0)}, \widehat{\sigma}_{LTS,\mathcal{F}}^{(0)}, \widehat{\nu} \right) \cdot r_{it,\mathcal{F}}^2(\beta), \quad (29)$$

where the estimated cutoff point  $\widehat{\nu}$  is defined as the  $(1 - \widehat{d}_{NT})$ th quantile of the distribution  $F^+$ :  $\widehat{\nu} = \min \{ \nu \mid F^+(\nu) \geq 1 - \widehat{d}_{NT} \}$ . This estimator preserves the BDP properties of the initial robust estimator and achieves the asymptotic efficiency of the normal errors. Last, Čížek (2010) and Aquaro and Čížek (2013) propose a reweighted LTS (RLTS) estimator. The weights in Eq.(26) are constructed with the data-dependent cutoff point  $\widehat{\nu}$  but these new weights are used with the LTS estimator instead of the least squares estimator:

$$\widehat{\beta}_{RLTS,\mathcal{F}} = \arg \min_{\beta} \sum_{h=1}^{\widehat{H}} [r_{h,\mathcal{F}}^2(\beta)]_{h:NT} = \arg \min_{\beta} \sum_{h=1}^{\widehat{H}} [(y_{h,\mathcal{F}} - X_{h,\mathcal{F}}\beta)^2]_{h:NT}, \quad (30)$$

where

$$\widehat{H} = \sum_{i=1}^N \sum_{t=1}^{T(\mathcal{F})} I \left( \left| \frac{r_{h,\mathcal{F}}(\widehat{\beta}_{LTS,\mathcal{F}}^{(0)})}{\widehat{\sigma}_{LTS,\mathcal{F}}^{(0)}} \right| < \widehat{\nu} \right) = \sum_{i=1}^N \sum_{t=1}^{T(\mathcal{F})} \widehat{w}_{it} \left( \widehat{\beta}_{LTS,\mathcal{F}}^{(0)}, \widehat{\sigma}_{LTS,\mathcal{F}}^{(0)}, \widehat{\nu} \right). \quad (31)$$

This estimator also preserves the BDP properties of the initial robust estimator. It is also asymptotically independent of the initial robust estimator and allows the asymptotic efficiency of the Gaussian errors. Aquareo and Čížek (2013) conducted Monte Carlo simulations across different samples sizes ( $N = 50, 100, 200, 400$ ,  $T = 3$ ) and ( $N = 50$ ,  $T = 4, 6, 12, 24$ ). They show that robust estimators based on the median transformation are not consistent for a fixed number of time periods (i.e.,  $T = 3$ ).<sup>10</sup> When  $T > 3$ , the best mean square error results are obtained for the pairwise differences transformation.<sup>11</sup>

## 14.3 ROBUST IV ESTIMATORS

---

Classic estimators such as ordinary least squares (OLS), generalized least squares (GLS), two-stage least squares (2SLS) and generalized method of moments (GMM) have a breakdown point of zero. When outliers are suspected and the orthogonality conditions are not fulfilled, we need robust consistent instrumental variables (IV) estimators. Outliers generate two kinds of problems for IV estimators. If the orthogonality conditions required for consistency are valid for the bulk of the data, this may not be the case for the outliers or bad leverage points. Second, outliers may invalidate the rank condition for identification. For instance, outliers can cause the projections of explanatory variables on the instruments to be perfectly collinear and lead to a break down of the 2SLS estimator.

Wagenvoort and Waldmann (2002) proposed two  $B$ -robust estimation procedures for the linear static panel data model: the two-stage generalized  $M$ -estimator (2SGM) and the robust generalized method of moments estimator (RGMM) (see also Ronchetti and Trojani (2001)). These two estimators are  $B$ -robust in the sense that their influence function is bounded, consistent, and asymptotically normally distributed. As the 2SGM-estimator does not fully exploit all the moment conditions, an alternative is the RGMM-estimator for linear panel data model with correlated and/or heteroskedastic errors.

The two-stage generalized MS-estimate (2SGM) is obtained as follows:

*Stage 1:* suppose that there are  $M$  ( $\geq K$ ) instrumental variables  $A_{it}$  which are correlated with the explanatory variables  $X_{it}$  but independent of the error term  $r_{it}$ . The errors are assumed to be *iid* with zero mean and variance  $\sigma^2$ . The explanatory variable  $X_k$  (the  $k^{th}$  column of  $X$ ) is regressed on the instrumental variables  $A$ :  $X_{it,k} = A_{it}\eta_k + \xi_{it,k}$ . The high breakpoint generalized  $M$ -estimate (GM) and the prediction of the  $k^{th}$  column of  $X$  is computed according to:

$$\widehat{X}_k = A (A' W_A(A) W_r(r_{1,k}) A)^{-1} A' W_A(A) W_r(r_{1,k}) X_k, \quad (32)$$

where  $W_A(A)$  and  $W_r(r_{1,k})$  are the diagonal matrices comprising the weight functions  $W_A(A_{it})$  and  $W_r(r_{1it,k})$ .  $r_{1,k}$  are the first stage GM residuals associated with  $X_k$ .

$(r_{1,k} = X_k - A\hat{\eta}_k)$  and  $W_r(r_{1,k})$  differs for every distinct column of  $X$ . Thus  $K$  separate GM regressions are performed.

Stage 2: the explanatory variables of the original equation are replaced by their robust projection on  $A$ . This returns the high breakpoint generalized  $M$ -estimator, called the 2SGM estimator:

$$\hat{\beta}_{2SGM} = (\hat{X}' W_X(\hat{X}) W_r(r_2) \hat{X})^{-1} \hat{X}' W_X(\hat{X}) W_r(r_2) \hat{y}, \quad (33)$$

where  $W_Z(\hat{Z})$  and  $W_r(r_2)$  are diagonal matrices containing the second step GM weights and  $r_2$  are the second stage GM residuals ( $r_2 = y - \hat{X}\hat{\beta}_{2SGM}$ ). This estimator is consistent and asymptotically distributed but leads to a complicated expression of a consistent and relatively efficient variance-covariance matrix of the 2SGM estimator (see Wagenvoort and Waldmann 2002 and Baltagi and Bresson 2012).

The previous estimation method can be extended to the general case with any finite error covariance matrix  $\Sigma$ . A more efficient estimator can be obtained using the relevant moment conditions. In the IV case, the empirical moment condition:

$$X' A (A' \Sigma A)^{-1} A' (y - X\beta) = 0 \quad (34)$$

is an unreliable approximation of the theoretical moment condition when the data are contaminated. So, Wagenvoort and Waldmann (2002) suggested robustifying Eq.(34) using the weights  $W_A(.)$  as in the case of the 2SGMS estimator. Moreover, they proposed to multiply all the variables  $y_{it}$  and  $X_{it}$  by the square root of the 2SGMS weight functions  $W_X(.)$  and  $W_r(.)$ . The robust normal equations becomes:

$$\begin{aligned} & X' \sqrt{W_X(X)} \sqrt{W_r(r)} W_A(A) A (A' W_A(A) \Omega W_A(A) A)^{-1} \times \\ & A' W_A(A) \sqrt{W_X(X)} \sqrt{W_r(r)} (y - X\beta) = 0, \end{aligned} \quad (35)$$

where the appropriate variance-covariance matrix of the weighted errors is:

$$\Omega = E \left[ \sqrt{W_r(r)} \sqrt{W_X(X)} \varepsilon \varepsilon' \sqrt{W_X(X)} \sqrt{W_r(r)} \right]. \quad (36)$$

The estimated variance-covariance matrix  $\hat{\Omega}$  is block-diagonal since they assume no correlation between cross-sectional units. Then, for individual  $i$ , the diagonal elements  $\hat{\omega}_{i,t}$  of this  $(T \times T)$  matrix  $\hat{\Omega}_i$  are given by the squared weighted 2SGMS residuals  $\hat{\omega}_{i,t} = r_{it}^2 \cdot W_X(X_{it}) W_r(y_{it} - X_{it} \hat{\beta}_{2SGMS})$ . The non diagonal elements  $\hat{\omega}_{i,ts}$  of  $\hat{\Omega}_i$  are computed as:

$$\hat{\omega}_{i,ts} = \frac{1}{N} \sum_{i=1}^N r_{it} \sqrt{W_r(r_{it}) W_X(X_{it})} r_{is} \sqrt{W_r(r_{is}) W_X(X_{is})}. \quad (37)$$

Solving the first order condition of Eq.(35) with respect to  $\beta$  leads to the *RGMM* estimator:

$$\widehat{\beta}_{RGMM} = \begin{bmatrix} X' \sqrt{W_X(X) W_r(r)} W_A(A) A (A' W_A(A) \widehat{\Omega} W_A(A) A)^{-1} \times \\ \sqrt{W_X(X) W_r(r)} X \\ X' \sqrt{W_X(X) W_r(r)} W_A(A) A (A' W_A(A) \widehat{\Omega} W_A(A) A)^{-1} \times \\ \sqrt{W_X(X) W_r(r)} y \end{bmatrix}^{-1}. \quad (38)$$

This estimator leads to a complicated expression for the consistent and efficient estimation of the variance-covariance matrix (see Wagenvoort and Waldmann 2002). But, this estimator is *B*-robust in the sense that it generates reliable parameter estimates even if there are small disturbances in the specified distribution of the error terms. The usual GMM method does not have this appealing property. Wagenvoort and Waldmann (2002) extend this *RGMM* estimator to the case of a linear panel data with fixed effects and measurement errors. In a Monte Carlo study, they use the *B*-robust *RGMM* estimator and make corrections for conditional heteroskedasticity, serial correlation in the regression errors, measurement errors in the independent variables and outlying observations for  $N = 100$  and  $T = 4$ . They show that the *RGMM* estimator still provides accurate estimates when outliers corrupt the data while the usual *GMM* estimator shows significant bias and reduced efficiency.

### 14.3.1 Robust Hausman-Taylor Estimator

The Hausman-Taylor panel data estimator deals with the empirical fact that some of our explanatory variables are time varying, while others are time invariant. In addition, some are correlated with the individual effects and some are not. Hausman and Taylor (1981) (hereinafter HT) proposed a two-step instrumental variable procedure that is (i) more efficient than the Within estimator and (ii) recaptures the effects of time invariant variables which are wiped out by the Within transformation. But, this two-step method may lead to bias and inefficient estimates when the data set is contaminated. Baltagi and Bresson (2012) (hereinafter BB) propose a robust HT estimator that deals with the possible presence of outliers. This entails two modifications of the classical HT estimator. The first modification uses the Bramati and Croux (2007) robust WMS-estimator instead of the Within estimator in the first stage of the HT estimator. The second modification uses a two-stage generalized MS-estimate (2SGMS)—in the spirit of the robust estimator of Wagenvoort and Waldmann (2002)—instead of the 2SLS estimate in the second step of the HT estimator.

Consider the HT model where some of the explanatory variables are time varying ( $X_{it}$ ), while others are time invariant ( $Z_i$ ):

$$y_{it} = X'_{it}\beta + Z'_i\gamma + \mu_i + \nu_{it}, \quad i = 1, \dots, N, \quad t = 1, \dots, T \quad (39)$$

$\mu_i$  is IID( $0, \sigma_\mu^2$ ),  $v_{it}$  is IID( $0, \sigma_v^2$ ) independent of each other and among themselves. HT allowed *some* of the  $X$  and  $Z$  variables to be correlated with the individual effects  $\mu_i$ . This is in contrast to the fixed effects estimator where *all* the regressors are correlated with the individual effects, and the random effects estimator where *none* of the regressors are correlated with the individual effects. Using the HT notation:  $X = [X_1, X_2]$  and  $Z = [Z_1, Z_2]$  where  $X_1$  is  $(NT \times k_1)$ ,  $X_2$  is  $(NT \times k_2)$ ,  $Z_1$  is  $(NT \times g_1)$  and  $Z_2$  is  $(NT \times g_2)$ .  $X_1$  and  $Z_1$  are assumed exogenous in that they are not correlated with  $\mu_i$  and  $v_{it}$ , while  $X_2$  and  $Z_2$  are endogenous because they are correlated with  $\mu_i$ , but not  $v_{it}$ .

HT proposed the following two-step consistent estimator of  $\beta$  and  $\gamma$ :

1. Perform the fixed effects (FE) or Within estimator obtained by regressing  $\tilde{y}_{it} = (y_{it} - \bar{y}_i)$ , where  $\bar{y}_i = \sum_{t=1}^T y_{it}/T$ , on a similar within transformation on the regressors. Note that the Within transformation wipes out the  $Z_i$  variables since they are time invariant, and we only obtain an estimate of  $\beta$  which we denote by  $\tilde{\beta}_W$ .
  - Then, HT average the within residuals over time

$$\hat{d}_i = \bar{y}_i - \bar{X}'_i \tilde{\beta}_W \quad (40)$$

- To get an estimate of  $\gamma$ , HT suggest running 2SLS of  $\hat{d}_i$  on  $Z_i$  with the set of instruments  $A = [X_1, Z_1]$ . This yields

$$\hat{\gamma}_{2SLS} = (Z' P_A Z)^{-1} Z' P_A \hat{d}_i, \quad (41)$$

where  $P_A = A(A'A)^{-1}A'$ . It is clear that the order condition has to hold ( $k_1 \geq g_2$ ) for  $(Z' P_A Z)$  to be nonsingular. In fact, if  $k_1 = g_2$ , then the model is just-identified and one stops here.

2. If  $k_1 > g_2$ , HT suggest estimating the variance-components as follows:

$$\hat{\sigma}_v^2 = (y_{it} - X'_{it} \tilde{\beta}_W)' Q (y_{it} - X'_{it} \tilde{\beta}_W) / N(T-1) \quad (42)$$

and

$$\hat{\sigma}_1^2 = (y_{it} - X'_{it} \tilde{\beta}_W - Z'_i \hat{\gamma}_{2SLS})' P (y_{it} - X'_{it} \tilde{\beta}_W - Z'_i \hat{\gamma}_{2SLS}) / N, \quad (43)$$

where  $\sigma_1^2 = T\sigma_\mu^2 + \sigma_v^2$ .  $P = I_N \otimes \bar{J}_T$  and  $\bar{J}_T = J_T/T$ , with  $I_N$  being a matrix of dimension  $N$ , and  $J_T$  is a matrix of ones of dimension  $T$ .  $P$  is a matrix which averages the observation across time for each individual.  $Q = I_{NT} - P$ . Once the variance-components estimates are obtained, the model is transformed using  $\hat{\Omega}^{-1/2}$  where

$$\hat{\Omega}^{-1/2} = \frac{1}{\sigma_1} P + \frac{1}{\sigma_v} Q, \quad (44)$$

see Baltagi (2008). Note that  $y^* = \hat{\sigma}_v \hat{\Omega}^{-1/2} y$  has a typical element  $y_{it}^* = y_{it} - \hat{\theta} \bar{y}_i$  where  $\hat{\theta} = 1 - (\hat{\sigma}_v / \hat{\sigma}_1)$  and  $X_{it}^*$  and  $Z_i^*$  are defined similarly. In fact, the

transformed regression becomes:

$$\widehat{\sigma}_v \widehat{\Omega}^{-1/2} y_{it} = \widehat{\sigma}_v \widehat{\Omega}^{-1/2} X_{it} \beta + \widehat{\sigma}_v \widehat{\Omega}^{-1/2} Z_i \gamma + \widehat{\sigma}_v \widehat{\Omega}^{-1/2} u_{it}, \quad (45)$$

where  $u_{it} = \mu_i + v_{it}$ . The asymptotically efficient HT estimator is obtained by running 2SLS on this transformed model using  $A_{HT} = [\tilde{X}, \bar{X}_1, Z_1]$  as the set of instruments. In this case,  $\tilde{X}$  denotes the within transformed  $X$  and  $\bar{X}_1$  denotes the time average of  $X_1$ . More formally, the HT estimator under over-identification is given by:

$$\begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix}_{HT} = \left[ \begin{pmatrix} X^{*'} \\ Z^{*'} \end{pmatrix} P_{A_{HT}}(X^*, Z^*) \right]^{-1} \begin{pmatrix} X^{*'} \\ Z^{*'} \end{pmatrix} P_{A_{HT}} y^*, \quad (46)$$

where  $P_{A_{HT}}$  is the projection matrix on  $A_{HT} = [\tilde{X}, \bar{X}_1, Z_1]$ , see also Breusch, Mizon, and Schmidt (1989).

To robustify the HT estimator for the possible presence of outliers, two MS-estimators are successively used. In the first step, Baltagi and Bresson (2012) use the WMS proposed by Bramati and Croux (2007) and compute the residuals  $\hat{r}_{it}$ . Then, they use a 2SGMS-estimator for the estimation of  $\gamma$ . As in the second step of HT, they compute the variance components estimates  $\widehat{\sigma}_v$  and  $\widehat{\sigma}_1$  and use a 2SGMS-estimator for the estimation of both  $\beta$  and  $\gamma$ .

*First step.* Compute  $\tilde{\beta}_{WMS}$ , the WMS estimator given in Eqs.(17–18). Instead of averaging the within residuals over time as HT suggest, BB obtain the median of the resulting residuals over time:

$$\hat{r}_i = \text{med}_t(y_{it} - X'_{it} \tilde{\beta}_{WMS}) \quad (47)$$

and instead of the 2SLS procedure suggested by HT, BB propose a 2SGMS estimator following Wagenvoort and Waldmann (2002).

*Stage 1:* suppose that there are  $m_1$  instrumental variables  $A_{it}$  which are correlated with the explanatory factors  $Z_i$  but independent of the error term  $\varepsilon_{it}$  ( $= \mu_i + v_{it}$ ). The explanatory variable  $Z_k$  (the  $k^{th}$  column of  $Z$ ) is regressed on the instrumental variables  $A = [X_1, Z_1]$ :  $Z_{it,k} = A_{it} \eta_k + \xi_{it,k}$ . The high breakpoint generalized  $M$ -estimate and the prediction of the  $k^{th}$  column of  $Z$  is computed according to:

$$\widehat{Z}_k = A \left( A' W_A(A) W_r(r_{1,k}) A \right)^{-1} A' W_A(A) W_r(r_{1,k}) Z_k \quad (48)$$

where  $W_A(A)$  and  $W_r(r_{1,k})$  are the diagonal matrices comprising the weight functions  $W_A(A_{it})$  and  $W_r(r_{1it,k})$ .  $r_{1,k}$  are the first stage GM residuals associated with  $Z_k$  ( $r_{1,k} = Z_k - A \widehat{\eta}_k$ ) and  $W_r(r_{1,k})$  differs for every distinct column of  $Z$ . Thus  $(g_1 + g_2)$  separate GM regressions are performed if  $\dim(Z) = g_1 + g_2$ . Contrary to Wagenvoort and Waldmann (2002), BB suggest using the residuals  $(r_{1,k})$  to estimate a new robust scale estimator of the residuals  $\widehat{\sigma}_S$ , which is then used to re-estimate a new weight function  $W_r(r_{1it,k})$ , and so on. Following the suggestion of Maronna and Yohai (2006), BB compute this iterated MS procedure using a maximum of 20 iterations.

*Stage 2:* replacing the explanatory variables of the original equation by their robust projection on  $A$ . This returns the high breakpoint 2SGMS-estimator:

$$\tilde{\gamma}_{2SGMS} = (\widehat{Z}' W_Z(\widehat{Z}) W_r(r_2) \widehat{Z})^{-1} \widehat{Z}' W_Z(\widehat{Z}) W_r(r_2) \widehat{r}_2, \quad (49)$$

where  $W_Z(\widehat{Z})$  and  $W_r(r_2)$  are diagonal matrices containing the second step GMS weights and  $r_2$  are the second stage GMS residuals ( $r_2 = y - \widehat{Z}\tilde{\gamma}_{2SGMS}$ ).

*Second step.* The variance-components estimates are obtained as follows:

$$\tilde{\sigma}_v^2 = (y_{it} - X'_{it}\tilde{\beta}_{WMS})' Q(y_{it} - X'_{it}\tilde{\beta}_{WMS}) / N(T-1) \quad (50)$$

and

$$\tilde{\sigma}_1^2 = (y_{it} - X'_{it}\tilde{\beta}_{WMS} - Z'_i\tilde{\gamma}_{2SGMS})' P(y_{it} - X'_{it}\tilde{\beta}_{WMS} - Z'_i\tilde{\gamma}_{2SGMS}) / N, \quad (51)$$

where  $\sigma_1^2 = T\sigma_\mu^2 + \sigma_v^2$ . Once the variance-components estimates are obtained, BB compute

$$y_{it}^* = y_{it} - \tilde{\theta}\bar{y}_i, \quad (52)$$

where  $\tilde{\theta} = 1 - \tilde{\sigma}_v/\tilde{\sigma}_1$  and  $X_{it}^*$  and  $Z_i^*$  are defined similarly. The 2SGMS procedure applied to this transformed model can be described as follows:

*Stage 1:* each explanatory variable of  $V = [X_{it}^*, Z_i^*]$  is regressed on the  $m_2$  instrumental variables  $A_{HT} = [\tilde{X}, \bar{X}_1, Z_1]$ . The  $k^{th}$  explanatory variable  $V_k$  is regressed on the instrumental variables:  $V_{it,k} = A_{HT,it}\delta_k + \xi_{it,k}$ . This returns the GM estimate, and the prediction of the  $k^{th}$  column of  $V = [X_{it}^*, Z_i^*]$  is computed according to:

$$\begin{aligned} \widehat{V}_k &= A_{HT} (A'_{HT} W_{A_{HT}}(A_{HT}) W_r(r_{1,k}) A_{HT})^{-1} \times \\ &\quad A'_{HT} W_{A_{HT}}(A_{HT}) W_r(r_{1,k}) V_k. \end{aligned} \quad (53)$$

$W_{A_{HT}}(A)$  and  $W_r(r_{1,k})$  are the diagonal matrices comprising the weight functions  $W_{A_{HT}}(A_{HT,it})$  and  $W_r(r_{1it,k})$ .  $r_{1,k}$  are the first stage GM residuals associated with  $V_k$  ( $r_{1,k} = V_k - A_{HT}\widehat{\delta}_k$ ), and  $W_r(r_{1,k})$  differs for every distinct column of  $V$ .

Thus ( $K = k_1 + k_2 + g_1 + g_2$ ) separate GM regressions are performed if  $\dim(V) = K$ . With these residuals ( $r_{1,k}$ ), BB estimate a new robust scale estimator of the residuals  $\widehat{\sigma}_S$  which is used to re-estimate a new weight function  $W_r(r_{1it,k})$ , and so on. Following the suggestion of Maronna and Yohai (2006), BB compute this iterated MS procedure up to a maximum of 20 iterations.

*Stage 2:* replacing the explanatory variables of the original equation by their robust projection on  $A_{HT}$  and applying the GM technique one more time provides the 2SGMS-estimates:

$$\tilde{\lambda}_{2SGM} = \left( \begin{array}{c} \tilde{\beta} \\ \tilde{\gamma} \end{array} \right)_{2SGMS} = \frac{(\widehat{V}' W_V(\widehat{V}) W_r(r_2) \widehat{V})^{-1} \times}{\widehat{V}' W_V(\widehat{V}) W_r(r_2) y^*}. \quad (54)$$

$W_V(\widehat{V})$  and  $W_r(r_2)$  are diagonal matrices containing the second step GMS weights and  $r_2$  are the second stage GMS residuals ( $r_2 = y^* - \widehat{V}\tilde{\lambda}_{2SGMS}$ ).

As for the 2SGM-estimator of Wagenvoort and Waldmann (2002), the variance-covariance matrix of the robust HT estimator has a complicated expression (see Baltagi and Bresson 2012). Monte Carlo simulations were conducted for  $N = (100, 200)$  and  $T = (5, 10)$ . When no outliers are present, the robust HT shows small loss in mean square error (MSE) relative to the standard HT estimator. Contrasting that to the various types of 5% and 10% contaminations considered, the gain in absolute as well as relative MSE is huge for the robust HT estimator compared to the classical HT estimator. When the outliers are block-concentrated vertical outliers, BB get similar results but when the outliers are block-concentrated leverage points, the gain in MSE of the robust HT estimate becomes more pronounced. Whatever the sampling scheme, the MSE of the parameter of the  $Z_i$  variables is always more affected than that of the other parameters but the robust version yields better results than HT no matter what type of contamination. The use of the robust HT estimator on the Cornwell and Rupert (1988) Mincer wage equation shows that the returns to education are roughly the same as in the classical HT case but the gender effect becomes significant and the race effect becomes smaller as compared to the classical HT estimates.

### 14.3.2 Robust Estimators for the Linear Dynamic Panel Data Model

Robust GMM estimators have been applied to dynamic linear panel data models. Lucas, Van Dijk, and Kloek (1994) seem to be the first to propose a variant of the GMM estimator which is less sensitive to contaminated data (see also Lucas, Van Dijk, and Kloek 2007). The usual Arellano and Bond (1991) GMM estimator was not designed to be outlier-robust. It is, in fact, highly sensitive to even a small number of outliers. Lucas, Van Dijk, and Kloek (2007) show that the Arellano-Bond estimator has an unbounded influence function. They propose an estimator that replaces the standard moment conditions by observation weighted moment conditions. Using Monte-Carlo simulations, they show that the standard GMM estimator is very sensitive to aberrant observations. They also compare robust and non robust GMM estimators for the determinants of capital structure of 715 American non financial firms over the period 1987–1992. They find that the robust GMM estimator is stable, whereas the traditional, nonrobust GMM estimator is not stable. Moreover, the over-identifying restrictions test supports the moment conditions used by the robust GMM estimator, but rejects the moment conditions used by the nonrobust estimator. Later, Janz (2003) applies the same robust GMM estimator to an Euler equation model of firm investment for a panel of 96 German non financial stock companies observed over the period 1987–1992. The robust GMM can be used as a diagnostic to check whether the results obtained with the usual GMM estimator are driven by outliers or bad leverage points. Janz finds that the usual GMM estimator yields empirical results which contradict the theory implying

a negative discount rate and a positive price elasticity of demand, whereas the robust GMM results yield a positive discount rate and a negative price elasticity of demand.

Consider the standard linear dynamic panel data model with individual specific effects, see chapter 4 in this handbook:

$$\begin{aligned} y_{it} &= \gamma y_{it-1} + X_{it}\beta + \alpha_i + r_{it} \\ &= Z_{it}\delta + u_{it}, \quad i = 1, \dots, N, t = 1, \dots, T, \end{aligned} \quad (55)$$

where  $Z_{it} = [y_{it-1}, X_{it}]$ ,  $\delta' = (\gamma, \beta')$  and the disturbance  $u_{it}$  is the sum of the individual specific effect  $\alpha_i \sim iid(0, \sigma_\alpha^2)$  and a remainder term  $r_{it} \sim iid(0, \sigma_r^2)$ . Let  $F^D$  be the  $((T-1) \times T)$  first difference filter matrix which transforms the  $(T \times 1)$  vector  $y_i$  into a vector of first difference:  $F^D y_i = (\Delta y_{i2}, \dots, \Delta y_{iT})$ , and let  $A = diag(A_1, \dots, A_N)$  be the instrumental variable matrix such that the  $((T-1) \times M)$  sub-matrix  $A_i$  may either be defined as in Arellano and Bond (1991) ( $A_{i,AB}$ ) or Ahn and Schmidt (1995) ( $A_{i,AS}$ ):

$$A_{i,AB} = \begin{pmatrix} [y_i^0, X_i^1] & 0 & 0 \\ 0 & [y_i^1, X_i^2] & 0 \\ & & \ddots \\ 0 & & [y_i^{T-2}, X_i^{T-1}] \end{pmatrix} \quad (56)$$

$$\text{or } A_{i,AS} = \begin{pmatrix} y_i^0 & 0 & 0 & X_i^T & 0 & 0 \\ 0 & y_i^1 & 0 & 0 & X_i^T & 0 \\ & & \ddots & & & \ddots \\ 0 & 0 & y_i^{T-2} & 0 & 0 & X_i^T \end{pmatrix},$$

where  $y_i^t = (y_{i1}, \dots, y_{it})$  and  $X_i^t = (X_{i1}, \dots, X_{it})$ . The Ahn and Schmidt ( $A_{i,AS}$ ) IV matrix is based on a strict exogeneity assumption for the  $X_i$  rather than a weak exogeneity assumption as for the Arellano and Bond IV matrix (see Baltagi 2008). Lucas, Van Dijk, and Kloek (1994, 2007) proposed a robust GMM estimator which minimizes the following quadratic criterion function:

$$\min_{\delta} (y - Z\delta)' \left( I_N \otimes F^D \right)' W_A(A) W_r(r) A \Omega_N A' W_A(A) W_r(r) \left( I_N \otimes F^D \right) (y - Z\delta), \quad (57)$$

where  $I_N$  is the identity matrix of dimension  $N$  and  $\otimes$  is the Kronecker product.  $\Omega_N$  is an asymptotically non stochastic, positive definite weighting matrix of dimension  $M$ , the number of instruments. As the robust GMM estimator is nonlinear, it has to be computed iteratively. The robust first-step GMM estimator is given by:

$$\begin{aligned} \hat{\delta}^{(0)} &= \left( Z' \left( I_N \otimes F^D \right)' W_A(A) A \Omega_N^{(0)} A' W_A(A) \left( I_N \otimes F^D \right) Z \right)^{-1} \times \\ &\quad Z' \left( I_N \otimes F^D \right)' W_A(A) A \Omega_N^{(0)} A' W_A(A) \left( I_N \otimes F^D \right) y \end{aligned} \quad (58)$$

with  $\Omega_N^{(0)} = \left( N^{-1} \sum_{i=1}^N A_i' W_A(A_i) \Theta W_A(A_i) A_i \right)^{-1}$  where  $\Theta$  is the Toeplitz matrix built by the  $(T-1)$ -dimensional vector  $(2, -1, 0, \dots, 0)'$  (see Baltagi 2008). At iteration  $j$ , Lucas, and Van Dijk, and Kloek (1994, 2007) and Janz (2003) update the weights matrix  $W_r(r)$  whose  $(i, t)$  diagonal element  $W_r(r_{it})$  is a function of  $(r_{it}/\sigma)$  (see Eq.(8)). To update the scale  $\sigma$ , they use the MAD estimator:  $\hat{\sigma}^{(j)} = \text{med}(\left| \hat{r}_{it}^{(j)} - \text{med}(\hat{r}_{it}^{(j)}) \right|)$ . Then, the weighting matrix  $\Omega_N^{(j)}$  is computed as:

$$\Omega_N^{(j)} = \left( N^{-1} \sum_{i=1}^N A_i' W_A(A_i) W_r(\hat{r}_i^{(j-1)}) \hat{r}_i^{(j-1)} \hat{r}_i^{(j-1)'} W_r(\hat{r}_i^{(j-1)}) W_A(A_i) A_i \right)^{-1} \quad (59)$$

allowing them to get the robust GMM estimates  $\hat{\delta}^{(j)}$ .

Dhaene and Zhu (2009) proposed a robust estimator for the linear dynamic panel data model with fixed effects in short panels (i.e.,  $N$  large and  $T$  small). In the case of a pure AR(1) process with fixed effects:  $y_{it} = \gamma y_{it-1} + \alpha_i + r_{it}$ , they show that the estimator

$$\hat{\gamma} = 1 + 2 \cdot \text{med}\left( \frac{\Delta y_{it}}{\Delta y_{it-1}} \right) \quad (60)$$

is highly robust as compared to the Blundell and Bond (1998) estimator or the maximum likelihood estimator based on differenced data as suggested by Hsiao, Pesaran, and Tahmisioglu (2002). Moreover, it has a reasonably high BDP when the data are corrupted.

The use of heavy-tailed distributions is a valuable tool in developing robust Bayesian procedures, limiting the influence of outliers on posterior inference (see also Section 14.5). Juarez and Steel (2010) propose a non-Gaussian Bayesian autoregressive model with individual specific effects to accommodate heavy tails, skewness and/or outliers. They focus on an AR(1) process for an unbalanced panel:

$$y_{it} = \gamma y_{it-1} + \alpha_i (1 - \gamma) + \lambda^{-\frac{1}{2}} r_{it}, \quad i = 1, \dots, N, \quad t = 1, \dots, T_i \quad (61)$$

Prior distributions have to be specified for the model parameters  $(\gamma, (\alpha_1, \dots, \alpha_N), \lambda)$ , then inference can be carried out using Markov Chain Monte Carlo (MCMC) techniques. A common choice is to consider a standard Gaussian distribution for the errors  $r_{it} \sim IIN(0, 1)$ . In order to take into account heavy tails, skewness and/or outliers, the authors use a family of skewed versions of normal and Student- $t$  distributions. As the posteriors are not of a known form, they implement the Metropolis-Hastings within Gibbs samplers to estimate the parameters. They apply these non-Gaussian Bayesian models on three data sets. The first one concerns yearly average earnings of production workers in 14 metropolitan areas in California between 1945 and 1977. The second is drawn from the PSID and is related to annual labor earnings for 515 males over 1967–1991. The last one analyses annual growth of real GDP per capita for 25 OECD countries for the period 1950–2000. The application to average earnings in California strongly favours skewed specifications with moderate tails. In contrast, for the other

data sets, evidence against normality mainly focuses around heavy tails. Juarez and Steel (2010) also test the poolability of the dynamics across the whole panel. This is rejected for the two applications on individual earnings and GDP growth.

## 14.4 ROBUST ESTIMATORS FOR NONLINEAR PANEL DATA MODELS

---

Recently, Čížek (2008), Sutradhar and Bari (2007, 2010) and Sinha (2012a, 2012b) proposed robust analysis of longitudinal models for binary and count panel data with nonignorable missing response. The main framework is the generalized linear longitudinal mixed model (GLLMM) (see Breslow and Clayton 1993; Cantoni and Ronchetti 2001) which includes binary and Poisson mixed models.

Define the linear predictor as:

$$\eta_{it} = X_{it}\beta + Z_{it}\alpha_i, \quad (62)$$

where  $X_{it}$  and  $Z_{it}$  are observed covariates and  $\beta$  and  $\alpha_i$  represent the fixed (population) effects and the random specific effects respectively. The expected conditional value  $\mu_{it} = E[y_{it} | X_{it}, Z_{it}, \alpha_i]$  of the binary or count outcome variable  $y_{it}$  is related to the linear prediction by a monotonic link function<sup>12</sup>:  $\mu_{it} = g^{-1}(\eta_{it})$ . The GLLMM assumes  $y_{it}$  to be conditionally independent given the observed covariates  $X_{it}$  and  $Z_{it}$  with conditional density:

$$\zeta(y_{it} | X_{it}, Z_{it}, \alpha_i) = c(y_{it}, \phi) \cdot \exp\left(\frac{y_{it}\eta_{it} - \psi(\eta_{it})}{s(\phi)}\right) \quad (63)$$

for some functions  $c$ ,  $s$  and where  $\phi$  is a dispersion parameter and  $\mu_{it} = g^{-1}(\eta_{it}) = \partial\psi(\eta_{it})/\partial\eta_{it}$ . Assuming that  $\alpha_i$  are mutually *i.i.d.* with density function  $f(\alpha_i | \theta)$ , where  $\theta$  denotes the unknown parameters in the density, the likelihood function then becomes:

$$L = \sum_{i=1}^N \int \sum_{t=1}^T \zeta(y_{it} | X_{it}, Z_{it}, \alpha_i) f(\alpha_i | \theta) d\alpha_i. \quad (64)$$

Estimation and prediction methods use either the Laplace approximation, Gaussian quadrature, Monte Carlo integration or hybrid methods (see Lai and Shih 2003). From this general framework, Sutradhar and Bari (2007, 2010) propose robust inference in the presence of outliers for two particular specifications of the binary and count dynamic panel data models.<sup>13</sup> They develop a robust generalized quasi-likelihood approach. They introduce the fully standardized Mallow's type quasi-likelihood approach (FSMQL) for consistent regression estimation with outliers. As, for the likelihood or the quasi-likelihood approaches, the existing outliers robust Mallow's type quasi-likelihood (MQL) method can also generate biased regression estimators. Using a fully standardized score function in the MQL estimating equation,

they show that this new robust estimator is almost unbiased and ensures a higher consistency performance.<sup>14</sup>

Using the same generalized linear longitudinal mixed models framework (see Eq.(64)), Sinha (2012a, 2012b) proposes a robust methodology for analyzing nonignorable missing response in panel data. Missing data analysis based on the likelihood approach has been extensively studied in the statistics literature. Many authors suggest the use of EM algorithms for maximum likelihood estimation in generalized linear models for data with nonignorable missing covariates or missing responses. But full likelihood analysis of longitudinal data under nonignorable missingness requires intensive computation. To overcome this problem, pseudo-likelihood approaches were proposed (see Cantoni and Ronchetti 2001) but maximum likelihood and maximum pseudo-likelihood estimators are sensitive to potential outliers in the data. To cope with this, Sinha (2012a, 2012b) proposed a robust method in the framework of the pseudo-likelihood of Troxel, Lipsitz and Harrington (1998).<sup>15</sup> Sinha (2012a, 2012b) shows that the robust pseudo-likelihood (RPL) estimator is almost as efficient as ML for non corrupted data but, when the data are contaminated, the RPL provides the smallest bias and MSE for the estimated parameters. One disadvantage is the intensive computation as the number of individuals  $N$  becomes large.

It is clear that much work remains to be done to extend the robust linear panel data literature to non-linear panel data models with fixed and random effects. While some progress has been made in statistics using the GLLMM approach, this has not caught on in the standard non-linear panel data methods discussed in the econometrics literature, see Wooldridge (2010).

## 14.5 DETECTION OF INFLUENTIAL OBSERVATIONS AND OUTLIERS IN PANEL DATA

---

Since the seminal article of Cook (1977) on detection of influential observations in linear regression on cross-sections, considerable research has been done to extend Cook's distance to detecting influential observations in more complex data.<sup>16</sup> For instance, Banerjee and Frees (1997) consider the role of influence diagnostics in the case of longitudinal models similar to Eq.(62). They use the concept of partial influence to propose influence measures in the longitudinal setting under both fixed and random covariates. They extend existing partial influence diagnostics to GLS estimates that handle serial correlation in longitudinal models. Preisser and Qaqish (1996) developed Cook's distance for generalized estimating equations which are an extension of the generalized linear model to accommodate correlated data. More recently, Zhu, Ibrahim, and Cho (2012) have shown that deleting subsets with different numbers of observations introduces different degrees of perturbation to the model fitted to the data. In this

case, Cook's distance is associated with the degree of perturbation. Zhu, Ibrahim, and Cho (2012) suggest scaled Cook's distances provide important information about the relative influential level of each deleted subset.

Outlier rejection in Bayesian analysis was first described by De Finetti (1961). Theoretical results were given by O'Hagan (1988) who considered general Bayesian modeling based on Student- $t$  distributions (see also O'Hagan 1979, 1990, for a survey). There are several methods in the Bayesian framework for detecting outliers in panel data or longitudinal data. For instance, Empirical Bayes estimates using multi-level models have been suggested for detection and subsequent deletion of outliers (see Ecob and Der 2003). Peruggia, Santer, and Ho (2004) suggest detecting outliers within a hierarchical Bayesian modelling for panel data models with random coefficients and serial correlation defined as:

$$y_{it} = X_{it}\beta_i + r_{it}, \quad r_{it} = \sum_{j=1}^p \phi_j r_{it-j} + \varepsilon_{it}, \quad \beta_{ik} \sim N\left(Z_{ik}\beta_k^0, \sigma_{\beta_k}^2\right), \quad (65)$$

where the normal *prior* distributions for the  $\beta_{ik}$  depend upon specific covariates  $Z_{ik}$ . Peruggia, Santer, and Ho (2004) propose several diagnostics for detecting location-shift outliers on an extended version of Eq. (65). They want to know whether or not the regression contains measurement error outliers, and whether or not the coefficient vector  $\beta_i$  is inconsistent with the regression coefficient vector of the remaining individuals. They suggest the following specification:

$$y_{it} = X_{it}\beta_i + \delta_i^y c_{it} + r_{it}, \quad r_{it} = \sum_{j=1}^p \phi_j r_{it-j} + \varepsilon_{it}, \quad \beta_{ik} \sim N\left(Z_{ik}\beta_k^0 + \delta_i^\beta d_{ik}, \sigma_{\beta_k}^2\right), \quad (66)$$

where  $\delta_i^y$  and  $\delta_i^\beta$  take the values 0 or 1.  $c_{it}$  is a measurement-shift value whose *prior* distribution is  $c_i (= (c_{i1}, \dots, c_{iT})') \sim N(\mu_c e_T, \sigma_c^2 I_T)$  where  $e_T$  is a vector of ones of dimension  $T$ .  $d_{ik}$  is a shift in the regression coefficient  $\beta_{ik}$  whose *prior* distribution is  $d_{ik} \sim N(\mu_{dk}, \sigma_{dk}^2)$ . Once the *posteriors* of Eq.(65) and Eq.(66) are computed with MCMC methods, the diagnostics use the *posterior* distributions of outliers indicators and the location shift and compute the posterior probabilities  $\Pr(\delta_i^y = 1)$  and  $\Pr(\delta_i^\beta = 1)$ . The  $i$ th individual is judged to have one or more measurement outlying components when  $\Pr(\delta_i^y = 1)$  is high, and is also judged to have one or more regression outlying coefficients when  $\Pr(\delta_i^\beta = 1)$  is high. Marshall and Spiegelhalter (2007) propose identifying outliers in a Bayesian hierarchical framework using a simulation-based approach. In particular, they suggest a diagnostic test based on the measure of conflict between the predictive *prior* of a parameter given the remainder of the data and its likelihood. This is a practical approach for detecting outliers which can be widely implemented using the MCMC software. The main applications have been in biostatistics. In the Bayesian literature in order to investigate robustness with respect to the functional form of a base prior distribution  $\pi_0$ , the  $\varepsilon$ -contamination model of prior distributions  $\Gamma = \{\pi : \pi = (1 - \varepsilon)\pi_0(\theta | \lambda) + \varepsilon q, q \in Q\}$ , has been proposed.

Here  $\pi_0(\theta | \lambda)$  is the base elicited prior,  $\lambda$  is a vector of fixed hyperparameters,  $q$  is a contamination belonging to some suitable class  $Q$  and  $\varepsilon$  reflects the amount of error in  $\pi_0(\theta | \lambda)$  (see Berger 1984 and Berger and Berliner 1986, for a survey). This has been applied more recently by Singh and Chaturvedi (2012) to panel data.

## 14.6 CONCLUSION

---

Large panel data sets are a blessing and a curse. They are a blessing because panels help control for heterogeneity among the individuals, better able to estimate dynamic behaviour, to mention a few of the benefits listed in Baltagi (2008). They may also be a curse because with more data there is more likelihood of having outliers. If one is willing to sacrifice some efficiency in favor of protection against the adverse effects of outliers, one can use some of the robust panel data methods surveyed in this chapter. We hope that this will spur more applications of robust methods in empirical work using panel data. Prominent examples are the huge number of applications of growth studies or purchasing power parity studies using a panel of countries. These robust methods may naturally warn the researcher on what countries display distinctly different behavior from other countries and perhaps should not be pooled or assumed to have the same model as the remainder countries.

## NOTES

---

1. The term “*robust*” was introduced by Box (1953) and Box and Andersen (1955). Tukey (1960) showed the non robustness of the arithmetic mean under small deviations from normality. The seminal paper of Huber (1964) “became the founding paper of “the stability theory of statistical procedures” that by a historical accident was called “robust statistics” (Hampel, 1992, p. 1). By now, an extensive theory of robustness has been developed and is still growing in statistics, (see the books of Hampel et al. 1986; Huber 1981; Huber and Ronchetti 2009; Olive 2008; Rousseeuw and Leroy 2003 to mention a few).
2. The  $\alpha$ -trimmed mean is the arithmetic mean when the  $\alpha\%$  percent of the ends are discarded from the sample. The major advantage of the trimmed mean is the robustness and the higher efficiency for mixed distributions and for heavy-tailed distributions.
3. The three crucial equivariance properties include: regression equivariance, scale equivariance and affine equivariance (see Rousseeuw and Leroy 2003 and Bramati and Croux 2007).  $\theta(\Omega) \equiv \theta(\{y_{it}, X_{it}\})$  is *regression equivariant* if  $\theta(\{y_{it} + X_{it}\delta, X_{it}\}) = \theta(\{y_{it}, X_{it}\}) + \delta$  where  $\delta$  is a  $(K \times 1)$  vector of constants.  $\theta(\{y_{it}, X_{it}\})$  is *scale equivariant* if  $\theta(\{cy_{it}, X_{it}\}) = c\theta(\{y_{it}, X_{it}\})$  for any scalar  $c$ .  $\theta(\{y_{it}, X_{it}\})$  is *affine equivariant* if  $\theta(\{y_{it}, X_{it}A\}) = (A')^{-1}\theta(\{y_{it}, X_{it}\})$  for any non singular  $(K \times K)$  matrix  $A$ .
4. Once  $\widehat{\beta}_{LTS}$  is estimated, one gets the estimated fixed effects as  $\widehat{\alpha}_i(\widehat{\beta}_{LTS}) = \text{median}_t(y_{it} - X_{it}\widehat{\beta}_{LTS}), i = 1, \dots, N$ .
5. More precisely, if  $\widehat{\beta}_{LTS}^{(0)}$  and  $\widehat{\sigma}_{LTS}^{(0)}$  are the regression and scale estimators of the initial LST estimator, then, observations having large standardized residuals  $[r_{it}(\widehat{\beta}_{LTS}^{(0)})/\widehat{\sigma}_{LTS}^{(0)}]$

are downweighted using the Tukey biweight  $W_{r,it}^{(0)} = W_r(r_{it}(\hat{\beta}_{LTS}^{(0)})/\hat{\sigma}_{LTS}^{(0)})$ . So, we can perform weighted least squares to obtain  $\hat{\beta}_{LTS}^{(reweighted)} = (\tilde{X}' W_r^{(0)} \tilde{X})^{-1} \tilde{X}' W_r^{(0)} \tilde{y}$  where  $W_r^{(0)}$  is the  $(NT \times NT)$  matrix with diagonal elements given by  $W_{r,it}^{(0)}$ .

6. In the context of cross-section data, Dehon, Gassner, and Verardi (2012) proposed a Hausman-type test to determine whether a robust  $S$ -estimator is more appropriate than the ordinary least squares in case of contaminated data.
7. From equation (14), the algorithm calculates the hyperplane of  $K$  observations that fits all points perfectly. For each subset, the residuals are defined as the vertical distance separating each observation from the hyperplane. Using these residuals, a scale estimate  $\hat{\sigma}_S$  is obtained as in (13) for each  $p$ -subset. Salibian-Barrera and Yohai (2006) proposed the following number of generated sub-samples  $N_{sub}$ :

$$N_{sub} = \left\lceil \frac{\log(1 - P)}{\log(1 - (1 - \nu)^K)} \right\rceil$$

where  $\nu$  is the maximal expected proportion of outliers.  $P$  is the desired probability of having at least one  $p$ -subset without outliers among the  $N_{sub}$  subsamples and  $\lceil x \rceil$  is the ceiling operator of  $x$ , *i.e.*, the smallest integer not less than  $x$ . The number of sub-samples is chosen to guarantee that at least one  $p$ -subset without outliers is selected with high probability (see Salibian-Barrera and Yohai 2006; Maronna and Yohai 2006; Croux and Verardi 2008). As Croux and Verardi (2008) warn, sub-sampling algorithms can easily lead to collinear sub-samples if various dummies are present. A rough solution is to use subsets of size a little bit larger than  $K$  but an exact solution is given by Maronna and Yohai (2006) who introduce the  $MS$ -estimator that alternates an  $S$ -estimator (for continuous variables) and an  $M$ -estimator (for dummy ones) until convergence.

8. The robust estimates  $\mu_x$  and  $V_x$  can be obtained using the Minimum Covariance Determinant (MCD) estimator (see Rousseeuw 1984). The MCD method looks for the  $H (> NT/2)$  observations whose classical covariance matrix has the lowest possible determinant. The raw MCD estimate of location  $\mu_x$  is then the average of these  $H$  points, whereas the raw MCD estimate of the scatter  $V_x$  is their covariance matrix, multiplied by a consistency factor. The MCD estimates can resist  $(NT - H)$  outliers and a value of  $H = 0.75NT$  is recommended to obtain a high finite-sample efficiency. The computation of the MCD estimator is non-trivial. Rousseeuw and Van Driessen (1999) suggest a fast resampling algorithm (FAST-MCD). Several other algorithms have been proposed (see Olive 2008, chap. 10 for a discussion).
9. One of the advantages of pairwise differences is that in addition to removing the individual effects  $\alpha_i$  like the first difference, it uses a larger sample size  $NT(T - 1)/2$  instead of  $N(T - 1)$ .
10. For instance, the mean square error reported is 0.511 for *WGM* and 0.581 for *REWLS* when  $(N, T) = (50, 3)$ . However, this drops to 0.045 for *WGM* and 0.044 for *REWLS* when  $(N, T) = (50, 4)$ .
11. For instance, when  $(N, T) = (50, 4)$ , the mean square error of the estimates decreases from 0.045 to 0.028 to 0.019 for *WGM* when we use respectively the median, the first difference and the pairwise differences transformations. For *REWLS*, we get 0.044, 0.026 and 0.017, respectively.

12. In the case of binary panel data with a logistic link:  $\Pr[y_{it} = 1 | X_{it}, Z_{it}, \alpha_i] = g^{-1}(\eta_{it}) = [1 + \exp(-\eta_{it})]^{-1}$ . In case of count panel data with a Poisson distribution:  $\mu_{it} = g^{-1}(\eta_{it}) = \exp(\eta_{it})$  and

$$\Pr[y_{it} = k | X_{it}, Z_{it}, \alpha_i] = \exp(-\mu_{it}) \frac{\mu_{it}^k}{k}, k = 0, 1, 2, 3, \dots$$

Chib and Carlin (1999) (resp. Chib, Greenberg, and Winkelmann (1998)) proposed a Bayesian framework using MCMC in the case of a binary panel data (res. count panel data) model where the linear predictor is defined as in Eq.(62). But they did not deal with the case of contaminated data.

13. The binomial AR(1) panel model is:  $y_{it} | y_{it-1} \sim \text{Bin}(\mu_{it} + \rho(y_{it-1} - \mu_{it-1}))$  with  $y_{i1} \sim \text{Bin}(\mu_{i1})$  and  $\alpha_i = 0, \forall i$ . The count panel AR(1) model is:  $y_{it} = \rho y_{it-1} + \text{Poisson}(\mu_{it} - \rho \mu_{it-1})$  with  $y_{i1} \sim \text{Poisson}(\mu_{i1})$  and  $\alpha_i = 0, \forall i$ .
14. As the model used by Sutradhar and Bari (2007, 2010) deviates a bit from our topic of panel data models with individual specific effects, we do not give more insights to this estimator. The reader interested in this approach can refer to the quoted articles.
15. This pseudo-likelihood method was developed under the assumption that the outcomes are independent over time, and yields asymptotically unbiased estimators when the marginal model for the response at each time period and the model for missingness have been correctly specified.
16. For a cross-section regression model:  $y = X\beta + \varepsilon$ , the Cook's distance, for the  $i$ th observation  $X_i$  for an estimate  $\hat{\beta}$ , is given by:

$$C_i = \frac{t_i^2}{K} \cdot \frac{p_i}{1-p_i} \text{ with } t_i = \left( \frac{y_i - X_i \hat{\beta}}{\hat{\sigma}_\varepsilon \sqrt{1-p_i}} \right), p_i = X_i (X'X)^{-1} X_i'$$

$\hat{\beta}$  is the OLS estimator,  $t_i$  is the studentized residual and  $p_i$  is the  $i$ th diagonal element of the hat matrix.  $t_i^2$  is a measure of the degree to which the  $i$ th observation can be considered as an outlier from the assumed model and the ratio  $(p_i / (1 - p_i))$  defines the relative sensitivity of the estimate  $\hat{\beta}$  to potential outlying observations at each data point. Cook (1977) suggested that each  $C_i$  be compared with the quantiles of the central  $F(p, N - p)$  distribution. One must note that several other influence diagnostics have been proposed, for instance, the Belsley, Kuh, and Welsch covariance ratio, the Kuh and Welsch distance, the partial influence i.e., when an observation can be an outlier and/or influential only in one dimension or a few dimensions but not for all the regression coefficients, see Chatterjee and Hadi 1986).

## REFERENCES

- Ahn, S.C. and P. Schmidt, 1995, Efficient estimation of models for dynamic panel data, *Journal of Econometrics*, 68, 5–27.
- Andersen, R., 2008, *Modern Methods for Robust Regression*, SAGE publications, Thousand Oaks, CA.
- Aquaro, M. and P. Čížek, 2013, One-step robust estimation of fixed-effects panel data models, *Computational Statistics and Data Analysis*, 57, 1, 536–548.

- Arellano, M. and S. Bond, 1991, Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations, *Review of Economic Studies*, 58, 277–297.
- Baltagi, B.H., 2008, *Econometric Analysis of Panel Data*, 4th edition, John Wiley & sons, Chichester.
- Baltagi, B. and G. Bresson, 2012, A robust Hausman-Taylor estimator, in *Advances in Econometrics: Essays in Honor of Jerry Hausman*, vol. 29, (Baltagi, B.H., Carter Hill, R., Newey, W.K. and H.L. White, eds.), Emerald Group Publishing Limited, 175–214.
- Banerjee, M. and E.W. Frees, 1997, Influence diagnostics for linear longitudinal models, *Journal of the American Statistical Association*, 92, 439, 999–1005.
- Barnett, V. and T. Lewis, 1994, *Outliers in Statistical Data*, John Wiley, New York.
- Bari, W. and B.C. Sutradhar, 2010, On bias reduction in robust inference for generalized linear models, *Scandinavian Journal of Statistics*, 37, 1, 109–125.
- Belsley, D.A., Kuh, E. and R.E. Welsch, 1980, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, John Wiley, New York.
- Berger, J.O., 1984, The robust Bayesian viewpoint, in *Robustness in Bayesian Statistics*, (Kadane, J. ed.), Elsevier Science Publishers, Amsterdam, 63–124.
- Berger, J.O. and L.M. Berliner, 1986, Robust Bayes and empirical Bayes analysis with  $\varepsilon$ -contaminated priors, *Annals of Statistics*, 14, 461–486.
- Blundell, R. and S. Bond, 1998, Initial conditions and moment restrictions in dynamic panel data models, *Journal of Econometrics*, 87, 1, 115–143.
- Box, G.E.P, 1953, Non-normality and tests on variances, *Biometrika*, 40, 318–335.
- Box, G.E.P. and S.L. Andersen, 1955, Permutation theory in the derivation of robust criteria and the study of departures from assumption, *Journal of the Royal Statistical Society, B*, 17, 1–34.
- Bramati, M.C. and C. Croux, 2007, Robust estimators for the fixed effects panel data model, *Econometrics Journal*, 10, 521–540.
- Breslow, N.E. and D.G. Clayton, 1993, Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association*, 88, 9–25.
- Breusch, T.S., Mizon, G.E., and P. Schmidt, 1989, Efficient estimation using panel data, *Econometrica*, 57, 695–700.
- Cantoni, E. and E. Ronchetti, 2001, Robust inference for generalized linear models, *Journal of the American Statistical Association*, 96, 1022–1030.
- Chatterjee, S. and A.S. Hadi, 1986, Influential observations, high leverage points, and outliers in linear regression, *Statistical Science*, 1, 3, 379–393.
- Chib, S. and B.P. Carlin, 1999, On MCMC sampling in hierarchical longitudinal models, *Statistics and Computing*, 9, 17–26.
- Chib, S., Greenberg, E. and R. Winkelmann, 1998, Posterior simulation and Bayes factors in panel count data models, *Journal of Econometrics*, 86, 33–54.
- Čížek, P., 2008, Robust and efficient adaptive estimation of binary-choice regression models, *Journal of the American Statistical Association*, 103, 687–696.
- Čížek, P., 2010, Reweighted least trimmed squares: an alternative to one-step estimators, CentER Discussion paper 2010-91, Tilburg University, The Netherlands.
- Cook, R.D., 1977, Detection of influential observation in linear regression, *Technometrics*, 19, 1, 15–18.
- Cornwell, C. and P. Rupert, 1988, Efficient estimation with panel data: an empirical comparison of instrumental variables estimators, *Journal of Applied Econometrics*, 3, 149–155.

- Croux, C. and V. Verardi, 2008, Robust regression in Stata, *The Stata Journal*, 9, 439–453.
- De Finetti, B., 1961, The Bayesian approach to the rejection of outliers, *Proceedings of the Fourth Berkeley Symposium on Probability and Statistics*, Vol. 1, Berkeley: University of California Press, 99–210.
- Dehon, C., Gassner, M. and V. Verardi, 2009, Beware of ‘good’ outliers and overoptimistic conclusions, *Oxford Bulletin of Economics and Statistics*, 71, 437–452.
- Dehon, C., Gassner, M. and V. Verardi, 2012, Extending the Hausman test to check for the presence of outliers, in *Advances in Econometrics: Essays in Honor of Jerry Hausman*, vol. 29, (Baltagi, B.H., Carter Hill, R., Newey, W.K. and H.L. White, eds.), Emerald Group Publishing Limited, 435–453.
- Dhaene, G. and Y. Zhu, 2009, Median-based estimation of dynamic panel models with fixed effects, Working paper, Faculty of Business and Economics, Catholic University of Leuven, Belgium.
- Donoho D.L. and P.J. Huber, 1983, The notion of breakdown point, in *A Festschrift for Erich L. Lehmann*, (Bickel, P.J., Doksum, K.A. and J. L. Hodges, eds.), Belmont, Wadsworth, 157–184.
- Ecob, R. and G. Der, 2003, An iterative method for the detection of outliers in longitudinal growth data using multilevel models, in *Multilevel Modeling: Methodological Advances, Issues and Applications*, Lawrence Erlbaum Publishers, Mahwah, New Jersey.
- Geary, R.C., 1947, Testing for normality, *Biometrika*, 34, 3-4, 209–242.
- Gervini, D. and V.J. Yohai, 2002, A class of robust and fully efficient regression estimators, *The Annals of Statistics*, 30, 2, 583–616.
- Giavazzi, F., Jappelli, T. and M. Pagano, 2000, Searching for non-linear effects of fiscal policy: Evidence from industrial and developing countries, *European Economic Review*, 44, 1259–1289.
- Hadi, A.S. and J.S. Simonoff, 1993, Procedures for the identification of multiple outliers in linear models, *Journal of the American Statistical Association*, 88, 424, 1264–1272.
- Hampel, F.R., 1968, Contributions to the theory of robust estimation, PhD thesis, *University of California*, Berkeley.
- Hampel, F.R., 1973, Robust estimation: a condensed partial survey, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 27, 87–104.
- Hampel, F.R., 1985, The breakdown points of the mean combined with some rejection rules, *Technometrics*, 27, 95–107.
- Hampel, F.R., 1992, Introduction to Huber (1964) robust estimation of a location parameter, in *Breakthroughs in Statistics II: Methodology and Distributions*, (Kotz, S. and L.N Johnson, eds.), Springer-Verlag, New York, 492–518.
- Hampel, F.K., 2002, Robust Inference, in *Encyclopedia of Environmetrics*, (El-Shaarawi, A.H. and W.W. Piegorsch, eds.), John Wiley, Chichester, 1865–1885.
- Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and W.A. Stalel, 1986, *Robust Statistics: The Approach Based on Influence Functions*, John Wiley, New York.
- Hausman, J.A. and W.E. Taylor, 1981, Panel data and unobservable individual effects, *Econometrica*, 49, 1377–1398.
- Hinloopen, J. and J.L.M. Wagenvoort, 1997, On the computation and efficiency of a HBP-GM estimator: Some simulation results, *Computational Statistics and Data Analysis*, 25, 1–15.
- Honoré, B.E. and J.L. Powell, 2005, Pairwise difference estimation of nonlinear models, in *Identification and Inference for Econometric Models. Essays in Honor of Thomas Rothenberg*, (Andrews, D.W.K. and J.H. Stock, eds.), Cambridge University Press, 520–553.

- Hsiao, C., Pesaran, M.H. and A. Tahmisioglu, 2002, Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods, *Journal of Econometrics*, 109, 1, 107–150.
- Huber, P., 1964, Robust estimation of a location parameter, *Annals of Mathematical Statistics*, 35, 73–101.
- Huber, P., 1965, A robust version of the probability ratio test, *Annals of Mathematical Statistics*, 36, 1753–1758.
- Huber, P.J., 1981, *Robust Statistics*, Series in Probability and Mathematical Statistics, 1st Edition, John Wiley, New York.
- Huber, P.J., 1996, *Robust Statistical Procedures*, 2nd edition, SIAM, Philadelphia.
- Huber, P.J. and E.M. Ronchetti, 2009, *Robust Statistics*, Series in Probability and Mathematical Statistics, 2nd edition, John Wiley, New York.
- Janz, N., 2003, Robust GMM estimation of an Euler equation investment model with German firm level panel data, in *Contributions to Modern Econometrics: From Data Analysis to Economic Policy*, (Klein, I. and S. Mitnik, eds), Series: Dynamic Modeling and Econometrics in Economics and Finance, Vol. 4, Kluwer Academic Publishers, Dordrecht, 87–102.
- Juarez, M.A. and M.K.F. Steel, 2010, Non-Gaussian dynamic Bayesian modelling for panel data, *Journal of Applied Econometrics*, 25, 7, 1128–1154.
- Krasker, W.S. and R.E. Welsch, 1982, Efficient bounded influence regression estimation, *Journal of the American Statistical Association*, 77, 595–604.
- Krasker, W.S., Kuh, E. and R.E. Welsch, 1983, Estimation for dirty data and flawed models, in *Handbook of Econometrics* (Griliches, I.Z. and M.D. Intriligator, eds.), North-Holland, Amsterdam, chapter 11, 651–698.
- Lai, T.L. and M.C. Shih, 2003, A hybrid estimator in nonlinear and generalized linear mixed effects models, *Biometrika*, 90, 859–879.
- Lucas, A., 1996, *Outlier robust unit root analysis*, Thesis Publishers, Vrije Universiteit, Amsterdam, the Netherlands.
- Lucas, A., van Dijk, R. and T. Kloek, 1994, Outlier robust GMM estimation of leverage determinants, Discussion paper TI 94-132, Tinbergen Institute, Rotterdam, the Netherlands.
- Lucas, A., van Dijk, R. and T. Kloek, 2007, Outlier robust GMM estimation of leverage determinants in linear dynamic panel data models, Unpublished manuscript, Vrije Universiteit, Amsterdam, the Netherlands.
- Mallows, C.L., 1975, On some topics in robustness, Technical Memorandum, Bell Telephone Laboratories, Murray Hill, New Jersey.
- Maronna, R., and V. Yohai, 2006, Robust regression with both continuous and categorical predictors, *Journal of Statistical Planning and Inference*, 89, 197–214.
- Maronna, R.A., Martin, R.D. and V.J. Yohai, 2006, *Robust Statistics*, John Wiley & Sons, New York.
- Marshall, E.C. and D.J. Spiegelhalter, 2007, Identifying outliers in Bayesian hierarchical models: a simulation-based approach, *Bayesian Analysis*, 2, 2, 409–444.
- O'Hagan, A., 1979, On outlier rejection phenomena in Bayes inference, *Journal of the Royal Statistical Society, B*, 41, 358–367.
- O'Hagan, A., 1988, Modelling with heavy tails, in *Bayesian Statistic III*, (Bernardo, J.M., DeGroot,M.H., Lindley, D.V. and A.F.M Smith, eds.), Clarendon Press, Oxford, 345–359.
- O'Hagan, A., 1990, Outliers and credence for location parameter inference, *Journal of the American Statistical Association*, 85, 172–176.

- Olive, D.J., 2008, *Applied Robust Statistics*, Department of Mathematics, Southern Illinois University Press.
- Peruggia, M., Santner, T. and Y.-Y. Ho, 2004, Detecting stage-wise outliers in hierarchical Bayesian linear models of repeated measures data, *Annals of the Institute of Statistical Mathematics*, 56, 415–433.
- Preisser, J.S. and B.F. Qaqish 1996, Deletion diagnostics for generalized estimating equations, *Biometrika*, 83, 551–562.
- Ronchetti, E. and F. Trojani, 2001, Robust inference with GMM estimators. *Journal of Econometrics*, 101, 37–69.
- Rousseeuw, P.J., 1984, Least median of squares regression, *Journal of the American Statistical Association*, 79, 871–880.
- Rousseeuw, P.J. and A.M. Leroy, 2003, *Robust Regression and Outlier Detection*, John Wiley, New York.
- Rousseeuw, P.J. and K. Van Driessen, 1999, A fast algorithm for the minimum covariance determinant estimator, *Technometrics*, 41, 212–223.
- Rousseeuw, P.J. and V. Yohai, 1987, Robust Regression by means of S-estimators, in *Robust and Nonlinear Time Series Analysis*, (Franke, J., Härdle, W. and D. Martin, eds.), Springer Verlag, Berlin, 256–272.
- Salibian-Barrera, M. and V. Yohai, 2006, A fast algorithm for S-regression estimates, *Journal of Computational and Graphical Statistics*, 15, 414–427.
- Singh, A. and A. Chaturvedi, 2012, Robust Bayesian analysis of autoregressive fixed effects panel data model, working paper, Department of mathematics, University of Allahabad, India.
- Sinha, S.K., 2012a, Robust analysis of longitudinal data with nonignorable missing responses, *Metrika*, 75, 913–938.
- Sinha, S.K., 2012b, Robust inference for incomplete binary longitudinal data, *Canadian Journal of Statistics*, 4, 2, 128–147.
- Sutradhar, B.C. and W. Bari, 2007, On generalized quasi-likelihood inference in longitudinal model for count data, *Sankhya: The Indian Journal of Statistics*, 69, 671–689.
- Sutradhar, B.C. and W. Bari, 2010, Robust inferences in longitudinal models for binary and count panel data in presence of outliers, *Sankhya: The Indian Journal of Statistics*, 72, 11–37.
- Troxel, A.B., Lipsitz, S.R. and D.P. Harrington, 1998, Marginal models for the analysis of longitudinal measurements subject to nonignorable nonmonotone missing data, *Biometrika*, 85, 661–672.
- Tukey, J.W., 1960, A survey of sampling from contaminated distributions, in *Contributions to Probability and Statistics*, (Olkin, I., Ghurye, S.G., Hoeffding, W., Madow, W.G. and H.B. Mann, eds.), Stanford University Press, 448–485.
- Verardi, V. and J. Wagner, 2011, Productivity premia for German manufacturing firms exporting to the Euro-area and beyond: First evidence from robust fixed effects estimations, *The World Economy*, 35, 6, 694–712.
- Wagenvoort R. and R. Waldmann, 2002, On B-robust instrumental variable estimation of the linear model with panel data, *Journal of Econometrics*, 106, 297–324.
- Wooldridge, J.M., 2010, *Econometric Analysis of Cross-Section and Panel Data*, MIT Press, Boston.
- Yohai, V. 1987, High breakdown-point and high efficiency estimates for regression, *The Annals of Statistics*, 15, 642–665.
- Zhu, H., Ibrahim, J.G. and H. Cho, 2012, Perturbation and scaled Cook's distance, *The Annals of Statistics*, 40, 2, 785–811.

P A R T II

---

PANEL DATA  
APPLICATIONS

---



## CHAPTER 15

---

# THE ANALYSIS OF MACROECONOMIC PANEL DATA

---

JÖRG BREITUNG

### 15.1 INTRODUCTION

---

*The recent work on cross-country regressions is also like looking at a black cat in a dark room. Whether or not all this work has accomplished anything on the substantive economic issues is a moot question.*

G. S. Maddala (1999)

MACROECONOMIC panel data typically involve aggregate data from various countries such as output (GDP), employment, inflation rates, wages, etc. Is it appropriate to employ traditional panel data methods like the within-group estimator to such cross-country data? In this chapter we argue that some adjustments and refinements of the usual panel data toolbox are required to find the black cat in the dark room. In recent years, the econometric toolbox has been supplemented with a variety of statistical methods accommodating typical features encountered in the analysis of cross-country data.

What are the main differences when using macroeconomic data instead of individual (survey) data? In contrast to the “large  $N$ , small  $T$ ” framework, the two dimensions of a typical macroeconomic data set are more balanced, often providing a comparable number of time periods and countries (regions). Although this is inconsequential for the analysis based on the linear static panel data framework, it becomes crucial when estimating a dynamic model. Furthermore, in the analysis of macroeconomic data cross-section dependence among countries is an important issue. In many cases this dependence cannot be accommodated by a simple function of the geographical distance but also depends on trade relations and the level of economic development. Another important difference is that cross-country data often exhibit a much richer

pattern of heterogeneity that cannot be represented just by letting the intercept vary across countries. While it is infeasible to allow for individual specific regression coefficients in a “large  $N$ , small  $T$ ” panel data set, this may be a suitable option when analyzing macroeconomic data. Finally, potential problems arising from possibly non-stationarity of the variables become more relevant if the number of time periods is comparable to the number of countries. This issue is considered in more detail in Chapter 2.

In this chapter we discuss how standard panel data methods can be adapted to accommodate typical features of macroeconomic panel data. In particular we will focus on an asymptotic framework where  $N$  and  $T$  tend to infinity at the same rate, that is,  $N/T \rightarrow c$  with  $0 < c < \infty$ . To this end, we confine ourselves to various alternative methods for coping with a particular deviation from the classical panel data model, leaving the other model assumptions intact. Although in empirical practice it is likely to encounter several deviations from the classical panel data model at the same time, most methods presented here can be straightforwardly combined to accommodate more general macroeconomic panel data models.

Let us introduce some notation used in this chapter. Sometimes it will be more convenient to write the model adapting a country-wise matrix notation:

$$y_i = \mu_i \iota_T + X_i \beta + u_i, \quad (15.1)$$

where  $y_i = [y_{i1}, \dots, y_{iT}]'$ ,  $X_i = [x_{i1}, \dots, x_{iT}]'$  is a  $T \times k$  regressor matrix of the  $i$ 'th panel group (country),  $\mu_i$  denotes the unit-specific intercept and  $\iota_T$  is a  $T \times 1$  vector of ones. The error vector is defined conformably. In Section 15.4 it is more convenient to use the period-wise matrix notation

$$y_t = \mu + X_t \beta + u_t, \quad (15.2)$$

where  $y_t = [y_{1t}, \dots, y_{Nt}]'$ ,  $X_t = [x_{1t}, \dots, x_{Nt}]'$  ( $N \times k$ ), and  $\mu = [\mu_1, \dots, \mu_N]'$ . The within-group estimator is given by

$$\hat{\beta} = \left( \sum_{i=1}^N X_i' M_T X_i \right)^{-1} \sum_{i=1}^N X_i' M_T y_i, \quad (15.3)$$

where the matrix  $M_T = I_T - T^{-1} \iota_T \iota_T'$  produces deviations from the mean. As usual a bar indicates the mean of the respective group of observations, that is,  $\bar{x}_i = T^{-1} \sum_{t=1}^T x_{it}$  or  $\bar{x}_t = N^{-1} \sum_{i=1}^N x_{it}$ . Deviations from the individual specific mean are denoted by  $\tilde{x}_{it} = x_{it} - \bar{x}_i$ , resp.  $\tilde{y}_{it} = y_{it} - \bar{y}_i$  and the corresponding matrix of the transformed regressors is  $\tilde{X}_i$  ( $\tilde{y}_i$ ) or  $\tilde{X}_t$  ( $\tilde{y}_t$ ). With this notation the within-group estimator is written as

$$\begin{aligned}\hat{\beta} &= \left( \sum_{i=1}^N \tilde{X}'_i \tilde{X}_i \right)^{-1} \sum_{i=1}^N \tilde{X}'_i \tilde{y}_i \\ &= \left( \sum_{t=1}^T \tilde{X}'_t \tilde{X}_t \right)^{-1} \sum_{t=1}^T \tilde{X}'_t \tilde{y}_t.\end{aligned}$$

## 15.2 DYNAMIC MODELS

This section addresses the problem of estimating a dynamic panel data model when the two panel dimensions  $N$  and  $T$  are of similar magnitude. We first study the properties of the usual within-group estimator for the case that the underlying relationship is dynamic, that is, we analyze the effect of dynamic misspecification. It is argued that although the dynamic model possesses a static representation that can be estimated consistently by the within-group estimator as  $T$  tends to infinity, it is difficult to provide a useful interpretation of this static form. Specifically, the static coefficients mix-up the original dynamic multipliers and the parameters characterizing the data-generating process of the regressors. It may even happen that the dynamic multipliers  $\partial E(y_{i,t+h}|x_{it})/\partial x_{it}$  are positive for all horizons  $h = 0, 1, \dots$  but the estimated coefficient of the static representation nevertheless converges in probability to a negative limit. Another problem is that, in general, the regressors can no longer be treated as strictly exogenous. Accordingly, the within-group estimator suffers from an  $O(T^{-1})$  bias.

If  $T$  is “sufficiently large” the bias resulting from the mean-adjusted lagged dependent variable can be neglected. Adopting an asymptotic framework with  $N/T \rightarrow c$  Hahn and Kuersteiner (2002) and Alvarez and Arellano (2003) show that in a pure auto-regressive model  $y_{it} = \mu_i + \alpha y_{i,t-1} + \varepsilon_{it}$ , the within-group estimator of the auto-regressive coefficient  $\alpha$  possesses the limiting distribution

$$\sqrt{NT}(\hat{\alpha} - \alpha) \xrightarrow{d} \mathcal{N}(-\sqrt{c}(1+\alpha), 1-\alpha^2). \quad (15.4)$$

Accordingly, although the estimator is  $\sqrt{NT}$ -consistent, statistical inference based on the usual  $t$  or  $F$  statistics remains invalid. For example, consider the  $t$ -statistic

$$t_\alpha = \frac{\hat{\alpha} - \alpha}{\sqrt{1-\alpha^2}/\sqrt{NT}} \xrightarrow{d} \mathcal{N}\left(-\frac{\sqrt{c}(1+\alpha)}{\sqrt{1-\alpha^2}}, 1\right).$$

If, for example,  $\alpha = \alpha_0 = 0.8$ ,  $N = 50$ , and  $T = 100$  the expectation of the  $t$ -statistic is approximately  $-2.12$ . This implies that using the two-sided critical value  $\pm 1.96$  the test will reject with a probability of more than 50 percent even if the null hypothesis  $\alpha = 0.8$  is correct; although one might be tempted to argue that  $T$  is quite large so that the Nickell bias can be ignored.

For dynamic panels with large  $N$  and small  $T$ , variants of the GMM estimator have become popular that provide consistent parameter estimates and valid inference (see Chapter 3). If  $T$  gets large relative to  $N$ , however, the instrument count (as a squared function of  $T$ ) becomes very large which may result in poor small sample properties of the estimator and hypothesis tests. To overcome this problem, various methods for cutting down the number of instruments were proposed, but it is not clear whether these techniques are successful in condensing the instruments. A more appealing approach is to adopt a bias-adjustment to ML type estimators that are asymptotically efficient as  $T$  tends to infinity. The within-group estimator is equivalent to the (Gaussian) ML estimator for the dynamic model treating the initial values  $y_{i1}$  ( $i = 1, \dots, N$ ) as fixed constants. Since the initial values become irrelevant if  $T$  gets large, this estimator approaches the efficiency bound as  $N$  and  $T$  tend to infinity (cf. Hahn and Kuersteiner 2002). Thus, there is reason to expect that bias corrected within-group estimators performs best in dynamic panel data models with moderate or large  $T$ . If  $T$  is small and the auto-regressive coefficient comes close to unity, the initial values have an important effect and, therefore, a (transformed) maximum likelihood that takes into account the distributional assumptions of the initial values may be superior.

In recent years GMM estimation has become very popular for empirical research in micro- and macroeconomics, whereas maximum likelihood (ML) and bias-corrected estimators found itself in a state of shadowy existence. In this chapter we argue that at least for moderate and large  $T$  this is not warranted. While GMM estimation suffers from the problem of instrument proliferation leading to small sample bias and unreliable inference, ML and bias-corrected estimators can be seen as just-identified (nonlinear) moment estimators that approach the lower variance bound as  $T$  gets large no matter of the ratio  $N/T$ . These attractive features come at the expense of more restrictive model assumptions such as strictly exogenous regressors, homoskedastic errors, or stationary initial conditions. Moreover, the availability of the GMM methodology (including a variety of useful specification tests) in many statistical software packages including Stata, Eviews, SAS, and R facilitates the application of GMM methodology, whereas ML and bias-corrected estimators are provided only sporadically.<sup>1</sup>

### 15.2.1 The Static Representation of a Dynamic Panel Data Model

Following Baltagi and Griffin (1984), Pirotte (1999), and Egger and Pfaffermayr (2005) we consider a linear dynamic panel data model of the form

$$\alpha(L)y_{it} = \mu_i + \beta(L)x_{it} + \varepsilon_{it}, \quad (15.5)$$

where  $\alpha(L) = 1 - \alpha_1 L - \cdots - \alpha_p L^p$  with  $|\alpha(z)| \neq 0$  for all  $|z| \leq 1$ ,  $\beta(L) = \beta_0 + \beta_1 L + \cdots + \beta_q L^q$  and  $\varepsilon_{it}$  is *i.i.d.* with  $E(\varepsilon_{it}^2) = \sigma^2$ . We further assume that  $x_{it}$  is a strictly exogenous stationary process with  $E(x_{it}) = 0$  and autocorrelation function  $\rho_k = E(x_{it}x_{i,t+k})/E(x_{it}^2)$ . Using  $E(x_{i,t-k}|x_{it}) = \rho_k x_{it}$  we obtain the following static representation

$$y_{it} = \mu_i^* + \theta x_{it} + u_{it}, \quad (15.6)$$

where

$$\begin{aligned} \theta &= \sum_{k=0}^{\infty} \gamma_k \rho_k \\ u_{it} &= \alpha(L)^{-1} \varepsilon_{it} + \sum_{k=1}^{\infty} \gamma_k v_{i,t-k} \\ v_{i,t-k} &= x_{i,t-k} - \rho_k x_{it} \end{aligned}$$

$\mu_i^* = E(y_{it}) = \mu_i/\alpha(1)$ , and  $\beta(L)/\alpha(L) = \gamma_0 + \gamma_1 L + \gamma_2 L^2 + \cdots$ . Since

$$E(x_{i,t-k} u_{it}) = \sum_{j=k}^{\infty} \gamma_j (\rho_{j-k} - \rho_j \rho_k)$$

it follows that  $E(x_{it} u_{it}) = 0$  but  $E(x_{i,t-k} u_{it}) \neq 0$  for  $k \geq 1$  and, therefore,  $x_{it}$  is no longer strictly exogenous with respect to  $u_{it}$ . Since

$$E\left[\frac{1}{T} \sum_{t=1}^T (x_{it} - \bar{x}_i)(u_{it} - \bar{u}_i)\right] = -\frac{1}{T} E\left[\left(\frac{1}{\sqrt{T}} \sum_{t=1}^T x_{it}\right) \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T u_{it}\right)\right] = O(T^{-1})$$

it can be seen that the bias of the within-group estimator of the parameter  $\theta$  in (15.6) is  $O(T^{-1})$ .

To illustrate these results let us now analyze the probability limit for fixed  $T$  and  $N \rightarrow \infty$  for the special case of a first order dynamic model given by

$$\begin{aligned} y_{it} &= \mu_i + \alpha y_{i,t-1} + \beta_0 x_{it} + \beta_1 x_{i,t-1} + \varepsilon_{it} \\ x_{it} &= \rho x_{i,t-1} + \eta_{it}. \end{aligned} \quad (15.7)$$

We impose mean stationary starting values of the form  $(1 - \alpha)y_{i1} = \mu_i + \beta_0 x_{i1} + \varepsilon_{i1}$ . In matrix notation the model can be written as

$$\begin{bmatrix} 1 - \alpha & 0 & 0 & \cdot & 0 & 0 \\ -\alpha & 1 & 0 & \cdot & 0 & 0 \\ 0 & -\alpha & 1 & \cdot & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & -\alpha & 1 \end{bmatrix} y_i = \mu_i \iota_T + \begin{bmatrix} \beta_0 & 0 & 0 & \cdot & 0 & 0 \\ \beta_1 & \beta_0 & 0 & \cdot & 0 & 0 \\ 0 & \beta_1 & \beta_0 & \cdot & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & \beta_1 & \beta_0 \end{bmatrix} X_i + \varepsilon_i$$

$$\text{or } Ay_i = \mu_i \iota_T + BX_i + \varepsilon_i \\ y_i = \mu_i^* \iota_T + A^{-1}BX_i + A^{-1}\varepsilon_i.$$

As  $N \rightarrow \infty$  the probability limit of the within-group estimator results as

$$\operatorname{plim}_{N \rightarrow \infty} \hat{\theta} = \frac{\operatorname{tr}(M_T A^{-1} B \Sigma_x)}{\operatorname{tr}(M_T \Sigma_x)}, \quad (15.8)$$

where  $\Sigma_x$  is the covariance matrix of  $[x_{i1}, \dots, x_{iT}]'$  with typical element  $\rho^{|r-s|} E(x_{it}^2)$ . In general, this probability limit depends on  $\alpha$ ,  $\beta$ , and  $\rho$  in a complex way. Table 15.1 presents the probability limits of the within-group estimator for various combinations of these three parameters as well as different  $T$ . As  $T \rightarrow \infty$  the probability limit of the within-group estimator converges to  $\theta$  which in our model results as  $\theta = \beta(\rho)/\alpha(\rho)$ .

Is there any natural interpretation of the parameter  $\theta$ ? Obviously, the probability limit is different from the short-run effect  $\beta(0)/\alpha(0) = 1$  and the long-run multiplier  $\gamma(1) = \alpha(1)/\beta(1)$  which is also presented in Table 15.1. Only in the special case when the regressor is white noise (i.e.,  $\rho = 0$ ) or a random walk (i.e., if the regression is a cointegrating relationship) the parameter  $\theta$  converges to the short-run and long-run multipliers, respectively. In all other cases, the parameter  $\theta$  is difficult to interpret. If the regressor is negatively autocorrelated, it may happen that all multipliers  $\gamma_k$  are positive but the within-group estimator nevertheless converges to a negative limit.<sup>2</sup>

In the special case of common factors of the form  $\alpha(L) = \beta(L)/\beta_0$ , the dynamic representation boils down to the static representation

$$y_{it} = \mu_i^* + \gamma x_{it} + u_{it},$$

where the residuals are represented by the auto-regressive process  $\alpha(L)u_{it} = \varepsilon_{it}$ . The common factor restrictions can be tested within a dynamic panel data model by testing the nonlinear hypotheses  $\beta_0 \alpha_k = -\beta_k$  for  $k = 1, \dots, p$  (e.g., Sargan 1980).

Let us conclude our discussion. We found that the linear dynamic model (15.5) possesses a static representation (15.6), where the errors  $u_{it}$  are autocorrelated. Although  $x_{it}$  is assumed to be strictly exogenous with respect to  $\varepsilon_{it}$ , using  $x_{it} - \bar{x}_i$  as a regressor gives rise to a severe bias if  $T$  is small. Moreover, the regression coefficient  $\theta$  depends on the data-generating process of  $x_{it}$  and is therefore difficult to interpret. Hence, the static regression (15.6) is meaningless if the underlying relationship is dynamic and the within-group estimator of the static coefficient  $\theta$  is biased if  $T$  is fixed.

### 15.2.2 GMM Estimation When $T$ is Large

If  $T$  is small and  $N$  is large the GMM estimators suggested by Arellano and Bond (1991), Blundell and Bond (1998), and Ahn and Schmidt (1997) can be used to consistently estimate the dynamic model. As in Chapter 3 we consider the dynamic model

of the form

$$y_{it} = \mu_i + \alpha y_{i,t-1} + \beta' x_{it} + u_{it}, \quad t = 1, \dots, T. \quad (15.9)$$

The GMM estimator proposed by Arellano and Bond (1991) is based on the moment conditions

$$E[y_{is}(\Delta y_{it} - \alpha \Delta y_{i,t-1} - \beta' \Delta x_{it})] = 0 \quad \text{for } s = 0, \dots, t-2, \quad (15.10)$$

**Table 15.1 Probability limits of  $\hat{\theta}$  as  $N \rightarrow \infty$**

$\rho = 0$					
	$\beta_0 = 1, \beta_1 = 0.5$	$\beta_0 = 1, \beta_1 = 2$	$\beta_0 = 1, \beta_1 = -0.5$		
	$\alpha = 0.4$	$\alpha = 0.8$	$\alpha = 0.4$	$\alpha = 0.8$	$\alpha = 0.4$
$T = 10$	0.923	0.846	0.691	0.385	1.077
$T = 50$	0.984	0.954	0.934	0.816	1.016
$T = 100$	0.992	0.976	0.967	0.904	1.008
$T = \infty$	1.000	1.000	1.000	1.000	1.000
$\gamma(1)$	2.5	7.5	5.0	15	0.833
$\rho = 0.8$					
	$\beta_0 = 1, \beta_1 = 0.5$	$\beta_0 = 1, \beta_1 = 2$	$\beta_0 = 1, \beta_1 = -0.5$		
	$\alpha = 0.4$	$\alpha = 0.8$	$\alpha = 0.4$	$\alpha = 0.8$	$\alpha = 0.4$
$T = 10$	1.564	1.910	2.327	2.200	1.055
$T = 50$	1.962	3.312	3.532	5.728	0.916
$T = 100$	2.012	3.596	3.582	6.464	0.899
$T = \infty$	2.059	3.889	3.823	7.222	0.882
$\gamma(1)$	2.5	7.5	5.0	15	0.833
$\rho = -0.8$					
	$\beta_0 = 1, \beta_1 = 0.5$	$\beta_0 = 1, \beta_1 = 2$	$\beta_0 = 1, \beta_1 = -0.5$		
	$\alpha = 0.4$	$\alpha = 0.8$	$\alpha = 0.4$	$\alpha = 0.8$	$\alpha = 0.4$
$T = 10$	0.517	0.498	-0.363	-0.261	1.104
$T = 50$	0.468	0.403	-0.435	-0.337	1.107
$T = 100$	0.461	0.385	-0.445	-0.351	1.065
$T = \infty$	0.454	0.366	-0.454	-0.366	0.061
$\gamma(1)$	2.5	7.5	5.0	15	0.833

Note: Probability limits for the within-group estimator of the static representation of the dynamic model (15.7) based on (15.8).  $\gamma(1) = \beta(1)/\alpha(1)$  is the long-run multiplier.

whereas the “system estimator” proposed Arellano and Bover (1995) and Blundell and Bond (1998) adds the  $T - 1$  additional moment conditions

$$E[\Delta y_{i,t-1}(y_{it} - \alpha y_{i,t-1} - \beta' x_{it})] = 0, \quad t = 2, 3, \dots, T$$

that result from assuming mean-stationary initial values (see Chapter 3 for more details). These additional instruments are useful in particular if the dependent variable is highly persistent (Blundell and Bond 2000 and Blundell, Bond, and Windmeijer 2000). Further instruments are available if it is assumed that  $u_{it}$  is temporally homoskedastic (Ahn and Schmidt 1997). The problem with using these estimators for macroeconomic panels is that the number of moment conditions ( $m$ ) is  $m_{AB} = (T - 1)T/2 + k$  for the Arellano-Bond estimator and  $m_{sys} = (T - 1)(T + 2)/2 + k$  for the system estimator. For example, if  $T = 20$  and  $k = 3$  strictly exogenous regressors this implies 193 (resp. 212 for the system estimator) instruments. In many macroeconomic applications  $N$  is fairly small, say 30, implying that the number of instruments is much larger than  $N$ . The original asymptotic theory for the GMM estimator assumes that  $T$  is fixed and  $N \rightarrow \infty$  so that  $m/N \rightarrow 0$ . As shown by Bekker (1994), however, the asymptotic theory for GMM estimators breaks down if  $m/N$  tends to a nonzero constant. In our case, however, the situation is even worse as  $m/N \rightarrow \infty$  as  $N/T \rightarrow c$ , which is considered to be the relevant scenario in the analysis of macroeconomic panels.

Alvarez and Arellano (2003) study the asymptotic properties of the GMM estimator for the pure auto-regressive model if both dimensions  $N$  and  $T$  gets large. They show that the GMM estimator is consistent if  $(\log T)^2/N \rightarrow 0$ ; however, the estimator involves an asymptotic bias proportional to  $\sqrt{T/N}$ . Accordingly, reliable inference based on the usual GMM statistics requires that  $N$  is large relative to  $T$ . This is in stark contrast to the asymptotic bias of the within-group estimator which is  $O(\sqrt{N/T})$  and, therefore, valid asymptotic inference for the within-group estimator requires that  $T$  is large relative to  $N$ . Interestingly, if  $N/T \rightarrow c$  the asymptotic variances of both estimators are identical to the asymptotic variance of the pooled OLS estimator without individual specific effects. If  $c \rightarrow \infty$  the asymptotic bias term of the GMM estimator disappears and the asymptotic variance approaches the Cramer-Rao lower bound.

### 15.2.3 Dealing with the Problem of Instrument Proliferation

As argued in the previous section, the usual GMM techniques (in particular the two-stage estimation procedures) proposed by Arellano and Bond (1991) and Blundell and Bond (1998) suffer from the “many instruments problem” whenever the instrument count is of comparable magnitude as  $N$ . In recent years, various strategies were developed to cut down the number of instruments. Judson and Owen (1999) suggested to

restrict the instruments to a fixed numbers of  $m$  lags, that is, for period  $t$  the GMM estimator employs  $\min(m, t - 2)$  lags instead of employing all potential  $t - 2$  instruments. Indeed if  $\alpha$  is substantially smaller than unity, it makes sense to assume that the information of the lagged dependent variable diminishes as the lag order increases. Using Monte Carlo simulations, Judson and Owen (1999) found that limiting the number of lags may improve the small sample properties of the GMM estimator in sample sizes typically encountered in macroeconomic applications.

Breitung (1994) and Roodman (2009a) proposed to “collapse” the instruments yielding the  $(T - 1) \times (T - 1)$  instrumental variable matrix

$$Z'_i = \begin{bmatrix} y_{i0} & y_{i1} & y_{i2} & \cdot & y_{i,T-2} \\ 0 & y_{i0} & y_{i1} & \cdot & y_{i,T-3} \\ 0 & 0 & y_{i0} & \cdot & y_{i,T-4} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & y_{i0} \\ \Delta x_{i2} & \Delta x_{i3} & \Delta x_{i4} & \cdot & \Delta x_{iT} \end{bmatrix}$$

such that  $E(Z'_i \Delta u_i) = 0$  with  $\Delta u_i = [\Delta u_{i2}, \dots, \Delta u_{iT}]'$ . If  $y_{it}$  is a stationary process, then  $y_{i,t-\ell} \Delta u_{it}$  has the same distribution for all  $t$  and fixed  $\ell$ . Thus, there is no reason to attach different weights to the moments. Accordingly, it is natural to sum up the corresponding moments yielding the collapsed moment condition  $E(\sum_{t=\ell+1}^T y_{i,t-\ell} \Delta u_{it}) = 0$  for  $\ell = 2, 3, \dots, T - 1$ . Note also that the instruments of a strictly exogenous regressor are treated in a similar manner (the “IV style” according to Roodman 2009b).

Another popular approach is to apply the method of principal components to the sample covariance matrix of all possible instruments. To this end a smaller number of linear combinations of the instrumental variables are selected to represent the main information in the set of instruments. The relevant linear combinations are obtained from the eigenvectors associated with the largest eigenvalues of the sample covariance (or correlation) matrix of all instrumental variables. This reduction technique has a long tradition starting with Kloeck and Mennes (1960) and was adopted to the estimation of dynamic panel data models by Doran and Schmidt (2006), Bai and Ng (2010), Kapetanios and Marcellino (2010), and Bontempi and Mammi (2012). Although this method is intuitively appealing, there is no guarantee that the first principal components of the instrumental variables comprises the most relevant information in the instrument set since the principal components just capture the co-movement among the instruments but fail to exploit the relationship between the instruments and the regressors. The asymptotic variance of the IV estimator, however, depends on the correlation between the instruments and regressors which is not exploited by the method of principal components.<sup>3</sup>

Bai and Ng (2009) consider methods for instrumental variable selection if the number of potential instruments is large. In some sense the problem of choosing good instruments is related to the problem of selecting optimal predictors from a large set

of candidates. From the two-stage interpretation of the IV estimator it follows that the variance of the IV estimator is inversely related to the  $R^2$  from a regression of the endogenous regressor on the instruments. Thus, the problem is to find those instrument that best “predict” the lagged dependent variable. Ng and Bai (2008) propose to apply a boosting algorithm that introduces one instrument after the other until the AIC criterion suggests that no further improvement is possible. A somewhat related method is “hard thresholding” where the instruments are ranked according to their  $t$ -statistic in the first-step regression. When adapting these methods to the GMM estimator for the dynamic panel data model, it is important to note that the first-step regression involves a particular variable transformation. Consider the regression model  $y = X\beta + u$ , where  $E(uu') = \Omega$ . The instrumental variable matrix is denoted by  $Z$ . The two-stage interpretation for the GMM estimator is obtained as

$$\begin{aligned}\Omega^{-1/2}y &= \Omega^{-1/2}X\beta + \Omega^{-1/2}u \\ \Omega^{-1/2}X &= \Omega^{1/2}Z\Pi + \nu.\end{aligned}$$

It is easy to see that replacing  $X^* = \Omega^{-1/2}X$  by  $\widehat{X}^* = \Omega^{1/2}Z\widehat{\Pi}$ , where  $\widehat{\Pi}$  is the least-squares estimator from a regression of  $X^*$  on  $Z^* = \Omega^{1/2}Z$ , the usual GMM estimator results. It is important to note that the instrumental variable matrix  $Z$  is transformed by  $\Omega^{1/2}$  instead of  $\Omega^{-1/2}$ . Accordingly, some set of instruments with high explanatory power for  $X$  need not necessarily imply a good fit of  $X^* = \Omega^{-1/2}X$ . Hence, the instrumental variable selection should be based on the transformed variables  $X^*$  and  $Z^*$ .

#### 15.2.4 Bias-Corrected Estimators

The GMM methodology avoids the Nickell bias by employing instruments that are uncorrelated with the error term of the differenced model. An alternative approach invokes a bias correction of the ordinary within-group estimator that has been shown to possess much better small sample properties than the GMM estimators if  $T$  is large relative to  $N$  (e.g., Kiviet 1995; Judson and Owen 1999). Such estimators can be motivated by the results of Lancaster (2002), who showed that a likelihood conditioned on the ML estimate of an orthogonalized effect leads to bias-corrected scores and respective method-of-moments estimators.

To focus on the main issues, let us consider the pure auto-regressive model  $y_{it} = \mu_i + \alpha y_{i,t-1} + \varepsilon_{it}$ . Assuming normally distributed errors and treating  $y_{i0}$  as fixed constants, the first order condition of the ML estimator is

$$\zeta_{NT}(\widehat{\alpha}) = \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T (y_{i,t-1} - \bar{y}_{i,-1})(y_{it} - \widehat{\alpha}y_{i,t-1}) = 0,$$

where  $\bar{y}_{i,-1} = T^{-1} \sum_{t=1}^T y_{i,t-1}$ . The probability limit of the scores evaluated at the true values results as

$$\operatorname{plim}_{N \rightarrow \infty} \zeta_{NT}(\alpha) = -\sigma^2 b_T(\alpha) \quad (15.11)$$

where

$$b_T(\alpha) = \frac{1}{T} \sum_{t=1}^T \sum_{j=0}^{T-t-1} \alpha^j \quad (15.12)$$

(cf. Chapter 4). Furthermore, assuming stationary initial conditions

$$y_{i0} \sim \mathcal{N}(\mu_i/(1-\alpha), \sigma^2/(1-\alpha^2))$$

we obtain

$$\begin{aligned} E\left[\frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T (y_{i,t-1} - \bar{y}_{i,-1})^2\right] &= \frac{\sigma^2}{1-\alpha^2} \left( T - \sum_{j=-T+1}^{T-1} \alpha^j (T-|j|) \right) \\ &\equiv \sigma^2 T \varphi_T(\alpha) \end{aligned}$$

(cf. Chapter 4). Since  $b_T(\alpha) = 1/(1-\alpha) + O(T^{-1})$  and  $\varphi_T(\alpha) = 1/(1-\alpha^2) + O_p(T^{-1})$  it follows that

$$\operatorname{plim}_{N \rightarrow \infty} \hat{\alpha} - \alpha = -\frac{1+\alpha}{T} + O(T^{-2}). \quad (15.13)$$

Accordingly, Hahn and Kuersteiner (2002) suggest the simple bias corrected estimator

$$\tilde{\alpha}_{bc} = \hat{\alpha} + \frac{1}{T}(1+\hat{\alpha}) = \frac{T+1}{T}\hat{\alpha} + \frac{1}{T}.$$

Obviously, this estimator involves a trade-off between bias and variance, since the factor  $(T+1)/T$  leads to a higher variance. Alternatively, the bias correction may be computed iteratively by inserting the bias-corrected estimator such that

$$\tilde{\alpha}_{bc}^\infty = \frac{T}{T-1}\hat{\alpha} + \frac{1}{T-1}$$

(cf. Juodis 2013b).

A bias-correction suitable for small  $T$  was proposed by Kiviet (1995), which is based on a very accurate approximation of the bias in the dynamic model with strictly exogenous regressors. Instead of inserting the (biased) within-estimator, Kiviet (1995) suggested to plug in a consistent initial estimator (e.g., a GMM estimator) for estimating the bias term. An important problem is, however, that the estimation error of the bias term affects the asymptotic distribution of the bias-corrected estimator in a complex way. Thus, standard errors and  $t$ -statistics have to be computed by using bootstrap methods.

Bun and Carree (2005) propose a bias-corrected estimator<sup>4</sup> by iteratively solving the equation:

$$\widehat{\alpha}_{bc} = \widehat{\alpha} + \frac{b_T(\widehat{\alpha}_{bc})}{T\varphi(\widehat{\alpha}_{bc})}.$$

The estimator proposed by Dhaene and Jochmans (2012) applies a bias correction to the profile scores obtained from the gradients of the concentrated log-likelihood

$$\begin{aligned} s_{NT}(\alpha) &= \frac{1}{\widehat{\sigma}^2 NT} \sum_{i=1}^N \sum_{t=1}^T y_{i,t-1} (u_{it} - \bar{u}_i) \\ &= \frac{\sum_{i=1}^N \sum_{t=1}^T y_{i,t-1} (u_{it} - \bar{u}_i)}{\sum_{i=1}^N \sum_{t=1}^T (u_{it} - \bar{u}_i)^2}. \end{aligned}$$

The adjusted (centered) profile scores are defined as

$$s_{NT}^A(\alpha) = s_{NT}(\alpha) - \lim_{N \rightarrow \infty} E[s_{NT}(\alpha)].$$

A consistent estimator can be obtained from setting the adjusted profile scores equal to zero. As shown by Breitung (2013), the solution is equivalent to a method of moment estimator based on the nonlinear moment

$$m_{NT}(\alpha) = \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T \left( y_{i,t-1} + \frac{b_T(\alpha)}{T-1} u_{it} \right) (u_{it} - \bar{u}_i), \quad (15.14)$$

where  $b_T(\alpha)$  is defined as in (15.12). As  $N/T \rightarrow c$  the limiting distribution of the profile score estimator is obtained as  $\sqrt{NT}(\widehat{\alpha}_{ps} - \alpha) \xrightarrow{d} \mathcal{N}(0, 1 - \alpha^2)$  and, therefore, the estimator eliminates the finite sample bias as well as the asymptotic bias without affecting the asymptotic variance.

This bias-correction method can easily be extended to models with exogenous regressors by expanding the moments as  $m_{NT}(\alpha, \beta) = [m_{NT,\alpha}(\alpha, \beta), m_{NT,\beta}(\alpha, \beta)']'$  with

$$m_{NT,\alpha}(\alpha, \beta) = \sum_{t=1}^T \left( y_{i,t-1} + \frac{b_T(\alpha)}{T-1} u_{it} \right) (u_{it} - \bar{u}_i) \quad (15.15)$$

$$m_{NT,\beta}(\alpha, \beta) = \sum_{t=1}^T x_{it} (u_{it} - \bar{u}_i), \quad (15.16)$$

where the notation suppresses the dependence of the residuals  $u_{it}$  on the coefficients  $\alpha$  and  $\beta$ . The resulting profile score estimator is obtained from solving the moment condition  $m_{NT}(\widehat{\alpha}_{ps}, \widehat{\beta}_{ps}) = 0$ . It is important to note that these moment conditions imply a polynomial equation in  $\alpha$  giving rise to multiple solutions. As shown by Dhaene and Jochmans (2012) the parameters are locally identified within a fixed interval around

the true values. Accordingly, using the within-group estimator as initial value for the gradient method, the nonlinear method of moments estimator eventually converges to the consistent solution when  $T$  is sufficiently large.

The bias-correction method considered so far is based on the within-group estimator. At least for the auto-regressive model with stationary initial conditions and without exogenous variables, much simpler estimators are obtained by employing alternative methods to eliminate the individual effects. For example it is not difficult to verify (cf. Phillips and Han 2008 and Breitung 1994) that the pooled OLS estimator of  $\alpha$  in the two equations

$$\begin{aligned} (2\Delta y_{it} + \Delta y_{i,t-1}) &= \alpha \Delta y_{i,t-1} + e_{it} \\ (2\tilde{y}_{it} - \tilde{y}_{i,t-1}) &= \alpha \tilde{y}_{i,t-1} + v_{it} \end{aligned}$$

are consistent for  $\alpha$ , where  $\tilde{y}_{it} = y_{it} - y_{1t}$ . Since the errors of these regressions are autocorrelated, the OLS estimator is inefficient. Han, Phillips, and Sul (2011) suggest a transformation based on “X-differences” that combines the usual first differences and long differences in order to cancel out the individual effects and obtaining uncorrelated errors at the same time. For example, it can be shown that the regression

$$\Delta y_{it} + \Delta y_{i,t-1} + \Delta y_{i,t-2} = y_{it} - y_{i,t-3} = \alpha \Delta y_{i,t-1} + e_{it} \quad (15.17)$$

yields an asymptotically unbiased estimator with  $E(\Delta y_{i,t-1} e_{it}) = 0$  and serially uncorrelated errors. Unfortunately it is not clear how to adapt this appealing transformations to a dynamic panel data model with exogenous regressors.

### 15.2.5 ML Estimation

The ML estimator of the transformed model proposed by Breitung (1994), Hsiao, Pesaran, and Tahmisioglu (2002) and Kruiniger (2008) is another attractive approach whenever  $T$  is moderately large.<sup>5</sup> The idea is to maximize the implied likelihood function for the transformed vector  $\Delta \tilde{u}_i = [v_{i1}, \Delta u_{i2}, \dots, \Delta u_{iT}]'$ , where  $\Delta u_{it} = \Delta y_{it} - \alpha \Delta y_{i,t-1} - \beta' \Delta x_{it}$  for  $t = 2, \dots, T$  and  $v_{it}$  is the residual of the initial condition discussed below. Note that the dynamic model implies a system of simultaneous equations of the form

$$A \Delta y_i = \Delta X_i \beta + \Delta u_i, \quad (15.18)$$

where  $\Delta y_i = [\Delta y_{i1}, \dots, \Delta y_{iT}]'$ ,  $\Delta X_i = [\Delta x_{i1}, \dots, \Delta x_{iT}]'$

$$A = \begin{bmatrix} 1 & 0 & 0 & \cdot & 0 & 0 \\ -\alpha & 1 & 0 & \cdot & 0 & 0 \\ 0 & -\alpha & 1 & \cdot & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & -\alpha & 1 \end{bmatrix}.$$

The covariance matrix of the errors depends on the initial condition but has the general form

$$\Omega = \begin{bmatrix} \omega_{11} & \omega_{12} & \omega_{13} & \omega_{14} & \cdot & \omega_{1,T-1} & \omega_{1T} \\ \omega_{12} & 2 & -1 & 0 & \cdot & 0 & 0 \\ \omega_{13} & -1 & 2 & -1 & \cdot & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \omega_{1T} & 0 & 0 & 0 & \cdot & -1 & 2 \end{bmatrix}.$$

The likelihood function of this model corresponds to a system of simultaneous equations with cross-equation restrictions and covariance restrictions. Accordingly, the maximization of the log-likelihood function is not trivial. An important problem with this approach is that an initial condition needs to be imposed for  $\Delta y_{i1}$ . Note that the dynamic model implies

$$v_{i1} = \Delta y_{i1} - \beta' \Delta x_{i1} = \alpha \Delta y_{i0} + \Delta u_{i1}.$$

If the process starts in the remote past, the contributions of the individual effect sum up to  $\mu_i/(1-\alpha)$  in  $y_{i0}$  and  $y_{i,-1}$  so that the initial condition will not depend on  $\mu_i$ , at least if  $x_{i0}, x_{i,-1}, \dots$  do not depend on the individual effect. Since it is assumed that the regressors are strictly exogenous,  $x_{i1}$  is not correlated with  $\Delta u_{i2}, \dots, \Delta u_{iT}$ . This implies that  $\omega_{1t} = 0$  for  $t \geq 3$ ,  $\omega_{11} = \text{var}(v_{i1})$  is an unrestricted parameter and  $\omega_{12} = -\sigma^2$ .<sup>6</sup> Assume, on the other hand, that the process is started with  $y_{i0} = 0$ . In this case the individual effect does not cancel from  $\Delta y_{i1}$  and, therefore, additional assumptions on the individual effect in  $\Delta y_{i1}$  are required.

Obviously, the difference transformation is not the only transformation that eliminates the individual effect. As an alternative consider the following transformation (see also Alvarez and Arellano (2004) and Grassetto (2011))

$$y_{it} - y_{i1} = \alpha(y_{i,t-1} - y_{i0}) + \beta'(x_{it} - x_{i1}) + u_{it} - u_{i1} \quad t = 3, 4, \dots, T \quad (15.19)$$

$$y_{it}^* = \xi_i + \alpha y_{i,t-1}^* + \beta' x_{it}^* + u_{it}, \quad (15.20)$$

where  $y_{it}^* = y_{it} - y_{i1}$ ,  $x_{it}^* = x_{it} - x_{i1}$ , and  $\xi_i = -u_{i1}$  is uncorrelated with all  $u_{i2}, \dots, u_{iT}$ . Interestingly, this transformation converts the original dynamic model with fixed effect into a dynamic random effects model. The ML estimation of the latter model is considered in Nerlove (1971) and Bhargava and Sargan (1983).

Bai (2013) proposes a factor analytic interpretation of the estimation problem. For simplicity consider the pure auto-regressive model. Define the  $(T-1) \times (T-1)$  matrix  $W_N = (w_{ts,N})$  ( $t, s \in \{1, \dots, T-1\}$ ) with typical element

$$w_{ts,N} = \frac{1}{N} \sum_{i=1}^N (y_{i,t+1} - \alpha y_{it})(y_{i,s+1} - \alpha y_{is})$$

and expectation

$$\Sigma(\theta) = E(W_N) = \iota_{T-1} \iota_{T-1}' \pi_N + \sigma^2 I_{T-1},$$

where  $\pi_N = N^{-1} \sum_{i=1}^N \mu_i^2$  and  $\theta = [\alpha, \sigma^2, \pi_N]'$ .<sup>7</sup> For estimation of the parameters the discrepancy function

$$Q_N(\theta) = \log|\Sigma(\theta)| + \text{tr}[W_N \Sigma(\theta)^{-1}]$$

is minimized which is equivalent to ( $-N/2$  times) the standard log-likelihood function. It is important to appreciate that instead of the individual parameters  $\mu_i$ , the objective function only depends on the second moment of the individual effects  $\pi_N$ . Essentially, by treating the individual effects as random components this approach sidesteps the incidental parameters problem. As shown by Bai (2013) the resulting estimator of  $\alpha$  is asymptotically efficient for fixed  $T$  or  $T \rightarrow \infty$ . In the model with strictly exogenous regressors  $x_{it}$ , the individual effect is specified as  $\mu_i = c_0 + c'_1 x_{i1} + \dots + c'_T x_{iT} + \eta_i$  (the so-called Mundlak-Chamberlain model), where  $\eta_i$  is the remaining individual effect that is uncorrelated with  $x_{it}$ . Alternatively, the transformation (15.20) can be used to transform the fixed effect into a random effect.

### 15.2.6 VAR Models

Holtz-Eakin, Newey, and Rosen (1988) proposed a GMM estimator for the vector autoregressive model<sup>8</sup>

$$y_{it} = \mu_i + A_1 y_{i,t-1} + \dots + A_p y_{i,t-p} + u_{it}, \quad (15.21)$$

where  $y_{it} = [y_{1,it}, \dots, y_{m,it}]'$  is an  $m$ -dimensional time series vector,  $\mu_i = [\mu_{1,i}, \dots, \mu_{m,i}]'$  and  $E(u_{it} u'_{it}) = \Sigma$ . We may include further predetermined or strictly exogenous variables  $x_{it}$  which gives rise to the VARX model. A comprehensive review of theoretical and empirical work based on VAR or VARX models is provided by Canova and Ciccarelli (2013).

A natural approach for estimating the system of equations (15.21) is to apply the single-equation GMM estimator (allowing for predetermined regressors) to each of the  $m$  equations. In contrast to the pure time series case with  $N = 1$ , the single equation estimator is not asymptotically efficient. Holtz-Eakin, Newey, and Rosen (1988) suggest to combine the set of moment conditions for all equations yielding a system-estimator for the entire system. Furthermore, it should be noted that the regressors are predetermined but not strictly exogenous and, therefore, the set of moment conditions are given by  $E[\text{vec}(y_{is} \Delta u'_{it})] = 0$  for  $s = 1, \dots, t-2$ .

It is apparent that the pitfalls of the GMM estimator for large  $T$  also apply to the GMM estimator adapted to a VAR system, in particular, as the regressors also have to be instrumented by their lags. Hahn and Kuersteiner (2002) derived the bias of the within-group estimator applied to each equation with  $p = 1$ . It is not difficult to generalize the

results of Section 15.2.4 to the first order VAR yielding

$$\begin{aligned} E \left[ \frac{1}{N} \sum_{i=1}^N \sum_{t=3}^T (y_{i,t-1} - \bar{y}_{i,-1})(u_{it} - \bar{u}_{it})' \right] \\ = (I_m - A)^{-1} \Sigma + O(T^{-1}) \end{aligned} \quad (15.22)$$

$$\begin{aligned} E \left[ \frac{1}{N} \sum_{i=1}^N \sum_{t=3}^T (y_{i,t-1} - \bar{y}_{i,-1})(y_{i,t-1} - \bar{y}_i^1)' \right] \\ = T \left[ \Sigma + \sum_{\ell=1}^{\infty} A^\ell \Sigma (A')^\ell \right] + O(1). \end{aligned} \quad (15.23)$$

With these results a bias corrected estimator is obtained by adjusting the moments (15.22) accordingly (see also Juodis (2013a) for a bias correction that is appropriate for small  $T$ ). Binder, Hsiao, and Pesaran (2005) propose a (transformed) ML estimator for the panel VAR model which is a multivariate extension of the single equation ML estimator considered in Section 15.2.5. Another strand of literature employs Bayesian methods to allow for individual specific and temporal variations in the auto-regressive components (cf. Canova and Ciccarelli 2013).

An interesting application within a panel VAR framework is testing for (Granger) causality which is equivalent to a significance test of the corresponding elements of the matrix  $A$  (or  $A_1, \dots, A_p$ ). As argued in Section 15.2.2 the asymptotic bias of the within-group estimator renders such tests invalid even if  $T$  is large. Therefore, a GMM estimator or bias-corrected estimator is required for valid inference on Granger causality. By using the usual identification schemes (e.g., Lütkepohl and Krätsig 2004) it is straightforward to identify structural shocks from the covariance matrix  $\Sigma$ .

It is important to notice that the panel VAR relies on a number of restrictive assumptions that may be considered as unrealistic in many empirical applications. First, it is assumed that the dynamic impulse responses are similar for all panel units and the respective shocks have identical variances and covariances. Moreover, the panel units are assumed to be independent of each other such that a shock in one panel unit does not affect any of the other units. In other words, countries are considered as “clones” of each other with an identical dynamic propagation mechanism but the shocks that hit the economies are independent across countries. A much more general dynamic framework was studied by Groen and Kleibergen (2003) which allows for heterogeneous dynamics and dynamic spill-overs from one country to another. On the other hand such a multi-country VAR involves a substantial number of additional parameters that may even exceed the number of observations. A compromise between an overly restrictive panel VAR and an (actually infeasible) interdependent multi-country VAR is the “Global VAR” which is considered in Section 15.4.6.

### 15.2.7 Random Versus Fixed Effect

In the dynamic analysis of panel data, it has become standard to adopt a fixed-effect framework. If the individual effects are uncorrelated with the regressors, however, additional (and potentially powerful) instruments are available for improving the efficiency of the estimator. Another drawback of the fixed-effect framework is that time-constant variables are “differenced out.” To get an idea of the efficiency gain, consider first a simple static panel data model of the form  $y_{it} = \mu_i + \beta x_{it} + u_{it}$ . The relative efficiency crucially depends on the ratio of the between-group variance and the within-group variance of  $x_{it}$

$$\lambda = \frac{\frac{1}{N} \sum_{i=1}^N (\bar{x}_i - \bar{x})^2}{\frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i)^2}.$$

The relative efficiency results as

$$\frac{var(\hat{\beta}_{gls})}{var(\hat{\beta})} = \frac{1}{1 + \varrho_T \lambda},$$

where  $\hat{\beta}_{gls}$  is the random effects GLS estimator,  $\varrho_T = \sigma^2 / (T\sigma_\mu^2 + \sigma^2)$  and  $\sigma_\mu^2 = E(\mu_i^2)$ . Since  $\varrho_T = O(T^{-1})$  it follows that the efficiency gain vanishes as  $T \rightarrow \infty$ . For moderate values of  $T$ , however, the efficiency gain may be sizable, in particular if the between-group variance is large relative to the within-group variance (measured by  $\lambda$ ).

If  $\mu_i$  and  $\bar{x}_i$  are uncorrelated, then additional instruments are available (e.g., Arellano 2003, p. 137). For the dynamic model the additional  $k$  moment conditions are given by

$$E[\bar{x}_i(\bar{y}_{i,0} - \alpha \bar{y}_{i,-1} - \beta' \bar{x}_i)] = 0$$

and a Hausman type test for the null hypothesis  $E(\mu_i|\bar{x}_i)$  is available (cf. Arellano 1993). A bias-corrected estimator based on the adjusted profile scores is proposed by Breitung (2013).

## 15.3 POOLED VERSUS INDIVIDUAL SPECIFIC PARAMETERS

---

In classical panel data analysis, unobserved heterogeneity is captured by introducing individual specific constants, whereas all other parameters are assumed to be identical across individuals. There are several reasons for this assumption. For instance, economic theory often implies that rational agents respond similarly to a changes in

economically relevant variables. Moreover, in panels with small  $T$  and large  $N$  it is not suitable to treat all coefficients as individual specific since estimation and inference are unreliable (or even infeasible) and it is not clear how to interpret the large set of individual specific parameters. Indeed the main appeal of panel data analysis is the possibility of pooling individual data in a large data set, assuming at least some similarity among individuals.

This reasoning, however, is less convincing when analyzing macroeconomic panel data as the institutional and cultural background, preferences, and the level of economic development may vary substantially among different countries (e.g., Mairesse and Griliches 1990). In macroeconomic applications it is therefore important to consider more general forms of heterogeneity. Since typical macroeconomic panels provide a relative large number of time periods, individual specific estimates of all parameters are feasible and sufficiently reliable. Alternative tests for the hypothesis of heterogeneous slope parameters are considered in Section 15.3.1.

There are two popular approaches to obtain a common parameter value that represents the central tendency among heterogeneous responses.<sup>9</sup> First, the (possibly transformed) observations of all panel groups may be pooled in a large data-set in order to estimate a common model. Second, the model may be estimated separately for all panel groups and averaged afterwards to obtain a joint estimator. The latter method is called the *mean-group estimator* (cf. Pesaran and Smith 1995; and Pesaran, Smith, and Im, 1996). It is obvious that under the usual *i.i.d.* assumptions the pooled (within-group) estimator for a static panel data model is more efficient than the mean-group estimator. The situation is less clear, however, if the slope parameters are heterogeneous as in this case the errors are heteroskedastic and autocorrelated. Swamy (1970) proposed a GLS estimator based on the estimated covariance matrix of the random coefficient vectors (see Chapter 13). Notwithstanding the desirable asymptotic properties of this estimator, it has not become very popular in practice. In Section 15.4.2 it is argued that the mean-group estimator is remarkably robust. In some sense it can be seen as a fixed-effects variant of the random coefficient model. A Hausman test can be employed to choose among the two approaches.

### 15.3.1 Testing for Parameter Heterogeneity

A straightforward approach to test the hypotheses  $\beta_1 = \dots = \beta_N = \beta$  (poolability) in the heterogeneous panel data model

$$\begin{aligned} y_{it} &= \mu_i + \beta'_i x_{it} + u_{it} \\ \tilde{y}_{it} &= \beta'_i \tilde{x}_{it} + \tilde{u}_{it} \end{aligned}$$

is the  $F$ -statistic that results in a comparison of the residual sum of squares from a within-group estimation (i.e., a regression under the null hypothesis)

$$\text{SSE}_0 = \sum_{i=1}^N \sum_{t=1}^T (\tilde{y}_{it} - \hat{\beta}' \tilde{x}_{it})^2$$

and the sum of the squared residuals from separate estimations of all panel units,  $\sum_{i=1}^N \tilde{y}_i' \tilde{M}_i \tilde{y}_i$ , where  $\tilde{M}_i = I_T - \tilde{X}_i(\tilde{X}_i' \tilde{X}_i)^{-1} \tilde{X}_i'$ . The usual  $F$ -statistic

$$F_H = \frac{NT - (k+1)N}{k(N-1)} \left( \frac{\text{SSE}_0 - \sum_{i=1}^N \tilde{y}_i' \tilde{M}_i \tilde{y}_i}{\sum_{i=1}^N \tilde{y}_i' \tilde{M}_i \tilde{y}_i} \right)$$

is  $F$  distributed with  $k(N-1)$  and  $NT - (k+1)N$  degrees of freedom as  $T \rightarrow \infty$  and fixed  $N$  (e.g., Baltagi 2008, section 4.1). An important practical problem is, however, that this test statistic requires large  $T$  relative to  $N$  as otherwise the test suffers from severe size distortions. Furthermore, as shown by Bun (2004), the  $F$ -test suffers from size distortions in dynamic panel data models if  $T/N \rightarrow \infty$ .

A test statistic with much better size properties was suggested by Pesaran and Yamagata (2008). Assuming individual specific variances  $E(u_{it}^2) = \sigma_i^2$  the pooled estimator is the weighted least-squares estimator

$$\hat{\beta}_p = \left( \sum_{i=1}^N \frac{1}{\hat{\sigma}_i^2} \tilde{X}_i' \tilde{X}_i \right)^{-1} \sum_{i=1}^N \frac{1}{\hat{\sigma}_i^2} \tilde{X}_i' \tilde{y}_i,$$

where  $\hat{\sigma}_i^2 = (\tilde{y}_i - \tilde{X}_i \hat{\beta}_i)'(\tilde{y}_i - \tilde{X}_i \hat{\beta}_i)/(T-k-1)$ .

The dispersion statistic is based on the (squared Mahalanobis) distance between the panel specific OLS estimates  $\hat{\beta}_i$  and the pooled estimate  $\hat{\beta}_p$ :

$$\tilde{S} = \sum_{i=1}^N \frac{1}{\hat{\sigma}_i^2} (\hat{\beta}_i - \hat{\beta}_p)' \tilde{X}_i' \tilde{X}_i (\hat{\beta}_i - \hat{\beta}_p)$$

and the standardized version of this statistic is obtained as

$$\hat{\Delta} = \sqrt{N} \frac{N^{-1} \tilde{S} - k}{\sqrt{2k}}. \quad (15.24)$$

If  $N$  and  $T$  tend to infinity and  $\sqrt{N}/T \rightarrow 0$ , this test statistic has a standard normal limiting distribution (cf. Pesaran and Yagamata 2008).

Juhl and Lugovskyy (2014) and Breitung, Salish, and Roling (2013) treat  $\beta_i$  as random coefficients and propose an LM statistics for the null hypothesis  $\text{var}(\beta_i - \beta) = 0$  against  $\text{var}(\beta_i - \beta) = \bar{\omega} I_k$  and  $\text{var}(\beta_i - \beta) = \text{diag}(\omega_1, \dots, \omega_k)$ , respectively. The latter test is based on the likelihood scores

$$\tilde{s}_i = \begin{bmatrix} \sum_{t=1}^T \sum_{s=1}^T \tilde{u}_{it} \tilde{u}_{is} x_{it,1} x_{is,1} - \sum_{i=1}^N \hat{\sigma}^2 x_{it,1}^2 \\ \vdots \\ \sum_{t=2}^T \sum_{s=1}^{t-1} \tilde{u}_{it} \tilde{u}_{is} x_{it,k} x_{is,k} - \sum_{t=1}^T \hat{\sigma}^2 x_{it,k}^2 \end{bmatrix}.$$

where  $\hat{\sigma}^2$  denotes an asymptotically unbiased estimator of the residual variance. Juhl and Lugovskyy (2013) and Breitung, Salish, and Roling (2013) propose modified versions of the LM test that are robust against non-normal error distributions and heteroskedasticity.

### 15.3.2 Alternative Estimators for Heterogeneous Panels

There are two general principles to construct a panel data estimator. First, we may pool the data and estimate the parameter from the (possibly transformed) data set and, second, the parameter may be estimated for each panel group and in a second step the individual specific estimators are pooled to obtain a single estimator. Let us focus on a panel regression with a single regressor ( $k = 1$ ). Denote the individual specific OLS estimators by  $\hat{\beta}_1, \dots, \hat{\beta}_N$  treating the individual effect  $\mu_i$  as a constant constants. It is not difficult to see that the within-group estimator can be written as

$$\hat{\beta} = \sum_{i=1}^N w_i \hat{\beta}_i, \quad (15.25)$$

where

$$w_i = \frac{s_{xi}^2}{\sum_{j=1}^N s_{xj}^2}$$

and  $s_{xi}^2 = T^{-1} \sum_{t=1}^T (x_{it} - \bar{x}_i)^2$ . The mean-group estimator of Pesaran and Smith (1995) results by letting  $w_i = 1/N$ . This estimator can also be written as a weighted within-group estimator, where the variables are divided by  $s_{xi}$ . The weights for the GLS estimator are obtained as

$$w_i = \frac{(\sigma^2 + \sigma_\beta^2 s_{xi}^2)^{-1} s_{xi}^2}{\sum_{j=1}^N (\sigma^2 + \sigma_\beta^2 s_{xj}^2)^{-1} s_{xj}^2},$$

where the deviation  $\beta_i - \beta$  is assumed to be random with  $\sigma_\beta^2 = \text{var}(\beta_i - \beta)$ . Obviously, all estimators are similar if the regressor variance  $s_{xi}^2$  does not depend on  $i$ .

It is interesting to compare these estimators if the coefficients  $\beta_i$  vary systematically. For concreteness assume that  $\beta_i$  linearly depends on some variable  $z_i$  such that

$$\beta_i = \beta + \delta z_i + \nu_i, \quad (15.26)$$

where  $\nu_i$  is an *i.i.d.* random effect. Following the influential work of Mundlak (1978) we may assume  $z_i = (\bar{x}_i - \bar{x})$ , where the overall-mean  $\bar{x} = (NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T x_{it}$  is subtracted to ensure that  $E(N^{-1} \sum_{i=1}^N \beta_i) = \beta$ . Since  $\hat{\beta}_i$  is an unbiased estimator, it follows that as  $N \rightarrow \infty$ , the mean-group estimator converges in probability to  $E(\beta_i) = \beta$ . In contrast, the within-group estimator (and also the GLS estimator) is inconsistent whenever  $z_i$  is correlated with  $w_i$  (resp.  $s_{xi}^2$ ). In this sense, the mean-group estimator is more robust against systematically varying slopes. On the other hand, the robustness of the mean-group estimator comes at the expense of a higher variance, in particular, in comparison with the GLS estimator.

Pesaran, Smith, and Im (1996) propose a Hausman test based on the difference between the within-group estimator and the mean-group estimator. Since the within-group estimator is inconsistent whenever  $\beta_i$  is correlated with  $s_{xi}^2$ , whereas the mean-group estimator remains consistent, a rejection of the Hausman test indicates that  $\beta_i$  varies systematically with  $s_i^2$ . This suggests that the within-group (or random effects GLS) estimator is preferred whenever the Hausman test is insignificant, whereas the mean-group estimator is chosen if the Hausman test rejects.

Pesaran and Smith (1995) studied various estimation methods in the dynamic panel data model with random coefficients given by

$$y_{it} = \mu_i + \alpha_i y_{i,t-1} + \beta_i x_{it} + u_{it}.$$

Let  $\alpha_i = \alpha + \nu_{1i}$  and  $\beta_i = \beta + \nu_{2i}$ , where  $\nu_{1i}$  and  $\nu_{2i}$  are *i.i.d.* disturbances so that

$$y_{it} = \mu_i + \alpha y_{i,t-1} + \beta x_{it} + e_{it}$$

with  $e_{it} = y_{i,t-1} \nu_{1i} + x_{it} \nu_{2i} + u_{it}$ . If  $x_{it}$  is autocorrelated, then  $E(y_{i,t-1} e_{it}) \neq 0$  since  $y_{i,t-1}$  depends on the errors  $\nu_{1i}$  and  $\nu_{2i}$ . It follows that the within-group estimator (as well as other estimators) is inconsistent even if  $T \rightarrow \infty$ .

## 15.4 CROSS-SECTION DEPENDENCE

---

Analyzing regional data, it is natural to assert that (for example due to international trade and migration) all regions depend on each other, where the correlation decreases with the geographic or economic distance among the regions. Some of this spatial dependence may be captured by explanatory variables but it is likely that some cross-correlation remains among the regression errors. Alternative approaches to model cross-section dependence are considered in Chapter 1 and Section 15.4.1. If the number of countries is small relative to the number of time periods, the SUR-GLS approach

is appealing as it does not impose any restriction on the contemporaneous covariance matrix. If  $T/N$  is small (e.g.,  $< 10$ ), however, GLS estimation and inference suffer from poor small sample properties. In this case the within-group estimator equipped with “panel corrected standard errors” (PCSE) is a reasonable alternative (e.g., Beck and Katz 1995). Another possibility is to approximate the cross-section dependence by a factor structure.

Various alternative test procedures are available to test against cross-section dependence. Most tests are based on a (possibly weighted) sum of squared error covariances  $E(u_{it}u_{jt})$  for all  $i \neq j$ . It should be noted, however, that valid estimation and inference based on the within-group estimator does not require mutually uncorrelated errors. As long as the cross-covariances do not vary systematically with the cross products of the regressors, cross-correlation does not render inference based on the within-group estimator invalid.

In recent years factor-augmented panel data models (or the interactive effect model) have become popular. This class of models is considered briefly in Section 15.4.5 (see Chapters 1 and 4 for a more detailed discussion). A convenient approach proposed by Pesaran (2006) just includes the cross-section means of all variables in order to capture the cross-dependence due to a few common factors. More sophisticated (but not necessarily better methods) are based on a principal component analysis.

### 15.4.1 Modeling Cross-Section Dependence

The simplest method to cope with cross-section dependence is to introduce “time effects” (that is period-specific dummy variables). Let  $\xi_t$  and  $\varepsilon_{it}$  denote *i.i.d.* error components, where  $\xi_t$  is common to all individuals and consider the composed error term  $u_{it} = \xi_t + \varepsilon_{it}$ . Since  $E(u_{it}u_{jt}) = E(\xi_t^2)$  for  $i \neq j$  it is obvious that including time effects accounts for a constant contemporaneous correlation among the panel units. Although this approach is simple and very popular in empirical practice, it is overly restrictive in most macroeconomic applications.

The opposite extreme is to assume an unrestricted covariance matrix for the error vector  $u_t = [u_{1t}, \dots, u_{Nt}]'$  such that  $\Omega = E(u_t u_t')$  is any symmetric and positive definite  $N \times N$  matrix. A serious drawback of this approach is that it involves  $N(N+1)/2$  covariance parameters which raises severe problems if  $N$  is large relative to  $T$ . Accordingly, the respective ML estimator suffers from poor small sample properties in typical macroeconomic applications.

If the error correlation is related to the (geographic or economic) distance between the panel units, the correlation may be specified by a spatial lag model given by

$$u_t = \rho W_N u_t + \varepsilon_t, \quad (15.27)$$

where the off-diagonal elements  $w_{ij,N}$  with  $i \neq j$  are decreasing functions of the distance between the  $i$ 'th and  $j$ 'th panel group. If it is assumed that  $E(\varepsilon_t \varepsilon_t') = \sigma_\varepsilon^2 I_N$ , the

covariance matrix  $\Omega = \sigma_\epsilon^2 (I_N - \rho W_N)^{-1} (I_N - \rho W'_N)^{-1}$  involves only two parameters. This model is considered in more detail in Chapter 12.

A natural generalization of the model with common time effects is to allow for individual specific time effects of the form  $\gamma_i f_t$ , where for identification it is assumed that  $E(f_t^2) = 1$  and the regression error is specified as  $u_{it} = \gamma_i f_t + \varepsilon_{it}$ . This specification is referred to as the factor augmented panel data model or interactive effects model (cf. Bai 2009). If it is assumed that  $\varepsilon_{it}$  is independently distributed across  $i$  and  $t$  and  $E(\varepsilon_{it}^2) = \sigma_i^2$ , this specification implies a contemporaneous correlation of the form  $\gamma_i \gamma_j$  and the error covariance matrix results as  $\Omega = E(u_t u'_t) = \gamma \gamma' + D$  with  $D = \text{diag}(\sigma_1^2, \dots, \sigma_N^2)$  and  $\gamma = [\gamma_1, \dots, \gamma_N]'$ . The straightforward generalization to  $r$  factors yields

$$u_{it} = \sum_{j=1}^r \gamma_{i,j} f_{t,j} + \varepsilon_{it},$$

where  $f_{t,1}, \dots, f_{t,r}$  are mutually uncorrelated factors with unit variances. The idea is to capture the cross correlation among  $u_{it}$  by including a sufficient number of factors such that the cross-section dependence is sufficiently represented by the common component  $\chi_{it} = \sum_{j=1}^r \gamma_{i,j} f_{t,j}$ . Estimation of factor augmented panel data model is considered in Section 15.4.5.

Following Chudik, Pesaran, and Tosetti (2011), it is useful to distinguish weak and strong cross-section dependence in panel data (see Chapter 1). This concept can be related to the eigenvalues of the covariance matrix  $\Omega = E(u_t u'_t)$ . Note that for the largest eigenvalue  $\lambda_1(\Omega)$  it holds for the respective eigenvector

$$\lambda_1(\Omega) = \max_{v' v = 1} v' \Omega v.$$

Since  $w = N^{-1/2} v$  satisfies the granularity conditions of Chapter 1 we have

$$\lambda_1(\Omega) = \max_{v' v = 1} \text{var}(v' u_t) = N \max_{w' w = 1/N} \text{var}(w' u_t). \quad (15.28)$$

This implies that if  $\lambda_1/N \rightarrow 0$ , then  $u_t$  is characterized by weak cross-section dependence, whereas  $u_t$  is strongly dependent if the largest eigenvalue is  $O(N)$ . Another related criterion for weak cross-section dependence is employed Stock and Watson (2002) and Bai and Ng (2002). Let  $\tau_{ij} = \sup_t |E(u_{it} u_{jt})|$ . Then  $u_t$  is weakly cross-section dependent if

$$\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \tau_{ij} \leq M < \infty.$$

Note that the sum of  $N^2$  non-negative terms is divided by  $N$ . Thus, this concept allows, for example, that the units are correlated with other units in the neighborhood but any pervasive correlation affecting all elements is ruled out. Accordingly, the spatial lag model (15.27) with  $|\rho| < 1$  gives rise to weakly dependent errors, whereas the interactive effects (or factor) model implies that the errors are strongly dependent.

The distinction between weak and strong cross-section dependence is important for the asymptotic properties of the estimator. If the errors exhibit strong dependence the efficiency loss from using the OLS (within-group) estimator instead of the GLS estimator increases with  $N$ , whereas under weak dependence the efficiency loss is bounded as  $N \rightarrow \infty$ .

### 15.4.2 Tests for Cross-Section Dependence

Tests for correlation among a large number of random variables already have a long history (cf. Moscone and Tosetti 2009 and Chapter 1). The LR statistic is given by

$$\text{LR} = -T^* \left( \log |\widehat{\Omega}| - \sum_{i=1}^N \log (\widehat{\sigma}_i^2) \right),$$

where  $\widehat{\sigma}_i^2 = T^{-1} \sum_{t=1}^T \widehat{u}_{it}^2$  and  $\widehat{\Omega} = T^{-1} \sum_{t=1}^T \widehat{u}_t \widehat{u}_t'$ . For small  $T$  the small sample correction  $T^* = T - (2N+5)/6$  is employed (cf. Bartlett 1951). Under the null hypothesis of no cross-section correlation,  $T \rightarrow \infty$  and fixed  $N$ , the test statistic is  $\chi^2$  distributed with  $(N-1)N/2$  degrees of freedom. The LM (score) statistic is obtained as

$$\text{LM} = T \sum_{i=2}^N \sum_{j=1}^{i-1} \widehat{\rho}_{ij}^2, \quad (15.29)$$

where  $\widehat{\rho}_{ij}$  denotes the sample correlation coefficient between  $\widehat{u}_{it}$  and  $\widehat{u}_{jt}$ . This test statistic has the same limiting null distribution as the LR statistic.<sup>10</sup>

Pesaran (2004) demonstrates that the LM statistic suffers from severe size distortions whenever  $N$  is large relative to  $T$ . This is due to the fact that in small samples the distribution of  $T\widehat{\rho}_{ij}^2$  is poorly approximated by a  $\chi_1^2$  distribution and summing over all squared correlations cumulates the approximation error. On the other hand, the first two moments of  $T\widehat{\rho}_{ij}$  are well approximated by zero and one even if  $T$  is small. By invoking the central limit theorem Pesaran (2004, 2012) shows that the standardized sum of all possible cross-correlations

$$\lambda_{CD} = \sqrt{\frac{2T}{N(N-1)}} \sum_{i=2}^N \sum_{j=1}^{i-1} \widehat{\rho}_{ij}$$

possesses a standard normal limiting distribution.<sup>11</sup> It should be noted that this test may lack power if some of the correlations are positive and some others are negative so that the sum of the correlations sum up to zero. In empirical practice, however, it often makes sense to assume that the correlation between the panel units are (at least in most cases) positive, in particular, if correlations among the errors are well represented by a time effect.

To improve the small sample properties of the original LM test, Pesaran, Ullah, and Yamagata (2008) derived higher order approximations of the mean and variances of  $\widehat{\rho}_{ij}^2$ . The simulation results presented in Pesaran, Ullah, and Yamagata (2008) and Demetrescu and Homm (2014) suggest that the adjusted LM test performs well even if  $T$  is small relative to  $N$ .

Baltagi, Feng, and Kao (2011) adapt John's (1972) test for "sphericity" which is based on the statistic

$$U = \frac{N^{-1} \text{tr}(\widehat{\Omega}^2)}{[N^{-1} \text{tr}(\widehat{\Omega})]^2} - 1.$$

By noting that

$$N^{-1} \text{tr}(\widehat{\Omega}^2) - [N^{-1} \text{tr}(\widehat{\Omega})]^2 = 2 \sum_{i=2}^N \sum_{j=1}^{i-1} \widehat{\sigma}_{ij}^2$$

it turns out that the standardized version of statistic  $U$  is closely related to the LM statistic. The main difference is that the LM statistic is based on the square of all cross-correlations whereas John's statistic is based on the covariances. Baltagi, Feng, and Kao (2011) provide a correction which yields reliable size properties even for very small values of  $T$ . Baltagi, Feng, and Kao (2012) derive the limiting distribution of the original LM statistic when  $N/T \rightarrow c$  with  $0 < c < \infty$  and suggest a bias corrected and standardized test statistic

$$\text{LM}^* = \sqrt{\frac{1}{N(N-1)}} \sum_{i=2}^N \sum_{j=1}^{i-1} (T\widehat{\rho}_{ij}^2 - 1) - \frac{N}{2(T-1)}.$$

Under the null hypothesis of uncorrelated errors, this test statistic has a standard normal limit distribution.

Demetrescu and Homm (2014) propose a "directed test" which is equivalent to an information matrix test for the null hypothesis that the covariance matrix of the estimated regression coefficients is correctly specified. Accordingly, this test focus on cross-section dependence that result in a bias of the within-group standard errors for the regression coefficients. If the test statistic rejects, then the usual standard errors should be replaced by HAC standard errors (PCSE) or a GLS estimation procedure should be employed.

Ng (2006) proposes a test constructed from the ordered sequence  $\phi_{[1]} \leq \dots \leq \phi_{[n]}$  of all possible  $n = N(N-1)/2$  transformed correlation coefficients  $0.5 \leq \Phi(|\widehat{\rho}_{ij}|) \leq 1$ , where  $\Phi(\cdot)$  denotes the c.d.f. of the standard normal distribution. The " $p$ -spacings" are defined as  $\delta_j^p = \phi_{[j+p]} - \phi_{[j]}$ . Under the null hypothesis the spacings  $\delta_j^p$  form an exchangeable set of beta-distributed random variables. Ng (2006) proposed a variance ratio statistic, where the variance of the  $p$ -spacings are compared to the  $p$  times the variance of the 1-spacings. If the variance ratio statistic exceeds the critical value obtained from the asymptotic normal distribution, then the hypothesis of independent panel units is rejected. The spacing can also be used to partition the sample correlation coefficients into subsets of weak and strong correlations.

Sarafidis, Yamagata, and Robertson (2009) propose a test for cross-section dependence in dynamic panel data model. A Sargan's difference test is constructed comparing two sets of moment conditions. The first moment conditions (using the exogenous variables as instruments) is valid no matter of a possible cross-section dependence of the errors. The second set of moment conditions (involving instruments derived from the lagged dependent variable) is only valid under the null hypothesis of no cross-section dependence. If the difference between the Sargan-Hansen statistics for both GMM estimators is too large (with respect to the appropriate  $\chi^2$  distribution), the null hypothesis of no cross-section dependence is rejected.

### 15.4.3 Panel Corrected Standard Errors

Residual cross-correlation gives rise to inefficient estimation and invalid inference (e.g., biased  $t$ -statistics). As the number of observations in a panel data set is typically large, inefficiency of the estimator is not a major concern. We therefore focus on the problem of valid inference. Let us consider the static fixed-effect panel data regressions with serially uncorrelated but cross-sectionally dependent errors. We adapt the period-wise notation

$$\tilde{y}_t = \tilde{X}_t \beta + \tilde{u}_t,$$

where  $\tilde{X}_t$  is defined as in Section 15.1. We further assume that  $E(u_t u_t') = \Omega$  and  $E(u_t u_s') = 0$  for  $t \neq s$  and  $X_t$  is strictly exogenous with respect to  $u_t$ . The covariance matrix of the within-group estimator is given by

$$\text{var}(\hat{\beta}) = \left( \sum_{t=1}^T \tilde{X}_t' \tilde{X}_t \right)^{-1} \left( \sum_{t=1}^T \tilde{X}_t' \Omega \tilde{X}_t \right) \left( \sum_{t=1}^T \tilde{X}_t' \tilde{X}_t \right)^{-1}.$$

It is not difficult to see that the usual covariance matrix estimator  $\hat{\sigma}^2 (\sum_{t=1}^T \tilde{X}_t' \tilde{X}_t)^{-1}$  is biased if the elements  $\sigma_{ij}$  of the covariance matrix  $\Omega$  systematically vary with the elements of the matrix  $S_i = \sum_{t=1}^T (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)'$ .

This problem can easily be fixed by employing "panel corrected standard errors" (PCSE) (cf. Arellano 1987 and Beck and Katz 1995) given by

$$\begin{aligned} \hat{V}_1 &= \left( \sum_{i=1}^N S_i \right)^{-1} \left( \sum_{t=1}^N \tilde{X}_t' \hat{u}_t \hat{u}_t' \tilde{X}_t \right) \left( \sum_{i=1}^N S_i \right)^{-1} \\ \text{or } \hat{V}_2 &= \left( \sum_{i=1}^N S_i \right)^{-1} \left( \sum_{t=1}^T \tilde{X}_t' \hat{\Omega} \tilde{X}_t \right) \left( \sum_{i=1}^N S_i \right)^{-1}, \end{aligned}$$

where  $\hat{u}_t = y_t - X_t \hat{\beta}$  and  $\hat{\Omega} = T^{-1} \sum_{t=1}^T \hat{u}_t \hat{u}_t'$ . Note that the estimator  $\hat{V}_2$  can only be computed for balanced panels. To analyze the asymptotic properties, consider the case of a single regressor ( $k = 1$ ). We have

$$\begin{aligned} \frac{1}{NT} \sum_{t=1}^T \tilde{X}_t' \hat{u}_t \hat{u}_t' \tilde{X}_t &= \frac{1}{T} \sum_{t=1}^T \left[ \frac{1}{\sqrt{N}} \sum_{i=1}^N \tilde{x}_{it} u_{it} - \left( \frac{1}{N} \sum_{i=1}^N x_{it}^2 \right) \sqrt{N} (\hat{\beta} - \beta) \right]^2 \\ &= \frac{1}{T} \sum_{t=1}^T \left( \frac{1}{\sqrt{N}} \sum_{i=1}^N \tilde{x}_{it} u_{it} \right)^2 + O_p(T^{-1}) \end{aligned}$$

where we use  $\hat{\beta} - \beta = O_p(N^{-1/2} T^{-1/2})$ . If the errors and regressors are weakly dependent we obtain (under some regularity conditions for the appropriate law of large numbers)

$$\frac{1}{T} \sum_{t=1}^T \left( \frac{1}{\sqrt{N}} \sum_{i=1}^N \tilde{x}_{it} u_{it} \right)^2 \xrightarrow{p} \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E \left( \frac{1}{N} \tilde{X}_t' \Omega \tilde{X}_t \right).$$

Accordingly, the estimator is consistent for  $T \rightarrow \infty$  and fixed  $N$  (as well as  $N \rightarrow \infty$ ).

It is interesting to analyze the asymptotic properties of the PCSE approach if it is assumed that the errors are strongly dependent. For concreteness consider the single factor model  $u_{it} = \gamma_i f_t + \varepsilon_{it}$ , where  $f_t \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$  and  $\varepsilon_{it} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ . It is not difficult to see that if  $\tilde{X}_t' \gamma = O_p(N^{1/2})$  with  $\gamma = [\gamma_1, \dots, \gamma_N]'$  (i.e., the regressor is weakly dependent according to the granularity conditions), the asymptotic properties are similar as in the case of weak cross-section dependence. If, however, the regressor is also strongly dependent such that  $\tilde{X}_t' \gamma = O_p(N)$ , then the asymptotic properties of the within-group estimator are different. In this case we obtain

$$\begin{aligned} \frac{1}{N^2 T} \sum_{t=1}^T E \left[ \tilde{X}_t' (\gamma \gamma' + \sigma^2 I_N) \tilde{X}_t \right] &= \frac{1}{T} \sum_{t=1}^T E \left[ (N^{-1} \tilde{X}_t' \gamma) (N^{-1} \gamma' \tilde{X}_t) \right] \\ &\quad + \frac{\sigma^2}{NT} \sum_{t=1}^T E \left( \frac{1}{N} \tilde{X}_t' \tilde{X}_t \right). \end{aligned}$$

Note that  $\tilde{X}_t' \tilde{X}_t = \sum_{i=1}^N \tilde{x}_{it}^2$  is  $O_p(N)$  and, therefore, the last term disappears as  $N \rightarrow \infty$ . It follows that the rank of the asymptotic covariance matrix is equal to the minimum of the number of common factors and the number of regressors and, thus the asymptotic covariance matrix, may be singular. Moreover, the estimation error is  $\hat{\beta} - \beta = O_p(1/\sqrt{T})$  instead of  $\hat{\beta} - \beta = O_p(1/\sqrt{NT})$  in the case of weak cross-section dependence. Accordingly,

$$\begin{aligned} \frac{1}{N^2 T} \sum_{t=1}^T \tilde{X}_t' \hat{u}_t \hat{u}_t' \tilde{X}_t &= \frac{1}{T} \sum_{t=1}^T \left[ \frac{1}{N} \sum_{i=1}^N \tilde{x}_{it} u_{it} - \left( \frac{1}{N} \sum_{i=1}^N x_{it}^2 \right) (\hat{\beta} - \beta) \right]^2 \\ &= \frac{1}{T} \sum_{t=1}^T \left( \frac{1}{N} \sum_{i=1}^N \tilde{x}_{it} u_{it} \right)^2 + O_p(T^{-1}) \end{aligned}$$

and, therefore, under the conditions of the weak law of large numbers for independent random variables the PCSE estimator remains consistent if  $T \rightarrow \infty$ .

For all these results we assumed that  $u_{it}$  is independent, across time periods. To cope with serial and contemporaneous correlation, Driscoll and Kraay (1998) propose a generalized PCSE estimator. Rewrite

$$E\left[\frac{1}{NT} \left( \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i) u_{it} \right) \left( \sum_{i=1}^N \sum_{t=1}^T u_{it} (x_{it} - \bar{x}_i)' \right) \right] = E\left(\frac{1}{T} \sum_{i=1}^N q_t q_t' \right),$$

where  $q_t = N^{-1/2} \sum_{i=1}^N (x_{it} - \bar{x}_i) u_{it}$ . If  $u_{it}$  is weakly dependent, then  $q_t$  converges to a well-behaved random variable as  $N \rightarrow \infty$ . The serial correlation of  $q_t$  can be accounted for by using the usual estimators for the long-run variance (e.g., Newey and West 1987). Alternative robust covariance matrix estimators were proposed by Bester, Conley, and Hansen (2011) and Vogelsang (2012).

An alternative approach was suggested by Dufour and Khalaf (2002). The idea is to employ test statistics that are invariant to the covariance matrix of the errors. Let  $\widehat{\Omega}_0$  denote the estimated covariance matrix under the null hypothesis. Then the likelihood ratio statistic

$$\begin{aligned} \text{LR} &= T \log(|\widehat{\Omega}_0|/|\widehat{\Omega}|) \\ &= T \log(|\widehat{\Psi}_0|/|\widehat{\Psi}|) \end{aligned}$$

is invariant to the covariance matrix  $\Omega$ , where  $\widehat{\Psi}_0 = \Omega^{-1/2} \widehat{\Omega}_0 \Omega^{-1/2}$  and  $\widehat{\Psi} = \Omega^{-1/2} \widehat{\Omega} \Omega^{-1/2}$ . Dufour and Khalaf (2002) consider various test statistics that are (as the LR statistic) invariant with respect to the covariance matrix and are functions of the eigenvalues of the generalized eigenvalue problem  $|\widehat{\Omega} - \lambda \widehat{\Omega}_0|$  including the Wilks and the Lawley-Hotelling statistic. If  $T/N^2$  is small, however, the asymptotic test criteria are seriously biased towards overrejection. Dufour and Khalaf (2002) propose a general method for constructing exact tests based on Monte Carlo simulations.

#### 15.4.4 The SUR-GLS Estimator

It is well known that the within-group estimator may involve a dramatic loss of efficiency relative to the GLS (or ML) estimator for seemingly unrelated regressions (SUR) given by<sup>12</sup>

$$\tilde{\beta} = \left( \sum_{t=1}^T \tilde{X}'_t \Omega^{-1} \tilde{X}_t \right)^{-1} \sum_{t=1}^T \tilde{X}'_t \Omega^{-1} \tilde{y}_t, \quad (15.30)$$

where the notation is explained at the end of the introduction. In practice, the unknown covariance matrix  $\Omega$  is replaced by the within-group estimator  $\widehat{\Omega} = T^{-1} \sum_{t=1}^T \widehat{u}_t \widehat{u}_t'$  with  $\widehat{u}_t = \tilde{y}_t - \tilde{X}_t \widehat{\beta}$ .

To analyze the efficiency of the SUR-GLS estimator relative to the OLS (within-group) estimator, assume that  $\tilde{x}_{it}$  is *i.i.d.* across  $i$  and  $t$ . The covariance matrix of the errors possesses the spectral decomposition  $\Omega = V\Lambda V'$ , where  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_N)$  is a diagonal matrix of the  $N$  eigenvalues (arranged in a descending order) and  $V$  is the associate matrix of orthonormal eigenvectors. In this example the asymptotic relative efficiency results as

$$\lim_{T \rightarrow \infty} \frac{\text{var}(\hat{\beta})}{\text{var}(\tilde{\beta})} = \left( \frac{1}{N} \sum_{i=1}^N \frac{1}{\lambda_i} \right) \left( \frac{1}{N} \sum_{i=1}^N \lambda_i \right) \geq 1$$

by Jensen's inequality. A second order Taylor expansion around  $\bar{\lambda} = N^{-1} \sum_{i=1}^N \lambda_i$  yields

$$\lim_{T \rightarrow \infty} \frac{\text{var}(\hat{\beta})}{\text{var}(\tilde{\beta})} \approx 1 + \frac{1}{4\bar{\lambda}^2} \left[ \frac{1}{N} \sum_{i=1}^N (\lambda_i - \bar{\lambda})^2 \right]$$

which shows that the relative efficiency depends on the dispersion of eigenvalues.

In practice there are two practical problems with the GLS estimator. First, the GLS estimator requires  $T \geq N$  as otherwise the matrix  $\widehat{\Omega}$  is not invertible. Moreover, for statistical inference the number of time periods should be much larger than  $N$ . This is due to the fact that the first order asymptotic theory ignores the estimation error in the estimated covariance matrix  $\widehat{\Omega}$ . Specifically, hypotheses on the coefficients are based on the estimated covariance matrix  $\widehat{\text{var}}(\tilde{\beta}) = (\sum_{t=1}^T \tilde{X}'_t \widehat{\Omega}^{-1} \tilde{X}_t)^{-1}$ . We decompose the estimated covariance matrix as

$$\widehat{\Omega} = \Omega + \left( T^{-1} \sum_{t=1}^T u_t u'_t - \Omega \right) + O_p(N^{-1/2} T^{-1}),$$

where the  $O_p(N^{-1/2} T^{-1})$  term represents the effect of the estimation error  $\hat{\beta} - \beta$ . The inverse can be represented as

$$\widehat{\Omega}^{-1} = \Omega^{-1} + T^{-1/2} A_{NT} + N^{-1/2} T^{-1} B_{NT},$$

where all elements of the  $N \times N$  matrices  $A_{NT}$  and  $B_{NT}$  are stochastically bounded. It follows that

$$\frac{1}{N} \tilde{X}'_t \widehat{\Omega}^{-1} \tilde{X}_t = \frac{1}{N} \tilde{X}'_t \Omega^{-1} \tilde{X}_t + T^{-1/2} \frac{1}{N} \tilde{X}'_t A_{NT} \tilde{X}_t + N^{-1/2} T^{-1} \frac{1}{N} \tilde{X}'_t B_{NT} \tilde{X}_t$$

and since  $\tilde{X}'_t A_{NT} \tilde{X}_t = \sum_{i=1}^N \sum_{j=1}^N a_{NT,ij} \tilde{x}_{it} \tilde{x}'_{jt} = O_p(N^2)$  for  $E(\tilde{x}_{it} \tilde{x}'_{jt}) \neq 0$  we obtain

$$\frac{1}{NT} \sum_{t=1}^T \tilde{X}'_t \widehat{\Omega}^{-1} \tilde{X}_t = \frac{1}{NT} \sum_{t=1}^T \tilde{X}'_t \Omega^{-1} \tilde{X}_t + O_p\left(\frac{N}{\sqrt{T}}\right).$$

Therefore, for reliable small sample inference  $T$  must be large relative to  $N$ . In contrast, the PCSE estimator is reliable for large  $T$  no matter of the magnitude of  $N$ .

A second problem with the SUR-GLS estimator is that the usual estimator of the covariance matrix  $\widehat{\Omega}$  requires a balanced panel. To sidestep this problem the covariance matrix may be estimated from the “pseudo-balanced” panel where time periods involving missing observations are discarded. Obviously, this approach may lead to a severe loss of information as a full vector of  $N$  observations is ignored if only one observation is missing. To exploit all available information, an ML estimator under missing observations may be used. Denote the vector of available observations by the  $N_t \times 1$  vector  $\tilde{u}_t = S_t u_t$ , where  $S_t$  is a  $N_t \times N$  selection matrix with  $N_t \leq N$ . The Gaussian log-likelihood function is given by

$$\mathcal{L}(\Omega) = \text{const} - \frac{1}{2} \sum_{t=1}^T \log |S_t \Omega S_t' - \frac{1}{2} \sum_{t=1}^T \tilde{u}_t' (S_t \Omega S_t')^{-1} \tilde{u}_t.$$

By inserting  $\tilde{u}_t = S_t \tilde{y}_t - S_t \tilde{X}_t \tilde{\beta}$ , this log-likelihood function can be maximized to obtain the ML estimator of  $\Omega$ .

### 15.4.5 Common Factors

If  $N$  is large relative to  $T$ , the SUR-GLS approach based on an unrestricted error covariance matrix exhibit poor small sample properties. Furthermore, a substantial gain in efficiency is possible when accounting for strong dependence. It is therefore desirable to impose some specific structure on the covariance matrix  $\Omega$ . A flexible and convenient approach is to assume a factor structure given by

$$u_{it} = \sum_{j=1}^r \gamma_{i,j} f_{t,j} + \varepsilon_{it} \quad (15.31)$$

$$= \gamma_i' f_t + \varepsilon_{it}, \quad (15.32)$$

where  $\gamma_i = [\gamma_{i,1}, \dots, \gamma_{i,r}]'$  and  $f_t = [f_{t,1}, \dots, f_{t,r}]'$ . Assuming  $E(f_t f_t') = I_r$ ,  $E(\varepsilon_{it}^2) = \sigma_i^2$ ,  $E(\varepsilon_{it} \varepsilon_{jt}) = 0$ , and  $E(\varepsilon_{it} f_s) = 0$  the covariance matrix results as

$$\Omega = \Gamma \Gamma' + \Sigma_\varepsilon, \quad (15.33)$$

where  $\Gamma = [\gamma_1, \dots, \gamma_N]'$  and  $\Sigma_\varepsilon = \text{diag}(\sigma_1^2, \dots, \sigma_N^2)$ . Note that the covariance matrix involve only  $(r+1)N$  instead of  $N(N+1)/2$  parameters implied by an unrestricted covariance matrix.

Let us first consider a single factor model with  $r = 1$  and assume that  $u_{it} = y_{it} - \mu_i - x_{it}' \beta$  is observed such that  $u_{it} = \lambda_i f_t + \varepsilon_{it}$ . Taking cross-section means yields

$$\bar{u}_t = \bar{\lambda} f_t + \bar{\varepsilon}_i \quad (15.34)$$

and, therefore, by assuming  $\bar{\lambda} \neq 0$  we obtain a simple estimator for the common factor:

$$\frac{\bar{u}_t}{\bar{\lambda}} = a_0 + b_0 \bar{y}_t + c'_0 \bar{x}_t = f_t + O_p(T^{-1/2}),$$

where  $a_0 = -(\bar{\lambda}N)^{-1} \sum_{i=1}^N \mu_i$ ,  $b_0 = 1/\bar{\lambda}$ ,  $c_0 = -\beta/\bar{\lambda}$ . Replacing the unknown factor by its estimate yields

$$y_{it} = \mu_i^* + \beta' x_{it} + b_i \bar{y}_t + c'_i \bar{x}_t + e_{it}, \quad (15.35)$$

where  $\mu_i^* = \mu_i + \lambda_i a_0$ ,  $b_i = \lambda_i b_0$ ,  $c_i = \lambda_i c_0$  and  $e_{it} = \varepsilon_{it} + (\lambda_i/\bar{\lambda}) \bar{\varepsilon}_t$ . Pesaran (2006) suggests to estimate the augmented panel model by a within or mean-group estimator. He shows that the estimator is consistent and asymptotically normally distributed as  $N \rightarrow \infty$ ,  $T \rightarrow \infty$ , and  $T/N \rightarrow 0$ . Note that if  $r = 1$  we may improve the efficiency of the estimator by imposing the parameter restrictions and estimate the nonlinear model

$$y_{it} = \mu_i^* + \beta' x_{it} + \lambda_i^* (\bar{y}_t - \beta' \bar{x}_t) + e_{it}, \quad (15.36)$$

where  $\lambda_i^* = \lambda_i/\bar{\lambda}$ .

So far we have assumed that  $\bar{\lambda} \neq 0$ . If  $\bar{\lambda} = 0$ , then  $\bar{u}_t = \bar{\varepsilon}_t$  in (15.34) and, therefore, the least-squares estimators of  $c_i$  tend to zero in probability. Thus, the augmentation does not cancel out the factor. It is nevertheless possible to obtain a consistent estimator for  $\beta$  by assuming that  $\lambda_i$  is independent of  $\bar{x}_i$ . However inference is biased as the errors are cross-correlated due to the remaining factor. Furthermore, as demonstrated by Westerlund and Urbain (2013a), the CCE estimator can be severely biased if the rank condition is violated (which boils down to  $\bar{\lambda} = 0$  in the single factor model) and  $\lambda_i$  is correlated with  $\bar{x}_i$  (that is, if the loadings of the factors in the data generation process of  $x_{it}$  and  $\lambda_i$  are correlated).

Pesaran (2006) also shows that employing the unrestricted regression (15.35) may account for up to  $r \leq k + 1$  factors whenever there exist a nonsingular matrix  $\Xi$  with  $rk(\Xi) = r$  such that  $f_t = \Xi \underset{N \rightarrow \infty}{\text{plim}} \bar{z}_t$ , where  $\bar{z}_t = [\bar{y}_t, \bar{x}_{t,1}, \dots, \bar{x}_{t,k}]'$ . See Chapter 1 for more details. Bai (2009) and Moon and Weidner (2013) propose a Principal Component (PC) analysis to account for common factors. This estimator minimizes the total sum of squares

$$S(\beta, \Lambda, F) = \sum_{t=1}^T (\tilde{y}_t - \tilde{X}_t \beta - \Lambda \tilde{f}_t)' (\tilde{y}_t - \tilde{X}_t \beta - \Lambda \tilde{f}_t)$$

subject to  $T^{-1} \sum_{i=1}^T \tilde{f}_i \tilde{f}_i' = T^{-1} \tilde{F}' \tilde{F} = I_r$ , where  $\tilde{f}_t = f_t - \bar{f}$ ,  $\Lambda = [\lambda_1, \dots, \lambda_N]'$ , and  $\Lambda' \Lambda$  is a diagonal matrix.<sup>13</sup> The minimization problem can be solved by a sequential estimation procedure that cyclically switches between the estimation of  $\beta$  given  $[F', \Lambda']$  (which is a simple linear least-squares problem) and the estimation of  $[F', \Lambda']$  given  $\beta$  (performed by an ordinary PC analysis), until convergence. If the idiosyncratic errors are heteroskedastic and autocorrelated, the efficiency of this estimation procedure may be improved by applying a GLS version as proposed by Breitung and

Tenhofen (2011). As shown by Bai (2009) statistical inference can be performed by estimating the transformed model

$$\widehat{Y} = \widehat{X}_{(1)}\beta_1 + \cdots + \widehat{X}_{(k)}\beta_k + \widetilde{\epsilon}, \quad (15.37)$$

where  $\widehat{Y} = M_{\widehat{F}}\widetilde{Y}M_{\widehat{\Lambda}}$ ,  $M_{\widehat{F}} = I_T - \widehat{F}(\widehat{F}'\widehat{F})^{-1}\widehat{F}'$ ,  $M_{\widehat{\Lambda}} = I_N - \widehat{\Lambda}(\widehat{\Lambda}'\widehat{\Lambda})^{-1}\widehat{\Lambda}'$  and  $\widetilde{Y} = (y_{it} - \bar{y}_i)$  arranges all observations in a  $T \times N$  matrix. The matrix  $\widehat{X}_{(j)} = M_{\widehat{F}}\widetilde{X}_{(j)}M_{\widehat{\Lambda}}$  is similarly defined as  $\widetilde{Y}$  using a regressor specific  $T \times N$  matrix of observations  $\widetilde{X}_{(j)} = (x_{it,j} - \bar{x}_{i,j})$  ( $j = 1, \dots, k$ ). Note that it is not sufficient to cancel out the factors by just including  $\widehat{f}_t$  as an additional regressor. Since a Taylor series expansion yields  $\widehat{\lambda}'_i \widehat{f}_t \simeq \lambda'_i f_t + \lambda'_i (\widehat{f}_t - f_t) + f'_t (\widehat{\lambda}_i - \lambda_i)$ , the estimated factors and the factor loadings need to be augmented for valid asymptotic inference. Hence, valid  $t$ -statistics for  $\beta$  are obtained from the augmented regression

$$y_{it} - \bar{y}_i = \beta'(x_{it} - \bar{x}_i) + \gamma'_i \widehat{f}_t + \delta'_i \widehat{\lambda}_i + e_{it}, \quad (15.38)$$

which is equivalent to the representation (15.37). If  $N/T \rightarrow 0$  or  $T/N \rightarrow \infty$  and  $\varepsilon_{it}$  is *i.i.d.* then standard inference based on  $t$  or  $F$  statistics is asymptotically valid. If the idiosyncratic errors are heteroskedastic the estimator possess an asymptotic bias as  $T/N \rightarrow c$  with  $0 < c < \infty$  (cf. Bai 2009). As shown by Greenaway-McGrevy, Han, and Sul (2012) this bias disappears as  $N/T^3 \rightarrow \infty$ , since under this condition the factor augmented within-group estimator is asymptotically equivalent to the infeasible estimator based on known factors.

Westerlund and Urbain (2013b) compare the CCE and PC approaches for estimating panel data models with common factors. They found that both approaches suffer from an asymptotic bias whenever  $N/T \rightarrow c$  which disappears if  $c = 0$ . In fairly small samples the CCE estimator tends to outperform the PC estimator in terms of bias and size distortions of the corresponding  $t$ -statistics. Pesaran and Kapetanios (2007) draw similar conclusions based on a particular Monte Carlo design and recommend the use of the CCE estimator.

Ahn, Lee, and Schmidt (2013) propose a GMM estimation procedure that is based on the assumption that  $T$  is fixed and  $N \rightarrow \infty$ . Consider the single factor model where  $f_t$  and  $\lambda_i$  are scalars. We normalize the factors the factor space as  $g_t = f_t/f_1$  such that  $f_1 = 1$  and  $\gamma_i^* = \gamma_i f_1$ . In matrix notation the factor model is written as

$$y_i = X_i \beta + \begin{bmatrix} 1 \\ g \end{bmatrix} \gamma_i^* + \varepsilon_{it},$$

where  $g = [g_2, \dots, g_T]$ . Let  $H(g) = [g, -I_{T-1}]$  denote the orthogonal complement of the vector  $[1, g']$  such that  $H(g)[1, g']' = 0$ . If  $x_{it}$  is strictly exogenous the following  $(T-1)Tk$  dimensional (nonlinear) moment conditions hold

$$m_i(g, \beta) = E([H(g)y_i - H(g)X_i\beta] \otimes \text{vec}(X_i)) = 0.$$

A necessary condition for the identification of the parameters is that the number of moment conditions is larger than the number of parameters  $(T-1+k)$ . Ahn, Lee,

and Schmidt (2013) propose sequential tests and information criteria to determine the number of factors.

### 15.4.6 The Global VAR

An important drawback of panel data models with cross-correlated error term is that SUR type models rule out dynamic spillovers among different countries or regions. Pesaran, Schuermann, and Weiner (2004) propose a more general framework, called the Global VAR (GVAR), that combines individual specific vector error-correcting models, in which the dynamic interdependence between panel units is represented by country-specific foreign variables computed as a weighted average of all other country variables. The GVAR framework can be motivated in two different ways. Dees et al. (2007) derive the GVAR approach as an approximation to a global common factor model, whereas Chudik and Pesaran (2011) obtain the GVAR approach as an approximation to a high-dimensional VAR system.

Consider the country-specific models VARX(2,2) model

$$y_{it} = \Phi_{i1}y_{i,t-1} + \Phi_{i2}y_{i,t-2} + \Lambda_{i0}x_{it}^* + \Lambda_{i1}x_{i,t-1}^* + \Lambda_{i2}x_{i,t-2}^* + u_{it},$$

where for notational simplicity we assume that  $y_{it}$  and  $x_{it}^*$  are  $m \times 1$  vectors. Furthermore, we suppress deterministic terms like constants, trends or dummy variables. The vector of foreign variables  $x_{it}^*$  is defined as

$$x_{it}^* = \sum_{j=1}^N w_{ij}y_{jt},$$

where  $\{w_{ij}\}$  are a suitable set of weights (e.g., trade shares) with  $\sum_{j=1}^N w_{ij} = 1$  and  $w_{ii} = 0$ . The model can be written as a vector error correction model (VECM):

$$\Delta y_{it} = -\alpha_i \beta_i' z_{i,t-1} + \Lambda_{i0} \Delta x_{it}^* + \Gamma_i \Delta z_{i,t-1} + u_{it}, \quad (15.39)$$

where  $z_{it} = (y_{it}', x_{it}'')'$ . In this VECM  $x_{it}^*$  is treated as weakly exogenous with respect to the parameters of the model. All variables of the system are comprised in the vector  $Y_t = (y_{1t}', \dots, y_{Nt}')'$  such that

$$z_{it} = \begin{pmatrix} y_{it} \\ x_{it}^* \end{pmatrix} = W_i Y_t,$$

where the  $2m \times Nm$  matrix  $W_i$  is a known matrix of the corresponding country weights  $w_{ij}$ . The VAR representation of the VECM system results in

$$A_{i0} z_{it} = A_{i1} z_{i,t-1} + A_{i2} z_{i,t-2} + u_{it},$$

where  $A_{i0} = (I, -\Lambda_{i0})$ ,  $A_{i1} = [\Phi_{i1}, \Lambda_{i1}]$ , and  $A_{i2} = [\Phi_{i2}, \Lambda_{i2}]$ . The system representation (including the reference country  $i = 0$ ) is given by

$$A_{i0} W_i Y_t = A_{i1} W_i Y_{t-1} + A_{i2} W_i Y_{t-2} + u_{it} \quad \text{for } i = 0, 1, \dots, N$$

$$G_0 Y_t = G_1 Y_{t-1} + G_2 Y_{t-2} + U_t$$

where

$$G_0 = \begin{pmatrix} A_{00} W_0 \\ A_{10} W_1 \\ \vdots \\ A_{N0} W_N \end{pmatrix} \quad G_1 = \begin{pmatrix} A_{01} W_0 \\ A_{11} W_1 \\ \vdots \\ A_{N1} W_N \end{pmatrix} \quad G_2 = \begin{pmatrix} A_{02} W_0 \\ A_{12} W_1 \\ \vdots \\ A_{N2} W_N \end{pmatrix}$$

This gives rise to the global VAR representation

$$Y_t = G_0^{-1} G_1 Y_{t-1} + G_0^{-1} G_2 Y_{t-2} + G_0^{-1} U_t \quad (15.40)$$

$$= F_1 Y_{t-1} + F_2 Y_{t-2} + U_t^*. \quad (15.41)$$

Note that although  $F_1, F_2$  are  $Nm \times Nm$  matrices, most of the elements are known (for example 0, 1, or  $w_{ij}$ ). Each of the country-specific VECM models (15.39) are estimated by (partial) maximum likelihood. The specification is selected and evaluated by applying information criteria and a wide range of specification tests. The global VAR system (15.41) is obtained by merging the  $N$  country-specific VAR models into the system representation. To study the dynamic interaction among the countries impulse response functions are computed from the MA representation of the GVAR, where Dees et al. (2007) propose to identify the shocks by (country-wise) generalized impulse response functions (Pesaran and Shin 1998). Eickmeier and Ng (2011) apply sign restrictions for identifying country-specific shocks. A variety of recent applications of the GVAR methodology can be found in diMauro and Pesaran (2013).<sup>14</sup>

## 15.5 CONCLUDING REMARKS

This chapter reviews a wide range of recently developed econometric tools for analyzing macroeconomic panel data. Notwithstanding the important achievements so far, the analysis of macroeconomic panels is still in its infancy. Modeling observable and unobservable interactions across sectors, regions, and economies is important for valid statistical inference. For example, business cycles across countries exhibit pronounced patterns of co-movement for variables like output, inflation, interest rates, and real equity prices. Incorporating such dynamic spill-over effects across countries remains a grand challenge for future research. The GVAR approach seems to be a

promising direction of future research as this modeling framework accommodates typical macroeconomic data features like nonstationarity, parameter heterogeneity, and cross-section dependence.

---

## NOTES

---

1. Kiviet's (1995) bias-corrected estimator is implemented in the external STATA procedure XTLSDVC developed by Bruno (2005).
2. Consider, for example,  $\beta_0 = 1$  and  $\beta_1 = 2$ . If  $\rho < -0.5$  it follows that  $\theta < 0$  although the coefficients in the polynomial  $\gamma(L) = \beta(L)/\alpha(L)$  are all positive. (See the lower panel of Table 15.1.)
3. Hall and Peixe (2003) consider a canonical correlation analysis of the instruments which seems to be more promising at least if the number of instruments is substantially smaller than the number of panel units  $N$ .
4. Bun and Carree (2005) include additional exogenous variables that we ignore for the ease of exposition. Bun and Carree (2006) and Juodis (2013a) allow for some forms of heteroskedasticity.
5. A similar estimator was first employed by MacCurdy (1982) to estimate a univariate ARMA process with panel data.
6. Other initial conditions are discussed in Breitung (1994) and Hsiao, Pesaran, and Tahmisioglu (2002).
7. For the ease of exposition we ignore the time effects and temporal heteroskedasticity in the model of Bai (2013).
8. Holtz-Eakin, Newey, and Rosen (1988) consider a more general model with time dependent individual effects.
9. Pesaran and Smith (1995) also study the between-group regression and the time series regression using aggregate data. Since these estimators are less efficient than the pooled or mean-group estimators, we do not consider these alternatives here.
10. Moscone and Tosetti (2009) consider a variant of the test based on the absolute values  $|\rho_{ij}|$  whereas Frees (1995) employ Spearman rank correlations leading to improved small sample properties of the test. Halunga, Orme, and Yamagata (2011) suggest a version of the test statistic that is robust against temporal heteroskedasticity.
11. The test statistic is available from the external STATA routine XTCSD.
12. A generalization for autocorrelated errors is proposed by Parks (1967).
13. These restrictions are imposed just for computational convenience. Note that  $\tau^2$  normalization restrictions are required to obtain a unique factor representation.
14. A Matlab Toolbox called GVAR is provided by Vanessa Smith (Center for Financial Analysis & Policy, Cambridge).

---

## REFERENCES

---

- Ahn, S.C. and P. Schmidt (1997), Efficient estimation of dynamic panel data models: Alternative assumptions and simplified estimation, *Journal of Econometrics*, 76, 309–321.

- Ahn, S.C., Y.H. Lee, and P. Schmidt (2013), Panel data models with multiple time-varying individual effects, *Journal of Econometrics*, 174, 1–14.
- Alvarez, J. and M. Arellano (2003), The Time Series and cross-section asymptotics of dynamic panel data estimators. *Econometrica*, 71, 1121–1159.
- Alvarez, J. and M. Arellano (2004), Robust likelihood estimation of dynamic panel data models, mimeo.
- Arellano, M. (1987), Computing robust standard errors for within-group estimators, *Oxford Bulletin of Economics and Statistics*, 49, 431–434.
- Arellano, M. (1993), On the testing of correlated effects with panel data, *Journal of Econometrics*, 59, 87–97.
- Arellano, M. (2003), *Panel data econometrics*, Oxford: Oxford University Press.
- Arellano, M. and S. Bond (1991), Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations, *Review of Economic Studies*, 58, 277–297.
- Arellano, M. and Bover (1995), Another look at the instrumental variable estimation of error-component models, *Journal of Econometrics*, 68, 29–51.
- Bai, J. (2009), Panel data models with interactive fixed effects, *Econometrica*, 77, 1229–1279.
- Bai, J. (2013), Fixed-effects dynamic panel models, a factor analytical method, *Econometrica*, 81, 285–314.
- Bai, J. and S. Ng (2002), determining the number of factors in approximate factor models, *Econometrica*, 70, 191–221.
- Bai, J. and S. Ng (2009), Selecting instrumental variables in a data rich environment, *Journal of Time Series Econometrics*, 1, 1–32.
- Bai, J. and S. Ng (2010), Instrumental variable estimation in a data rich environment, *Econometric Theory*, 26, 1577–1606.
- Baltagi, B.H. (2008), *Econometric analysis of panel data*, 4th ed., New York: John Wiley.
- Baltagi, B.H. and J.H. Griffin (1984), Short and long run effects in pooled models, *International Economic Review*, 25, 631–645.
- Baltagi, B.H., Q. Feng and C. Kao (2011), Testing for sphericity in a fixed effects panel data model, *Econometrics Journal*, 14, 25–47.
- Baltagi, B.H., Q. Feng, and C. Kao (2012), A Lagrange multiplier test for cross-sectional dependence in a fixed effects panel data model, *Journal of Econometrics*, 170, 164–177.
- Bartlett, M.S. (1951), The effect of standardization on a chi-square approximation in factor analysis, *Biometrika*, 38, 337–344.
- Beck, N. and J.N. Katz (1995), What to do (and not to do) with time-series cross-section data, *American Political Science Review*, 89, 634–647.
- Bekker, P.A. (1994), Alternative approximations to the distributions of instrumental variables estimators, *Econometrica*, 63, 657–681.
- Bester, A., T. Conley, and C. Hansen (2011) Inference with dependent data using cluster covariance estimators, *Journal of Econometrics*, 165, 137–151.
- Binder, M., C. Hsiao, and M.H. Pesaran (2005), Estimation and inference in short panel vector autoregressions with unit roots and cointegration, *Econometric Theory*, 21, 795–837.
- Bhargava, A. and J.D. Sargan (1983), Estimating dynamic random effects models from panel data covering short time periods, *Econometrica*, 51, 1635–1659.

- Blundell, R. and S. Bond (1998), Initial conditions and moment restrictions in dynamic panel data models, *Journal of Econometrics*, 87, 115–143.
- Blundell, R. and S. Bond (2000), GMM estimation with persistent panel data: An application to production functions, *Econometric Reviews*, 19, 321–340.
- Blundell, R., S. Bond, and F. Windmeijer (2000), Estimation in dynamic panel data models: Improving on the performance of the standard GMM estimator, in B. Baltagi (ed.), *Non-stationary Panels, Panel Cointegration, and Dynamic Panels, Advances in Econometrics*, Vol. 15, Amsterdam: JAI Press, 161–178.
- Bontempi, M.E. and I. Mammi (2012), A strategy to reduce the count of moment conditions in panel data GMM, *Discussion Paper*.
- Breitung, J. (1994), Estimating dynamic panel data models: A comparison of different approaches, *Tinbergen Institute Discussion Paper No. TI 94-1*, <http://www.ect.unibonn.de/publikationen/panel.pdf>.
- Breitung, J. (2013), Bias-corrected estimators for various dynamic panel data models, University of Bonn, mimeo.
- Breitung, J. and J. Tenuhofen (2011), GLS estimation of dynamic factor models, *Journal of the American Statistical Association*, 106, 1150–1166.
- Breitung, J., N. Salish, and C. Roling (2013), LM-type tests for slope homogeneity in panel data models, University of Bonn, mimeo.
- Bruno, G.S.F. (2005), Estimation and inference in dynamic unbalanced panel-data models with a small number of individuals, *Stata Journal*, 5, 473–500.
- Bun, M.J.G. (2004), Testing poolability in a system of dynamic regressions with nonspherical disturbances, *Empirical Economics*, 29, 89–106.
- Bun, M.J.G. and M.A. Carree (2005), Bias-corrected estimation in dynamic panel data models, *Journal of Business & Economic Statistics*, 23, 200–210.
- Bun, M.J.G. and M.A. Carree (2006), Bias-corrected estimation in dynamic panel data models with heteroskedasticity, *Economics Letters*, 92, 220–227.
- Canova, F. and M. Ciccarelli (2013), Panel vector auto-regressive models: A survey, forthcoming in: *Advances in Econometrics*, Vol. 31.
- Chudik, A. and M.H. Pesaran (2011), Infinite-dimensional VARs and factor models, *Journal of Econometrics*, 163, 4–22.
- Chudik, A., M.H. Pesaran, and E. Tosetti (2011), Weak and strong cross-section dependence and estimation of large panels, *Econometrics Journal*, 14, C45–C90.
- Dees, S., F. diMauro, M.H. Pesaran, and L.V. Smith (2007), Exploring the international linkages of the euro area: A Global VAR analysis, *Journal of Applied Econometrics*, 22, 1–38.
- Demetrescu, M. and U. Homm (2014), Tests for No Cross-Sectional Error Correlation in Large-N Panel Data Models, mimeo.
- Dhaene, G. and K. Jochmans (2012), An adjusted profile likelihood for non-stationary panel data models with incidental parameters, University of Leuven, mimeo.
- DiMauro, F. and M.H. Pesaran (2013), *The GVAR Handbook: Structure and applications of a macro model of the global economy for policy analysis*, Oxford: Oxford University Press.
- Doran, H.E. and P. Schmidt (2006), GMM estimators with improved finite sample properties using principle components of the weighting matrix, with an application to the dynamic panel data model, *Journal of Econometrics*, 133, 387–409.
- Driscoll, J.C., and A.C. Kraay (1998), Consistent covariance matrix estimation with spatially dependent panel data, *Review of Economics and Statistics*, 80, 549–560.

- Dufour, J.-M. and Khalaf, L. (2002), Simulation based finite and large sample tests in multivariate regressions, *Journal of Econometrics* 111, 303–322.
- Eickmeier, S. and T. Ng (2011), How do credit supply shocks propagate internationally? A GVAR approach, CEPR Discussion Paper, 8720.
- Egger, P. and M. Pfaffermayr (2005), Long run and short run effects in static panel models, *Econometric Reviews*, 23, 199–214.
- Engle, R. F., D.F. Hendry, and J.F. Richard (1983), Exogeneity, *Econometrica*, 51, 277–304.
- Frees, E.W. (1995), Assessing cross-sectional correlation in panel data, *Journal of Econometrics*, 69, 393–414.
- Grassetti, L. (2011), A note on transformed likelihood approach in linear dynamic panel models, *Statistical Methods & Applications*, 20, 221–240.
- Greenaway-McGrevey, R., C. Han and D. Sul (2012), Asymptotic distribution for factor augmented estimators for panel regression, *Journal of Econometrics*, 169, 48–53.
- Groen, J.J.J. and F. Kleibergen (2003), Likelihood-based cointegration analysis in panels of vector error correction models, *Journal of Business & Economic Statistics*, 21, 295–318.
- Hahn, J., and G. Kuersteiner (2002), Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and T are Large, *Econometrica*, 70, 1639–1657.
- Hall, A.R. and F.P.M. Peixe (2003), A consistent method for the selection of relevant instruments, *Econometric Reviews*, 22, 269–287.
- Halunga, A., C.D. Orme, and T. Yamagata (2011), A heteroskedasticity robust Breusch-Pagan test for contemporaneous correlation in dynamic panel data models, University of Manchester, Discussion Paper, EDP-1118.
- Han, C., P.C.B. Phillips, and D. Sul (2011), Uniform asymptotic normality in stationary and unit root autoregression, *Econometric Theory*, 27, 1117–1151.
- Holtz-Eakin, D., W. Newey, and H.S. Rosen (1988), Estimating vector autoregressions with panel data, *Econometrica* 56, 1371–1395.
- Hsiao, C., M.H. Pesaran, and A.K. Tahmisioglu (2002), Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods, *Journal of Econometrics*, 109, 107–150.
- John, S. (1972), The distribution of a statistic used for testing sphericity of normal distributions, *Biometrika*, 59, 169–173.
- Judson, R.A. and Owen, A.L. (1999), Estimating dynamic panel data models: A guide for macroeconomists, *Economics Letters*, 65, 9–15.
- Juhl, T. and O. Lugovskyy (2014), A test for slope heterogeneity in fixed effects model, *Econometric Reviews*, 33, 906–935.
- Juodis, A. (2013a), A note on bias-corrected estimation in dynamic panel data models, *Economics Letters*, 118, 435–438.
- Juodis, A. (2013b), Iterative bias correction procedures revisited: A Monte Carlo study, University of Amsterdam, mimeo.
- Kapetanios, G. and M. Marcellino (2010), Factor-GMM estimation with large sets of possibly weak instruments, *Computational Statistics and Data Analysis*, 54, 2655–2675.
- Kiviet, J.F. (1995), On bias, inconsistency, and efficiency of various estimators in dynamic panel data models, *Journal of Econometrics*, 68, 53–78.

- Kloek, T., and L.B.M. Mennes (1960), Simultaneous equations estimation based on principal components of predetermined variables, *Econometrica*, 28, 45–61.
- Kruiniger, H. (2008), Maximum likelihood estimation and inference methods for the covariance stationary panel AR(1) unit root model, *Journal of Econometrics*, 144, 447–464.
- Lancaster, T. (2002), Orthogonal parameters and panel data, *Review of Economic Studies*, 69, 647–666.
- Lütkepohl, H. and M. Krätsig (2004), *Applied Time Series Econometrics*, Cambridge: Cambridge University Press.
- MacCurdy, T. (1982), The use of time series processes to model the time structure of earnings in a longitudinal data analysis, *Journal of Econometrics*, 18, 83–114.
- Maddala, G.S. (1999), On the use of panel data methods with cross-country data, *Annales d'Economie et de Statistique*, 55–56, 430–448.
- Mairesse, J. and Z. Griliches (1990), Heterogeneity in panel data: Are there stable production functions?, in P. Champsaur et al. (Eds.), *Essays in honor of Edmund Malinvaud*, MIT Press, Cambridge.
- Moon, H.R. and M. Weidner (2013), Linear regression for panel with unknown number of factors as interactive fixed effects, Working Paper.
- Moscone, F. and E. Tosetti (2009), A review and comparison of tests of cross-section independence in panels, *Journal of Economic Surveys*, 23, 528–561.
- Mundlak, Y. (1978), On the pooling of time series and cross-section data, *Econometrica*, 46, 69–85.
- Nerlove, M. (1971), Further evidence on the estimation of dynamic economic relations from a time series of cross-sections, *Econometrica*, 39, 359–382.
- Newey, W.K. and K.D. West (1987), A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica* 55, 703–708.
- Ng, S. (2006), Testing cross-section correlation in panel data using spacings, *Journal of Business and Economics Statistics*, 24, 12–23.
- Parks, R. (1967), Efficient estimation of a system of regression equations when disturbances are both serially and contemporaneously correlated, *Journal of the American Statistical Association*, 62, 500–509.
- Pesaran, M.H. (2004), General diagnostic tests for cross-section dependence in panels. CESifo Working Paper No. 1229.
- Pesaran, M.H. (2006), Estimation and inference in large heterogeneous panels with multifactor error structure, *Econometrica*, 74, 967–1012.
- Pesaran, M.H. (2012), Testing weak cross-sectional dependence in large panels, IZA Discussion Paper No. 6432, March 2012.
- Pesaran, M.H. and G. Kapetanios (2007), Small Sample Properties of Cross Section Augmented Estimators for Panel Data Models with Residual Multi-factor Structures; with M. H. Pesaran, in *The Refinement of Econometric Estimation and Test Procedures: Finite Sample and Asymptotic Analysis*, Garry Phillips and Elias Tzavalis (eds.), Cambridge: Cambridge University Press.
- Pesaran, M.H. and Y. Shin (1998), Generalized impulse response analysis in linear multivariate models, *Economics Letters*, 58, 17–29.
- Pesaran, M.H., T. Schuermann, and S.M. Weiner (2004), Modeling regional interdependences using a global error-correcting macroeconomic model, *Journal of Business and Economic Statistics*, 22, 129–162.

- Pesaran, M.H., Y. Shin, and R. Smith (1999), Pooled mean group estimation of dynamic heterogeneous panels, *Journal of the American Statistical Association*, 94, 621–634.
- Pesaran, M.H. and R. Smith (1995), Estimation of long-run relationships from dynamic heterogeneous panels, *Journal of Econometrics*, 68, 79–114.
- Pesaran, M.H., R. Smith, and K.S. Im (1996), Dynamic linear models for heterogeneous panels, Chapter 8 in L. Matyas and P. Sevestre (eds.), *The econometrics of panel data: A handbook of the theory with applications*, Dordrecht: Kluwer Academic Publishers, 145–195.
- Pesaran, M.H., A. Ullah, and T. Yamagata (2008), A bias-adjusted LM test of error cross-section independence, *Econometrics Journal*, 11, 105–127.
- Pesaran, M.H. and T. Yamagata (2008), Testing slope homogeneity in large panels, *Journal of Econometrics*, 142, 50–93.
- Phillips, P.C.B. and C. Han (2008), Gaussian inference in AR(1) time series with or without a unit root, *Econometric Theory*, 24, 631–650.
- Phillips, P.C.B. and D. Sul (2007), Bias in dynamic panel estimation with fixed effects, incidental trends and cross-section dependence, *Journal of Econometrics*, 137, 162–188.
- Pirotta, A. (1999) Convergence of the static estimation toward long run effects of the dynamic panel data models, *Economics Letters*, 63, 151–158.
- Reed, W.R. and R. Webb (2010), The PCSE estimator is good—Just not as good as you think, *Journal of Time Series Econometrics*, 2, Article 8.
- Roodman, D. (2009a), A note on the theme of too many instruments, *Oxford Bulletin of Economics and Statistics*, 71, 135–158.
- Roodman, D. (2009b), How to do xtabond2: An introduction to difference and system GMM in Stata, *Stata Journal*, 9, 86–136.
- Sarafidis, V. and T. Wansbeek (2012), Cross-sectional dependence in panel data analysis, *Econometric Reviews*, 31, 483–531.
- Sarafidis, V., T. Yamagata, and D. Robertson (2009), A test of cross-section dependence for a linear dynamic panel model with regressors, *Journal of Econometrics*, 148, 149–161.
- Sargan, J.D. (1980), Some tests of dynamic specification for a single equation, *Econometrica*, 48, 879–897.
- Stock, J.H., and M.W. Watson (2002), Forecasting using principal components from a large number of predictors, *Journal of the American Statistical Association*, 97, 1167–1179.
- Swamy, P.A.V.B. (1970), Efficient inference in a random coefficient regression model, *Econometrica*, 38, 311–323.
- Vogelsang, T.J. (2012) Heteroskedasticity, autocorrelation, and spatial correlation robust inference in linear panel models with fixed-effects, *Journal of Econometrics*, 166, 303–319.
- Westerlund, J. and J.-P. Urbain (2013a), On the estimation and inference in factor-augmented panel regressions with correlated loadings, *Economics Letters*, 119, 247–250.
- Westerlund, J. and J.-P. Urbain (2013b), Cross-sectional averages versus principal components, mimeo.

## CHAPTER 16

---

# COHORT DATA IN HEALTH ECONOMICS

---

STEPHANIE VON HINKE KESSLER SCHOLDER AND  
ANDREW M. JONES

### 16.1 INTRODUCTION

---

A cohort is a group of individuals who share a common characteristic or experience within a defined period of time. For example, we can define a cohort as all those born in the year 1975, or all World War II veterans. A cohort study is a longitudinal study that follows up a cohort over time to gather repeated measures of their life experiences and environments. Data from cohort studies have historically been used particularly by those in medicine, epidemiology, and social policy. Currently, however, interest has expanded to embrace a much wider range of disciplines, including those in economics. For example, cohort studies have been used to examine the economic returns to education (Blundell et al. 2000) and to compare the effects of selective and nonselective secondary education on children's test scores (Bonhomme and Sauder 2011). Within economics, however, partly due to the historical focus on the collection of data on health and health behaviors, cohort studies are of particular interest to health economists. Indeed, birth cohort studies typically collect detailed data on circumstances around birth and on early health conditions of the cohort members, which are then linked to physical and mental health, human capital accumulation, attitudes, family, and parenting later in life, allowing the researcher to explore the links between early life health and environment and later life development. In addition, the majority of studies now include biomedical information. There is an increasing interest among economists and social scientists to use objective information on, for example, cotinine concentrations, measures of cortisol, as well as genetic information. The availability of such data allow researchers to investigate biological factors that contribute to, and interact with, health,

education, or social conditions. Combined with the longitudinal data on cohort members, it will shed light on the complex interplay between biology, behavior, and environment over the life course, and its influences on health and well-being later in life.

Cohort studies, however, are only one of many different longitudinal research designs. Common designs include age or area (birth) cohorts, where all those of a certain age, or those living in a specific area, are sampled; household or family panels; economic short-term panels; and record linkage studies. This chapter considers birth cohorts, with a particular focus on the four—soon to be five—nationally representative UK birth cohorts, though we also discuss some of the rich UK area cohorts.<sup>1</sup> We focus on the United Kingdom, due to its unique history of longitudinal birth cohort data. Indeed, the British birth cohorts that follow nationally representative samples of individuals from birth to adulthood with direct contact with the cohort members from infancy rather than via registers, were, until relatively recently, unique.

The first nationally representative birth cohort, the National Survey of Health and Development, started in 1946, followed by the 1958, 1970, and 2000 cohort studies. Recruitment of pregnant women for the pilot phase of the next cohort study is due to start in 2014. In addition to these nationally representative birth cohorts, the United Kingdom has several area birth cohorts. Although these are not nationally representative, they may be substantially richer in the type of data that is collected. Until the Millennium Cohort Study, the studies were initially driven by medical interests in the conditions around birth. Subsequently, however, funding has come from different agencies, leading to different emphases in the studies' coverage (Bynner and Joshi 2007). However, all studies contain similar key variables including health, education, employment, and family life. The combination of these different (area) birth cohorts allows one to chart the changing nature of the life course in Britain, in relation to changing economic and political circumstances. They cover individual developmental processes and are particularly useful in linking early life, social, psychological, economic, and health experiences to outcomes later in life.

No other country has comparable social science and medical research resources for understanding the development of the human life course. The closest to the UK cohorts are the area-based New Zealand birth cohorts based in Dunedin and Christchurch (Bynner and Joshi 2007). More recently, however, cohort studies have been set up in other countries, including Norway, Denmark, and Ireland. In the United States, the pilot phase of the National Children's Study (NCS) started in 2009 and will follow children from birth to age 21. The French Longitudinal Study of Children (*Étude Longitudinale Française depuis d'Enfance*, or ELFE), started in 2011 and will follow 20,000 infants from birth (Bynner et al. 2007).

Before we discuss the UK birth cohorts, we consider the scientific rationale for studying birth cohorts. We argue that cohort studies are particularly interesting to health economists, linking this to the recent interest in biomedical and genetic data within economics. We will then describe the existing UK (area) birth cohort studies and refer

to some key papers in economics that use these data. Finally, we will end with a review of some of the econometric methods that have been applied to these cohort studies.

## 16.2 BIRTH COHORTS AND AREA BIRTH COHORTS

---

One of the main advantages of birth cohort studies is that they allow early life social, psychological, economic, and health conditions to be linked to outcomes later in life. The longer the individuals are followed up over time, the richer the potential of the data for the analysis of lifetime outcomes. They allow researchers to study the effects of exposures to different biological, economic, and environmental influences on the human life course, using a large sample of individuals born at the same time. As the first information about the cohort members is generally collected at, or very close to, birth, there is less scope for recall bias. Furthermore, birth cohort studies collect data not only about the cohort member but also about their carer (typically the mother). Hence, by construction, the data include information on two generations, allowing for potential intergenerational analysis, giving insights into the intergenerational transfer of health and resources and into cycles of deprivation and achievement (Martin et al. 2006). With a sufficiently long follow-up, many of the initial cohort members may also be observed when they have their own children, allowing for intergenerational analysis across three, and for some variables, even four generations. Johnson et al. (2010a, 2010b), for example, investigate intergenerational class mobility across three generations, and Johnston et al. (2012) examine the intergenerational correlation in mental health between three generations. Similarly, Emanuel et al. (1992) study the relationship between grandparents' socioeconomic indicators, and parents' and children's birth weights.

Furthermore, the long history of UK birth cohorts can add value by being used in combination. When multiple birth cohorts are combined, they allow the researcher to investigate changes in the human environment and its impact on the life course. For example, Schoon and Parsons (2003) show an increase in alcohol consumption since the 1946 cohort, followed by a plateau during the recession in the 1980s. With a series of cross-sections, the examination of such differentials can only be tracked at the aggregate level. The use of cohort data allows for an investigation into the individual trajectories underlying the trends, and allow for an analysis of smaller sub-groups within the data, such as the unemployed, or those on low incomes. Indeed, Schoon and Parsons (2003) show that, although men drank considerably more than women, the male 1958 cohort members tended to drink less as they got older, particularly those in unskilled occupations. The female 1958 cohort members, on the other hand, tended to drink more as they got older, particularly those in the lowest social class.

In addition, intergenerational analyses and cohort comparisons can be combined. Blanden et al. (2004) compare intergenerational mobility in earnings between the 1958 and the 1970 cohorts, showing that intergenerational mobility decreased substantially in this 12-year period. Much of this fall can be explained by factors such as non-cognitive skills, labor market attachment, and by the fact that the expansion of the higher education system has benefited people from well-off families more than those from poorer families (Blanden et al. 2004, 2007).<sup>2</sup>

There are, of course, also limitations to cohort studies. In particular, although individuals are followed up over time, they are generally not able to deal with short-term life course dynamics due to the relative infrequency of data collection. To deal with this, studies may use retrospective data collection, or increase the frequency of data collection during periods of rapid development, such as in infancy, early childhood, and in old age (Bynner and Joshi 2007). However, this may mean that less predictable transitions, such as the formation of partnerships, losing your job, or becoming seriously ill, are more difficult to be captured as well in cohort studies compared to other longitudinal designs such as annual panels.

In addition, the time window between subsequent waves often differs from one wave to the next. For example, the 1970 birth cohort observes individuals at ages 0, 5, 10, 16, 26, 34, 38, and 42. The unequally spaced intervals between subsequent waves introduce additional complications for the estimation of any dynamic models. For example, the differencing approach used in the dynamic panel data literature to remove the individual effects can no longer be applied (for a discussion of error component models with AR(1) disturbances in the context of unequally spaced panels, see, e.g., Baltagi and Wu 1999; and for an extension of this to “pseudo-panels,” see McKenzie 2001).

Another disadvantage of many of the existing cohort studies is that they tend to only collect data on the cohort member and—to a lesser extent—their parent, but fail to collect information on other family members, such as siblings. The more recent cohort studies, however, broaden the data collection to other family members. Furthermore, cohort studies will have attrition, reducing numbers for longitudinal analysis and potentially biasing the analysis. However, this is not specific to cohort studies, but extends to any longitudinal research design. As such, it can be dealt with using (for example) re-weighting or multiple imputation methods. Indeed, the high response rates at the start of the UK cohort studies means a lot is known about those who left the study, which can be used to construct weights. Alternatively, one can use “pseudo panels,” replacing individual observations with cohort means (Deaton 1985). As repeated cross-sections suffer much less from typical panel data problems such as attrition, and are very often substantially larger, both in number of individuals/households and in the time period that they span, pseudo panels are immune to attrition bias and can incorporate long time periods (Deaton 1985; Verbeek 2008).

Area-based birth cohorts—one of the alternatives to nationally representative birth cohorts—do not necessarily provide population estimates. However, they do allow

for easier linkage of data to local institutions such as school, hospital, and GP records. In addition, children and families can be visited more frequently than in studies based on national samples, or they can be invited to specially designated study centers, where more extensive tests can be done that are not possible in home visits. As such, these studies are likely to contain much richer data.

A potential disadvantage of birth cohorts is their definition of eligibility: they either include all those born in a small geographically defined region, or those born in a larger area, but within a short period of time, eliminating much of the potential exogenous variation relating to the date or place of birth. The rise in the UK minimum school leaving age to age 16, for example, occurred in 1973, making the 1958 cohort the first year group required to stay on at school for an extra year. As the cohort members were all born in one week in March, there is no variation in who is, and who is not, affected by this policy. With that, designs that exploit the change in the school leaving age can generally not be applied to these cohorts. Some studies, however, have used similar methods to identify exogenous variation in the years of education of the *parents* of the cohort members to examine the effect of parental education on children's outcomes (see, e.g., Lindeboom et al. 2009).

For nationally representative birth (though not area) cohorts, however, one may be able to exploit any geographical or spatial variation. Using the 1958 cohort, for example, Kelly (2011) examines the effects of *in utero* exposure to the 1957 Asian influenza pandemic on physical and cognitive development, identifying the effects using geographical variation at birth in the intensity of the influenza outbreak across Local Authorities.

### 16.3 THE BIRTH COHORTS

---

We next include a brief description of the nationally representative birth cohorts available in the United Kingdom. As they all cover a similar wide variety of information on health, education, employment, family life, socioeconomic, and socio-demographic characteristics, we do not list these in detail. Instead, table 16.1 gives a brief summary of the cohorts, with details on where to find more information and how to access them.

Although we do not aim to give a comprehensive overview of the many papers that use each of the cohort studies, we discuss some of the key papers in (health) economics. A full list of publications can be found on the cohorts' websites. We then describe some of the UK area-based birth cohorts. For all cohort studies we discuss, data collection is ongoing and the cohorts will continue to gain value as they are extended and the length of follow-up increases. This has particular relevance for economic models that take a life-cycle perspective and focus on long-term consequences and outcomes.

**Table 16.1 Overview of the British birth (area) cohorts**

Cohort	The Sample	How to Access the Data	CLOSER <sup>a</sup>	HALCyon <sup>b</sup>	More Information	
MRC NSHD	All singleton births to married mothers born in one week of spring 1946 in England, Wales, and Scotland	Contact study team	Yes	Yes	<a href="http://www.nshd.mrc.ac.uk">http://www.nshd.mrc.ac.uk</a>	
NCDS	All those born in Britain in one week of spring 1958	UK Data Archive <sup>c</sup>	Yes	Yes	<a href="http://www.esds.ac.uk/longitudinal/access/ncds">http://www.esds.ac.uk/longitudinal/access/ncds</a> <a href="http://www.cls.ioe.ac.uk/ncds">http://www.cls.ioe.ac.uk/ncds</a> <a href="http://www2.le.ac.uk/projects/birthcohort">http://www2.le.ac.uk/projects/birthcohort</a>	
BCS	All those born in Britain in one week of spring 1970	UK Data Archive <sup>c</sup>	Yes	No	<a href="http://www.esds.ac.uk/longitudinal/access/bcs70">http://www.esds.ac.uk/longitudinal/access/bcs70</a> <a href="http://www.cls.ioe.ac.uk/bcs70">http://www.cls.ioe.ac.uk/bcs70</a>	
MCS	A sample from the population of all UK births over a 12-month period from September 1, 2000, in England and Wales, and December 1, 2000, in Scotland and Northern Ireland	UK Data Archive <sup>c</sup>	Yes	No	<a href="http://www.esds.ac.uk/longitudinal/access/mcs">http://www.esds.ac.uk/longitudinal/access/mcs</a> <a href="http://www.cls.ioe.ac.uk/mcs">http://www.cls.ioe.ac.uk/mcs</a>	
Life Study	Will follow over 100,000 UK babies and their families through pregnancy, birth, and early years	Pilot phase will start in 2014	Yes	No	<a href="http://www.lifestudy.ac.uk">http://www.lifestudy.ac.uk</a>	
The Hertfordshire Studies	The HAS includes those born between 1920–1930 and still resident in Hertfordshire in the 1990s. The HCS includes those born between 1931–1939 and still resident in Hertfordshire in 1998	Contact study team	Yes (HCS)	Yes	<a href="http://www.mrc.soton.ac.uk/herts">http://www.mrc.soton.ac.uk/herts</a>	
The Lothian Birth Cohorts	The LBC21 and LBC36 are follow-up studies of the Scottish Mental Surveys of 1932 and 1947, respectively. Further data have been collected at ages 79, 83, 87, 90 (LBC21), and 69.5, 73, 76 (LBC36)	Contact study team	No	Yes (LBC21)	<a href="http://www.lothianbirthcohort.ed.ac.uk">http://www.lothianbirthcohort.ed.ac.uk</a>	
Boyd Orr Cohort	This is based on a follow-up of the 1936/37 study of Family Diet and Health in Pre-War Britain	Contact group <sup>d</sup>	steering	No	Yes	<a href="http://www.bris.ac.uk/social-community-medicine/projects/boyd-orr">http://www.bris.ac.uk/social-community-medicine/projects/boyd-orr</a>
ALSPAC	Includes children of women with an expected delivery date between April 1, 1991, and December 31, 1992, living in the Bristol area	Contact executive committee	Yes	No	<a href="http://www.bristol.ac.uk/alspac">http://www.bristol.ac.uk/alspac</a>	
Born-in-Bradford	All pregnant women who visited Bradford Royal Infirmary for a routine antenatal appointment between March 2007–December 2010	Contact executive committee	No	No	<a href="http://www.borninbradford.nhs.uk">http://www.borninbradford.nhs.uk</a>	
GMS	All infants born to mothers resident in Gateshead between June 1999 and May 2000	Contact study team	No	No	<a href="http://www.research.ncl.ac.uk/gms">http://www.research.ncl.ac.uk/gms</a>	
Gemini	Twins born in England and Wales between March–December 2007	Contact study team	No	No	<a href="http://www.geministudy.co.uk">http://www.geministudy.co.uk</a>	

<sup>a</sup>CLOSER (Cohort and Longitudinal Studies Enhancement Resources) includes nine longitudinal studies, aiming to promote cross-cohort analysis and comparisons (see <http://www.closer.ac.uk>).

<sup>b</sup>The HALCyon collaboration is an interdisciplinary group, investigating healthy aging in nine UK cohorts (see Kuh et al. 2012; <http://www.halcyon.ac.uk>).

<sup>c</sup>Non-commercial users can download the data free of charge and on completion of an end use licence from the UK Data Archive.

<sup>d</sup>Data from ESRC-funded research have been deposited in the UK Data Archive.

### 16.3.1 MRC National Survey of Health and Development (1946 Cohort)

The first of the national birth cohort studies, the MRC National Survey of Health and Development (NSHD), started as a survey of maternity services, gathering information in anticipation of the introduction of the National Health Service (NHS) in 1948. The target sample included all singleton births to married mothers born in one week of spring 1946 in England, Wales, and Scotland. Although the study was initially cross-sectional, when the children were 2 years old, a selection of the cohort members' mothers were re-interviewed (Wadsworth et al. 2006). Since then, data have been collected at ages 4, 6, 7, 8, 9, 10, 11, 13, 15, 16, 19, 20, 22, 24, 26, 31, 36, 43, 53, and most recently included an intensive clinic-based data collection at ages 60–64 (Kuh et al. 2011).

Although the initial sample consisted of over 13,000 births, at age 2, only a selection of the initial cohort members ( $N = 5,362$ ) were re-interviewed. By age 53 (in 1999), 3,035 cohort members were still participating. Despite this attrition, in most respects, the sample represents the national population of a similar age (Wadsworth et al. 2006; Kuh et al. 2011).

With the focus of the MRC NSHD team on biomedical research, the data have been used less in economics. However, many of the research questions that have been examined with the NSHD are of interest to (health) economists. For example, Kuh et al. (1997, 2002) investigate the lifetime influences on male and female earnings, Bukodi and Goldthorpe (2011) examine social mobility across cohorts, and Neuburger et al. (2011) explore cross-cohort changes in gender pay differences.

### 16.3.2 National Child Development Study (1958 Cohort)

The National Child Development Study (NCDS), the second nationally representative British birth cohort, follows up all those born in one week of spring 1958. The study began as the Perinatal Mortality Survey (PMS), aiming to identify social and obstetric factors linked to stillbirth and neonatal death. Although the survey was planned to be cross-sectional, when the cohort members were aged 7, it was converted into a longitudinal study to supply evidence for an enquiry into primary education. The cohort members were subsequently monitored at ages 11, 16, 23, 33, 41/42, 46, and 50/51. The next survey is planned for 2013. Information from the 1971 and 1981 censuses has been linked to the NCDS. Furthermore, at age 33, the study included a random one in three sample of cohort members' children, and in 1999/2000, the questionnaires to

the 1958 and 1970 cohort were integrated to facilitate comparisons between the generations. Finally, when the cohort members were 44–45 years old, in 2002–2004, the research team gathered biomedical and genetic data.

The PMS, or the first wave of the NCDS, included 17,416 cohort members. At age 42, in 2000, the sample still included 71% of those eligible (i.e., excluding those who had died or emigrated, see Plewis et al. 2004). The distribution of birth characteristics in this sample is similar to the distribution of the same characteristics in the sample at birth (Power and Elliott 2006).

The first key findings from the NCDS emerged from the PMS, showing the adverse effects of smoking during pregnancy on birth weight and perinatal mortality (Butler, Goldstein, and Ross 1972). Since then, the study has been widely used, including by economists.

Although, at first glance, there is no exogenous variation in birth circumstances to exploit, such as the timing of births, several papers do use “natural experiments” to explore the long-term effects of childhood health or circumstances on later life outcomes. Indeed, Kelly (2011) exploits geographical variation in the intensity of *in utero* exposure to the 1957 Asian influenza pandemic on physical and cognitive development. Similarly, Sacerdote (2002) compares adoptees in high versus those in lower socioeconomic status (SES) families, arguing that, in certain cases, adoption can be thought of as a natural experiment in which children are randomly assigned to different family backgrounds.

Looking at the effects of early health on later outcomes, Currie and Hyson (1999) find that low birth weight has significant long-term effects on self-reported health status, educational attainment, and labor market outcomes later in life. Similarly, Case, Fertig, and Paxson (2005) show that, controlling for parental income, education, and social class, children who experience poor health have significantly lower educational attainment, poorer health, and lower social status as adults. Rather than looking at the effects of health on education, other studies examine the effects of education on health. Jones et al. (2011, 2012), for example, show that educational attainment and school quality are important predictors of health and health-related behavior later in life.

Many studies compare the 1958 cohort to one or more of the later UK cohorts, or compare the UK cohorts to similar data from the United States. Using the NCDS, BCS, and MCS, where the mothers’ knowledge about the harms of prenatal smoking varied substantially, Fertig (2010) examines the importance of selection in the relationship between prenatal smoking and birth outcomes. Similarly, Galindo-Rueda and Vignoles (2005) use the 1958 and 1970 cohorts to explore whether Britain’s substantial expansion of education during the past decades have been associated with changes in the role of cognitive ability and parental background in determining educational achievement. Case and Paxson (2008) use these cohorts, as well as data from the United States to study the relationship between height and earnings, and Cutler and Lleras-Muney (2010) investigate the possible explanations for the relationship between education and health behaviors in the United Kingdom and the United States.

### 16.3.3 Birth Cohort Study (1970 Cohort)

The Birth Cohort Study (BCS) includes all individuals born in Britain in one week in spring 1970. Unlike the NSHD and the NCDS, the BCS was always planned to be longitudinal. It aimed to examine the social and biological characteristics of the mother in relation to neonatal morbidity and to compare the results with those of the NCDS. In 1975, 1980, and 1986, the parents were interviewed by health visitors, and information was gathered from the child's class teacher, head teacher, from the school health service, and from the children themselves. In total, the cohort members were followed up at ages 5, 10, 16, 26, 34, 38, and 42. The 2004/05 data collection also includes information on (and from) a 50% sample of cohort members' children, focusing on their health and a number of ability scales.

Up to age 16, cohort members were traced through their schools. However, industrial action by teachers, who were responsible for the educational tests, led to incomplete data collection in 1986 (age 16). The integration with the NCDS in 1999/2000 was key in restoring the BCS sample and establishing the scientific content of the adult surveys (Elliott and Shepherd 2006). At age 30, in 2000, the sample included 70% of those eligible (i.e., excluding those who had died or emigrated; see Plewis et al. 2004). The distribution of characteristics in this sample is similar to the distribution of characteristics at birth (Power and Elliott 2006).

Various studies in the economics literature have used data from the 1970 cohort study. Many of these use data from more than one UK birth cohort (discussed above). Other studies use only the BCS, such as Kline and Tobias (2008), who examine the effect of body mass index (BMI) on earnings, allowing for nonlinearities in the relationships between BMI and log wages by letting the potentially endogenous BMI enter the log wage equation non-parametrically; and Sturgis and Sullivan (2008), who use a latent class growth analysis framework to identify different social class trajectory groups between 1980 and 2000.

### 16.3.4 Millennium Cohort Study

The 12-year interval between the previous birth cohorts was broken when there was no new cohort in the 1980s and 1990s. But with the newly elected government in 1997 placing increasing emphasis on evidence-based policy, interest in a new birth cohort was renewed. With ESRC funding and contributions from government departments, the new cohort study started in 2000.

In contrast to the one week time-window of births in the 1946, 1958, and 1970 cohorts, the Millennium Cohort Study (MCS) includes a sample from the population of all births in the United Kingdom over a 12 month-period from September 1, 2000, in England and Wales, and December 1, 2000, in Scotland and Northern Ireland. In addition, stratified sampling ensured that ethnic minorities and the economically deprived appear in sufficient numbers for any subgroup analysis.

Reflecting its social science funding, the study aimed to be multipurpose, not just medical, and was set up to understand the social and economic circumstances of British children born at the beginning of the twenty-first century. The data was collected at 9 months, and ages 3, 5, and 7. The age 11 survey took place during 2012, when the children were in their final year of primary school. In addition to collecting data on the cohort members, it includes information on the children's siblings and parents. In addition, some routinely collected (administrative) health and educational records are linked to the study, such as the National Pupil Database.

The initial sample includes almost 19,000 babies at age 9 months. The third and fourth wave included 79% and 72% of cohort members, respectively (Ketende et al. 2008; Hansen et al. 2010). Plewis et al. (2008) show that residential mobility is an important predictor of sample loss over the first two waves.

Many studies that use the MCS compare it to one or more of the previous cohorts. As the MCS is still a relatively young cohort, they tend to examine birth or early childhood outcomes (e.g., Fertig 2010), or intergenerational mobility (Blanden and Machin 2007), though other work has examined, for example, ethnic differences in birth outcomes (Dearden et al. 2006), socioeconomic gradients in children's cognitive development (Waldfogel and Washbrook 2010), and parental preferences and school choice (Burgess et al. 2009).

### **16.3.5 Life Study**

Life Study, the UK's next nationally representative birth cohort, is an interdisciplinary research study, tracking social, health, development, and biological information for over 100,000 UK babies and their families through pregnancy, birth, and their early years. The study will cover the social diversity of the next generation of UK citizens, including information on family structures, ethnic identities, and socioeconomic circumstances. It is the first UK-wide study to start recruitment in pregnancy, rather than at birth, allowing researchers to examine prenatal, as well as postnatal, risk factors that affect the health and development of young children. Recruitment of pregnant women for its pilot phase is due to start in 2014.

The study will link to routine health and administrative data, and children and their families will be invited to attend specially designed study centers, where more extensive tests and measures can be done that are not possible in home visits. Hence, the data will allow researchers to explore the complex relationships between biology, behavior, and environment during early child development, and to investigate how these influence the future health and well-being of children and their parents.

## 16.4 OTHER BRITISH BIRTH (AREA) COHORTS

### 16.4.1 The Hertfordshire Studies

In 1911, Ethel Margaret Burnside assembled a team of midwives and nurses charged with improving the health of children in Hertfordshire. A midwife attended women during childbirth and recorded the birth weight of their offspring. A health visitor subsequently went to each baby's home throughout its infancy and recorded its illnesses, vaccinations, development, and method of infant feeding and weaning. The baby was then weighed again at age 1, and all data was transcribed into ledgers. These ledgers cover all births in Hertfordshire from 1911 until the NHS was formed in 1948 (Syddall et al. 2005).

In the early 1990s, surviving men and women who were born between 1920 and 1930 and still resident in Hertfordshire were contacted and underwent detailed physiological tests. In total, 717 individuals were included in the first follow-up of the Hertfordshire Ageing Study (HAS). In 2004, a second follow-up was carried out. The main objective of the HAS is to examine life course influences on healthy aging.

In 1998, a larger and younger cohort of individuals, born in Hertfordshire between 1931 and 1939 and still resident in Hertfordshire, were recruited for the Hertfordshire Cohort Study (HCS). Out of the 24,130 recorded single births between 1931 and 1939, 7,106 were traced as still living in Hertfordshire in 1998 and registered with a GP. A final sample of 3,225 (2,997) agreed to a home interview (and clinic). HCS cohort participants were generally comparable to those in the Health Survey for England (Syddall et al. 2005). The HCS cohort members have been followed up through primary care and hospital records, and are flagged with the NHS Central Register for notification of deaths.

Some of the main findings from the Hertfordshire ledgers came from their linkage to mortality records. Osmond et al. (1993) were the first that used individual-level (rather than aggregated) data to show the relationship between risk of death from cardiovascular disease and low birth weight/low weight at 1 year. Further research has shown the relationship between small size at birth and in infancy and a wide range of diseases such as type II diabetes (Hales et al. 1991). These findings subsequently led to the 'developmental origins' hypothesis, arguing that the nourishment a baby received from its mother during pregnancy, and its nutrition and illnesses in infancy determine its susceptibility to disease later in life.

### 16.4.2 The Lothian Birth Cohort Studies

The Lothian Birth cohorts of 1921 and 1936 (LBC21 and LBC36) are follow-up studies of the Scottish Mental Surveys of 1932 and 1947, respectively. Both surveys consisted

of intelligence tests taken by 11-year-old children in Scotland. On June 1, 1932, the Scottish Mental Survey took place simultaneously across schools in Scotland, testing a total of 87,498 children. Almost 70 years later, those born in 1921 and who had potentially taken part in the survey were invited to participate in the LBC21. In total, 550 participants with an average age of 79 completed the first follow-up between 1999 and 2001, who have also been contacted at ages 83, 87, and 90 (Deary et al. 2011).

The 1947 Scottish Mental Survey used the same intelligence test as the 1932 Survey and took place on June 4, 1947, testing 70,805 children. Between 2004 and 2007, those born in 1936 and who might have taken part in the 1947 Survey were invited to participate in the LBC36. With an average age of 69.5, 1,091 participants joined. The LBC36 members have since completed an additional follow-up at mean age 73 and 76 (Deary et al. 2011).

The LBC21 and LBC36 were set up to examine individual differences in cognitive changes from childhood to old age, and within old age. Cognitive tests include measures of reasoning, processing speed, executive function, and memory. In addition, detailed medical history and physical information are collected at each wave, as well as other background variables (see Deary et al. 2007, 2009, 2011). Although we are not aware of any studies within economics that have used these data, they have been used, for example, in intergenerational studies in psychology, examining class mobility across three generations (see, e.g., Johnson et al. 2010a, 2010b).

### **16.4.3 Boyd Orr Cohort**

The Boyd Orr Study is a historical cohort study, based on the 65 year follow-up of 4,999 children who were surveyed in the Carnegie United Kingdom Trust's study of Family Diet and Health in Pre-War Britain. The aim of the initial 1937–1939 study was to relate the quality of diet to family income and to the health of the children living in the household. As well as requiring families to keep food diaries, the study recorded the method of infant feeding, and other family background characteristics. Of the 4,999 children surveyed in 1936/37, 1,648 completed a detailed follow-up questionnaire in 1997/98, who were also contacted in 2002/03.

We are aware of only one study within the economics literature that uses these data. Frijters et al. (2010) investigate the importance of childhood socioeconomic conditions in predicting life expectancy. They find that socioeconomic conditions during childhood are significant predictors of longevity, but that their role differs for different causes of death, with household income being a significant predictor of death from smoking-related cancer.

### 16.4.4 Avon Longitudinal Study of Parents and Children

Those eligible for enrollment in the Avon Longitudinal Study of Parents and Children (ALSPAC), also known as the “Children of the 90s,” were women with an expected delivery date between April 1, 1991, and December 31, 1992, who were living in the Bristol area. With that, it is often referred to as the “missing cohort,” bridging the 30-year gap between the BCS and MCS. The initial sample includes children from 14,541 pregnancies.

ALSPAC is unique in that it has collected a vast amount of information on the cohort members; much more than any previous British cohort study. When the cohort members were aged 18, the follow-up included 59 child-based questionnaires (completed by the child, mother/carer, or teacher), and nine clinical assessment visits. In addition, mothers and partners have completed up to 20 questionnaires about themselves. The response varies with the questionnaires, though at age 140 months, the sample included around 55–60% of the original sample. The resource comprises a wide range of information on the cohort members, linked to administrative records on health, educational, economic, criminal, and neighborhood data, as well as an extensive biobank (Boyd et al. 2012; Fraser et al. 2012).

ALSPAC has been widely used by economists, such as those examining the effects of income on cognitive ability, socio-emotional outcomes, and physical health (Gregg et al. 2008), the mechanisms through which family income affects child health and the effects of maternal mental health (Propper et al. 2007), the effects of early maternal employment on child cognitive development (Gregg et al. 2005), and of breastfeeding on children’s cognitive development (Iacovou and Sevilla-Sanz 2010). Its genetic data have also been used within the economics literature, examining the effects of height and fat mass on educational attainment, behavioral problems, self-esteem, and symptoms of depression in an Instrumental Variables (IV) setup (von Hinke Kessler Scholder et al. 2011, 2013).

### 16.4.5 Born-in-Bradford

Bradford is a very diverse city with a population of about 500,000. Around 50% of the 6,000 babies born each year are of South Asian origin. It includes areas that are amongst the most deprived in Britain: 60% of babies born in Bradford are born into the poorest 20% of the population in England and Wales. Its infant mortality has been consistently above the national average, peaking in 2003 at 9.4 deaths per 1,000 live births; almost double the national 2003 average of 5.5 deaths per 1,000 births. One of the aims of the Born in Bradford study is to explore and understand these health inequalities and to help identify people at risk.

All pregnant women who visited Bradford Royal Infirmary for a routine antenatal appointment between March 2007 and December 2010 were asked to join the Born

in Bradford cohort study, currently consisting of 12,453 women, 13,776 children, and 3,448 partners (see Wright et al. 2012). Data were collected from mothers at pregnancy ( $\sim 28$  weeks), and babies at birth, and at many points during childhood. Data linkage across health and education sectors allows this cohort study to track key outcomes including mortality, morbidity, and educational attainment over time.

### **16.4.6 Gateshead Millennium Study**

All infants born to mothers resident in Gateshead—an urban district in northeast England—between June 1999 and May 2000 were invited to join the Gateshead Millennium Study (GMS). From the 1,252 eligible mothers, 1,011 joined the study. The sample is comparable to the northern region of England in terms of socioeconomic deprivation, though it is slightly underrepresented among the most affluent.

The first interview occurred shortly after each baby was born, followed up by data collected by midwives and health visitors at hospital discharge, at 6 and 10 days, and age 3 months. Parents were sent an additional four postal questionnaires within the first year, and were invited to a clinic appointment at 13 months. Further questionnaire data were collected at 30 months, 5–6 years, 6–8, 8–10 years, and 11–13 years (Parkinson et al. 2011).

### **16.4.7 Gemini**

Gemini is a birth cohort study of young twins designed to assess genetic and environmental influences on early childhood weight trajectories, with a focus on infant appetite and the family environment. A total of 2,042 families with twins born in England and Wales between March and December 2007 agreed to participate and returned completed baseline questionnaires. Since then, families have been followed up at age 15, 20, 24, 28, 40, 44, and 60 months. It is the first twin birth cohort to focus on childhood weight gain with detailed and repeated measures of children's appetite, food preferences, activity behavior, and parental feeding styles, with repeated collection of anthropometrics (van Jaarsveld et al. 2009).

## **16.5 ECONOMETRIC METHODS APPLIED TO COHORTS**

---

Studies based on panel data sets can exploit the regular repeated observations to account for time invariant unobserved heterogeneity through fixed or random effects specifications, or by explicitly modeling any dynamic relationships. The scope for

doing this with birth cohorts is more limited as the number of waves tends to be smaller, they are widely and unequally spaced and different information is collected at different points in time. These limitations are offset by the length of follow-up and the rich set of information that is collected on the respondents, their parents, and, sometimes, their grandparents and own children. This includes information such as direct tests of cognitive ability, measures of non-cognitive skills, genetic information, and other biomarkers that can be used to proxy factors that would typically have to be treated as unobservables in panel data sets. As such, many of the econometric specifications that have been estimated with the birth cohorts are effectively cross-sectional regressions but with right-hand side variables that include a comprehensive set of measures of the individual's personal and family history. Direct regression approaches have been augmented by matching methods, and when more structural approaches have been adopted, these have drawn on past events or outcomes as a source of exogenous variation.

Using the ALSPAC data, Gregg et al. (2005) examine how cognitive development between age 4 and 7 is influenced by the timing of a mother's return to work. They specify linear models for cognitive ability at different ages as a function of maternal employment and other controls, recognizing that this relationship may be susceptible to unobserved heterogeneity bias if there are third factors that influence both cognitive development and the return to work. The models are estimated by OLS and, in the absence of suitable instrumental variables or potential natural experiments, the problem of unobserved heterogeneity is addressed by adding additional controls as proxy variables to "mop up" the residual variation. Identification therefore relies on a selection on observables strategy.

A similar strategy is adopted by Propper et al. (2007), also using the ALSPAC data. To analyze the impact of family background, in particular family income and maternal mental health, on childhood health outcomes, they specify a joint model that includes a child health production function and an equation for parental income. Child health is specified as a linear function of parental income, initial health, and a child fixed effect, along with other controls, while the income equation includes a parental fixed effect. Any correlation between the child and parent fixed effects introduces endogeneity bias in the estimate of parental income on child health. The authors argue that differencing out the fixed effects, a strategy often applied to panel data on adult outcomes, is invalid, as such "fixed" individual characteristics are still developing during childhood and may do so at different rates. Instead the equations are estimated separately, relying on a rich set of controls to capture as much of the unobserved heterogeneity as possible. As in Gregg et al. (2005), this takes advantage of the wide range of background information gathered in ALSPAC, but means that caution is required in giving a causal interpretation to the results.

Lindeboom et al. (2009) attempt to go beyond measures of association and estimate a causal effect of parental education on childhood health outcomes, exploiting the raising of the UK school leaving age from 14 to 15 in 1947 as a quasi-experiment. They use the NCDS, exploiting variation in the years of cohort members' parents' education due

to the reform. This is used to instrument for each parent's years of schooling, examining the effects on child birth weight, illness during the first week of life, and other childhood health outcomes. The fuzzy regression-discontinuity design identifies a local treatment effect in the sense that it relates to the impact of an extra year of schooling only among those parents who were induced to stay in school by the reform. They find little effect on child health.

Case and Paxson (2008) explore the well-known association between height and earnings, arguing it is largely attributable to the association between the former and cognitive ability. They suggest that height in adolescence is effectively a marker for cognitive ability, and that it is the latter that attracts higher earnings. Their model assumes a common unobserved factor or endowment, such as genetic or environmental factors, that influence both cognitive ability and heights. Both are assumed to be linear functions of the endowment, and adult wages are assumed to be a function of cognitive ability. The implication of their model is that a regression of wages on height at different ages, not controlling for the endowment, will show the largest coefficients for the ages at which the association with the endowment is highest, for example, during periods of adolescent growth spurts. Empirical testing of this model requires data that include height and cognitive ability in childhood, as well as earnings in adulthood: the kind of longitudinal span offered by birth cohorts. In addition to US data, they base their empirical analysis on the NCDS and BCS. This allows them not only to estimate adult earnings as a function of height in childhood, as well as adolescence, but also to explore the association between physical growth and cognitive ability that are collected as part of the cohort studies.

Following people from birth through to their middle age in the NCDS, Case et al. (2005) take a broader perspective on the long-term impacts of childhood health and family background on adult health and socioeconomic outcomes. Their empirical approach begins by exploring the association between the adult outcomes and a comprehensive set of childhood and family characteristics. For example, they specify linear regressions for educational qualifications at age 16, probit and ordered probit models for self-assessed health at ages 23, 33, and 42, as well as for employment and socioeconomic status at ages 33 and 42. The observed positive associations between health in childhood and education, health, employment, and social status later in life motivate a model in which linear equations for both socioeconomic status and health in middle age depend on their lagged values. The latter are then substituted out to give reduced form equations that depend on prenatal and childhood characteristics. In the absence of suitable instruments or sufficient panel data to allow for fixed effects, estimation of the structural equations relies on having a sufficiently rich set of controls.

A strength of cohort studies is that they offer the possibility of analyzing very long-term consequences of childhood circumstances. Indeed, Frijters et al. (2010) use linkage of the Boyd Orr cohort with official death records up to 2005 to analyze differences in life expectancy. They estimate mixed proportional hazard models for mortality in which the hazard function is a product of three components: first, a regressor function that depends on observed socioeconomic characteristics of the household;

second, the baseline hazard, which is assumed to be piecewise constant; and third, time invariant unobserved heterogeneity, modeled by a set of discrete mass points. The results show that socioeconomic conditions in childhood are a strong predictor of longevity and that this varies according to the cause of death, with family income more strongly (negatively) associated with deaths from smoking-related cancers than with other causes of death.

As the UK birth cohorts span from birth to adulthood, they have also been used to analyze the impact of education on later life outcomes. Cutler and Lleras-Muney (2010) draw on a range of US and UK data sets, including the NCDS, to describe and decompose the association between education and a range of self-reported health behaviors, such as smoking, diet, drinking, and use of preventive medical care. They estimate linear models for the health behaviors as a function of education, controlling for a set of standard socioeconomic controls. They then examine the percentage decline in the coefficient for education as additional controls, intended to proxy the pathways through which education may be associated with health, are added to the models. Together these proxy variables account for 60–80% of the measured association.

Jones et al. (2012) also use the NCDS but focus on inequality of opportunity in the relationship between education and adult health and use measures of the type and quality of school attended rather than individual attainment. This requires analysis of the full conditional distribution of the health outcomes as a function of individual circumstances. Inequality of opportunity is assessed by testing for stochastic dominance in these distributions using both nonparametric tests and parametric conditional distributional regression models.

Blundell et al. (2000) focus on university degrees and other qualifications obtained from higher education and use the NCDS to evaluate their impact on employment and earnings at age 33. As there is selection into higher education, a matching strategy is adopted to find suitable controls who share similar observable characteristics to those who enter higher education, thereby relying on these observed proxies to account for systematic differences in unobservables that may drive any selection. The comparison groups are narrowed down by restricting the analysis to those who have at least one A level. Observed heterogeneity in the treatment effects is accommodated by interacting the educational qualifications with socioeconomic characteristics.

Matching methods also appear in Jones et al. (2011) who analyze the long-term impact of the type and quality of secondary schooling experienced by the NCDS members on their health behaviors and outcomes later in life. As they attended secondary schools during the early 1970s, the respondents were exposed to a major educational reform which gradually replaced a selective system, made up of grammar and secondary modern schools, with a nonselective system based on comprehensive schools. Variation over time and geographic areas means that different respondents attended different types of school. Descriptive analysis shows systematic differences in the pre-schooling characteristics of those who went on to attend schools in the selective and comprehensive systems (see also Manning and Pischke 2006). To reduce the imbalance in these characteristics, Jones et al. (2011) use a matching approach, combining

coarsened exact matching on key control variables that measure cognitive and non-cognitive skills, along with propensity score and Mahalanobis exact matching on a broader set of pre-schooling characteristics. This is implemented in two steps. First, those who attended comprehensive schools are matched with a control group who attended selective schools. Then, to explore heterogeneity in the effects of school quality, those who attended grammar schools and those who attended secondary moderns are each matched with comparable controls who went to comprehensives. The matched samples are then coupled with parametric regression modeling of the health outcomes.

Bonhomme and Sauder (2011) also exploit the timing in the implementation of comprehensive schooling, analyzing test scores for cognitive ability at ages 11 and 16 in the NCDS. The timing of the test scores corresponds to the periods before and after secondary schooling, and the treatment is whether or not a child attended a selective school. A potential outcomes model is developed that allows for an unobserved initial endowment that can influence both cognitive ability and selection into the types of school. This generalizes the standard linear difference-in-differences specification, while maintaining the assumption of additivity of the initial endowment, and identifies the entire counterfactual distribution for the potential outcomes. They show that most of the difference in performance between pupils who attended selective or comprehensive schools is attributable to selection.

An explicit structural model based on potential outcomes is central to the research program, summarized in Conti and Heckman (2010) and Conti et al. (2010), that explores the early-life origins of health disparities and in particular the relationship between health and education. The model distinguishes causal effects from selection effects and allows for essential heterogeneity in treatment effects, so that a distribution of treatment effects can be identified. The selection into schooling is modeled explicitly and jointly with the health outcomes. The full structural model includes endowments of cognitive ability, personality traits, and health that are treated as latent variables and proxied by observable indicators in a measurement model, in the spirit of structural equation modeling, using mixtures of multivariate normal distributions for the distribution of the latent variables. The model is applied to the 1970 BCS and suggests that there can be substantial heterogeneity in treatment effects that may be masked by only considering average effects.

Comparing multiple birth cohorts offers an opportunity to identify the evolution of broad socioeconomic trends, by making an explicit comparison across the cohorts. Galindo-Rueda and Vignoles (2005) do this using the NCDS and BCS and present evidence that cognitive ability measured early in childhood has become a poorer predictor of subsequent academic attainment while family background, including parental income, has become more important. The 1958 and 1970 cohorts are also compared by Blanden et al. (2007) in a study that shows that intergenerational income mobility has declined in Britain. Intergenerational mobility is estimated by specifying log-earnings for children as a linear function of the log-earnings of their parents. The overall intergenerational elasticity is then decomposed to account for pathways or mediating factors such as cognitive ability, non-cognitive or social skills, educational attainment,

and early labor market attachment. This is done by chaining together regressions of the child's earnings on the mediating factors with regressions of the mediating factors on parent's earnings.

Fertig (2010) draws on the NCDS, BCS, and MCS. Awareness of the harm associated with smoking during pregnancy has changed substantially over the decades spanned by these cohorts, as has the prevalence of prenatal smoking. Fertig (2010) finds that the impact of prenatal smoking on the probability of a low birth weight, given gestation, is substantially greater in the most recent cohort, for which a more selected group of mothers continue to smoke during pregnancy. Her analyses pool the three cohorts, using probit models to estimate what is effectively a difference-in-differences specification with indicators for prenatal smoking interacted with indicators for the cohorts.

## 16.6 CONCLUSION

---

The United Kingdom has a unique history of longitudinal birth cohort data, where samples of individuals are followed from birth to adulthood with direct contact with the cohort members from infancy. Most of the birth cohorts have historically been used particularly by those in medicine, epidemiology, and social policy. More recently, however, interest has expanded to embrace a much wider range of scientific programs, including those in economics. We argue that cohort studies are of particular interest to health economists, partly due to the historical focus on health and health behaviors, combining detailed data on early child health, linked to physical and mental health, human capital accumulation, attitudes, family, and parenting later in life.

This chapter starts by considering the scientific rationale for studying birth cohorts, comparing the use of birth cohorts to other longitudinal research designs. We discuss the five nationally representative British/UK birth cohorts, as well as seven area birth cohorts, showing the wide range of data available in these studies. We discuss some of the key papers in the economics literature that have used these cohorts, exploiting the rich data that cover multiple generations living in the United Kingdom, who were born between the early 1900s and the early 2000s.

## ACKNOWLEDGMENTS

---

We thank the ALSPAC executive, Ashley Adamson (Gateshead Millennium Study), Laura Basterfield (Gateshead Millennium Study), Ian Deary (Lothian Cohort Studies), Hayley Denison (Hertfordshire Studies), Jane Elliott (NCDS), David Gunnell (Boyd Orr Cohort), Diane Kuh (NSHD), Clare Llewellyn (Gemini), Rosie McEachan (Born in

Bradford Cohort), Louise McSeveny (Life Study), Kathryn Parkinson (Gateshead Millennium Study), Lucinda Platt (MCS), Jessica Reilly (Gateshead Millennium Study), Kate Smith (MCS), Alice Sullivan (BCS), Holly Syddall (Hertfordshire Studies), Jane Wardle (Gemini), and Julie Withey (NSHD) for earlier comments on the cohort descriptions. We gratefully acknowledge funding from the Economic and Social Research Council (ESRC) under the Large Grant Scheme, reference RES-060-25-0045 and from the Medical Research Council (MRC), grant number G1002345.

## NOTES

---

1. Strictly speaking, only the Millennium Cohort Study and the future Life Study cover the whole of the United Kingdom; the others are British. However, we here refer to the “UK cohort studies.”
2. These findings have been controversial in the sociological literature though (see, e.g., Goldthorpe and Jackson 2007).

## REFERENCES

---

- Baltagi, B., and P. Wu. 1999. “Unequally Spaced Panel Data Regressions with AR(1) Disturbances.” *Econometric Theory*, 15, 814–23.
- Blanden, J., and S. Machin. 2007. “Recent Changes in Intergenerational Mobility in Britain.” Report for Sutton Trust. <http://www.suttontrust.com/research/recent-changes-in-intergenerational-mobility-in-britain>.
- Blanden, J., P. Gregg, and L. Macmillan. 2007. “Accounting for Intergenerational Income Persistence: Noncognitive Skills, Ability and Education.” *The Economic Journal*, 117(519), C43–C60.
- Blanden, J., S. Machin, A. Goodman, and P. Gregg. 2004. “Changes in Intergenerational Mobility in Britain,” in M. Corak, ed., *Generational Income Mobility in North America and Europe*. Cambridge: Cambridge University Press.
- Blundell, R., L. Dearden, A. Goodman, and H. Reed. 2000. “The Returns of Higher Education in Britain: Evidence from a British Cohort.” *The Economic Journal*, 110(461), 82–99.
- Bonhomme, S., and U. Sauder. 2011. “Recovering Distributions in Difference-in-Differences Models: A Comparison of Selective and Comprehensive Schooling.” *The Review of Economics and Statistics*, 93(2), 479–94.
- Boyd, A., J. Golding, J. Macleod, D. Lawlor, A. Fraser, J. Henderson, L. Molloy, A. Ness, S. Ring, and G. Davey Smith. 2012. “Cohort Profile: The ‘Children of the 90s’—the Index Offspring of the Avon Longitudinal Study of Parents and Children.” *International Journal of Epidemiology*, doi:10.1093/ije/dys064.
- Bukodi, E., and J. Goldthorpe. 2011. “Class Origins, Education and Occupational Attainment: Cross-Cohort Changes among Men in Britain.” *European Societies*, 13, 347–75.
- Burgess, S., E. Greaves, A. Vignoles, and D. Wilson. 2009. “What Parents Want: School Preferences and School Choice.” CMPO WP 09/222.
- Butler, N., H. Goldstein, and E. Ross. 1972. “Cigarette Smoking in Pregnancy: Its Influence on Birth Weight and Perinatal Mortality.” *BMJ*, 2, 127–30.

- Bynner, J., and H. Joshi. 2007. "Building the Evidence Base from Longitudinal Data: The Aims, Contents, and Achievements of the British Birth Cohort Studies." *Innovation: The European Journal of Social Science Research*, 20(2), 159–79.
- Bynner, J., M. Wadsworth, H. Goldstein, B. Maughan, S. Purdon, R. Michael, K. Sylva, and J. Hall. 2007. "Scientific Case for a New Birth Cohort Study: Report to the Research Resources Board of the Economic and Social Research Council." <http://www.longviewuk.com/pages/reportsnew.shtml>.
- Case, A., and C. Paxson. 2008. "Stature and Status: Height, Ability, and Labor Market Outcomes." *Journal of Political Economy*, 116(3), 499–532.
- Case, A., A. Fertig, and C. Paxson. 2005. "The Lasting Impact of Childhood Health and Circumstance." *Journal of Health Economics*, 24(2), 365–89.
- Conti, G., and J. Heckman. 2010. "Understanding the Early Origins of the Education-Health Gradient: A Framework that Can Also be Applied to Analyse Gene–Environment Interactions." *Perspectives on Psychological Science*, 5(5), 585–605.
- Conti, G., J. Heckman, and S. Urzua. 2010. "The Education-Health Gradient." *American Economic Review: Papers and Proceedings*, 100, 234–38.
- Currie, J., and R. Hyson. 1999. "Is the Impact of Health Shocks Cushioned by Socioeconomic Status? The Case of Low Birth Weight?" *American Economic Review*, 89(2), 245–50.
- Cutler, D., and A. Lleras-Muney. 2010. "Understanding Differences in Health Behaviors by Education." *Journal of Health Economics*, 29, 1–28.
- Dearden, L., A. Mesnard, and J. Shaw. 2006. "Ethnic Differences in Birth Outcomes in England." *Fiscal Studies*, 27(1), 17–46.
- Deary, I., L. Whalley, and J. Starr. 2009. A Lifetime of Intelligence: Follow-up Studies of the Scottish Mental Surveys of 1932 and 1947. Washington, DC: American Psychological Association.
- Deary, I., A. Gow, A. Pattie, and J. Star. 2011. "Cohort Profile: The Lothian Birth Cohorts of 1921 and 1936." *International Journal of Epidemiology*, doi:10.1093/ije/dyr197.
- Deary, I., A. Gow, M. Taylor, J. Corley, C. Brett, V. Wilson, H. Campbell, L. Whalley, P. Visscher, D. Porteous, and J. Star. 2011. "The Lothian Birth Cohort 1936: A Study to Examine Influences on Cognitive Aging from Age 11 to Age 70 and Beyond." *BMC Geriatrics*, 7, 28.
- Deaton, A. 1985. "Panel Data from Time Series of Cross-Sections." *Journal of Econometrics*, 30, 109–26.
- Elliott, J., and P. Shepherd. 2006. "Cohort Profile: 1970 British Birth Cohort (BCS70)." *International Journal of Epidemiology*, 35, 836–43.
- Emanuel, I., H. Filakti, E. Alberman, and S. Evans. 1992. "Intergenerational Studies of Human Birth Weight from the 1958 Birth Cohort: Evidence for a Multigenerational Effect." *British Journal of Obstetrics and Gynaecology*, 99, 67–74.
- Fertig, A. 2010. "Selection and the Effect of Prenatal Smoking." *Health Economics*, 19(2), 209–26.
- Fraser, A., C. Macdonald-Wallis, K. Tilling, A. Boyd, J. Golding, G. Davey Smith, J. Henderson, J. Macleod, L. Molloy, A. Ness, S. Ring, S. Nelson, and D. Lawlor. 2012. "Cohort Profile: The Avon Longitudinal Study of Parents and Children: ALSPAC Mothers Cohort." *International Journal of Epidemiology*, doi:10.1093/ije/dys066.
- Frijters, P., T. Hatton, R. Martin, and M. Shields. 2010. "Childhood Economic Conditions and Length of Life: Evidence from the UK Boyd Orr Cohort, 1937–2005." *Journal of Health Economics*, 29(1), 69–47.

- Galindo-Rueda, F., and A. Vignoles. 2005. "The Declining Relative Importance of Ability in Predicting Educational Attainment." *The Journal of Human Resources*, 40(2), 335–53.
- Goldthorpe, J., and M. Jackson. 2007. "Intergenerational Class Mobility in Contemporary Britain: Political Concerns and Empirical Findings." *British Journal of Sociology*, 58(4): 525–46.
- Gregg, P., C. Propper, and E. Washbrook. 2008. "Understanding the Relationship between Parental Income and Multiple Child Outcomes: A Decomposition Analysis." CMPO WP 08/193.
- Gregg, P., E. Washbrook, C. Propper, and S. Burgess. 2005. "The Effects of Early Maternal Employment on Child Development in the UK." *The Economic Journal*, 115, F48–80.
- Hales, C., D. Barker, P. Clark, L. Cox, C. Fall, C. Osmond, and P. Winter. 1991. "Fetal and Infant Growth and Impaired Glucose Tolerance at Age 64." *BMJ*, 303, 1019–22.
- Hansen, K., E. Jones, H. Joshi, and D. Budge. 2010. "Millennium Cohort Study Fourth Survey: A User's Guide to Initial Findings." Centre for Longitudinal Studies, London.
- Hinke Kessler Scholder, S., von, G. Davey Smith, D. Lawlor, C. Propper, and F. Windmeijer. 2011. "Genetic Markers as Instrumental Variables." CMPO Working Paper, 11/274.
- Hinke Kessler Scholder, S., von, G. Davey Smith, D. Lawlor, C. Propper, and F. Windmeijer. 2013. "Child Height, Health and Human Capital: Evidence Using Genetic Markers." *European Economic Review*, 57, 1–22.
- Iacovou, M., and A. Sevilla-Sanz. 2010. "The Effect of Breastfeeding on Children's Cognitive Development." ISER WP 2010-40.
- Jaarsveld, C., van, L. Johnson, C. Llewellyn, and J. Wardle. 2009. "Gemini: A UK Twin Birth Cohort with a Focus on Early Childhood Weight Trajectories, Appetite and the Family Environment." *Twin Research and Human Genetics*, 13(1), 72–78.
- Johnson, W., C. Brett, and I. Deary. 2010a. "Intergenerational Class Mobility in Britain: A Comparative Look across Three Generations in the Lothian Birth Cohort 1936." *Intelligence*, 38, 268–81.
- Johnson, W., C. Brett, and I. Deary. 2010b. "The Pivotal Role of Education in the Association between Ability and Social Class Attainment: A Look across Three Generations." *Intelligence*, 38, 55–65.
- Johnston, D., S. Schurer, and M. Shields. 2012. "Evidence on the Long Shadow of Poor Mental Health across Three Generations." HEDG working paper 12/20.
- Jones, A., N. Rice, and P. Rosa Dias. 2011. "Long-term Effects of School Quality on Health and Lifestyle: Evidence from Comprehensive Schooling Reforms in England." *Journal of Human Capital*, 5, 342–76.
- Jones, A., N. Rice, and P. Rosa Dias. 2012. "Quality of Schooling and Inequality of Opportunity in Health." *Empirical Economics*, 42(2), 369–94.
- Kelly, E., 2011. "The Scourge of Asian Flu: In Utero Exposure to Pandemic Influenza and the Development of a Cohort of British Children." *Journal of Human Resources*, 46(4), 669–94.
- Ketende, S., H. Joshi, L. Calderwood, J. McDonald, and the MCS Team. 2008. "Millenium Cohort Study: Technical Report on Response." Centre for Longitudinal Studies, London.
- Kline, B., and J. Tobias. 2008. "The Wages of BMI: Bayesian Analysis of a Skewed Treatment-Response Model with Non-parametric Endogeneity." *Journal of Applied Econometrics*, 23, 767–93.
- Kuh, D., J. Head, R. Hardy, and M. Wadsworth. 1997. "The Influence of Education and Family Background on Women's Earnings in Midlife: Evidence from a British National Birth Cohort Study." *British Journal of Sociology of Education*, 18(3), 385–405.

- Kuh, D., R. Hardy, C. Langenberg, M. Richards, and M. Wadsworth. 2002. "Mortality in Adults Aged 26–54 Years Related to Socioeconomic Conditions in Childhood and Adulthood: Post was Birth Cohort Study." *BMJ*, 325, 1076–80.
- Kuh, D., M. Pierce, J. Adams, J. Deanfield, U. Ekelund, P. Friberg, A. Ghosh, N. Harwood, A. Hughes, P. Macfarlane, G. Mishra, D. Pellerin, A. Wong, A. Stephen, M. Richards, and R. Hardy. 2011. "Cohort Profile: Updating the Cohort Profile for the MRC National Survey of Health and Development: A New Clinic-Based Data Collection for Ageing Research." *International Journal of Epidemiology*, 40, e1–e9.
- Lindeboom, M., A. Llena-Nozal, and B. van der Klaauw. 2009. "Parental Education and Child Health: Evidence from a Schooling Reform." *Journal of Health Economics*, 28, 109–31.
- Manning, A., and S. Pischke. 2006. "Comprehensive Versus Selective Schooling in England and Wales: What Do We Know?" IZA Discussion Paper no. 2072, Bonn.
- Martin, J., J. Bynner, G. Kalton, H. Goldstein, P. Boyle, V. Gayle, S. Parsons, and A. Piesse. 2006. "Strategic Review of Panel and Cohort Studies: Report to the Research Resources Board of the Economic and Social Research Council." <http://www.longviewuk.com/pages/reportsnew.shtml>.
- McKenzie, D. 2001. "Estimation of AR(1) Models with Unequally Spaced Pseudo-Panels." *Econometrics Journal*, 4, 89–108.
- Neuburger, J., D. Kuh, and H. Joshi. 2011. "Cross-Cohort Changes in Gender Pay Differences in Britain: Accounting for Selection into Employment Using Wage Imputation, 1972–2004." *Longitudinal and Life Course Studies*, 2(3), 260–85.
- Osmond, C., D. Barker, P. Winter, C. Fall, and S. Simmonds. 1993. "Early Growth and Death from Cardiovascular Disease in Women." *BMJ*, 327, 1428–30.
- Parkinson, K., M. Pearce, A. Dale, J. Reilly, R. Drewett, C. Wright, C. Relton, P. McArdle, A. le Couteur, and A. Adamson. 2011. "Cohort Profile: The Gateshead Millennium Study." *International Journal of Epidemiology*, 40, 308–17.
- Plewis, I., L. Calderwood, D. Hawkes, and G. Nathan. 2004. "Changes in the NCDS and BCS70 Populations and Samples over Time." Centre for Longitudinal Studies, London.
- Plewis, I., S. Kentende, H. Joshi, and G. Hughes. 2008. "The Contribution of Residential Mobility to Sample Loss in a Birth Cohort Study: Evidence from the First Two Waves of the UK Millennium Cohort Study." *Journal of Official Statistics*, 24(3), 365–85.
- Power, C., and J. Elliott. 2006. "Cohort Profile: 1958 British Birth Cohort (National Child Development Study)." *International Journal of Epidemiology*, 35, 34–41.
- Propper, C., J. Rigg, and S. Burgess. 2007. "Child Health: Evidence on the Roles of Family Income and Maternal Mental Health from a UK Birth Cohort." *Health Economics*, 16, 1245–69.
- Sacerdote, B. 2002. "The Nature and Nurture of Economic Outcomes." *American Economic Review Papers & Proceedings*, 92(2), 344–48.
- Schoon, I., and S. Parsons. 2003. "Lifestyle and Health-Related Behaviour," in E. Ferri, J. Bynner, M. Wadsworth (eds.), *Changing Britain, Changing Lives: Three Generations at the Turn of the Century*. London: Institute of Education, University of London.
- Sturgis, P., and L. Sullivan. 2008. "Exploring Social Mobility with Latent Trajectory Groups." *Journal of the Royal Statistical Society, Series A*, 171(1), 65–88.
- Syddall, H., A. Aihie Sayer, E. Dennison, H. Martin, D. Barker, C. Cooper, and the Hertfordshire Cohort Study Group. 2005. "Cohort Profile: The Hertfordshire Cohort Study." *International Journal of Epidemiology*, 34, 1234–42.

- Verbeek, M. 2008. "Pseudo Panels and Repeated Cross-Sections," in L. Mátyás and P. Sevestre (eds.), *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*. Berlin: Springer-Verlag.
- Wadsworth, M., D. Kuh, M. Richards, and R. Hardy. 2006. "Cohort Profile: The 1946 National Birth Cohort (MRC National Survey of Health and Development)." *International Journal of Epidemiology*, 35, 49–54.
- Waldfogel, J., and E. Washbrook. 2010. "Low Income and Early Cognitive Development in the UK." London: Sutton Trust.
- Wright, J., N. Small, P. Raynor, D. Tuffnell, R. Bhopal, N. Cameron, L. Fairley, F. Lawlor, R. Parslow, E. Petherick, K. Pickett, D. Waiblinger, and J. West. 2012. "Cohort Profile: The Born in Bradford Multi-Ethnic Family Cohort Study." *International Journal of Epidemiology*, 42, 978–91.

## CHAPTER 17

---

# PANEL DATA AND PRODUCTIVITY MEASUREMENT

---

ROBIN C. SICKLES, JIAQI HAO AND CHENJUN SHANG

### 17.1 INTRODUCTION

---

THE chapter first discusses how productivity growth typically has been measured in classical productivity studies. We then briefly discuss how innovation and catch-up can be distinguished empirically. We next outline methods that have been proposed to measure productivity growth and its two main factors, innovation and catch-up. These approaches can be represented by a canonical form of the linear panel data model. A number of competing specifications are presented and model averaging is used to combine estimates from these competing specifications in order to ascertain the contributions of technical change and catch-up in world productivity growth. The chapter ends with concluding remarks and suggestions for the direction of future analysis.

The literature on productivity and its sources is vast in terms of empirical and theoretical contributions at the aggregate, industry, and firm level. The pioneering work of Dale Jorgenson and his associates<sup>1</sup> and Zvi Griliches and his associates,<sup>2</sup> the National Bureau of Economic Research,<sup>3</sup> the many research contributions made in U.S universities and research institutions, the World Bank and research institutes in Europe and other countries are not discussed here as our goal is by necessity rather narrow. We focus on work directly related to panel data methods that have been developed to address specific issues in specifying the production process and in measuring the sources of productivity growth in terms of its two main components of innovation (technical progress) and catch-up (efficiency growth), with emphasis given to one of the more important measures of the latter component and that is technical efficiency.

## 17.2 PRODUCTIVITY GROWTH AND ITS MEASUREMENT

---

### 17.2.1 Classical Residual based Partial and Total Factor Productivity Measurement

Total factor productivity (TFP) is measured by a ratio of a weighted average of outputs ( $Y_i$ ) to a weighted average of inputs ( $X_i$ ). For a single output the ratio is:

$$TFP = \frac{Y}{\sum a_i X_i}. \quad (1)$$

Historically, there are two common ways of assigning weights for this index. They are to use either an arithmetic or geometric weighted average of inputs. The *arithmetic weighted average*, due to Kendrick (1961), uses input prices as the weights while the *geometric weighted average* of the inputs, attributable to Solow (1957), uses input expenditure shares as the weights. The predominant *TFP* measure currently in use by the central governments in most countries is a variant of Solow's measure based on the Cobb-Douglas production function with constant returns to scale,  $Y = AX_L^\alpha X_K^{1-\alpha}$ , and leads to the *TFP* measure:

$$TFP = \frac{Y}{X_L^\alpha X_K^{1-\alpha}}. \quad (2)$$

At cost minimizing levels of inputs, the parameter  $\alpha$  describes the input expenditure share for labor. The *TFP* growth rate is the simple time derivative of *TFP* and is given by:

$$T\dot{F}P = \frac{dY}{Y} - \left[ \alpha \frac{dX_L}{X_L} + (1 - \alpha) \frac{dX_K}{X_K} \right].$$

*TFP* is simply a ratio of index numbers. Fisher (1927) discussed the optimal properties for index numbers and these are also explored in depth by Good, Nadiri, and Sickles (1997). Jorgenson and Griliches (1972) pointed out that the *TFP* index could be expressed as the difference between the log output and log input indices.

### 17.2.2 Modifications of the Neoclassical Model: The New Growth Theory

Endogenous growth models (Romer, 1986) were proposed to address the inflexibility and simplicity of exogenously driven ("manna from heaven") technical change (Scherer, 1971). This was of course not new as Griliches (1957) and Edwin Mansfield (1961), among others, addressed these issues using endogenous rates of penetration

and endogenous rates of imitation to explain technical change. In the endogenous growth theory capital was allowed to have non-diminishing rates of return due to external effects that spillover for a variety of reasons. The level of technology  $A$  can vary depending on the stock of some privately provided input  $R$  (such as knowledge) and the production function is formulated as

$$Y = A(R)f(K, L, R)$$

As for potential sources of spillovers that could shift the production function there are many explanations. Learning-by-doing was Arrow's (1962) explanation, while for Romer (1986) it was the stock of research and development, for Lucas (1988) it was human capital, for Coe and Helpman (1995) and Coe, Helpman, and Hoffmaister (1997) it was trade spillovers, and for Diao, Rattsø, and Stokke (2005) it was trade openness. However, efficiency is another explanation if one simply attaches another reason for the spillover, such as a loosening of constraints on the utilization of the technology.

Another comment about endogenous growth models and the need to address endogeneity issues in productivity analyses needs to be made here. The literature on structural modeling of productivity models is quite dense and, again, it is not within the scope of this chapter to discuss this very important literature. However, there is a particular literature within the broader structural modeling of static and dynamic productivity model (see, e.g., Olley and Pakes 1996) that speaks to the focused issues addressed in this chapter and that is the role of errors-in-variables, weak instrument bias, and stability in panel data modeling of production processes. These issues have been taken up by a number of researchers, especially those from the NBER and include studies by Griliches and Hausman (1986), Stoker et al. (2005), Griliches and Mairesse (1990, 1998), and Griliches and Pakes (1984), to name but a few.

### 17.2.3 Technical Efficiency in Production

Nontransitory production inefficiencies can be attributed to a number of factors, such as random mistakes, the existence of market power (Kutlu and Sickles 2012), and historical precedent (Alam and Sickles, 2000). Technical inefficiency concepts were developed by Debreu (1951), Farrell (1957), Shephard (1970), and Afriat (1972). Measuring the intrinsically unobservable phenomena of efficiency has proven to be quite challenging. Aigner, Lovell, and Schmidt (1977), Battese and Corra (1977), and Meeusen and Van den Broeck (1977) developed the econometric methods to measure efficiency in production, while linear programming methods were initially made feasible to utilize in the classic study by Charnes, Cooper, and Rhodes (1978). As relative efficiency is usually constructed from a normalized residual and such a residual is generated from an econometric model, theoretical consideration from the economic theory of the firm and assumptions of weak exogeneity are needed in order to identify

it as a factor that is distinct from innovation. Both efficiency and technology change are the main drivers of productivity growth, along with scale economies. Scale effects may have an important role at the firm level but not necessarily at the aggregate economy level we will consider in the empirical analysis we summarize towards the end of this Chapter. As efficiency estimators differ on which identifying restrictions are imposed, it should surprise no one that results from alternative estimators differ as well. Kumbhakar and Lovell (2000) and Fried, Lovell, and Schmidt (2008) provide excellent treatments of this literature and how various modeling assumptions utilized in the early panel productivity literature by Pitt and Lee (1981), Schmidt and Sickles (1984), and others provide substantial leverage in determining measured levels of efficiency.

#### **17.2.4 The Panel Stochastic Frontier Model**

Introducing efficiency into the dynamic of productivity growth requires that we introduce a frontier production process relative to which efficiency can be measured. In order for cross-sectional methods to be useful in such a setting, identification of the efficiency term often requires a parametric assumption about its distribution, an assumption not needed when using panel data. Panel data methods for time invariant efficiency measurement introduced by Pitt and Lee (1981) and Schmidt and Sickles (1984) were soon followed up by Cornwell, Schmidt, and Sickles (1990) and Kumbhakar (1990), Battese and Coelli (1992), and Lee and Schmidt (1993) who introduced methods that allowed the efficiency effects to vary over time and between cross-sectional units. Kim and Lee (2006) generalized the Lee and Schmidt (1993) model by considering different patterns for different groups, while Hultberg, Nadiri, and Sickles (1999, 2004) modified the neoclassical country growth convergence model to allow for heterogeneities in the efficiency catch-up rates. The Hultberg, Nadiri, and Sickles (1999, 2004) studies also are instructive as they relate a set of environmental factors, such as a country's political and social institutions, to the rate of catch-up, a factor which they found to determine up to 60% of the variation in efficiency. The firm level study by Bloom and Van Reeden (2007) found that productivity differences among firms (efficiency differences) were best explained by such arcane factors as shop floor operations, monitoring, targets, and incentives, factors typically overseen by management and also typically viewed as related to managerial efficiency.

#### **17.2.5 Index Number Approaches to Calculate Innovation and Efficiency Change**

Identifying the sources of *TFP* growth while imposing minimal parametric structure has obvious appeal on grounds of robustness. Sharpness of inferences may, however,

be comprised vis-à-vis parametric structural econometric models. There has been a long-standing tradition to utilize index number procedures as well as reduced form or structural econometric estimation to quantify *TFP* growth and its determinants. Space limits the coverage that this chapter can provide to such important index number approaches. The interested reader is directed to the panel data literature on productivity index numbers and to surveys (e.g., Good, Nadiri, and Sickles 1997; Fried, Lovell, and Schmidt 2008), particular advances in decomposing productivity change into technical and efficiency growth via the Malmquist index introduced into the literature by Caves, Christensen, and Diewert (1982) (Färe et al. 1994; Grifell-Tatjé and Lovell 1995; Färe et al. 1997), problems with such index number approaches and decompositions (Førsund and Hjalmarsson 2008), and numerical approaches via bootstrapping to construct inferential procedures to assess such measures (Simar and Wilson 2000; Jeon and Sickles 2004).

### 17.3 DECOMPOSITION OF ECONOMIC GROWTH-INNOVATION AND EFFICIENCY CHANGE IDENTIFIED BY REGRESSION

---

A relatively transparent way to see how a linear regression can be used to estimate technical change and efficiency change is based on the following derivation. Let the multiple output/multiple input technology be represented by a parametric output distance function (Caves, Christensen and Diewert 1982; Coelli and Perelman 1996). Consider an output distance function or single output production function that is linear in parameters. Standard parametric functional forms widely used in empirical work that are linear in parameters are the Cobb-Douglas, translog, generalized-Leontief and quadratic. The many different specifications we consider here and the way in which various forms of unobserved heterogeneity can be modeled can be motivated using the following model for a single output technology estimated with panel data assuming unobserved country (firm) effects:

$$y_{it} = x_{it}\beta + \eta_i(t) + \nu_{it} \quad (3)$$

where  $\eta_i(t)$  represents the country-specific fixed effect that may be time varying,  $x_{it}$  is a vector of regressors, some of which may be endogenous and correlated with the error  $\nu_{it}$  or the effects  $\eta_i(t)$ .

The regression model (3) can be derived by a relatively straightforward transformation and re-parameterization of the output distance function. A parsimonious representation of the  $m$ -output,  $n$ -input deterministic distance function  $D_o(Y, X)$  is given by the Young index (Balk 2008):

$$D_o(Y, X) = \frac{\prod_{j=1}^m Y_{it}^{\gamma_j}}{\prod_{k=1}^n X_{it}^{\delta_k}} \leq 1.$$

The output-distance function  $D_o(Y, X)$  is non-decreasing, linear homogeneous in outputs, and concave in  $Y$  and non-increasing and quasi-concave in  $X$ . If we take the natural logarithm of the inequality, add a symmetric disturbance term  $v_{it}$  to address the standard random error in a regression and a technical efficiency term  $\eta_i(t)$  to represent inefficiency then the observed value of the distance function for country  $i$  at time  $t$  can be written as:

$$-y_{1,it} = \sum_{j=2}^m \gamma_j y_{jit}^* + \sum_{k=1}^n \delta_k x_{kit}^* + \eta_i(t) + v_{it},$$

where  $y_{jit}^*, j=2, \dots, m = \ln(Y_{jit}/Y_{1,it})$  and  $x_{kit}^* = \ln(X_{kit})$ . After redefining a few variables the distance function can be written as

$$y_{it} = x_{it}\beta + \eta_i(t) + v_{it}.$$

The Cobb-Douglas distance function introduced by Klein (1953) not only assumes strong separability of outputs and inputs but also has a production possibility frontier that is convex instead of concave. This last drawback may not be as important as it seems, as pointed out by Coelli (2000), and the Cobb-Douglas remains a reasonable and parsimonious first-order local approximation to the true function. The Cobb-Douglas can be extended to the translog output distance function by adding second order terms to provide for flexibility in curvature possibilities and by allowing interactions among the output and inputs, thus avoiding the strong separability implied by the Cobb-Douglas output distance function. This functional form also can be transformed and re-parameterized to fit into the form of the linear panel data model given in equation (3). The translog output distance function is given by:

$$\begin{aligned} -y_{1,it} = & \sum_{j=2}^m \gamma_j y_{jit}^* + \frac{1}{2} \sum_{j=2}^m \sum_{l=1}^m \gamma_{jl} y_{jit}^* y_{lit}^* + \sum_{k=1}^n \delta_k x_{kit}^* + \frac{1}{2} \sum_{k=1}^n \sum_{p=1}^n \delta_{kp} x_{kit}^* x_{pit}^* \\ & + \sum_{j=2}^m \sum_{k=1}^n \theta_{jk} y_{jit}^* x_{kit}^* + \eta_i(t) + v_{it}. \end{aligned}$$

Since the model is linear in parameters, then after redefining a few variables the translog distance function also can be written as

$$y_{it} = x_{it}\beta + \eta_i(t) + v_{it}.$$

Transformations and re-parameterizations such as these can be used to put any output distance function that is linear in parameters into the canonical form of equation (3).

When there are multiple outputs then those that appear on the right hand side must be instrumented.

We will use this equation as the generic model for estimating efficiency change using frontier methods we will detail below. We will assume that technical innovations are available to all countries and interpret any country-specific error left over when we control for factor inputs as inefficiency. In so doing we can then decompose *TFP* growth into its two main components, innovation and catch-up. Innovation could be directly measured, for example using a distributed lag of R&D expenditures, patents, or any other direct measure of innovation. Baltagi and Griffin (1988) use time dummies to construct an innovation index. Exogenous or stochastic linear time trends have also been used (Bai, Kao, and Ng, 2009).

Below we examine a number of regression-based methods introduced into the literature to measure productivity growth and its decomposition into innovation and efficiency change using,

$$y_{it} = x_{it}\beta + \eta_i(t) + v_i,$$

which nests all multi-output/multi-input panel models that are linear in parameters and can be used to estimate productivity growth and decompose it into innovation and efficiency change. We will also assume that we have a balanced panel although this is done more for notational convenience than for substantive reasons. The methods we discuss also are appropriate when technical efficiency effects are not changing over time. After discussing the methods and how they are implemented we will discuss model averaging and how it can be used to evaluate world productivity growth from 1970 to 2000.

### 17.3.1 The Cornwell, Schmidt, and Sickles (1990) Panel Stochastic Frontier Model

Extensions of the panel data model to allow for heterogeneity in slopes as well as intercepts by Cornwell, Schmidt, and Sickles (CSS) (1990) allowed researchers to estimate productivity change that was specific to the cross-sectional unit (firms, industries, countries) that could change over time. A special parameterization of the CSS model that accomplishes this objective is:

$$y_{it} = x_{it}\beta + \eta_i(t) + v_{it},$$

where

$$\eta_i(t) = W_{it}\delta_i + v_{it}.$$

The  $L$  coefficients of  $W$ ,  $\delta_i$ , depend on different units  $i$ , representing heterogeneity in slopes. In their application to the US commercial airline industry, CSS specified  $W_{it} = (1, t, t^2)$ , although this was just a parsimonious parameterization useful for their application. It does not in general limit the effects to be quadratic in time.

A common construction can relate this model to standard panel data model. Let  $\delta_0 = E[\delta_i]$ , and  $\delta_i = \delta_0 + u_i$ . Then the model can be written as:

$$y_{it} = X_{it}\beta + W_{it}\delta_0 + \epsilon_{it}, \quad (4)$$

$$\epsilon_{it} = W'_{it}u_i + v_{it}. \quad (5)$$

Here  $u_i$  are assumed to be *i.i.d.* zero mean random variables with covariance matrix  $\Delta$ . The disturbances  $v_{it}$  are taken to be *i.i.d.* random variable with a zero mean and constant variance  $\sigma^2$ , and uncorrelated with the regressors and  $u_i$ . In matrix form, we have:

$$y = X\beta + W\delta_0 + \epsilon, \quad (6)$$

$$\epsilon = Qu + v, \quad (7)$$

where  $Q = \text{diag}(W_i)$ ,  $i = 1, \dots, N$  is a  $NT \times NL$  matrix, and  $u$  is the associated  $NL \times 1$  coefficients vector.

### 17.3.1.1 Implementation

Three different estimators can be derived based on differing assumptions made in regard to the correlation of the efficiency effects and the regressors, specifically, the correlation between the error term  $u$  and regressors  $X$  and  $W$ . They are the *within* (FE) estimator, which allows for correlation between all of the regressors and the effects, the *gls* estimator, which is consistent when no correlation exists between the technical efficiency term and the regressors (Pitt and Lee 1981; Kumbhakar 1990), and the *efficient instrumental variables* estimator, which can be obtained by assuming orthogonality of some of the regressors with the technical efficiency effects. The explicit formulas for deriving each estimator and methods for estimating the  $\delta_i$  parameters are provided in the CSS paper. Relative efficiencies, normalized by the consistent estimate of the order statistics identifying the most efficient country, are then calculated as:

$$\widehat{\eta}(t) = \max_j [\widehat{\eta}_j(t)]$$

and

$$RE_i(t) = \widehat{\eta}(t) - \widehat{\eta}_i(t),$$

where  $RE_i(t)$  is the relative efficiency of the  $i$ th country at time  $t$ . For this class of models the regressors  $X$  contain a time trend interpreted as the overall level of innovation. When it is combined with the efficiency term  $\widehat{\eta}_j(t)$ , we have a decomposition of TFP into innovation and catch-up. When the time trend and the efficiency term both enter the model linearly, then the decomposition is not identified using the within estimator but is for the *gls* and for selected variants of the efficient IV model, such as those used in the Cornwell et al. airline study. In the study of world productivity below we utilize the *gls* version of the CSS estimator (labelled CSSG) and the Efficiency IV estimator (labelled EIV).

### 17.3.2 The Kumbhakar (1990) Panel Stochastic Frontier Model

Here we consider a linear in log production function:

$$y_{it} = x_{it}\beta + \eta_i(t) + v_{it}, \quad (8)$$

where

$$\eta_i(t) = \gamma(t)\tau_i. \quad (9)$$

$v_{it}$  is assumed *i.i.d.* with distribution  $N(0, \sigma_v^2)$ ;  $\eta_i(t)$  is the inefficiency term with time-varying factor  $\gamma(t)$  and time-invariant characteristics  $\tau_i$ .  $\tau_i$  is assumed to be distributed as *i.i.d.* half-normal distributed  $\gamma(t)$  is specified as the logistic function

$$\gamma(t) = (1 + \exp(bt + ct^2))^{-1}$$

We can see that  $\gamma(t)$  is bounded between (0, 1) and that it accommodates increasing, decreasing, or time-invariant inefficiency behavior as the parameters  $b$  and  $c$  vary. Although the Kumbhakar model also estimates allocative efficiency from side conditions implied by cost-minimization (Schmidt and Lovell 1979), we will only examine the portion of his model that directly pertains to the technical inefficiency/innovation decomposition of productivity change.

#### 17.3.2.1 Implementation

Parametric maximum likelihood is used for estimation the model. Using the Kumbhakar notation let  $\theta_{it} = \gamma(t)\tau_i + v_{it}$ . Then the joint distribution of the composed error is  $f(\theta_i, \tau_i)$  and since both  $\tau_{it}$  and  $v_{it}$  are i.i.d and are independent of each other, the joint pdf is  $f(\theta_i, \tau_i) = f(\tau_i) \cdot (\prod_t f(v_{it})) = f(\tau_i) \prod_t f(\theta_{it} - \gamma(t)\tau_i)$ .

Marginalizing over  $\tau$ , one can derive the distribution of  $\theta$ . The the log-likelihood function is then defined as

$$\mathcal{L} = \sum_i \ln f(\theta_i)$$

and the parameters are given by the  $\arg \max(\mathcal{L})$ .

Consistent point estimates of the inefficiency term can be based on a method of moments estimator for the conditional mean of  $\tau_i|\theta_i$ . Since

$$f(\tau_i|\theta_i) = (2\pi\sigma_*^2)^{-1/2} \frac{\exp(-\frac{1}{2\sigma_*^2}(\tau - \mu_i^*)^2)}{\Phi(-\mu_i^*/\sigma_*)}, \tau_i \leq 0,$$

where  $\Phi$  is the distribution function for standard normal then  $E(\tau_i|\theta_i) = \mu_i^* - \sigma_* \frac{\phi(\mu_i^*/\sigma_*)}{\Phi(-\mu_i^*/\sigma_*)}$  and  $\widehat{E(\tau_i|\theta_i)} = \widehat{\tau}_i$ .  $\sigma_* = \frac{\sigma_v \sigma_\tau}{(\sigma_v^2 + \sigma_\tau^2 \sum_t \gamma^2(t))^{1/2}}$  and  $\mu_i^* = \frac{\sigma_\tau^2 \sum_t \gamma(t)\theta_{it}}{\sigma_v^2 + \sigma_\tau^2 \sum_t \gamma^2(t)}$ . The best predictor of technical efficiency is given by  $E(\exp\{\gamma(t)\tau_i|\theta_i\})$  and efficiency for each firm by  $\widehat{\eta}_i(t) = \gamma(t)\widehat{\tau}_i$ .

### 17.3.3 The Battese and Coelli Model (1992, 1995)

The production function is given by the generic model

$$y_{it} = x_{it}\beta + \eta_i(t) + v_{it}, \quad (10)$$

where the effects are specified as

$$\eta_i(t) = -\{\exp[-\eta(t-T)]\}u_i,$$

where  $v_{it}$  are assumed to be a *i.i.d.*  $N(0, \sigma_v^2)$  random variable and the  $u_{it}$  are assumed to follow an *i.i.d.* non-negative truncated  $N(\mu, \sigma^2)$  distribution.  $\eta$  is a scalar and the temporal movement of the technical efficiency effects depends on the sign of  $\eta$ . Time invariant technical efficiency corresponds to  $\eta = 0$ . To allow for a richer temporal path for firm efficiency effects that reflect more possibility of how firm effects change over time, one can also specify  $\eta(t-T)$  as

$$\eta_t(t-T) = 1 + a(t-T) + b(t-T)^2,$$

which permits the temporal pattern of technical efficiency effects to be convex or concave rather than simply increasing or decreasing at a constant rate.

#### 17.3.3.1 Implementation

The model is:

$$y_{it} = x_{it}\beta + \eta_i(t) + v_{it} \quad (11)$$

$$\eta_i(t) = e^{-\eta(t-T)}u_i, \quad (12)$$

where the  $u_i$ 's are assumed to follow the non-negative truncated  $N(\mu, \sigma^2)$  distribution whose density is

$$f_{U_i}(u_i) = \frac{\exp[-\frac{1}{2}(u_i - \mu)^2/\sigma^2]}{(2\pi)^{1/2}\sigma[1 - \Phi(-\mu/\sigma)]}, \quad u_i \geq 0$$

and where  $\Phi$  is the cumulative distribution function of the standard normal random variable. The  $v_{it}$ 's are assumed *i.i.d.*  $N(0, \sigma_v^2)$  and are independent of the  $u_i$ 's. Let  $y_i$  be the  $(T_i \times 1)$  vector of production level of firm  $i$ , and denote  $y = (y'_1, y'_2, \dots, y'_N)$ . Then the density function of  $y_i$  can be easily derived from the density of  $\epsilon_i$  and log-likelihood function  $L(\beta, \sigma_v^2, \sigma^2, \mu, \eta; y; x)$  for the model is given in Battese and Coelli (1992).

The minimum-mean-squared-error predictor of the efficiency for country (firm)  $i$  at time  $t$  is

$$E[\exp(-u_i t) | \epsilon_i] = \left\{ \frac{1 - \Phi[\eta_i \sigma_i^* - (\mu_i^*)/\sigma_i^*]}{1 - \Phi(-\mu_i^*/\sigma_i^*)} \right\} \exp\left[-\eta_i \mu_i^* + \frac{1}{2} \eta_i^2 \sigma_i^{*2}\right]$$

where  $\mu_i^* = \frac{\mu \sigma_v^2 - \eta_i' \epsilon_i \sigma^2}{\sigma_v^2 + n_i' n_i \sigma^2}$  and  $\sigma_i^* = \frac{\sigma^2 \sigma_v^2}{\sigma_v^2 + \eta_i' \eta_i \sigma^2}$ . Estimates of technical change due to innovation would be based on the coefficient of a time trend in the regression. The

effect of innovation as distinct from catch-up is identified by the non-linear time effects in the linear technical efficiency term and thus the decomposition of *TFP* growth into a technological change and efficiency change component is quite natural with this estimator. Cuesta (2000) generalized Battese and Coelli (1992) by allowing each country (firm, etc.) to have its own time path of technical inefficiency. Extensions of the Battese and Coelli model that allow for technical inefficiency to be determined by a set of environmental factors that differ from those that determine the frontier itself are given in Battese and Coelli (1995). These were also addressed by Reifschneider and Stevenson (1991) and by Good, Roeller, and Sickles (1995). Environmental factors that were allowed to partially determine the level of inefficiency and productivity were introduced in Cornwell, Schmidt, and Sickles (CSS) (1990) and in Good, Nadiri, Roeller, and Sickles (1993).

### 17.3.4 The Park, Sickles, and Simar (1998, 2003, 2006) Models

Park, Sickles, and Simar (PSS; 1998, 2003, 2007) considered linear stochastic frontier panel models in which the distribution of country-specific technical efficiency effects is estimated nonparametrically. They used methods developed in the statistics literature to estimate robust standard errors for semi-nonparametric models based on adaptive estimation techniques for semiparametric efficient estimators. They first consider models in which various types of correlations exist between the effects and the regressors (PSS 1998). These minimax-type estimators ensure that the variances of the estimators are the smallest within the set of variances based on the class of parametric sub-models built up from the basic parametric assumptions of the model and the use of nonparametric estimators (they utilize multi-variate kernel-based estimators) for the remaining portion of the model specified in terms of nuisance parameters. The nuisance parameters are the effects, the variances of the parametric disturbance terms, and the bandwidth parameters. In PSS (2003) they extend the basic model to consider serially correlated errors and in PSS (2006) consider dynamic panel data models. In our discussion of this class of estimators we will only consider the most basic model set up in PSS (1998). Details for the semiparametric efficient panel data efficiency model with serially correlated errors or with a dynamic structure can be referred to PSS (2003, 2006). In the empirical application to estimate world productivity growth we utilize three of the estimators discussed by PSS. PSS1 is the estimator outlined above. PSS2W is the within version of the semiparametric efficient estimator with serially correlated errors to control for potential misspecified dynamics while PSS2G is the corresponding random effects version of the estimator (PSS 2003).

The basic set up of the model is again the canonical linear panel data with cross-sectionally and time varying efficiency effects given by

$$y_{it} = X'_{it}\beta + \eta_i(t) + \nu_{it},$$

where  $v_{it}$ 's are the statistical noise that are independently and identically distributed with  $N(0, \sigma^2)$ ,  $\eta_i(t)$  are bounded above (or below for the cost frontier model). The  $(\eta_i(t), X_i)$ 's are assumed to be independently and identically distributed with some joint density  $h(\cdot, \cdot)$ .

PSS (1998) discuss three different cases for the dependency between the firm effects,  $\eta$ , and the other regressors,  $X$ . Their case 1 assumes no specific pattern of dependency between  $\eta$  and  $X$ , which leads to a semiparametric efficient estimator similar to the fixed effects estimator of Schmidt and Sickles (1984) and its extension to time-varying efficiency models of CSS (1990). Their case 2 assumes the firm effects are correlated with a subset of other explanatory variables,  $Z \in X$ . Case 3 assumes that  $\alpha$  affects  $Z$  only through its long run changes (average movements)  $\bar{Z}$ . The semiparametric efficient estimators of case 2 and 3 are analogous to those proposed in Hausman and Taylor (1981), and extended to the stochastic frontier literature in CSS (1990). Derivations of the semiparametric efficiency bound and of the adaptive estimators for these many specifications are too detailed for this chapter. The interested reader is referred to the PSS papers referenced above for details. Once the parameters have been estimated the method utilized in CSS (1990) can be used to estimate the technical efficiencies and their temporal changes. As with the CSS estimator, innovation change that is shared by all countries or firms and modeled using a time trend may be identified separately from technical efficiency based on the orthogonality conditions imposed in cases 2 and 3. In case 1, which collapses to the CSS within estimator, no such distinction can be made and, although TFP growth can be calculated, it cannot be decomposed into innovation and catch-up.

### 17.3.5 The Latent Class Models of Greene, Kumbhakar, and Tsionas

In stochastic frontier models the production or cost functional relationship is usually set uniformly for all countries or firms, implying that the same technology is used as the benchmark and that relative to that benchmark countries or firms perform with different levels of efficiency. Although other authors have questioned this assumption and have provided estimators that address this issue in part, Orea and Kumbhakar (2004), Tsionas and Kumbhakar (2004), Greene (2005b), were the first to address it in such a general fashion. Their logic is clear and the arguments compelling and relate to work on production heterogeneity by Mundlak (1961, 1978) and Griliches (1979), among others. Countries that have access to the world technology, or firms within a certain industry, have different sizes, innovation abilities, targeting groups, etc., and may operate with different technologies that can take advantage of different market niches. Imposing the same functional form in the model may misidentify differences in the technology applied as technical inefficiency when it fact in is due to the appropriate

use of the available technology to a different (or constrained) set of market conditions. We have discussed this issue in earlier sections. Such constrained conditions are nonetheless suboptimal to the benchmark we establish and estimate and the technical efficiency component of *TFP* growth remains silent on the source of the technical inefficiency. That said, it is important to find a way to empirically parse the sources of variation into one may regard is or is not technical inefficiency. A straightforward way to deal with this problem is to group countries or firms into different categories by some obvious criteria and then analyze their *TFP* growth separately. We do this below in our empirical analysis of world productivity growth. In general the grouping criteria can be information about certain characteristics of countries (e.g., region, level of development, etc., or some combination of these and many other characteristics) or firms (e.g., size, location etc.), or can be based on some statistical clustering algorithm. Were these analyses to be done separately, then it is clear that information represented by correlation between different groups would not be utilized. It may also be the case that the parameters of such models can not be identified by distinct categories and thus the suitability of the grouping criteria cannot be established empirically.

In the latent class stochastic frontier model there exist  $J$  unobserved classes in the panel data giving rise to a specification of the production (or distance) function as:

$$y_{it} = x'_{it}\beta_j + \eta_i(t)|j.$$

The observed dependent variable is characterized by a conditional density function:

$$g(y_{it}|x_{it}, \text{class}_j) = f(\Theta_j, y_{it}, x_{it}).$$

The functional form  $f(\cdot)$  is the same over the entire sample, while the parameter vector  $\Theta_j$  is class-specified and contains all of the parameters of the class specific parameterization of the function. The inefficiency terms are latent class-specific and take the form of  $\eta_i(t)|j$  and are assumed to distributed as half normal. The likelihood coming from country or firm  $i$  at time period  $t$  is

$$P(i, t|j) = f(y_{it}|x_{it}; \beta_j, \sigma_j, \lambda_j) = \frac{\Phi(\lambda_j \eta_i(t)|j)}{\Phi(0)} \frac{1}{\sigma_j} \phi\left(\frac{\eta_i(t)|j}{\sigma_j}\right),$$

where  $\eta_i(t)|j = y_{it} - x'_{it}\beta_j$ . Assuming the inefficiency terms to be i.i.d. draws over time the conditional likelihood for country or firm  $i$  is

$$P(i|j) = \prod_{t=1}^T P(i, t|j)$$

and the unconditional likelihood function is

$$P(i) = \sum_{j=1}^J \Pi(i, j) P(i|j) = \sum_{j=1}^J \Pi(i, j) \prod_{t=1}^T P(i, t|j).$$

Here  $\Pi(i,j)$  is a prior probability that establishes the distribution of firms in different classes. A relatively simple and noninformative prior is the uniform where  $\Pi(i,j) = \Pi(j)$ , for  $i = 1, \dots, N$ . In order to allow for heterogeneity in the mixing probabilities one can adopt the multinomial logit form,

$$\Pi(i,j) = \frac{\exp(\theta_i' \pi_j)}{\sum_{m=1}^J \exp(\theta_i' \pi_m)}, \quad \pi_J = 0.$$

The parametric log likelihood is

$$\log L = \sum_{i=1}^N \log P(i),$$

Although the specification outlined by Greene (2005b) assumed that inefficiency was independent over time, the latent class model proposed by Orea and Kumbhakar (2004) allows technical efficiency change over time by following a path given by an exponential function reminiscent of earlier estimators by Battese and Coelli (1992) and Kumbhakar (1990)

$$\eta_i(t)|j = \gamma_{it}(\eta_j) \cdot \zeta_{i|j} = \exp(z_{it}' \zeta_j) \cdot \zeta_{i|j}.$$

where  $z_{it} = (z_{1it}, \dots, z_{Hit})'$  is a vector of time-varying variables and  $\zeta_j = (\zeta_{1j}, \dots, \zeta_{Hj})'$  the associated parameters. With such a changing path, the individual likelihood in their model is defined directly over all time periods.

#### 17.3.5.1 Implementation

The parametric log likelihood function is maximized to solve for parameter vector  $\Theta_j$  and probability  $\pi_j$  simultaneously. Greene (2005b) employed an Expectation-Maximization (EM) algorithm. Alternatively, the model can also be estimated using Bayesian methods (see Tsionas and Greene, 2003b). Explicit derivations can be found in Greene (2005). After classifying firms into different groups, firm-specific parameters can be estimated. After the parameters of the underlying production or distance function are estimated and the time varying effects  $\eta_i(t)|j$  are identified for class  $j$  the decomposition of TFP into an innovation change component and catch-up or technical efficiency component is complete.

### 17.3.6 The Kneip, Sickles, and Song (2012) Model

Here we assume a linear semiparametric model panel data model that allows for an arbitrary pattern of technical change  $\eta_i(t)$  based on a factor model. The model takes the form

$$y_{it} = \beta_0(t) + \sum_{j=1}^p \beta_j x_{itj} + \eta_i(t) + v_{it}.$$

Here the  $\eta_i(t)$ 's are assumed to be smooth time-varying individual effects and identifiability requires that  $\sum_i \eta_i(t) = 0$ .  $\beta_0(t)$  is some average function (or common factor) shared by all of the cross-sectional units, such as countries or firms. For purposes of developing the estimator of  $\eta_i(t)$  we eliminate the common factor. However, once we have estimated the  $\beta_j$ 's and the  $\eta_i(t)$  terms, we can recover the common factor. For purposes of using this model as a vehicle for estimating *TFP* growth, the common factor will identify the common innovation that changes over time, while the  $\eta_i(t)$  term will, after the suitable normalization developed above for the CSS counterpart, provide us with relative efficiency levels and thus their growth rates to allow for *TFP* growth to be decomposed into its two constituent parts, innovation change and technical efficiency change. The centered form of the model is

$$y_{it} - \bar{y}_t = \sum_{j=1}^p \beta_j (x_{itj} - \bar{x}_{tj}) + \eta_i(t) + v_{it} - \bar{v}_i,$$

where  $\bar{y}_t = \frac{1}{n} \sum_i y_{it}$ ,  $\bar{x}_{tj} = \frac{1}{n} \sum_i x_{itj}$  and  $\bar{v}_i = \frac{1}{T} \sum_t v_{it}$ . Here  $\eta_i(t)$  is assumed to be a linear combination of a finite number of basis functions

$$\eta_i(t) = \sum_{r=1}^L \theta_{ir} g_r(t).$$

This construction is more flexible and realistic than parametric methods, which presume the change of individual effects follow some specified functional form, and the multiplicative effects models of Lee and Schmidt (1993), Ahn, Lee, and Schmidt (2007), Bai (2009), and Bai and Ng (2011). The model can be rewritten as

$$y_{it} - \bar{y}_t = \sum_{j=1}^p \beta_j (x_{itj} - \bar{x}_{tj}) + \sum_{r=1}^L \theta_{ir} g_r(t) + v_{it} - \bar{v}_i.$$

The authors introduce a suitable standardization to identify a specific basis they use in their model which results in a set of  $g_r$ 's that are orthogonal and  $\theta_{ir}$ 's that are empirical uncorrelated. Letting  $\eta_1 = (\eta_1(1), \dots, \eta_1(T))'$ , ...,  $\eta_n = (\eta_n(1), \dots, \eta_n(T))'$ , then the empirical covariance matrix of  $\eta_1, \dots, \eta_n$  is  $\Sigma_{n,T} = \frac{1}{n} \sum_i \eta_i \eta_i'$ . Let  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_T$  be the eigenvalues of the matrix, and  $\gamma_1, \gamma_2, \dots, \gamma_T$  be the corresponding eigenvectors. Then the basis functions will be

$$g_r(t) = \sqrt{T} \cdot \gamma_{rt} \quad \text{for all } r = 1, \dots; t = 1, \dots, T \quad (13)$$

$$\theta_{ir} = \frac{1}{T} \sum_t v_i(t) g_r(t) \quad \text{for all } r = 1, \dots; i = 1, \dots, n \quad (14)$$

$$\gamma_r = \frac{T}{n} \sum_i \theta_{ir}^2 \quad \text{for all } r = 1, \dots \quad (15)$$

And for all  $l = 1, 2, \dots$

$$\begin{aligned} \sum_{r=l+1}^T \gamma_r &= \sum_{i,t} (\eta_i(t) - \sum_{r=1}^l \theta_{ir} g_r(t))^2 \\ &= \min_{\tilde{g}_1, \dots, \tilde{g}_l} \sum_i \min_{\tilde{\vartheta}_{il}, \dots, \tilde{\vartheta}_{il}} \sum_t (\eta_i(t) - \sum_{r=1}^l \vartheta_{ir} \tilde{g}_r(t))^2. \end{aligned} \quad (16)$$

$\eta_i(t) \approx \sum_{r=1}^l \theta_{ir} g_r(t)$  will be the best  $l$ -dimensional linear estimate, and the dimension  $L$  naturally equals to  $\text{rank}(\Sigma_{n,T})$ . It can be shown that for selected values of  $L$  the normalizations imply basis functions that correspond to the standard fixed effect estimator, the CSS (1990) estimator and the Battese and Coelli (1992) estimator. Kneip, Sickles, and Song (2012) provide asymptotic results for large  $N$  and large  $T$ .

#### 17.3.6.1 Implementation

Since the  $\eta_i$ 's are assumed to be smooth trends, we can always find  $m$ -times continuously differentiable auxiliary functional variable  $v_i$ 's with domain  $[1, T]$  that can interpolate the  $T$  different values of  $\eta_i$ . Their method first estimates  $\beta$  and obtains the approximations  $v_i$  by smoothing splines (Eubank 1988). This then determines the estimates of the basis functions  $\hat{g}_r$  through the empirical covariance matrix  $\hat{\Sigma}_{n,T}$ , which is estimated by the  $(\hat{\eta}_1, \dots, \hat{\eta}_n) = (\hat{v}_1, \dots, \hat{v}_n)$ . The corresponding coefficients of the basis functions will be obtained by least squares. In the last step, they update the estimate of  $\eta_i$  by  $\sum_{r=1}^L \hat{\theta}_{ir} \hat{g}_r$ , which is proved to be more efficient than the approximations  $v_i$ . Returning to the non-centered model, the general average function  $\beta_0(t)$  is left unestimated. A non-parametric method similar to step 1 can be applied to get an approximation. An alternative is to assume  $\beta_0(t)$  also lies in the space spanned by the set of basis functions, that is,  $\beta_0(t) = \sum_{r=1}^L \bar{\theta}_r g_r(t)$ . The coefficients can then be estimated by a similar minimization problem as step 3 with objective function  $\sum_t (\bar{y}_t - \sum_{j=1}^p \hat{\beta}_j \bar{x}_{tj} - \sum_{r=1}^L \vartheta_r \hat{g}_r(t))^2$ . The common factor  $\beta_0(t)$  is interpreted as the shared technological innovation component and the  $\eta_i(t)$  the technical efficiency component whose growth constitute TFP growth.

#### 17.3.7 Ahn, Lee, and Schmidt (2013)

##### 17.3.7.1 Model

Ahn, Lee, and Schmidt (2013) generalize Ahn, Lee, and Schmidt (2007) and consider a panel data model with multiple individual effects that also change over time:

$$y_{it} = x'_{it} \beta + \sum_{j=1}^p \xi_{tj} \alpha_{ij} + \epsilon_{it}. \quad (17)$$

They focus on large  $N$  and finite  $T$  asymptotics. They develop a consistent estimator for the slope coefficients  $\beta$  when there is correlation between individual effects and the regressors. To emphasize this feature, the model interprets  $\xi_{it}$  as “macro shocks,” and  $\alpha_{ij}$  as “random coefficients” instead of “factors” and “factor loadings,” though the model itself resembles the factor models. This model takes the form of the canonical model considered above by other researchers as it can be written as

$$y_{it} = X'_{it}\beta + \eta_i(t) + \nu_{it}.$$

The model for individual  $i$  in matrix form is:

$$y_i = X_i\beta + u_i, \quad u_i = \eta_i + \epsilon_i = \Theta\alpha_i + \epsilon_i, \quad (18)$$

where  $y_i = (y_{i1}, \dots, y_{iT})'$  is the dependent variable vector,  $X_i = (x_{i1}, \dots, x_{iT})'$  is the  $T \times K$  matrix of regressors, and  $\beta$  is the dimension-comformable coefficients vector. The error term  $u_i$  is composed of the random noise  $\epsilon_i$  and individual effects  $\eta_i = \Theta\alpha_i$ .  $\Theta$  is a  $T \times p$  ( $T > p$ ) matrix containing  $p$  macro shocks that vary over time. The random noise  $\epsilon_{it}$  is usually assumed to be white noise to assure consistent estimates of coefficients in the case of large  $N$  and small  $T$ . This model relaxes this assumption in that it allows any kind of autocorrelation of  $\epsilon_i$  and only assumes that  $\epsilon_i$  is uncorrelated with regressors  $x_{it}$  while  $\alpha$  might be correlated with  $x_{it}$ . Then for identification, it is assumed there exist instrument variables that are correlated with  $\alpha_{ij}$  but not with  $\epsilon_{it}$ .

#### 17.3.7.2 Implementation

Due to the need for a particular rotation, it is not possible to separate the effects of  $\Theta$  and  $\alpha$ . For identification,  $\Theta$  is normalized such that  $\Theta = (\Theta'_1, -I_p)'$  with  $\Theta_1$  a  $(T - p) \times p$  matrix. With instruments, the GMM method proposed in Ahn et al. (2001) is extended to incorporate multiple time-varying effects and two methods are proposed to estimate the true number of individual effects. They first obtain consistent estimators of  $\beta$  and  $\Theta$  assuming the true number of effects  $p_0$  is known, and then estimate  $p$  using their new test statistic. Detailed assumptions and discussion can be found in the paper as well as how to extract the efficiency and innovation change measures for productivity measures.

#### 17.3.8 Additional Panel Data Estimators of in the Stochastic Frontier Literature

Space limits the possibility of dealing with the many other approaches that have been proposed to estimate the panel stochastic frontier and provide a decomposition of TFP growth into innovation and catch-up, or technical efficiency. Additional estimators that have been proposed for panel stochastic frontiers and that are also quite appropriate for general panel data problems are the *Bayesian Stochastic Frontier Model* (Liu,

Sickles, and Tsionas 2013), which builds on earlier work by Van den Broeck et al. (1994) and Tsionas (2006), the *Bounded Inefficiency Model* of Almanidis, Qian, and Sickles (2013) and related models of Lee (1996), Lee and Lee (2012), and Orea and Steinbuks (2012), and the “*True*” *Fixed Effects Model* of Greene (2005a,b). Kumbhakar, Parmeter, and Tsionas (2013) have recently considered a semiparametric smooth coefficient model to estimate the *TFP* growth of certain production technologies that addresses the *Skewness Problem* in classical SFA modeling considered by Feng, Horrace and Wu (2012), Almanidis and Sickles (2012) and Almanidis, Qian, and Sickles (2013). Recent work on spatial heterogeneity in SFA models has focused on new interpretations and measurement of spillovers in substitution possibilities, returns to scale, productivity change, and efficiency change that is spatially dimensioned instead of simply varying over time for particular firms, industries, or countries. The *Spatial Stochastic Frontier* shows great promise and has been pursued in recent work by Glass, Kenjegalieva, and Sickles (2013 a, b) based on the original contribution by Druska and Horrace (2004). Work on productivity measurement in the presence of spatial heterogeneity has also recently been pursued by Mastromarco and Shin (2013), Entur and Musolesi (2013), and Demetrescu and Homm (2013). Such spatial methods are alternatives to less structured approaches to address cross-sectional dependence in panel data models using methods such as those developed by Pesaran (2007). *Factor Models* continue to be pursued in the context of productivity modeling in panel data contexts and the space for such approaches is getting quite dense as pointed out by Kneip and Sickles (2012).

## 17.4 DISCUSSION ON COMBINING ESTIMATES

---

A solution to model uncertainty is to develop a consensus estimate by weighting or combining in some fashion estimates from various competing models. Sickles (2005) pursued this strategy in his examination of semiparametric and nonparametric panel frontier estimators. Burnham and Anderson (2002) provided a lucid and rather complete discussion of model selection criteria. However, they point out that the model selection exercise itself introduces uncertainty into the estimation process and any forecasts that result, a point also made by Hjorth (1994) and Leeb and Potscher (2005). As one can view all models as approximations and thus subject to misspecification, combining results from different models can be viewed as similar to constructing a diversified portfolio in order to reduce the risk of relying on on particular stock, of in our case, model.

Typical model selection from some encompassing supermodel can be viewed as a special case of weighting models which assigns the entire weight on one model and none on others. We do not pursue this approach in our empirical work below. Instead we utilize Insights from economics and from statistics to motivate several canonical methods to combine estimates and forecasts from a variety of potentially misspecified models. These insights are discussed in more depth in the cited works. We utilize

approaches to weighting outcomes from different models and estimators using the economic arguments of *majority voting* from the literature on social choice theory (see Moulin 1980) as well as the *contest function* of Tullock (1980); insights from statistics based on *model averaging*<sup>4</sup> in order to assess the proper weights to construct the weighted average; and the literature on optimal weights used in *combining forecasts*<sup>5</sup>. These studies provide the rationale for how we combine our many results into summary measures of weighted means and variances.

## 17.5 MODELING WORLD ECONOMIC GROWTH WITH THE UNIDO DATA

---

The proper measurement of nations' productivity growth is essential to understand current and future trends in world income levels, growth in per/capita income, political stability, and international trade flows. In measuring such important economic statistics it is also essential that a method that is robust to misspecification error is used. This section of the chapter addresses the robustification of productivity growth measurement by utilizing the various economic theories explaining productivity growth as well as various estimators consistent with those particular theories. We utilize the World Productivity Database from the UNIDO to analyze productivity during the period 1970–2000 and combine and consolidate the empirical findings from a number of the statistical treatments and various economic models of economic growth and productivity that we have discussed above.

We address the heterogeneity problem in part by grouping countries according to their geographical and, for the OECD countries, their development characteristics as well as by the use of various panel data techniques. We construct consensus estimates of world productivity *TFP* growth as well as confidence intervals and find that, compared to efficiency catch-up, innovation plays a much more important factor in generating *TFP* growth at this level of country aggregation.

### 17.5.1 UNIDO Data Description

The World Productivity Database (WPD) provides information on measures of the level and growth of *TFP* based on 12 different empirical methods across 112 countries over the period 1960–2000. Those interested in the data and variable construction should visit the UNIDO website <http://www.unido.org/statistics.html>. In our analysis we utilize two factor (capital and labor) aggregate production function determining a country's level of aggregate output.

### 17.5.2 Empirical Findings

Comparisons of productivity changes are made among Asian, Latin American, and OECD regions. The following methods are used to estimate *TFP* change and its decomposition into technological and technical efficiency change when possible: CSSG, EIV, BC, PSS1, PSS2W, PSS2G, two fixed-effect estimators and two random-effect estimators.

There are 10 different methods to estimate *TFP* growth and 6 different methods to estimate the decomposition of *TFP* growth into innovation and technical efficiency change. Data limitations forced us to use only three of the four possible capital measures, K06, Keff, and Ks along with the two labor measures, LF and EMP as well as data only from 1970 to 2000. The results are based on 60 different sets of estimates. The panel estimators are used to estimate productivity growth and its decomposition methodologies for countries in Asia (13 countries), Latin America (12 countries), and the OECD (24 countries). The specific countries in Asia are: Bangladesh, China, Hong Kong (SAR of China), India, Indonesia, Israel, Malaysia, Pakistan, Philippines, Singapore, Sri Lanka, Taiwan (Province of China), and Thailand. The countries in Latin America are: Argentina, Brazil, Chile, Colombia, Ecuador, Mexico, Guatemala, Jamaica, Panama, Peru, Trinidad and Tobago, and Venezuela. Finally, the countries in the OECD are: Australia, Austria, Belgium, Canada, Denmark, Finland, France, Greece, Iceland, Ireland, Italy, Japan, Republic of Korea, Luxembourg, Netherlands, New Zealand, Norway, Portugal, Spain, Sweden, Switzerland, Turkey, United Kingdom, and United States.

Our approach considers a Cobb-Douglas production function with two explanatory variables: Capital and Labor. The various measures we adopt to measure the two inputs are largely based on data limitations. K06 and K013 utilize a perpetual inventory method to measure capital services and differ based on differing but constant depreciation rates (6% and 13.3%, respectively, which correspond to about 12 and 6 year asset lives). A different way of measuring capital focuses on the profile of capital productivity and utilizes a time-varying depreciation rate. As the asset ages, its capital declines at an increasing rate. This leads to Keff. Labor input measurement involves two kinds of labor utilization rates for which labor force (LF) can be adjusted, variations in numbers employed and in hours worked. Again, for reasons of data limitations we use the second adjustment and also consider employment (EMP). Thus each region has 6 combinations of inputs. In addition to the 6 we have discussed above, we also include four simple panel data estimators (FIX1 is a fixed effect model including  $t$  as explanatory variable, FIX2 is a fixed effect model with  $t$  and  $t^2$  as explanatory variables. RND1 is a random effect model including  $t$  as explanatory variable, RND2 is a random effect model with  $t$  and  $t^2$  as explanatory variables). The estimation results (Table 1) are too numerous to include in this chapter and are available on the Sickles website at <http://rsickles.blogs.rice.edu/files/2014/04/Figures-and-Tables.pdf> as are summary results in Figures 1–8 referred to below.

We decompose *TFP* into technical efficiency change and innovation or technical change. Technical efficiency for each country is defined as the radial distance from the (possibly shifting) production frontier in a given period (Debreu 1951; Farrell 1957). The estimation methods for this component have been included in all standard stochastic frontier literature. Results are presented in Figure 1 for each of the three regions. We summarize the outcomes of technical efficiency by three different averages. The first two methods are simple average and geometric average. Since countries have different GDP sizes, instead of simply averaging in each period, it is natural to weigh the results by each country's GDP. The traditional fixed effect model and the random effect model do not estimate technical efficiency, therefore, there are 6 models for technical efficiency change in each region. From the figures, it is apparent that Asian countries' technical efficiency improvements have been on a decreasing trend since the late 1970s. Latin American countries' technical efficiency changes have been very small in magnitude. OECD countries' technical efficiency improvements increased until the mid- to late 1980s then started to decline. In the Asian countries, GDP weighted averages are somewhat larger than simple averages, which indicates that larger GDP countries (particularly China) have more technical improvements than smaller GDP countries. For OECD countries we have the opposite observation, which indicates smaller GDP countries on average have more technical efficiency improvements than larger GDP countries (such as the United States).

Technical innovation change is measured as the shift of the frontier between periods, or the time derivative of each model. In our study, we assume a constant rate of technological innovation, thus innovative progress is the coefficient of time variable. We have 60 estimates for each region as presented in Figure 2. Asian countries have the largest innovation changes among all regions on average, at around 1.56% per year. Latin American countries display very small magnitudes of innovation change. On average, the region has 0.3% increase of progress per year. OECD countries' average innovation improvement is about 0.73% per year.

*TFP* change is the sum of technical efficiency change and technical innovation change. As seen in Figure 3, Asian countries have the highest *TFP* improvements through the years, mainly because the innovation progress outperforms the declining trend of technical efficiency. Latin American countries have almost nonexistent improvements in productivity in most years. They even have negative *TFP* growth rates in a few years at the beginning and end of the sample period. OECD countries' *TFP* performances are between those of Asia and Latin America, although the trend has been decreasing throughout the periods. The overall *TFP* growth between 1972 and 2000 is 61.2% for Asian countries, 24.7% for OECD countries, and 7.46% for Latin American countries. We also used three averaging approaches to aggregate three regions to demonstrate the global trends of *TFP* growth, which are shown in Figure 4. These results appear to be comparable to other recent international studies based on index number approaches (Badunenko, Henderson, and Russell 2013).

Next we report the Solow Residual (hereinafter SR). The SR results based on GDP weighted growth rates across all the methods and combinations are presented in Figure 6. The average of SR is 0.78% for Asian countries, -0.07% for Latin American, countries and 0.37% for OECD countries. One of the major shortcomings of SR and growth accounting in general as pointed out by Chen (1997) is that the SR cannot differentiate disembodied technological change (similar to our definition of innovational progress) from embodied technological change (similar to our definition of efficiency change). Failure to separate different effects in addition to the input measurement problems makes *TFP* estimates using growth accounting somewhat difficult to interpret and decompose. We can decompose *TFP* into efficiency catch-up and innovation and provide a solution to this problem.

The last results we wish to discuss are the combined estimates (Figure 8). As discussed above, the motivation of employing a model averaging exercise is to obtain some consensus results based on all the competing models and data at hand. The simplest averaging is to take the arithmetic mean of all estimates, which implicitly assumes the equal importance of all models. The annual changes of technical efficiency, technical innovation, and *TFP* are -0.07%, 1.63%, and 1.56% for Asian countries, 0.01%, 0.24%, and 0.25% for Latin American countries, and -0.05%, 0.84% and 0.79% for OECD countries. The most crucial component of all “combining estimates methods” such as model averaging is how the weights are assigned. Besides simple averaging, we use four statistical criteria to assign weights. First, we simply assign weights according to R-square of each model. Since R-squares in our estimations are all close to each other, weighted results are very close to simple averaging results: technical efficiency, technical innovation, and *TFP* changes are -0.07%, 1.62%, and 1.55% for Asian countries, 0.02%, 0.22%, and 0.23% for Latin American countries, and -0.05%, 0.84%, and 0.79% for OECD countries. The second way is to set the weights as reciprocals of residual sum of squares (hereinafter RSS). RSS is a simple measure of how much the data are not explained by a particular model. Annual technical efficiency, technical innovation, and *TFP* changes are -0.04%, 1.52%, and 1.47% for Asia countries, 0.01%, 0.19%, and 0.20% for Latin American countries, and -0.04%, 0.75%, and 0.71% for OECD countries. The third method is to choose weights according to AIC. Since all the models in our study use the same variables on the same data set, we would have a simple expression of AIC, which only depends on RSS. So the results of the third method should be close to the second one. The annual technical efficiency, technical innovation, and *TFP* changes are -0.08%, 1.59%, and 1.52% for Asian countries, 0.02%, 0.18%, and 0.21% for Latin American countries, and -0.06%, 0.81%, and 0.75% for OECD countries. The last method is to use BIC as weights. BIC depends not only on RSS but also on the estimated variance of the error term. The annual technical efficiency, technical innovation, and *TFP* changes are -0.12%, 1.70% and 1.58% for Asian Countries, 0.01%, 0.20%, and 0.20% for Latin American Countries, and -0.15%, 0.88%, and 0.73% for OECD countries. As shown in the Figure 8, combined estimates of all criteria are rather similar. All of the methods we utilize tell us that

the during the 29 years span, the improvements of Asian countries and OECD countries' technical efficiencies are deteriorating. Even though Latin America countries have improved technical efficiency (very small in magnitude), because of its slower innovative progress, their *TFP* improvement has lagged behind not only Asian countries but also OECD countries. For inference purpose, the variances of combined estimates can also be calculated (Burnham et al. 2002; Huang and Lai 2012). Our results indicate significant positive *TFP* growth in Asian and the OECD while *TFP* growth in Latin America is not significantly different than zero.<sup>6</sup>

## 17.6 CONCLUSIONS AND SUGGESTIONS FOR FUTURE RESEARCH

---

In this chapter, we have focused on the role that panel data econometrics plays in formulating and estimating the most important contributors to productivity growth: innovation and catch-up. We have explained different theories on economic growth and productivity measurement and the econometric specifications they imply. Various index numbers and regression-based approaches to measuring productivity growth and its innovation and catch-up components have been discussed in detail. We have also discussed methods that can be used to combine results from the many different perspectives on how economic growth is modeled and estimated, focusing on methods used in model averaging and in the combination of forecasts. As this chapter is to provide the reader with an applied perspective, we have utilized these various panel data and model averaging methods in an analysis of world productivity growth using the World Productivity Database gathered by the United Nations Industrial Development Organization (UNIDO). We study Asian, Latin American, and OECD countries between 1970 and 2000 and find that Asian countries had the fastest *TFP* growth among the three regions, due largely to relatively rapid technical innovation. OECD countries made more moderate gains in *TFP* growth, again due largely to technical innovation as opposed to catch-up. Latin American countries overall had the slowest growth rate in *TFP*, although they had consistently managed positive improvements in both technical and technological efficiencies.

There are a number of research topics that we were not able to cover in this chapter. Allocative distortions as opposed to the radial technical inefficiency we have posited in our panel studies was not addressed, nor was the nascent literature on developing coverage intervals for relative efficiency levels and rankings of countries or firms. The models are of course linear and thus structural dynamic models that incorporate inefficiency as well as models that address, at a firm or industry level, the impact that deviations from neoclassical assumptions of perfectly coordinated allocations with no technical (or cost) inefficiency, may have on firm or industry level productivity has not been examined. These are areas for future research and we encourage those interested

in the intersection of more traditional productivity research, new productivity research that addresses imperfect decision making, and panel methods to pursue these topics.

## ACKNOWLEDGMENTS

---

The authors would like to thank Editor Badi Baltagi and two anonymous referees for their constructive suggestions that improved our chapter substantially. The usual caveat applies.

## NOTES

---

1. For a survey of some of Jorgenson's voluminous work on productivity, see Dale W. Jorgenson's *Productivity*, Vol. 1 and 2 (1995), Vol. 3 (2005).
2. For Griliches's work on this subject, the reader should consult the working papers of the NBER Productivity Program over the years before his untimely death in 1999, and over the years since. Mairesse (2003) contains a thoughtful overview of his many contributions to the field of productivity measurement.
3. The NBER Productivity, Innovation, and Entrepreneurship Program was led originally by Griliches who was followed by Ernst Berndt and is currently co-directed by Nicholas Bloom and Josh Lerner.
4. See, for example, Leeb and Potscher (2005), Buckland *et al.* (1997), Akaike (1973), Mallows (1973), Schwarz (1978), Hansen (2007), Carroll *et al.* (2006), Burnham and Anderson (2002), Claeskens and Hjort (2008), Raftery *et al.* (1997), Hoeting *et al.* (1999), and Koop *et al.* (2007), Timmermann (2006).
5. See, for example, Newbold and Harvey (2002), Bates and Granger (1969), Diebold and Lopez (1996), Lahiri, *et al.* (2011), Clemen (1989), Timmermann (2006), Lahiri and Shaeng (2010), Zarnowitz and Lambros (1987), Lahiri and Sheng (2010), Lahiri, Peng and Sheng. (2010), Davies and Lahiri (1995), and Lahiri, Teighland, and Zaprovski (1988).
6. A more detailed discussion of the Asian experience is discussed in Sickles, Hao, and Sheng (2014).

## REFERENCES

---

- Afriat, S. N. 1972. "Efficiency Estimation of Production Functions." *International Economic Review* 13(3):568–598.
- Ahn, S. C., Good, D. H., and Sickles, R. C. 2000. "Estimation of Long-run Inefficiency Levels: A Dynamic Frontier Approach." *Econometric Reviews* 19(4):461–492.
- Ahn, S. C., Lee, Y. H., and Schmidt, P. 2001. "GMM estimation of linear panel data models with time-varying individual effects." *Journal of Econometrics* 101: 219–255.
- Ahn, S. C., Lee, Y. H., and Schmidt, P. 2007. "Stochastic Frontier Models with Multiple Time-varying Individual Effects." *Journal of Productivity Analysis* 27(1):1–12.

- Ahn, S. C., Lee, Y. H., and Schmidt, P. 2013. "Panel Data Models with Multiple Time-varying Individual Effects." *Journal of Econometrics* 174(1):1–14.
- Aigner, D., Lovell, C., and Schmidt, P. 1977. "Formulation and Estimation of Stochastic Frontier Production Function Models." *Journal of Econometrics* 6(1):21–37.
- Akaike, H. 1973. "Information Theory and an Extension of the Maximum Likelihood Principle." In B.N. Petrov and F. Czàke, eds. Second International Symposium on Information Theory, Budapest: Akadémiai Kiadó.
- Alam, I. M. S., and Sickles, R. C. 2000. "A Time Series Analysis of Deregulatory Dynamics and Technical Efficiency: The Case of the U.S. Airline Industry." *International Economic Review* 41:203–218.
- Almanidis, P., Qian, J., and Sickles, R. C. 2013. "Bounded Stochastic Frontiers." Mimeo.
- Almanidis, P., and Sickles, R. C. 2012. "The Skewness Problem in Stochastic Frontier Models: Fact or Fiction?" chapter 10 of *Exploring Research Frontiers in Contemporary Statistics and Econometrics: A Festschrift in Honor of Leopold Simar*, Ingrid Van Keilegom and Paul Wilson, eds. New York: Springer Publishing, 201–227.
- Anders, I., Fethi, M., Hao, J., and Sickles, R. C. 2013. "World Productivity Growth." Mimeo, Rice University.
- Arrow, K. J. 1962. "The Economic Implications of Learning by Doing." *The Review of Economic Studies* 29(3):155–173.
- Badunenko, O., Henderson, D. J., and Russell, R. R. 2013. "Polarization of the Worldwide Distribution of Productivity." *Journal of Productivity Analysis*, October 2013, 40(2), pp 153–171 .
- Bai, J. 2009. "Panel Data Models with Interactive Fixed Effects." *Econometrica* 77(4): 1229–1279.
- Bai, J., and Ng, S. 2011. "Principal Components Estimation and Identification of the Factors." Department of Economics, Columbia University.
- Bai, J., Kao, C., and Ng, S. 2009. "Panel Cointegration with Global Stochastic Trends." *Journal of Econometrics* 149:82–99.
- Balk, B. M. 2008. *Price and Quantity Index Numbers: Models for Measuring Aggregate Change and Difference*. Cambridge: Cambridge University Press.
- Baltagi, B. H., and Griffin, J. M. 1988. "A General Index of Technical Change." *The Journal of Political Economy*, 96(1), 20–41.
- Bates, J. M., and Granger, C. W. 1969. "The Combination of Forecasts." *Journal of Operational Research Society*, 20(4): 451–468.
- Battese, G. E., and Coelli, T. J. 1992. "Frontier Production Functions, Technical Efficiency and Panel Data: With Application to Paddy Farmers in India." *Journal of Productivity Analysis* 3(1–2):153–169.
- Battese, G. E., and Coelli, T. J. 1995. "A Model for Technical Inefficiency Effects in a Stochastic Frontier Production Function for Panel Data." *Empirical Economics* 20:325–332.
- Battese, G. E., and Corra, G. S. 1977. "Estimation of a Production Frontier Model: With Application to the Pastoral Zone of Eastern Australia." *Australian Journal of Agricultural and Resource Economics* 21(3):169–179.
- Bloom, N., and Van Reeden, J. 2007. "Measuring and Explaining Management Practices across Firms and Countries." *Quarterly Journal of Economics* 122(4):1351–1408.
- Buchanan, J. M., Tollison, R. D., and Tullock, G. 1980. *Toward a Theory of the Rent-Seeking Society*. Number 4. Texas A & M University Press.

- Buckland, S. T., Burnham, K. P., and Augustin, N. H. 1997. "Model Selection: An Integral Part of Inference." *Biometrics*, 53(2), 603–618.
- Burnham, K. P., and Anderson, D. R. 2002. *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*. Second edition. Springer, New York, USA.
- Carroll, R. J., Midthune, D., Freedman, L. S., and Kipnis, V. 2006. "Seemingly Unrelated Measurement Error Models, with Application to Nutritional Epidemiology." *Biometrics* 62(1):75–84.
- Caves, D. W., Christensen, L. R., and Diewert, W. E. 1982. "The Economic Theory of Index Numbers and the Measurement of Input, Output, and Productivity." *Econometrica: Journal of the Econometric Society* 50(6), 1393–1414.
- Charnes, A., Cooper, W. W., and Rhodes, E. 1978. "Measuring the Efficiency of Decision Making Units." *European Journal of Operational Research* 2(6):429–444.
- Chen, E. K. 1997. "The Total Factor Productivity Debate: Determinants of Economic Growth in East Asia." *Asian-Pacific Economic Literature* 11(1):18–38.
- Claeskens, G., and Hjort, N. L. 2008. *Model Selection and Model Averaging*. Cambridge University Press, Cambridge.
- Clemen, R. T. 1989. "Combining Forecasts: A Review and Annotated Bibliography." *International Journal of Forecasting* 5(4):559–583.
- Coe, D. T., and Helpman, E. 1995. "International R&D Spillovers." *European Economic Review* 39(5):859–887.
- Coe, D. T., Helpman, E., and Hoffmaister, A. 1997. "North–South R&D Spillovers." *The Economic Journal* 107(440):134–149.
- Coelli, T. 2000. *On the Econometric Estimation of the Distance Function Representation of a Production Technology*. Center for Operations Research & Econometrics. Université Catholique de Louvain.
- Coelli, T., and Perelman, S. 1996. "Efficiency Measurement, Multiple-Output Technologie and Distance Functions: With Application to European Railways." *European Journal of Operational Research* 117:326–339.
- Coelli, T. J., Rao, D. P., O'Donnell, C. J., and Battese, G. E. 2005. *An Introduction to Efficiency and Productivity Analysis*. Springer.
- Cornwell, C., Schmidt, P., and Sickles, R. C. 1990. "Production Frontiers with Cross-Sectional and Time-Series Variation in Efficiency Levels." *Journal of Econometrics* 46(1): 185–200.
- Cuesta, R. A. 2000. "A Production Model with Firm-Specific Temporal Variation in Technical Inefficiency: With Application to Spanish Dairy Farms." *Journal of Productivity Analysis* 13(2):139–158.
- Davies, A., and Lahiri, K. 1995. "A New Framework for Analyzing Survey Forecasts Using Three-Dimensional Panel Data." *Journal of Econometrics* 68(1):205–227.
- Debreu, G. 1951. "The Coefficient of Resource Utilization." *Econometrica: Journal of the Econometric Society* 19(3):273–292.
- Demetrescu, M., and Homm, U. 2013. "A Directed Test of No Cross-Sectional Error Correlation in Large-N Panel Data Models." Mimeo.
- Diao, X., Rattsø, J., and Stokke, H. E. 2005. "International Spillovers, Productivity Growth and Openness in Thailand: An Intertemporal General Equilibrium Analysis." *Journal of Development Economics* 76(2):429–450.
- Diebold, F. X., and Lopez, J. A. 1996. *Forecast Evaluation and Combination*. National Bureau of Economic Research, Inc, Amsterdam: North-Holland.

- Druska, V., and Horrace W. C. 2004. "Generalized Moments Estimation for Spatial Panel Data: Indonesian Rice Farming." *American Journal of Agricultural Economics* 86: 185–198.
- Entur, C., and Musolesi, A. 2013. "Weak and Strong Cross Sectional Dependence: A Panel Data Analysis of International Technology Diffusion." Mimeo.
- Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression*. Marcel Dekker.
- Färe, R., Grosskopf, S., Grifell-Tatjé, E., and Lovell, C. A. K. 1997. "Biased Technical Change and the Malmquist Productivity Index." *Scandinavian Journal of Economics* 99(1): 119–127.
- Färe, R., Grosskopf, S., Norris, M., and Zhang, Z. 1994. "Productivity Growth, Technical Progress, and Efficiency Change in Industrialized Countries." *The American Economic Review*, 84(1):66–83.
- Farrell, M. J. 1957. "The Measurement of Productive Efficiency." *Journal of the Royal Statistical Society, Series A (General)* 120(3):253–290.
- Fisher, I. 1927. *The Making of Index Numbers: A Study of Their Varieties, Tests, and Reliability*. Boston, MA: Houghton Mifflin.
- Feng, Q., Horrace, W., and Wu, G. L. 2012. "Wrong Skewness and Finite Sample Correction in Parametric Stochastic Frontier Models." Mimeo.
- Førsund, F. R., and Hjalmarsson, L. 2008. "Dynamic Analysis of Structural Change and Productivity Measurement." Memorandum. Department of Economics, University of Oslo.
- Fried, H. O., Lovell, C. K., and Schmidt, S. S. 2008. *The Measurement of Productive Efficiency and Productivity Growth*. Oxford University Press, New York.
- Glass, A., Kenjegalieva, K., and Sickles, R. C. 2013a. "A Spatial Autoregressive Production Frontier Model for Panel Data: With an Application to European Countries." Mimeo.
- Glass, A., Kenjegalieva, K., and Sickles, R. C. 2013b. "Estimating Efficiency Spillovers with State Level Evidence for Manufacturing in the U.S." Mimeo.
- Good, D. H., Nadiri, M. I., and Sickles, R. C. 1997. "Index Number and Factor Demand Approaches to the Estimation of Productivity." chapter 1 of the *Handbook of Applied Economics*, Vol. 2: *Microeconomics*, M. H. Pesaran and P. Schmidt, eds. Oxford: Basil Blackwell.
- Good, D. H., Roeller, L. H., and Sickles, R. C. 1995. "Airline Efficiency Differences between Europe and the U.S.: Implications for the Pace of E.C. Integration and Domestic Regulation." *European Journal of Operational Research* 80:508–518.
- Good, D. H., Nadiri, M. I., Roeller, L. H., and Sickles, R. C. 1993. "Efficiency and Productivity Growth Comparisons of European and U.S. Air Carriers: A First Look at the Data." *Journal of Productivity Analysis Special Issue*, J. Mairesse and Z. Griliches, eds. 4: 115–125.
- Greene, W. 2005a. "Fixed and Random Effects in Stochastic Frontier Models." *Journal of Productivity Analysis* 23(1):7–32.
- Greene, W. 2005b. "Reconsidering Heterogeneity in Panel Data Estimators of the Stochastic Frontier Model." *Journal of Econometrics* 126(2):269–303.
- Grifell-Tatjé, E., and Lovell, C. A. K. 1995. "Note on the Malmquist Productivity Index." *Economics Letters* 47(2):169–175.
- Griliches, Z. 1957. "Hybrid Corn: An Exploration in the Economics of Technological Change." *Econometrica, Journal of the Econometric Society* 25(4): 501–522.
- Griliches, Z. 1979. "Issues in Assessing the Contribution of Research and Development to Productivity Growth." *The Bell Journal of Economics* 10(1):92–116.

- Griliches, Z., and Hausman, J. A. 1986. "Errors in Variables in Panel Data." *Journal of Econometrics* 31:93–118.
- Griliches, Z., and Mairesse, J. 1990. "Heterogeneity in Panel Data: Are There Stable Production Functions?" In *Essays in Honor of Edmond Malinvaud*. Cambridge, MA: MIT Press. Vol. 3:192–231.
- Griliches, Z., and Mairesse, J. 1998. "Production Functions: The Search for Identification." In *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial Symposium*, S. Ström, ed. Cambridge: University Press, 169–203.
- Griliches, Z., and Pakes, A. 1984. "Distributed Lags in Short Panels with an Application to the Specification of Depreciation Patterns and Capital Stock Constructs." *The Review of Economic Studies* 51(2):1175–1189.
- Hansen, B. E. 2007. "Least Squares Model Averaging." *Econometrica* 75(4):1175–1189.
- Hao, J. 2012. "Essays on Productivity Analysis." Unpublished dissertation, Rice University.
- Hausman, J. A., and Taylor, W. E. 1981. "Panel Data and Unobservable Individual Effects." *Econometrica* 49(6):1377–1398.
- Hjorth, J. U. 1994. *Computer Intensive Statistical Methods: Validation Model Selection and Bootstrap*. Chapman & Hall/CRC.
- Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. 1999. "Bayesian Model Averaging: A Tutorial." *Statistical Science* 14(4):382–417.
- Huang, C. J., and Lai, H. P. 2012. "Estimation of Stochastic Frontier Models Based on Multimodel Inference." *Journal of Productivity Analysis* 38(3):273–284.
- Hultberg, P. T., Nadiri, M. I., and Sickles, R. C. 1999. "An International Comparison of Technology Adoption and Efficiency: A Dynamic Panel Model." *Annales d'Economie et de Statistique* 449–474.
- Hultberg, P. T., Nadiri, M. I., and Sickles, R. C. 2004. "Cross-Country Catch-up in the Manufacturing Sector: Impacts of Heterogeneity on Convergence and Technology Adoption." *Empirical Economics* 29(4):753–768.
- Isaksson, A. 2007. "World Productivity Database: A Technical Description." *Research and Statistics Staff Working Paper* 10.
- Jeon, B. M., and Sickles, R. C. 2004. "The Role of Environmental Factors in Growth Accounting." *Journal of Applied Econometrics* 19(5):567–591.
- Jorgenson, D. W. 1995. *Productivity*, Vols. 1 and 2. Cambridge, MA: MIT Press.
- Jorgenson, D. W. 2005. *Productivity, Vol. 3: Information Technology and the American Growth Resurgence*. Cambridge, MA: MIT Press.
- Jorgenson, D., and Griliches, Z. 1972. "Issues in Growth Accounting: A Reply to Edward F. Denison." *Survey of Current Business* 52.5, Part II.
- Kendrick, J. W. 1961. "Front Matter, Productivity Trends in the United States." *Productivity Trends in the United States* NBER, 52–0.
- Kim, J. I., and Lau, L. J. 1994. "The Sources of Economic Growth of the East Asian Newly Industrialized Countries." *Journal of the Japanese and International Economies* 8(3): 235–271.
- Kim, J. I., and Lau, L. J. 1996. "The Sources of Asian Pacific Economic Growth." *The Canadian Journal of Economics/Revue Canadienne d'Economique* 29:S448–S454.
- Kim, S., and Lee, Y. H. 2006. "The Productivity Debate of East Asia Revisited: A Stochastic Frontier Approach." *Applied Economics* 38(14):1697–1706.
- Klein, L. 1953. *A Textbook of Econometrics*. Evanston, IL: Peterson.

- Kneip, A., and Sickles, R. C. 2012. "Panel Data, Factor Models, and the Solow Residual," chapter 5 of *Exploring Research Frontiers in Contemporary Statistics and Econometrics: A Festschrift in Honor of Leopold Simar*, Ingrid Van Keilegom and Paul Wilson, eds. New York: Springer Publishing, 83–114.
- Kneip, A., Sickles, R. C., and Song, W. 2012. "A New Panel Data Treatment for Heterogeneity in Time Trends." *Econometric Theory* 28:590–628.
- Koop, G., Poirier, D.J., and Tobias, J.L. 2007. *Bayesian Econometric Methods*. New York: Cambridge University Press.
- Krugman, P. 1994. "The Myth of Asia's Miracle." *Foreign Affairs* 62–78.
- Kumbhakar, S. C. 1990. "Production Frontiers, Panel Data, and Time-Varying Technical Inefficiency." *Journal of Econometrics* 46(1):201–211.
- Kumbhakar, S. C., and Lovell, C. A. K. 2000. *Stochastic Frontier Analysis*. Cambridge: Cambridge University Press.
- Kumbhakar, S. C., Parmeter, C., and Tsionas, E. G. 2013. "A Zero Inefficiency Stochastic Frontier Model." *Journal of Econometrics* 172(1):66–76.
- Kutlu, L., and Sickles, R. C. 2012. "Estimation of Market Power in the Presence of Firm Level Inefficiencies." *Journal of Econometrics* 168(1):141–155.
- Lahiri, K., Peng, H., and Sheng, X. 2010. "Measuring Aggregate Uncertainty in a Panel of Forecasts and a New Test for Forecast Heterogeneity."
- Lahiri, K., and Sheng, X. 2010. "Measuring Forecast Uncertainty by Disagreement: The Missing Link." *Journal of Applied Econometrics* 25(4):514–538.
- Lahiri, K., Peng, H., and Zhao, Y. 2011. "Forecast Combination in Incomplete Panels." Working paper.
- Lahiri, K., Teigland, C., and Zaporowski, M. 1988. "Interest Rates and the Subjective Probability Distribution of Inflation Forecasts." *Journal of Money, Credit and Banking* 20(2):233–248.
- Lee, J. W., and Barro, R. J. 2000. *International Data on Educational Attainment Updates and Implications*. National Bureau of Economic Research.
- Lee, Y. 1996. "Tail Truncated Stochastic Frontier Models." *Journal of Economic Theory and Econometrics* (2):137–152.
- Lee, Y., and Lee, S. 2012. "Stochastic Frontier Models with Threshold Efficiency." Mimeo.
- Lee, Y. H. 1991. "Panel Data Models with Multiplicative Individual and Time Effects: Applications to Compensation and Frontier Production Functions." Dissertation. Michigan State University.
- Lee, Y. H., and Schmidt, P. 1993. "A Production Frontier Model with Flexible Temporal Variation in Technical Efficiency." *The Measurement of productive efficiency: techniques and applications*. edited by Fried, H. O., Lovell, C. K., and Schmidt, S. S. New York: Oxford University Press.
- Leeb, H., and Pötscher, B. M. 2005. "Model Selection and Inference: Facts and Fiction." *Econometric Theory* 21(1):21–59.
- Liu, J., Sickles, R. C., and Tsionas, E. 2013. "Bayesian Treatment to Panel Data Models with Time-varying Heterogeneity." Mimeo.
- Lucas, Jr., R. E. 1988. "On the Mechanics of Economic Development." *Journal of Monetary Economics* 22(1):3–42.
- Mallows, C. L. 1973. "Some Comments on Cp." *Technometrics* 15(4):661–675.
- Mairesse, J. 2003. "In Memoriam: Zvi Griliches." *Econometric Reviews* 22(1):11–15.

- Mansfield, E. 1961. "Technical Change and the Rate of Imitation." *Econometrica* 29(4): 741–766.
- Mastromarco, C., and Shin, Y. 2013. "Modelling Technical Efficiency in Cross Sectionally Dependent Panels." Mimeo.
- Meeusen, W., and Broeck, J. van Den. 1977. "Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error." *International Economic Review* 18(2):435–444.
- Moulin, H. 1980. "On Strategy-Proofness and Single Peakedness." *Public Choice* 35(4): 437–455.
- Mundlak, Y. 1961. "Empirical Production Function Free of Management Bias." *Journal of Farm Economics* 43(1):44–56.
- Mundlak, Y. 1978. "On the Pooling of Time Series and Cross Section Data." *Econometrica: Journal of the Econometric Society* 46(1):69–85.
- Newbold, P., and Harvey, D. I. 2002. "Forecast Combination and Encompassing." *A Companion to Economic Forecasting* 268–283. A Companion to Economic Forecasting. Edited by Clements, M. P., and Hendry, D. F. John Wiley & Sons.
- Orea, L., and Kumbhakar, S. C. 2004. "Efficiency Measurement Using a Latent Class Stochastic Frontier Model." *Empirical Economics* 29(1):169–183.
- Olley, G. S., and Pakes, A. 1996. "The Dynamics of Productivity in the Telecommunications Equipment Industry." *Econometrica* 64(6):1263–1297.
- Orea, L., and Steinbuks, J. 2012. *Estimating Market Power in Homogenous Product Markets Using a Composed Error Model: Application to the California Electricity Market*. University of Cambridge, Faculty of Economics.
- Park, B. U., Sickles, R. C., and Simar, L. 1998. "Stochastic Panel Frontiers: A Semiparametric Approach." *Journal of Econometrics* 84(2):273–301.
- Park, B. U., Sickles, R. C., and Simar, L. 2003. "Semiparametric-efficient Estimation of AR (1) Panel Data Models." *Journal of Econometrics* 117(2):279–309.
- Park, B. U., Sickles, R. C., and Simar, L. 2007. "Semiparametric Efficient Estimation of Dynamic Panel Data Models." *Journal of Econometrics* 136(1):281–301.
- Pesaran, M. H. 2007. "A Simple Panel Unit Root Test in the Presence of Cross-section Dependence." *Journal of Applied Econometrics* 22:265–312.
- Pitt, M., and Lee, L. F. 1981. "The Measurement and Sources of Technical Inefficiency in the Indonesian Weaving Industry." *Journal of Development Economics* 9(1):43–64.
- Raftery, A. E., Madigan, D., and Hoeting, J. A. 1997. "Bayesian Model Averaging for Linear Regression Models." *Journal of the American Statistical Association* 92(437):179–191.
- Reifschneider D., and Stevenson R. 1991. "Systematic Departures from the Frontier: A Framework for the Analysis of Firm Inefficiency." *International Economic Review* 32:715–723.
- Reikard, G. 2005. "Endogenous Technical Advance and the Stochastic Trend in Output: A Neoclassical Approach." *Research Policy* 34(10):1476–1490.
- Romer, P. M. 1986. "Increasing Returns and Long-run Growth." *The Journal of Political Economy* 1002–1037.
- Sachs, J. D., Warner, A., Åslund, A., and Fischer, S. 1995. "Economic Reform and the Process of Global Integration." *Brookings Papers on Economic Activity* 1995(1):1–118.
- Scherer, F. M. 1971. *Industrial Market Structure and Economic Performance*. Chicago: Rand McNally.
- Schmidt, P., and Knox Lovell, C. 1979. "Estimating Technical and Allocative Inefficiency Relative to Stochastic Production and Cost Frontiers." *Journal of Econometrics* 9(3): 343–366.

- Schmidt, P., and Sickles, R. C. 1984. "Production Frontiers and Panel Data." *Journal of Business & Economic Statistics* 2(4):367–374.
- Schwarz, G. 1978. "Estimating the Dimension of a Model." *The Annals of Statistics* 6(2): 461–464.
- Shephard, W. G. 1970. *Market Power and Economic Welfare*. New York: Random House.
- Sickles, R. C. 2005. "Panel Estimators and the Identification of Firm-Specific Efficiency Levels in Semi-parametric and Non-parametric Settings." *Journal of Econometrics* 126:305–324.
- Sickles, R. C., Hao, J., and Shang, C. 2014. "Panel Data and Productivity Measurement: An Analysis of Asian Productivity Trends." *Journal of Chinese Economic and Business Studies*, 12(1), 211–231.
- Simar, L., and Wilson, P. W. 2000. "Statistical Inference in Nonparametric Frontier Models: The State of the Art." *Journal of Productivity Analysis* 13(1):49–78.
- Solow, R. M. 1957. "Technical Change and the Aggregate Production Function." *The Review of Economics and Statistics* 39(3):312–320.
- Stiroh, K. 2001. "Information Technology and the US Productivity Revival: What Do the Industry Data Say?" *FRB of New York Staff Report* 115.
- Stoker, T. M., Berndt, E. R., Ellerman, A. D., and Schennach, S. M. 2005. "Panel Data Analysis of U.S. Coal Productivity." *Journal of Econometrics* 127:131–164.
- Summers, R., Heston, A., and Aten, B. 2002. "Penn World Table Version 6.1." *Center for International Comparisons at the University of Pennsylvania*.
- Timmermann, A. 2006. "Forecast Combinations." *Handbook of Economic Forecasting* 1: 135–196.
- Tsionas, E. G. 2006. "Inference in Dynamic Stochastic Frontier Models." *Journal of Applied Econometrics* 21(5):669–676.
- Tsionas, E. G., and Greene, W. H. 2003. "Non-Gaussian Stochastic Frontier Models." Manuscript.
- Tsionas E. G., and Kumbhakar, S. C. 2004. "Markov Switching Stochastic Frontier Model." *Econometrics Journal, Royal Economic Society* 7(2):398–425.
- Tullock, G. 1980. "Efficient Rent-seeking." In *Toward a Theory of the Rent-Seeking Society*, edited by J. M. Buchanan, R. D. Tollison, and G. Tullock, 97–112. College Station, TX: A&M University Press.
- Van den Broeck, J., Koop, G., Osiewalski, J., and Steel, M. F. 1994. "Stochastic Frontier Models: A Bayesian Perspective." *Journal of Econometrics* 61(2):273–303.
- Young, A. 1992. "A Tale of Two Cities: Factor Accumulation and Technical Change in Hong Kong and Singapore." In *NBER Macroeconomics Annual 1992*, Vol. 7 13–64. Cambridge, MA: MIT Press.
- Young, A. 1995. "The Tyranny of Numbers: Confronting the Statistical Realities of the East Asian Growth Experience." *The Quarterly Journal of Economics* 110(3):641–680.
- Zarnowitz, V., and Lambros, L. A. 1987. *Consensus and Uncertainty in Economic Prediction*. National Bureau of Economic Research, Inc., New York.

## CHAPTER 18

---

# PANEL DATA DISCRETE CHOICE MODELS OF CONSUMER DEMAND

---

MICHAEL P. KEANE

### 18.1 INTRODUCTION

---

This chapter deals with the vast literature on panel data discrete choice models of consumer demand. One reason this area is so active is that high-quality data is readily available. Firms like Nielsen and IRI have, for over 30 years, been collecting panel data on households' purchases of consumer goods. This is known as "scanner data" because it is collected by check-out machine scanners. Available scanner data sets often follow households for several years and record all their purchases in several different product categories. The typical data set contains not only information on the universal product codes (UPC) of the goods that households buy on each shopping trip but also information on several exogenous forcing variables, such as price and whether the goods were displayed or advertised in various ways.

To my knowledge the first paper using scanner data to study the impact of price and other marketing variables on consumer demand was Guadagni and Little (1983) in *Marketing Science*. But few economists knew about scanner data until the mid- to late 1990s. Once they became aware of this treasure trove of data, they started to use it very actively. Today, estimation of demand models on scanner data has become a major part of the field of empirical industrial organization.

Thus, the consumer demand literature based on scanner data is unusual relative to other literatures discussed in this Handbook in two respects. First, it remains true that the majority of work in this area is by marketers rather than economists. Second, this is an uncommon case where the "imperial science" of economics (see, e.g., Stigler 1984) has experienced a substantial knowledge transfer from another area (i.e., marketing). Furthermore, it should be noted that discrete choice models of consumer demand are

also widely used in other fields like transportation research, agricultural and resource economics, environmental economics, and so on.

Given that the literature on panel data models of consumer demand is so large, I will make no attempt to survey all the important papers in the field. Instead, I will focus on the main research questions that dominate this area, and the progress that has been made in addressing them. Thus, I apologize in advance for the many important papers that are not cited.

The most salient feature of scanner panel data is that consumers exhibit substantial persistence in their brand choices. In the language of marketing, consumers show substantial “brand loyalty.” A second obvious aspect of the data is that, if we aggregate to the store level, then in most product categories the sales of a brand jump considerably when it is on sale (i.e., typically the price elasticity of demand is on the order of 3 to 5). Superficially, these two observations seem contradictory. If individual consumers are very loyal to particular brands, then why would demand for brands be very price sensitive in the aggregate?

In light of these empirical observations, the first main objective of the panel data demand literature has been to understand the underlying sources of persistence in brand choices. Based on work by Heckman (1981) on employment dynamics, it is now understood that persistence in brand choices may arise from three sources: (i) permanent unobserved heterogeneity in tastes, (ii) serial correlation in taste shocks, or (iii) “true” or “structural” state dependence.

Only the third source of persistence (i.e., state dependence) involves a causal effect of past choices on the current choice (and, likewise, an effect of the current choice on future choices). Uncovering whether state dependence exists is of great importance in both marketing and industrial organization. If it exists, then current marketing actions, such as price discounts, will affect not only current but also future demand. This has important implications for pricing policy, the nature of inter-firm competition, and so on.

The second major objective of the literature has been to distinguish alternative possible explanations for structural state dependence (assuming that it exists). Some of the potential explanations include habit persistence, learning about quality through trial, inventory behavior, variety seeking behavior, switching costs, and so on.

A third, but closely related, major objective of the literature has been to understand the dynamics of demand. Most important is to understand the sources of the observed increase in demand when a brand is on sale. The increase in sales may arise from three sources: (i) brand switching, (ii) category expansion, or (iii) purchase acceleration, also known as cannibalization. In everyday language, these correspond to (i) stealing customers from your competitors, (ii) bringing new customers into the category, or (iii) merely accelerating purchases by consumers who are loyal to a brand, and who would eventually have bought it at the regular price anyway.

The distinction among these three sources of increased demand is of crucial importance for pricing policy. For example, if most of the sales increase resulting from a price

discount is due to cannibalization of future sales, a policy of periodic price discounts makes no sense.

The estimation of discrete choice models with many alternatives is a difficult econometric problem. This is because the order of integration required to calculate choice probabilities in such a model is typically on the order of  $J - 1$ , where  $J$  is the number of choice alternatives. The development of simulation methods for the estimation of multinomial discrete choice models in the late 1980s was largely motivated by this problem (see McFadden 1989).

As discussed in Keane (1994), in the panel data case the required order of integration to construct the choice probabilities in discrete choice models is much higher. This is because it is the probability of a consumer's entire choice sequence that enters the likelihood function. Thus, the required order of integration is  $(J - 1) \cdot T$ , where  $T$  is the number of time periods. In typical scanner panels  $T$  is on the order of 50 to 200 weeks, so the order of integration is very high.

In Keane (1994), I developed a method of “sequential importance sampling” that makes estimation of panel data discrete choice models feasible. In the special case of the normal errors, which leads to the panel probit model, this method is known as the “GHK” algorithm. GHK is a highly accurate method for approximating multidimensional normal integrals. It is notable how development of simulation-based econometric methods has gone hand-in-hand with a desire to estimate demand models with large choice sets, many time periods, and complex error structures.

The outline of the remainder of the chapter is as follows. In Section 18.2, I describe a fairly general panel data discrete choice model. Section 18.3 discusses the econometric methods needed to estimate such models. Then, Section 18.4 discusses the theoretical issues involved in distinguishing state dependence from heterogeneity, while Section 18.5 discusses empirical work on state dependence and/or choice dynamics. Section 18.6 concludes.

## 18.2 THE TYPICAL STRUCTURE OF PANEL DATA DISCRETE CHOICE MODELS

---

Here I describe the typical structure of demand models used in marketing (and more recently in industrial organization). Let  $j = 1, \dots, J$  index alternatives,  $t = 1, \dots, T$  index time, and  $i = 1, \dots, N$  index people. Then the “canonical” brand choice model can be written as follows:

$$U_{ijt} = \alpha_{ij} + X_{ijt}\beta + \gamma d_{ij,t-1} + \varepsilon_{ijt} \quad \text{where } \varepsilon_{ijt} = \rho \varepsilon_{ij,t-1} + \eta_{ijt} \quad (1)$$

$$d_{ijt} = 1 \text{ if } U_{ijt} > U_{ikt} \text{ for all } k \neq j \quad \text{otherwise } d_{ijt} = 0 \quad (2)$$

Equation (1) expresses the utility that person  $i$  receives from the purchase of brand  $j$  at time  $t$ . Utility ( $U_{ijt}$ ) depends on a vector of product attributes  $X_{ijt}$  and the utility or attribute weights  $\beta$ . Utility also depends on consumer  $i$ 's intrinsic preference for brand  $j$ , which I denote by  $\alpha_{ij}$ . It is further assumed that utility depends on whether brand  $j$  was chosen by person  $i$  on the previous choice occasion ( $d_{ij,t-1} = 1$ ). Finally, there is a purely idiosyncratic person, time, and brand-specific taste shock  $\varepsilon_{ijt}$ . This is allowed to be serially correlated, with the fundamental shocks  $\eta_{ijt}$  being *iid*. Equation (2) simply says that person  $i$  chooses the brand  $j$  that gives him greatest utility at time  $t$ . Of course, in a discrete choice model we only observe choices and not utilities.

Before turning to the econometrics, it is important to give an economic interpretation to the terms in (1). A utility function that is linear in attributes is quite standard in the demand literature (for the classic exposition of attribute based utility, see Lancaster 1966). But in (1) we also assume the utility weights  $\beta$  are common across consumers (as in traditional logit and probit models). This is a strong assumption, but it is only for expositional convenience.<sup>1</sup> The simulation methods discussed below can easily accommodate heterogeneity in  $\beta$ .

I will focus attention on heterogeneity in the brand intercepts  $\alpha_{ij}$ . These capture consumer heterogeneity in tastes for attributes of alternatives that are not observed by the econometrician (see Berry 1994; Elrod and Keane 1995; Keane 1997). For example, for some products like cars, clothing, or perfume, different brands convey a certain “image” that is hard to quantify. Heterogeneous tastes for that “image” would be subsumed in the  $\alpha_{ij}$ . Of course, even mundane products have unobserved attributes (e.g., the “crispness” of different potato chips).

It is worth emphasizing that one of the attributes included in  $X_{ijt}$  is price, which we denote by  $p_{ijt}$ . The budget constraint conditional on purchase of brand  $j$  is  $C_{it} = I_{it} - p_{ijt}$ . As frequently purchased consumer goods are fairly inexpensive, it makes sense to assume the marginal utility of consumption of the outside good is a constant over the range  $[I_{it} - p_{max}, I_{it} - p_{min}]$ , where  $p_{max}$  and  $p_{min}$  are the highest and lowest prices ever observed in the category. This justifies making utility linear in consumption of the outside good. If we use the budget constraint to substitute for  $C_{it}$ , we obtain a conditional indirect utility function that is linear in income and price.

Furthermore, income is person-specific and not alternative-specific. Because income is the same across all alternatives  $j$  for an individual, it does not alter the utility differences between alternatives. As a result, income drops out of the model and we are left with only price. It is important to remember, however, that price only appears because we are dealing with an indirect utility function, and its coefficient is not interpretable as just another attribute weight. The price coefficient is actually the marginal utility of consumption of the outside good.

Thus, an important implication of consumer theory is that the price coefficient should be equal across all alternatives. However, it will generally vary across people, as the marginal utility of consumption is smaller for those with higher income. This can be accounted for by letting the price coefficient depend on income and other household characteristics.

The next important feature of (1) is the lagged choice variable  $d_{ij,t-1}$ . This captures an effect of lagged purchase of a brand on its current utility evaluation. Heckman (1981) calls this “structural” state dependence. Most papers use more elaborate forms of state dependence than just lagged purchase. For instance, Guadagni and Little (1983) used an exponentially smoothed weighted average of all the lagged  $d_{is}$  for  $s = 1, \dots, t - 1$ , and this specification is popular in the marketing literature. But I will focus on the first-order Markov model for expositional purposes.

There are many reasons why a structural effect of lagged purchase on current utility may exist; such as habit persistence, learning, inventories, variety seeking behavior, switching costs, and so on. I discuss efforts to distinguish among these sources of state dependence in Section 18.5.

But first, in Section 18.4, I will focus on the question of whether state dependence exists at all (whether  $\gamma \neq 0$ ). This question alone has been the focus of a large literature. The question is difficult to address, because failure to adequately control for heterogeneity and other serial correlation will lead to what Heckman (1981) called “spurious” state dependence. Furthermore, there are deep econometric and philosophical issues around the question of whether it is even possible to distinguish state dependence from heterogeneity (or serial correlation in general).

Finally, equation (1) includes idiosyncratic taste shocks  $\varepsilon_{ijt}$ . These may be interpreted in different ways, depending on one’s perspective. In the economic theory of random utility models (Bloch and Marschak 1960; McFadden 1974) choice is deterministic from the point of view of a consumer, who observes his/her own utility. Choice only *appears* to be random from the point of view of the econometrician, who has incomplete information about consumer preferences and brand attributes. As Keane (1997) discusses, the  $\varepsilon_{ijt}$  can be interpreted as arising from unobserved attributes of brands for which people have heterogeneous tastes that *vary* over time. This is in contrast to the brand intercepts  $\alpha_{ij}$ , which capture unobserved attributes of brands for which people have heterogeneous tastes that are *constant* over time. However, in psychology-based models of choice, the  $\varepsilon_{ijt}$  are interpreted as genuinely random elements of choice behavior. I am not aware of a convincing way to distinguish between these two perspectives.

If the  $\varepsilon_{ijt}$  arise from time-varying tastes, it is plausible that tastes show some persistence over time. This motivates the AR(1) specification  $\varepsilon_{ijt} = \rho \varepsilon_{ij,t-1} + \eta_{ijt}$ , where  $\eta_{ijt}$  is *iid* over time and people. If  $\rho > 0$ , then taste shocks exhibit temporal persistence. Of course, one could specify many other forms of serial correlation, but I focus on the AR(1) model for expositional purposes.

If the  $\eta_{ijt}$  are correlated across brands, it implies some brands are more similar than others on the unobserved attribute dimensions for which people have time-varying tastes. Similarly, if the intercepts  $\alpha_{ij}$  are correlated across brands, it implies some brands are more similar than others on the unobserved attribute dimensions for which people have time-invariant tastes. Brands that are more similar on the latent attribute dimensions, will, *ceteris paribus*, have more switching between them and higher cross-price elasticities of demand.

These ideas are the basis of the “market mapping” literature that uses panel data to determine the location of brands in a latent attribute space (see Elrod 1988; Elrod and Keane 1995; Keane 1997). For example, in a market map for cars, Mercedes and BMW would presumably lie close together in one part of the space, while Ford and Chevy trucks would also lie close together but in a very different part of the space. An estimated market map can, for example, help a firm to determine who its closest competitors are.

Note that the multinomial logit model assumes all errors are uncorrelated. This makes all brands “equally (dis)similar” (i.e., equally spread out in the market map) so that all cross-price elasticities of demand are equal. Hence, if one brand lowers its price, the sales of all other brands fall by a common percentage. This is known as independence of irrelevant alternatives (IIA).

A desire to escape this unrealistic assumption motivated work on simulation methods that make estimation of more general models (like multinomial probit) feasible—see, e.g., Lerman and Manski (1981), McFadden (1989). Simulation methods are the focus of the next section.

## 18.3 ESTIMATION OF PANEL DATA DISCRETE CHOICE MODELS

---

### 18.3.1 General Overview of the Model and Computational Issues

Here I discuss the computational problems that arise in estimating panel data discrete choice models. First, note that maximum likelihood estimation of the model in (1)–(2) requires distributional assumptions on the intercepts  $\alpha_{ij}$  and the errors  $\eta_{ijt}$ . The most common assumptions in the literature are that the intercepts are either multivariate normal ( $\alpha_i \sim N(0, \Sigma)$ ) or multinomial, while the  $\eta_{ijt}$  are either normal ( $\eta_{it} \sim N(0, \Omega)$ ) or *iid* type I extreme value. If both the  $\alpha_{ij}$  and  $\eta_{ijt}$  are normal we have the random effects panel probit model. If the  $\alpha_{ij}$  are normal while the  $\eta_{ijt}$  are extreme value and  $\rho = 0$ , we have a normal mixture of logits model (N-MIXL). If the  $\alpha_{ij}$  are multinomial, we have a discrete mixture of probits or logits. These are often called “latent class” models.

Estimation of the model in (1)–(2) requires some identifying normalizations. In discrete choice models, there is no natural scale for utility, and only utility differences among alternatives determine choice. Thus, one alternative (often but not always a “no purchase” option) is chosen as the base alternative, and its utility is normalized to zero. Hence, the error covariance matrices  $\Sigma$  and  $\Omega$  are of rank  $(J - 1)$  rather than  $J$ . The scale of utility is usually fixed by fixing the scale of one or more of the idiosyncratic

errors to a constant (e.g., letting  $\eta_{ij1}$  be *standard* normal or letting the  $\eta_{it}$  vector be *standard* type I extreme value).

Now, consider the panel probit case. In order to form the likelihood for a person  $i$ , we need to form the probability of his/her observed sequence of choices given the observed vector of covariates. That is, we need  $P(d_{ij(1),1}, \dots, d_{ij(T),T} | X_{i1}, \dots, X_{iT})$ , where  $j(t)$  denotes the index  $j$  of the option that the consumer actually chose at time  $t$ , while the  $X_{it} \equiv (x_{i1t}, \dots, x_{iJt})$  are vectors of covariates for all  $J$  alternatives at time  $t$ . The difficulty here is that, given the structure (1)–(2), this joint probability is very computationally difficult to construct.

First, consider the case where  $\gamma = \rho = 0$ . That is, there is no state dependence and the idiosyncratic errors  $\varepsilon_{ijt}$  are serially independent. Then the only source of persistence in choices over time is the brand-specific individual effects  $(\alpha_{i1}, \dots, \alpha_{iJ})$ . This gives an equicorrelated structure for the composite error terms  $v_{ijt} = \alpha_{ij} + \varepsilon_{ijt}$ , so we have a “random effects probit model.” Here, choice probabilities are independent over time *conditional* on the  $\alpha_{ij}$ , so we have:

$$P(d_{ij(1),1}, \dots, d_{ij(T),T} | X_{i1}, \dots, X_{iT}) = \int_{-\infty}^{\infty} \prod_{t=1}^T P(d_{ij(t),t} | X_{it}, \alpha) f(\alpha | \Sigma) d\alpha \quad (3)$$

Each conditional probability  $P(d_{ij(t),t} | X_{it}, \alpha_i)$  is a cross-section probit probability. As is well known, these are multivariate normal integrals of dimension  $J - 1$ . When  $J \geq 3$  or 4, it is necessary to use simulation methods like the GHK algorithm to evaluate these integrals. As the focus here is on panel data issues and not problems that already arise in cross-section discrete choice models, I will refer the reader to Geweke and Keane (2001) for further details.

The key problem in forming the choice probability in (3) is how to evaluate the integral over the density  $f(\alpha | \Sigma)$  of the multivariate normal  $(J - 1)$ -vector of individual effects  $\alpha$ . For the  $J = 2$  case, Butler and Moffitt (1982) proposed using Gaussian quadrature. Basic quadrature is a method to calculate areas under polynomials defined on intervals. Say one can write an integral as  $I = \int_{-1}^1 f(x) dx = \int_{-1}^1 \omega(x) Q(x) dx$  where  $Q(x)$  is a polynomial of degree  $2G - 1$  and  $\omega(x)$  is a known function. Then the exact value of the integral is  $I = \sum_{g=1}^G w_g Q(x_g)$  where the  $x_g$  and  $w_g$  are the appropriate quadrature points and weights, respectively.<sup>2</sup> Intuitively, if I have a 1<sup>st</sup> degree polynomial (i.e., a line) on  $[-1, 1]$ , and I know  $Q(0)$ , then the area under the line is exactly  $2Q(0)$ . I do not need to know the slope at  $Q(0)$ . This idea extends to higher degree polynomials. That is, for polynomials of degree  $2G - 1$  the quadrature rule based on  $G$  points is *exact*.

Quadrature requires a change in variables, to transform the domain to  $[-1, 1]$ , and a factorization  $f(x) = \omega(x)Q(x)$ . Different types of quadrature arise from different choices of  $\omega(x)$ :  $\omega(x) = 1$  is Gauss-Legendre,  $\omega(x) = \exp(-x^2)$  is Gauss-Hermite,  $\omega(x) = (1 - x^2)^{-1/2}$  is Gauss-Chebyshev.

The real usefulness of quadrature in econometrics arises in cases where a factorization  $f(x) = \omega(x)Q(x)$  cannot literally be achieved, but we can achieve a factorization

where  $Q(x)$  is well *approximated* by a polynomial (preferably of fairly low order). This encompasses a very general class of smooth functions. Butler and Moffitt (1982) note that in the  $J = 2$  case (where  $\alpha$  is a scalar) the density  $f(\alpha|\Sigma)$  in (1) is simply a univariate normal. Given a normal density kernel, Gauss-Hermite quadrature with  $\omega(x) = \exp(-x^2)$  is appropriate. So Butler and Moffitt replace the integral in (3) with a weighted sum over Gauss-Hermite quadrature points:

$$\hat{P}_{Q,G}(d_{ij(1),1}, \dots, d_{ij(T),T} | X_{i1}, \dots, X_{iT}) = \sum_{g=1}^G w_g \prod_{t=1}^T P(d_{ij(t),t} | X_{it}, \alpha_g) \quad (4)$$

Here the  $\alpha_g$  and  $w_g$  denote the points and weights, respectively. Butler and Moffitt find that very accurate evaluations of normal integrals can be obtained using a few points (i.e., only 6 or 7). This implies that  $\prod_{t=1}^T P(d_{ij(t),t} | X_{it}, \alpha)$  in (3) is well approximated by low order polynomials.

Can the quadrature approach be extended to calculate integrals of dimension 2 or higher? Mechanically, going from intervals to cubes or hypercubes is straightforward. One simply takes tensor products. For example, in the case of  $J = 3$  (i.e., two random effects  $(\alpha_1, \alpha_2)$ ) one needs two sets of quadrature points  $(\alpha_{g1}, \alpha_{g2})$ , and the single sum over  $g$  in (4) is replaced by a double sum over  $g_1$  and  $g_2$ . In general, a  $J - 1$  dimensional sum is required. But there are two main reasons this approach may not be advisable, especially for  $J > 3$ .

First, quadrature, like other numerical methods for evaluating integrals, suffers from the curse of dimensionality. To take the tensor product over points, we must evaluate the conditional probabilities  $\prod_{t=1}^T P(d_{ij(t),t} | X_{it}, \alpha)$  at  $G^{J-1}$  points for  $\alpha$ . Also, it is known that to achieve exact evaluation of a polynomial of degree  $2G - 1$  on a cube one needs  $G^2$  quadrature points.

Second, the theory of quadrature for two or more dimensions is far from fully developed. Mousavi, Xiao, and Sukumar (2010) explain why 2D is fundamentally different from 1D:

While the interval is the only connected compact subset of  $\mathbb{R}$ , regions in  $\mathbb{R}^2$  come in an infinite variety of shapes ... one might attempt to generalize one-dimensional quadrature ... using tensor products. However, this approach is far from optimal in terms of the number of quadrature nodes needed for ... precision. It seems likely that ... quadratures in higher dimensions have to be studied separately ... for different ... integration regions, and that each region should require a different set of rules. (p. 100)

This point is highly relevant to discrete choice. As we shall see below in describing the GHK algorithm, a given choice sequence occurs if the stochastic terms  $\eta$  fall in a *convex polyhedron* of the form  $\{\eta \in \mathbb{R}^d | A\eta \leq b\}$ . This is generally not a hypercube.

Given these issues, the usual advice for dealing with high dimensional integrals is to use Monte Carlo simulation. Given the speed of modern computers, a brute force “frequency” simulation approach is often feasible, even when  $J$  is very large. That is,

let  $\{\alpha_d\}_{d=1,\dots,D}$  denote  $D$  draws from the  $f(\alpha|\Sigma)$  density obtained using a random number generator. This gives:

$$\hat{P}_{F,D}(d_{ij(1),1}, \dots, d_{ij(T),T} | X_{i1}, \dots, X_{iT}) = \frac{1}{D} \sum_{d=1}^D \prod_{t=1}^T P(d_{ij(t),t} | X_{it}, \alpha_d) \quad (5)$$

The similarity between (4) and (5) is notable, as each involves evaluating the choice probabilities at a discrete set of  $\alpha$  values and summing the results. The difference is that the quadrature points are chosen analytically so as to provide an accurate approximation with as few points as possible, while in (5) the  $\alpha_d$  are simply drawn at random. This means the number of draws  $D$  needed to achieve reasonable accuracy may be quite large (at least a few hundred in most applications).

The virtue of simulation is that, unlike quadrature, it does not suffer from the curse of dimensionality. The error variance in simulation estimators of probabilities is of order  $1/D$ , regardless of the size of  $J$ . That is, in equation (5) we have  $\hat{P}_{F,D} - P \sim N(0, s^2/D)$  by the Central Limit Theorem, where  $s^2 = E(\hat{P}_{F,1} - P)^2$  and we assume the simulation errors are *iid* across draws  $d$ . Furthermore, by using the more sophisticated methods described below (like GHK) one can greatly improve on the accuracy of the crude frequency simulator (for given  $D$ ).

Another problem that arises in forming the choice probability in (3) is that the variance matrix  $\Sigma$  has  $(J-1) \cdot J/2$  unique elements. Even for modest  $J$  it is cumbersome to estimate so many parameters. And, although formally identified, estimation of large covariance matrices can create severe practical/numerical problems in discrete choice models (see Keane 1992; Keane and Wasi 2013). This issue has received far less attention than the high order integration problem.

A useful strategy when  $J$  is large is to impose a relatively low dimensional factor structure on  $\Sigma$ . Then the required order of integration in (3) is just the number of factors  $F$ , regardless of the size of  $J-1$ . And the number of parameters (i.e., factor loadings) to be estimated is  $F \cdot (J-1)$ . This is linear in  $J$ , thus mitigating the proliferation of parameters problem.

Lancaster (1963) discussed the idea that in a market with many products, those products may only be differentiated on a few attribute dimensions (e.g., there are hundreds of brands of cereal, but they differ on only a few attributes like sugar content, fiber content, etc.). Work on “market mapping” using scanner data finds that the unobserved attribute space for most products is well described by just a few factors (e.g.,  $F \leq 3$ ), even when  $J$  is very large (see Elrod 1984; Elrod and Keane 1995; Keane 1997; Andrews and Manrai 1999).

Before moving from frequency simulation to more sophisticated methods, it is important to make one general point: in *any* simulation estimation method it is vital to hold draws fixed as one iterates on the model parameters. Failure to do so creates two problems: (i) the simulated likelihood will “jump” when the draws change, so the change in the likelihood is not solely due to updating of parameters; (ii) such draw

induced changes in the simulated likelihood play havoc with the calculation of likelihood derivatives and the operation of parameter search algorithms. But holding the draws  $\{\alpha_d\}_{d=1,\dots,D}$  fixed would appear to be impossible in the random effects probit model, because as  $\Sigma$  changes it seems one must take new draws for  $\alpha$  from the new  $f(\alpha|\Sigma)$ .

A standard “trick” that can be used to hold draws fixed as  $\Sigma$  changes works as follows. First, let  $\Sigma = AA'$ , where  $A$  is the lower triangular Cholesky matrix. Then let  $\alpha = A\mu$  where  $\mu$  is a standard normal vector. The “trick” is to draw  $\mu$  rather than  $\alpha$ , and hold the draws  $\{\mu_d\}_{d=1,\dots,D}$  fixed as we iterate on the elements of  $A$ . Then the draws  $\{\alpha_d\}_{d=1,\dots,D}$  will vary smoothly as we vary  $A$ , causing  $\hat{P}_{F,D}$  in (5) to vary smoothly. This procedure has the added benefit that iteration on elements of  $A$  rather than the elements of  $\Sigma$  guarantees that  $\hat{\Sigma}$  will always be a positive definite covariance matrix (by definition of the Cholesky transform).

### 18.3.2 A More Sophisticated Probability Simulator— The GHK Algorithm

A more sophisticated way to simulate the integral in (3) is to use sequential importance sampling, as developed in Keane (1993, 1994). This approach, known as the “GHK” algorithm in the special case of importance sampling from the normal, is described in quite a few papers in the literature (see, e.g., Keane 1993, 1994; Hajivassiliou, McFadden, and Ruud 1996; Geweke, Keane, and Runkle 1994, 1997; Geweke and Keane 2001), so I just give a basic example here. Continue to consider the case of  $\gamma = \rho = 0$ , and define the composite error:

$$\nu_{ijt} = \alpha_{ij} + \varepsilon_{ijt} \quad (6)$$

Equation (1) implies a bound on  $\nu_{ijt} = \alpha_{ij} + \varepsilon_{ijt}$  such that option  $j$  is chosen at time  $t$ :

$$U_{ijt} > U_{ikt} \forall k \neq j \Rightarrow \nu_{ijt} \geq -X_{ijt}\beta + (X_{ikt}\beta + \nu_{ikt}) \forall k \neq j \quad (7)$$

To simplify even further, consider the case where  $J = 2$ . As we noted earlier, the utility of a base option (say #1) is normalized to zero, leaving a single utility index  $U_{it}$  for the other option (say #2). Hence, we do not need the  $j$  subscript in this case. We write that  $j = 2$  is chosen over  $j = 1$  iff:

$$U_{it} > 0 \Rightarrow \nu_{it} \geq -X_{it}\beta \quad (7')$$

Now, to be concrete, consider the problem of simulating the probability of a particular sequence ( $d_{i1} = 2, d_{i2} = 2, d_{i3} = 2$ ). That is,  $T = 3$  and the consumer chooses option 2 in all three periods.

To implement the GHK algorithm we divide the sequence probability into transition probabilities. That is, we have:

$$\begin{aligned} P(d_{i1} = 2, d_{i2} = 2, d_{i3} = 2 | X_{i1}, \dots, X_{i3}) &= P(d_{i1} = 2 | X_{i1}) \times \\ P(d_{i2} = 2 | d_{i1} = 2, X_{i1}, X_{i2}) \times P(d_{i3} = 2 | d_{i1} = 2, d_{i2} = 2, X_{i1}, X_{i2}, X_{i3}) \end{aligned} \quad (8)$$

A key point is that the transition probabilities in (8) depend on lagged choices and covariates despite the fact that we have assumed  $\gamma = 0$ , so there is no true state dependence (only serial correlation). This occurs because of a fundamental property of discrete choice models that I now describe.

Specifically, as we only observe choices and not the latent utilities, we cannot construct lagged values of the error term. For instance, if  $d_{i1} = 2$ , all this tells us is that  $v_{i1} \geq -X_{i1}\beta$ . Thus we cannot form the transition probability  $P(d_{i2} = 2 | v_{i1}, X_{i2})$ . We can only form:

$$P(d_{i2} = 2 | d_{i1} = 2, X_{i1}, X_{i2}) = P(d_{i2} = 2 | v_{i1} \geq -X_{i1}\beta, X_{i2}) \quad (9)$$

Notice that both the lagged choice and lagged covariates are informative about the distribution of  $v_{i2}$  as they enable us to infer its truncation (i.e.,  $v_{i1} \geq -X_{i1}\beta$ ). And, given that the errors are serially correlated, we have a conditional density of the form  $f(v_{i2} | v_{i1} \geq -X_{i1}\beta)$ .

The computational problem that arises in discrete choice panel data models becomes obvious when we move to period 3. Now, the fact that  $(d_{i1} = 2, d_{i2} = 2)$  tells us only that  $v_{i1} \geq -X_{i1}\beta$  and  $v_{i2} \geq -X_{i2}\beta$ . Therefore, we have that:

$$P(d_{i3} = 2 | d_{i1} = 2, d_{i2} = 2, X_{i1}, X_{i2}, X_{i3}) = P(d_{i3} = 2 | v_{i1} \geq -X_{i1}\beta, v_{i2} \geq -X_{i2}\beta, X_{i3}) \quad (10)$$

The point is that the history at  $t = 1$  still matters for the  $t = 3$  choice probability, because of the fact that  $v_{i1} \geq -X_{i1}\beta$  contains additional information about the distribution of  $v_{i3}$  beyond that contained in the  $t = 2$  outcome,  $v_{i2} \geq -X_{i2}\beta$ . Thus, the conditional density of  $v_{i3}$  has the form  $f(v_{i3} | v_{i1} \geq -X_{i1}\beta, v_{i2} \geq -X_{i2}\beta)$ . And the probability of the sequence (2, 2, 2) is:

$$\begin{aligned} &\int_{-X_{i1}\beta}^{\infty} \int_{-X_{i2}\beta}^{\infty} \int_{-X_{i3}\beta}^{\infty} f(v_3 | v_1 \geq -X_{i1}\beta, v_2 \geq -X_{i2}\beta) \\ &f(v_2 | v_1 \geq -X_{i1}\beta) f(v_1) dv_3 dv_2 dv_1. \end{aligned} \quad (11)$$

Thus, the probability of a three period sequence is a 3-variate integral. In general, the probability of a  $T$  period sequence is a  $T$ -variate integral, as the *entire* history matters for the choice probability in any period. If we consider  $J > 2$ , then the probability of a  $T$  period sequence is a  $T \cdot (J - 1)$  variate integral. This explains the severe computational burden of estimating panel probit models.

This problem is in sharp contrast to a linear model with serially correlated errors, such as:

$$y_{it} = x_{it}\beta + \varepsilon_{it} \quad \text{where} \quad \varepsilon_{ijt} = \rho \varepsilon_{ij,t-1} + \eta_{ijt} \quad \eta_{ijt} \sim iid \quad (12)$$

Here we can form  $E(y_{it}|x_{it}, \varepsilon_{i,t-1}) = x_{it}\beta + \rho\varepsilon_{i,t-1}$  because, conditional on any estimate of  $\beta$ , we observe the lagged error  $\varepsilon_{i,t-1} = y_{i,t-1} - x_{i,t-1}\beta$ . Similarly, if we could observe  $v_{i1}$  and  $v_{i2}$  in the probit model, then, letting  $v_1^*$  and  $v_2^*$  denote the observed values, equation (11) becomes:

$$\int_{-X_{i1}\beta}^{\infty} f(v_1)dv_1 \int_{-X_{i2}\beta}^{\infty} f(v_2|v_1 = v_1^*)dv_2 \int_{-X_{i3}\beta}^{\infty} f(v_3|v_1 = v_1^*, v_2 = v_2^*)dv_3 \quad (13)$$

Thus the sequence probability would simply be the product of three univariate integrals. The basic idea of the GHK algorithm is to draw values of the unobserved lagged  $v_t$ 's and condition on these, enabling us to use equations like (13) to evaluate sequence probabilities rather than (11).

Guided by the structure in (13), the GHK simulator of the sequence probability in (11) is:

$$\begin{aligned} & \hat{P}_{GHK,D}(d_{i1}=2, d_{i2}=2, d_{i3}=2|X_i) \\ &= \frac{1}{D} \sum_{d=1}^D \int_{-X_{i1}\beta}^{\infty} f(v_1)dv_1 \int_{-X_{i2}\beta}^{\infty} f(v_2|v_1 = v_1^d)dv_2 \\ & \quad \int_{-X_{i3}\beta}^{\infty} f(v_3|v_1 = v_1^d, v_2 = v_2^d)dv_3 \end{aligned} \quad (14)$$

where  $\{v_1^d, v_2^d\}_{d=1}^D$  are draws from the *conditional* distributions of  $v_1$  and  $v_2$  given that option 2 was chosen in both periods 1 and 2. So GHK replaces the 3-variate integral in (11) by three univariate integrals, and two draws from truncated normal distributions.

A key aspect of GHK is to draw the  $\{v_1^d, v_2^d\}_{d=1}^D$  sequences in (14) appropriately. The first step is to construct the Cholesky decomposition of the covariance matrix  $\Gamma$  of the error vector  $(v_{i1}, v_{i2}, v_{i3})$ . So, let  $\Gamma = AA'$ , where  $A$  is the Cholesky matrix with elements  $\alpha_{tt'}$ . Then we have:

$$\begin{pmatrix} v_{i1} \\ v_{i2} \\ v_{i3} \end{pmatrix} = \begin{pmatrix} 1 & & \\ a_{21} & a_{22} & \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} \eta_{i1} \\ \eta_{i2} \\ \eta_{i3} \end{pmatrix} \quad (15)$$

We set  $a_{11} = 1$  to impose that  $\sigma_{i1}^2 = 1$ , which is the identifying scale restriction on utility. It is straightforward to draw  $\eta_{i1} = v_{i1}$  from a truncated standard normal such that  $v_{i1} \geq -X_{i1}\beta$ . This can be done by drawing a uniform  $u_1^d$  on the interval  $[F(-X_{i1}\beta), 1]$  and then setting  $\eta_1^d = F^{-1}(u_1^d)$ .

Next, we have that  $v_{i2} = a_{21}\eta_1^d + a_{22}\eta_2$ . Thus, the truncation  $v_{i2} \geq -X_{i2}\beta$  implies truncation on  $\eta_2$  of the form  $\eta_{i2} \geq \frac{1}{a_{22}}[-X_{i2}\beta - a_{21}\eta_1^d]$ . So we now draw a uniform  $u_2^d$  on the interval  $F(\frac{1}{a_{22}}[-X_{i2}\beta - a_{21}\eta_1^d])$ , and set  $\eta_2^d = F^{-1}(u_2^d)$ . This process can be repeated multiple times for person  $i$  so as to obtain a set of draw sequences  $\{v_{i1}^d, v_{i2}^d\}_{d=1}^D$ . Consistent with our earlier discussion, it is the uniform draws  $\{u_{i1}^d, u_{i2}^d\}_{d=1}^D$  that should be held fixed as one iterates.

The GHK algorithm can be extended to more than three periods in an obvious way, by adding additional terms to (14). The bound on the time  $t$  draw is always of the form  $\eta_{it} \geq \frac{1}{a_{tt}}[-X_{it}\beta - a_{t1}\eta_1^d \cdots - a_{t,t-1}\eta_{t-1}^d]$ . With  $T$  periods one needs to evaluate  $T$  univariate integrals and draw  $T - 1$  truncated normals. These operations are extremely fast compared to  $T$ -dimensional integration.

Finally, GHK can be easily extended to more complex error structures. In the above example, the  $\nu_{it}$  had a random effects structure, so  $\Gamma$  was equicorrelated. But an important fact is that the algorithm does not change in any way if  $\Gamma$  has a more complex structure, such as that which would arise if the AR(1) parameter  $\rho$  were nonzero.

### 18.3.3 Alternatives to the Panel Probit Model

In this section, I consider some popular alternatives to the panel probit model. First, consider the case where the  $\alpha_{ij}$  are normal while the  $\eta_{ijt}$  are extreme value. This gives the normal mixture of logits model (N-MIXL). It has been studied extensively by Berry (1994), Berry, Levinsohn, and Pakes (1995), Harris and Keane (1999), McFadden and Train (2000), Train (2003), and others. Letting  $J$  be the base alternative, the choice probabilities have the form:

$$P(d_{ij(t),t}|X_{it},\alpha_i) = \exp(x_{ij(t),t}\beta + \alpha_{ij(t)}) / \left[ 1 + \sum_{j=1}^{J-1} \exp(x_{ijt}\beta + \alpha_{ij}) \right] \quad (16)$$

The probability simulator for this model is closely related to the frequency simulator in (5), except here we use a logit kernel rather than a probit kernel. As before, let  $\{\alpha_d\}_{d=1,\dots,D}$  denote  $D$  random vectors  $(\alpha_{1d}, \dots, \alpha_{J-1d})$  drawn from the  $f(\alpha|\Sigma)$  density, and form the frequency simulator:

$$\begin{aligned} \hat{P}_{MIXL,D}(d_{ij(1),1}, \dots, d_{ij(T),T} | X_{i1}, \dots, X_{iT}) \\ = \frac{1}{D} \sum_{d=1}^D \prod_{t=1}^T \exp(x_{ij(t),t}\beta + \alpha_{j(t),d}) / \left[ 1 + \sum_{j=1}^{J-1} \exp(x_{ijt}\beta + \alpha_{jd}) \right] \end{aligned} \quad (17)$$

One advantage of N-MIXL is that, in contrast to the random effects probit, once we condition on the individual effects  $\alpha_i = (\alpha_{i1}, \dots, \alpha_{i,J-1})$ , the choice probability integrals have a closed form given by the logit kernel  $\exp(\cdot)/[1 + \exp(\cdot)]$ . This makes simulation of the model rather fast and easy.

By introducing correlation across alternatives via the  $f(\alpha|\Sigma)$  distribution, N-MIXL relaxes the strong IIA assumption of multinomial logit. A number of papers consider more general distributions for  $\alpha$  than the normal. For instance, Geweke and Keane (1999), Rossi, Allenby, and McCulloch (2005), and Burda, Harding, and Hausman (2008) consider mixture-of-normals models. Indeed, an entire family of MIXL models can be obtained by different choices of the  $f(\alpha|\Sigma)$  distribution.

The next set of models that have been popular in the consumer demand literature are “latent class” models. In these models there are a discrete set of consumer types, each with its own vector of brand-specific individual effects. That is, we have  $(\alpha_{i1}, \dots, \alpha_{iJ-1}) \in (\alpha_1^c, \dots, \alpha_{J-1}^c)$  where  $c = 1, \dots, C$  indexes types or classes. One estimates both the  $\alpha^c$  vector for each class  $c$ , as well as the population proportion of each class,  $\pi^c$ . We then obtain unconditional choice sequence probabilities by taking the weighted sum over type-specific probabilities:

$$P_{LC}(d_{ij(1),1}, \dots, d_{ij(T),T} | X_{i1}, \dots, X_{iT}) = \sum_{c=1}^C \pi_c \prod_{t=1}^T P(d_{ij(t),t} | X_{it}, \alpha^c) \quad (18)$$

Here  $P(d_{ij(t),t} | X_{it}, \alpha^c)$  is typically a logit or probit kernel. We can interpret the latent class model as a special case of MIXL where the mixing distribution is discrete (in contrast to the normal mixing distributions we considered earlier). Note that the probability in (18) is analytical when a logit kernel is used (no simulation methods are needed).

To my knowledge, Kamakura and Russell (1989) was the first paper to apply the latent class approach in marketing. Work by Elrod and Keane (1995) showed that the latent class approach tends to underestimate the degree of heterogeneity in consumer preferences. I think it is fair to say that with the advent of simulation methods, latent class models have become relatively less widely used (at least in academic research) compared to probit and mixed logit models that allow for continuous heterogeneity distributions.

Recently Keane and Wasi (2013) used multiple data sets to compare the fit of latent class models to several alternative models with continuous heterogeneity distributions (including N-MIXL and a mixture-of-normals model). We found that models with continuous heterogeneity distributions typically provided a much better fit. Nevertheless, we also found that the simple structure of latent class models often gives useful insights into the structure of heterogeneity in the data, helping one to understand and interpret results from more complex models. Thus, it appears that latent class models still have a useful role to play in interpreting discrete choice demand data, even if they are outperformed by other models in terms of fit and predictive ability.

### 18.3.4 Extension to Serially Correlated Taste Shocks

So far, I have conducted the discussion of methods for estimating the model in equations (1)–(2) in the case where  $\gamma = \rho = 0$ . That is, there is no state dependence and the idiosyncratic errors (or taste shocks)  $\varepsilon_{ijt}$  are serially independent. Then the only source of serial correlation was brand-specific individual effects. I now consider generalizations of this model. As we discussed in Section 18.2, it is quite plausible that unobserved brand preferences vary over time rather than being fixed. An example is the AR(1) process in (1). Starting with Keane (1997) and Allenby and

Lenk (1994), a number of papers have added AR(1) errors to the random effects structure.

It is simple to discuss extension of the methods we have described to the case of serially correlated  $\varepsilon_{ijt}$ , as in every case the extension is either simple or practically impossible. For instance, the Butler and Moffitt (1982) quadrature procedure relies specifically on the random effects probit structure and it cannot be extended to serial correlation in  $\varepsilon_{ijt}$ .

Latent class models are designed specifically to deal with permanent unobserved heterogeneity, so they cannot handle serially correlated idiosyncratic errors. In principle one could have a model with both discrete types and serially correlated idiosyncratic errors. But the resultant model would no longer generate closed form choice probabilities as in (18). They would only be estimable using simulation methods.

On the other hand, the random effects probit model can easily be extended to include serially correlated idiosyncratic errors (like the AR(1) structure in (1)). To estimate this model using the GHK algorithm, one simply constructs the covariance matrix  $\Gamma$  in a way that incorporates the additional source of serial correlation. Then, construct the corresponding Cholesky matrix, and draw the  $\{\nu_1^d, \dots, \nu_{T-1}^d\}_{d=1}^D$  sequences in (14) accordingly. Unlike the random effects case, the  $\Gamma$  will no longer be equicorrelated. But the algorithm described in equations (13)–(15) does not change in any way if  $\Gamma$  has a more complex structure.

The frequency simulator in (5) can also be extended to allow for serially correlated  $\varepsilon_{ijt}$ . For instance, take the model  $U_{ijt} = \alpha_{ij} + X_{ijt}\beta + \varepsilon_{ijt}$ , where  $\varepsilon_{ijt} = \rho\varepsilon_{ij,t-1} + \eta_{ijt}$  and  $\alpha_i \sim N(0, \Sigma)$  and  $\eta_{it} \sim N(0, \Omega)$ . We can, *in principle*, simulate choice probabilities in this model just by drawing the  $\alpha_i$  and  $\{\eta_{i1}, \dots, \eta_{iT}\}$  from the appropriate distributions, constructing the implied utilities, and counting the frequency with which each option is preferred.

However, this approach is not practical, because if we draw the entire composite error  $\nu_{ijt} = \alpha_{ij} + \varepsilon_{ijt}$  the model will deterministically generate particular choice sequences that satisfy equation (7). So, equation (5) would become:

$$\hat{P}_{F,D}(d_{ij(1),1}, \dots, d_{ij(T),T} | X_{i1}, \dots, X_{iT}) = \frac{1}{D} \sum_{d=1}^D \prod_{t=1}^T I(d_{ij(t),t} | X_{it}, \alpha_d, \varepsilon_{dt}), \quad (19)$$

where  $I(d_{ij(t),t} | X_{it}, \alpha_d, \varepsilon_{dt})$  is an indicator function for the choice  $d_{ij(t),t}$  being observed at time  $t$  given the draws  $\alpha_d$  and  $\varepsilon_{dt}$ . The practical problem is that the number of possible sequences is  $J^T$ . As we noted in the introduction, this is a very large number even for modest  $J$  and  $T$ . As a result, most individual sequences have very small probabilities. Hence, even for large  $D$  the value of (19) will often be zero. As Lerman and Manski (1981) discussed, very large simulations sizes are needed to provide accurate simulated probabilities of low probability events.

A potential solution to this problem, proposed by Berkovec and Stern (1991) and Stern (1992), is to reformulate the model to “add noise” and “smooth out” the indicator functions in (19). For instance, we could recast the model as  $U_{ijt} = \alpha_{ij} + X_{ijt}\beta + \varepsilon_{ijt} + \omega_{ijt}$ , where all the serial correlation in the time-varying errors is captured by the  $\varepsilon_{ijt}$  process, while the  $\omega_{ijt}$  are *iid* random variables (perhaps normal or extreme value). Then (7) is replaced by the condition:

$$\begin{aligned} U_{ijt} > U_{ikt} \forall k \neq j \Rightarrow \omega_{ijt} &\geq -(X_{ijt}\beta + \alpha_{dj} + \varepsilon_{djt}) \\ &+ (X_{ikt}\beta + \alpha_{dk} + \varepsilon_{dkt} + \omega_{ikt}) \forall k \neq j \end{aligned} \quad (20)$$

These inequalities generate conditional probabilities  $P_\omega(d_{ij(t),t}|X_{it}, \alpha_d, \varepsilon_{dt})$ , where  $P_\omega(\cdot)$  depends on the distribution of the  $\omega_{ijt}$ . These probabilities are smooth functions of the model parameters provided the  $\omega_{ijt}$  are continuous random variables. Simply plug the  $P_\omega(\cdot)$  into (19) to obtain:

$$\hat{P}_{SF,D}(d_{ij(1),1}, \dots, d_{ij(T),T}|X_{i1}, \dots, X_{iT}) = \frac{1}{D} \sum_{d=1}^D \prod_{t=1}^T P_\omega(d_{ij(t),t}|X_{it}, \alpha_d, \varepsilon_{dt}) \quad (21)$$

Note that there are two ways to interpret (21). One could consider (20) the “true” model and view (21) as an unbiased probability simulator for this model. Alternatively, one could view the errors  $\omega_{ijt}$  as simply a smoothing device, and view (21) as a smoothed version of (19). Such ad hoc smoothing will induce bias in the simulator, as noted by McFadden (1989).

The normal mixture of logits model (N-MIXL), where the  $\alpha_{ij}$  are normal while the  $\eta_{ijt}$  are *iid* extreme value, can also be easily modified to accommodate serially correlated idiosyncratic shocks. Since the probability simulator for this model (equation (17)) is a frequency simulator, the procedure is exactly like what I just described, except in reverse. In this case the extreme value errors  $\omega_{ijt}$ , which are present in the basic model, play the role of the “noise” that smooths the simulated probabilities. It is the serially correlated shocks  $\varepsilon_{ijt}$  that are added.

### 18.3.5 Extension to Include State Dependence

Finally, consider including true state dependence ( $\gamma \neq 0$ ) in the model in (1)–(2). The difficulty here is that we must not only simulate the error terms but also lagged choices. Methods based on frequency simulation are not easily extended to this case. We can easily simulate entire choice histories from the model in (1)–(2) by drawing the  $\alpha_i$  and  $\eta_{it}$  from the appropriate distributions. In each period these draws imply that one choice is optimal, as it satisfies (2). This choice is then treated as part of the history when we move on to simulate data for the next period. So the frequency simulator in

(2) would become:

$$\begin{aligned} & \hat{P}_{F,D}(d_{ij(1),1}, \dots, d_{ij(T),T} | X_{i1}, \dots, X_{iT}) \\ &= \frac{1}{D} \sum_{r=1}^D \prod_{t=1}^T I(d_{ij(t),t} | X_{it}, \alpha_r, \varepsilon_{rt}, d_{r,t-1}), \end{aligned} \quad (22)$$

where  $I(d_{ij(t),t} | X_{it}, \alpha_r, \varepsilon_{rt}, d_{r,t-1})$  is an indicator function for the choice  $d_{ij(t),t}$  being observed at time  $t$  given the draws  $\alpha_r$  and  $\varepsilon_{rt}$  and the lagged simulated choice  $d_{r,t-1}$ . The practical problem here is the same as we discussed in the  $\gamma = 0$  case. The number of sequences is so large ( $J^T$ ) that we are unlikely to obtain draws that are consistent with a consumer's observed choice history. So in most cases (22) will simply be zero. But the smoothing methods discussed in Section 18.3.4 cannot be used here, because the simulated lagged choices  $d_{r,t-1}$  are indicator functions by definition.

In contrast, the GHK algorithm can be easily applied to estimate models that include individual effects, serial correlation, and structural state dependence without any modification to the procedure described earlier. This is because the central idea of the algorithm is to construct random draw sequences that are required to be consistent with a consumer's observed choice history. These are then used to simulate transition probabilities from the choice at  $t - 1$  to the choice at  $t$  (see equations (14)–(15) and the surrounding discussion).

In fact, Keane (1993, 1994) interpreted GHK as an importance sampling technique where stochastic terms are drawn in a constrained way such that they must be consistent with observed choice histories (see (7)). The admissible domain is a convex polyhedron in  $(J - 1)^T$  dimensional space, and it is not feasible to draw from such a region at random. Thus, draws are not taken at random from the distribution given by  $\alpha_i \sim N(0, \Sigma)$ ,  $\eta_{it} \sim N(0, \Omega)$  and  $\rho$ . Rather, this is only used as a source density to generate draws that satisfy the constraints implied by observed choices. Importance sampling weights are then applied to these draws when they are used to construct the probability simulator. That is, when taking the average over draws as in (14), sequences of draws that have greater likelihood under the correct distribution are given more weight. It turns out that in GHK the importance sampling weights simplify to transition probabilities as in (14).

There are also ways to use frequency simulation in conjunction with smoothing or importance sampling to construct feasible simulators in the presence of state dependence. For example Keane and Smith (2003) propose a method based on smoothing the lagged simulated choice  $d_{r,t-1}$ . Keane and Wolpin (2001) and Keane and Sauer (2010) develop an algorithm based on the idea that all discrete outcomes are subject to classification error. Then, any simulated draw sequence has a positive probability of generating any observed choice history. This is the probability of the set of misclassifications needed to reconcile the two histories. But this approach is not likely to be useful in most demand estimation contexts, as scanner data measure choices quite accurately.

## 18.4 TESTING FOR THE EXISTENCE STATE DEPENDENCE

---

A large part of the literature on panel data discrete choice models of consumer demand has been concerned with estimating the degree of true state dependence in choice behavior. Researchers have been concerned with the question of whether, and to what extent, the observed (substantial) persistence in choice behavior over time can be attributed to unobserved individual effects and/or serially correlated tastes, on the one hand, vs. true state dependence, on the other.

We can gain some valuable intuition into the nature of state dependence by considering the linear case. So we reformulate equation (1) to be:

$$U_{it} = \alpha_i + X_{it}\beta + \gamma U_{i,t-1} + \varepsilon_{it} \quad \text{where } \varepsilon_{it} = \rho \varepsilon_{i,t-1} + \eta_{it}, \quad (23)$$

where now  $U_{it}$  is an observed continuous outcome. I have suppressed the  $j$  subscripts to save on notation. By repeated substitution for the lagged  $U_{it}$ , we obtain:

$$\begin{aligned} U_{it} &= \alpha_i \left( \frac{1 - \gamma^t}{1 - \gamma} \right) + (X_{it} + \gamma X_{i,t-1} + \dots + \gamma^{t-1} X_{i,1})\beta \\ &\quad + \gamma^{t+1} U_{ij0} + (\varepsilon_{it} + \gamma \varepsilon_{i,t-1} + \dots + \gamma^t \varepsilon_{i,1}). \end{aligned} \quad (24)$$

Here  $U_{ij0}$  is the initial condition of the process. In conventional panel data analysis, with large  $N$  and small  $T$ , the treatment of initial conditions is often quite critical for the results. But in scanner data panels, where  $T$  is typically very large, results are usually not very sensitive to the treatment of initial conditions. Hence, I will not dwell on this topic here. (Wooldridge (2003a, b) has an excellent discussion of this topic in the large  $N$  small  $T$  case.)

The critical thing to note about (24) is that lagged  $X$ s matter for the current  $U$  iff  $\gamma \neq 0$ . Thus, the key substantive implication of structural state dependence is that lagged  $X$ s help to predict current outcomes. This point was emphasized by Chamberlain (1984, 1985). But, as Chamberlain (1985, p. 14) noted, “In order to make the distinction [between serial correlation and true state dependence] operational, there must be at least one variable which would not have a distributed lag response in the absence of state dependence.” That is, to test for state dependence we need at least one variable  $X_{it}^k$ , where we are sure that lagged  $X_{it}^k$  does not affect  $U_{it}$  directly, but only affects it indirectly through its effect on lagged  $U$ . This is analogous to saying that we have an  $X_{it-1}^k$  that is a valid instrument for  $U_{i,t-1}$  in equation (23).

To be concrete, in consumer demand applications using scanner data, the covariates in  $X$  are typically (i) the observed characteristics of the products in the choice set, which are usually time invariant; (ii) a set of brand intercepts, which capture intrinsic preferences for brands and/or mean preferences for the unobserved attributes of brands; and (iii) the “marketing mix” variables, such as price, promotion activity, and

advertising activity, which are time-varying. As only the marketing mix variables are time-varying, at least one of these (price, display, ad exposures, etc.) must play the role of  $X_{it-1}^k$  in our effort to identify true state dependence.

Is it plausible that a variable like price would affect current demand only through its effect on lagged demand? At first glance the answer may seem completely obvious: Why should the lagged price affect current demand? After all, it doesn't enter the consumer's current budget constraint. Isn't the only plausible story for why lagged price would predict current demand that it shifts lagged demand, which then affects current demand via some state dependence mechanism (like habit persistence, inventory, switching costs, etc.)?

But a closer examination of the issue reveals that there are subtleties. For example, Berry, Levinsohn, and Pakes (1995) argue that prices of different car models may be positively correlated with their unobserved (*to the econometrician*) quality. This would tend to bias price elasticities of demand toward zero. They proposed using exogenous instruments for price to deal with this problem. Notably, however, they considered data with only one or a few periods. In the scanner data context, where there are many periods, it is much more straightforward to use brand intercepts to capture unobserved attributes of brands. In the typical scanner data context, once one controls for brand intercepts, there is no reason to expect that prices are correlated with unobserved attributes of the alternatives.

In contrast to the brand intercepts, which capture mean preferences for the unobserved attributes of products, the  $\alpha_i$  are mean zero random variables, which are interpreted as capturing *heterogeneity* in tastes for unobserved attributes of products. In my view, it is also plausible that prices are uncorrelated with the  $\alpha_i$ . Why would the price of a product be correlated with person  $i$ 's intrinsic taste for that product? One person's tastes are too insignificant a part of total demand to affect the price. In general, I would argue that the random effects assumption:

$$E(\alpha_{ij}|X_{ij1}, \dots, X_{ijT}) = E(\alpha_{ij}) = 0 \quad (25)$$

is plausible when the Xs include only brand attributes and marketing mix variables like price.

Finally, consider the time-varying taste shocks  $\varepsilon_{ijt}$ . It seems highly implausible that idiosyncratic taste shocks of individuals could affect the price of a product. Thus, I would also argue it is quite plausible that the strict exogeneity assumption holds:

$$E(\varepsilon_{ijt}|X_{ij1}, \dots, X_{ijT}) = 0. \quad (26)$$

But this assumes the  $\varepsilon_{ijt}$  are independent across consumers. A source of potential concern is aggregate taste shocks that generate cross-sectional dependence. But again I would argue that, in weekly or daily data, it is implausible that unanticipated aggregate taste shocks could influence the current price. In most instances there is simply not enough time for retailers to assess the demand shift and alter prices so quickly. On the other hand, seasonal demand shocks are presumably anticipated long enough in

advance to be reflected in the price. But, given large  $T$ , one can control for these using seasonal dummies. Thus, I would argue that (26) is plausible even in the presence of aggregate shocks, provided the  $X$  vector includes seasonal dummies.

These arguments support treating price and other marketing mix variables as strictly exogenous in (24), and estimating this equation by random effects. If we further assume that price is a “variable which would not have a distributed lag response in the absence of state dependence,” then we can test for state dependence by testing the significance of lagged prices.

So far I have presented arguments that price is strictly exogenous with respect to idiosyncratic consumer tastes, but I have not yet confronted the question of whether lagged prices might have a direct effect on current demand  $U_{it}$ . In fact, there are a number of reasons to expect they might. I will describe three mechanisms that may generate such an effect:

- (i) *Reference price effects.* There is a large literature in marketing arguing that consumer demand does not depend on price itself but rather on how the price compares to a “reference price.” Key early work in this area was by Winer (1986). The reference price is typically operationalized as the average price of a product, or as some moving average of past prices. Reference price effects were originally motivated by psychological theories of choice. For instance, if the current price is higher than the reference price, the consumer may perceive the price as “unfair” and be unwilling to pay it. But regardless of how one rationalizes the reference price variable, its existence implies that all lagged prices help to predict current demand.
- (ii) *Inventory effects.* Erdem, Imai, and Keane (2003) argued that reference price effects could be motivated as resulting from inventory behavior. If a product is storable, consumers will try to time their purchases for when price is relatively low. This creates an economic rationale for consumers to care about current price relative to a reference price. More generally, consumers are more likely to buy if current price is low relative to expected future prices. Thus, lagged prices matter if they are useful for forecasting future prices.
- (iii) *Price as Signal of Quality.* Another mechanism for lagged prices to have a direct effect on current demand is if consumers have uncertainty about product attributes and use price as a signal of quality. Erdem, Keane, and Sun (2008) estimated a model of this form. In such a model, a history of high prices will cause relatively uninformed consumers to infer that a brand is high quality. As a result, willingness to pay for a product is increasing in its own lagged prices. Of course, such a mechanism becomes less important as consumers gain experience with a product category.

In all the above examples, the true model exhibits some form of dynamics but not what is generally known as true state dependence. As Chamberlain (1985, p. 12) states, “The intuitive notion is that if occupancy of a state affects an individual’s preferences

or opportunities, then there is state dependence.” This intuitive notion does not hold in the above three examples:

- (i) In the reference price model the actual purchase of a brand has no effect on its reference price. Only price realizations affect references prices.
- (ii) In the inventory model, lagged prices of a brand only matter because they affect expected future prices.<sup>3</sup>
- (iii) The signaling model resembles the reference price model in that higher lagged prices increase willingness to pay for a brand.

Conversely, there are plausible cases where lagged prices are insignificant in (24) but true state dependence nevertheless exists. A well-known class of structural models that generates true state dependence is the consumer learning model. In the learning model consumers have uncertainty about product attributes and learn about them over time through use experience, advertising, and other signals. Examples of structural learning models are Eckstein, Horsky, and Raban (1988), Roberts and Urban (1988), Erdem and Keane (1996), Ackerberg (2003), Crawford and Shum (2005), and Ching (2010). In the learning model of Erdem and Keane (1996), which Keller (2002) calls “the canonical *economic* model of brand equity,” consumers are risk-averse with respect to variability in brand quality. As a result, they are willing to pay a premium for familiar brands whose quality is relatively certain, as opposed to less familiar brands with equal expected quality but greater uncertainty. For this reason, lagged purchases affect the current utility evaluation, because they reduce one’s uncertainty about a product’s attributes.

Thus, if we estimate (24), and the true model is a learning model, we might expect to find that lagged prices matter because they influence lagged purchase decisions. But this is not so clear. In the simplest Bayesian learning model, with use experience as the only signal, the perceived variance of brand  $j$  at time  $t$  is:

$$\sigma_{ijt}^2 = [(1/\sigma_{ij0}^2) + N_{ij}(t)(1/\sigma_\varepsilon^2)]^{-1}. \quad (27)$$

Here,  $\sigma_{ij0}^2$  is consumer  $i$ ’s prior uncertainty about the quality of brand  $j$ , while  $\sigma_\varepsilon^2$  is the variability of experience signals.  $N_{ij}(t)$  is the total number of times that consumer  $i$  bought brand  $j$  prior to  $t$ . We would expect lower lagged prices to lead to higher  $N_{ij}(t)$  and hence lower  $\sigma_{ijt}^2$ . But, at the same time, a brand with relatively low  $\sigma_{ijt}^2$ ’s (across all consumers in the market) may charge relatively high prices because it has more brand equity. This leaves the correlation between lagged prices and current demand ambiguous.

This argument amounts to a statement that estimates of (24) may be unrevealing because prices and the  $\sigma_{ijt}^2$  are jointly determined in the learning model—rendering prices endogenous in (24). Fully structural estimation of the learning model resolves this problem by modeling the relationship between prices, the  $N_{ij}(t)$  and the  $\sigma_{ijt}^2$ . But, of course, this requires a strong set of maintained structural assumptions.

In light of the above arguments, I do not believe that the significance or insignificance of prices (or other marketing mix variables) in (24) provides a relatively “assumption free” test of whether true state dependence exists. If lagged prices are significant, it may be because of reference price, inventory, quality signaling, or other factors that cause lagged prices to directly influence current demand. Conversely, insignificance of lagged prices does not necessarily rule out the existence of state dependence, as illustrated by the example of the learning model.

Now consider the additional issues that arise in testing for state dependence in the case of a discrete dependent variable, as in (1)–(2). Recall from our discussion in Section 18.3, that in the case of a random effect but no state dependence (or other forms of serial correlation), we have:

$$P(d_{it}|d_{i1}, d_{i2}, \dots, d_{i,t-1}, X_{i1}, X_{i2}, \dots, X_{i,t-1}, X_{it}) \neq P(d_{it}|X_{it}). \quad (28)$$

Thus, the choice probability at time  $t$  depends on the whole history of the process  $\{d_{is}, X_{is}\}_{s=1}^{t-1}$ , and not just on  $X_{it}$ . In equation (10), we gave a simple intuition for why this occurs, based on a three-period case with only two alternatives, where the consumer chooses option 2 in all three periods:

$$P(d_{i3} = 2|d_{i1} = 2, d_{i2} = 2, X_{i1}, X_{i2}, X_{i3}) = P(d_{i3} = 2|\nu_{i1} \geq -X_{i1}\beta, \nu_{i2} \geq -X_{i2}\beta, X_{i3}) \quad (10')$$

That is, the reason the whole past history helps to predict  $d_{it}$  is that we cannot observe lagged utility, only lagged choices. But information on lagged choices, such as  $d_{i1} = d_{i2} = 2$ , implies conditions like  $\nu_{i1} \geq -X_{i1}\beta$  and  $\nu_{i2} \geq -X_{i2}\beta$ , which are informative about the distribution of the current error. In fact, as we noted earlier, the conditional density of  $\nu_{i3}$  in this case has the form  $f(\nu_{i3}|\nu_{i1} \geq -X_{i1}\beta, \nu_{i2} \geq -X_{i2}\beta)$ . This exact same argument holds regardless of whether the source of serial correlation in the errors is a random effect, serial correlation in the time-varying error component, or both.

As Heckman (1981) discussed, the fact that lagged choices help to predict the current error means that lagged choices will tend to be significant in a discrete choice model with serial correlation, even if there is no true state dependence. This phenomenon is known as “spurious state dependence.” The fact that the whole history matters when there is serial correlation makes it extremely difficult to distinguish true state dependence from serial correlation.<sup>4</sup>

Nevertheless, an important positive result about identification in the probit model is the following: Assume that the errors  $\eta_{ijt}$  in (1) are normal, and that  $\alpha_i \sim N(0, \Sigma)$ , giving a random effects probit. Then the coefficient  $\gamma$  on the lagged dependent variable is identified in (1). This is because, as Chamberlain (1984, p. 1279) notes: “the most general multivariate probit model cannot generate a Markov chain. So we can add a lagged variable and identify  $\gamma$ .” That is, if the multivariate distribution of the composite errors  $\nu_i = \{\nu_1, \dots, \nu_T\}$  is diagonal (no serial correlation), the probit (with  $\gamma = 0$ ) generates that choices are independent over time (conditional on  $X_i$ ). Alternatively, if the errors are serially correlated (but  $\gamma = 0$ ), then the whole history of choices prior to time  $t$  helps to predict the choice at time  $t$ . The intermediate case of a Markov process

cannot be attained, regardless of the specification of the error structure. It can only be attained by including a lagged dependent variable (i.e., allowing  $\gamma \neq 0$ ).

There are two practical implications of these results.

First, if one estimates a discrete choice model without adequately controlling for random effects and serial correlation, then one is likely to find spurious state dependence. Indeed, numerous studies since Guadagni and Little (1983) have found that the estimated strength of state dependence in consumer brand choices declines substantially when one controls for heterogeneity and serial correlation.

Second, within the probit framework, one can test if state dependence exists by including rich controls for heterogeneity and serial correlation and then testing the significance of lagged dependent variables. This approach was pursued in Keane (1997) and in a number of subsequent papers, such as Paap and Franses (2000), Smith (2005), Dubé, Hitsch, and Rossi (2010), and many others. This work consistently finds evidence for the existence of state dependence.

Chamberlain argued, however, that tests within the probit framework were suspect because of their reliance on the probit functional form—in particular, the fact that it is not possible within the probit framework to choose an error structure that generates a Markov chain. Chamberlain (1985, p. 14) went on to suggest that a test based on regressing the current choice on current and lagged  $X$ s (and controlling for heterogeneity) “should not be very sensitive to functional form.” However, we discussed tests based on lagged  $X$ s (especially price) earlier, and found that strong economic assumptions underlie such tests in the consumer demand context.

Chamberlain (1985) went on to argue that a completely non-parametric test for state dependence cannot exist, because one can always find a latent variable  $\alpha_i$  such that:

$$P(d_{it}|X_{i1}, \dots, X_{iT}, \alpha_i) = P(d_{it}|X_{it}, \alpha_i) = P(d_{it}|\alpha_i). \quad (29)$$

That is, one can always find a distribution of  $\alpha_i$  such that  $\{d_{i1}, \dots, d_{iT}\}$  is independent of  $\{X_{i1}, \dots, X_{iT}\}$ . He gives a simple example (p. 1281), where  $\alpha_i$  is simply a unique integer assigned to every different configuration of  $X$ s in the data. This is equivalent to a latent class model with a discrete distribution of types. Each type has its own vector of multinomial choice probabilities. And each configuration of  $X$ s in the data corresponds to a different type. Then, type summarizes all the information in the  $X$ s, giving independence of  $d$  and  $X$  conditional on  $\alpha$ .

Chamberlain defines a relationship of  $X$  to  $d$  as “static” conditional on  $\alpha$  if  $X$  is strictly exogenous (conditional on  $\alpha$ ) and if  $d_t$  is independent of  $\{X_{i1}, \dots, X_{i,t-1}\}$  conditional on  $X_t$  and  $\alpha$ . If a relationship is static, there is no structural state dependence. Equation (29) implies there *always* exists a specification of  $\alpha$  such that the relationship of  $X$  to  $d$  is static. Thus, we cannot test for structural state dependence without imposing some structure on  $P(\cdot | \cdot)$  and the distribution of  $\alpha$ .

However, I do not view this negative result as disturbing. As Koopmans et al. (1950) noted long ago, we cannot learn anything of substance from data without making some a priori structural assumptions (for discussion of this issue, see

Keane 2010a, b). So I would be very surprised if that were not true with regard to drawing inferences about state dependence. In other words, the fact that our inferences about the nature of state dependence, heterogeneity, and serial correlation in tastes are contingent on our modeling assumptions is not at all unique to this set of issues. It is the normal state of affairs throughout economics and the natural sciences as well.<sup>5</sup>

A good example of imposing structure is Chamberlain (1984)'s "correlated random effects probit model," hereinafter CRE. In this model,  $\alpha_i$  is constrained to be a linear function of the time-varying elements of  $X_i$ , which I denote by  $Z_i$ , plus a normal error term, giving:

$$\alpha_{ij} = Z_{ij1}\delta_{j1} + \dots + Z_{ijT}\delta_{jT} + \mu_{ij}. \quad (30)$$

Note that the effect of time-invariant elements of  $X_i$  on  $\alpha_i$  is not identified separately from the intercepts; letting a time-invariant element of  $X_i$  shift  $\alpha_i$  would be equivalent to letting it shift  $X_{it}\beta$  by a constant. Given (30), one can test for state dependence and strict exogeneity.

A CRE model combining (1)–(2) with (30) may be very useful if the Xs are individual characteristics, which obviously may be correlated with preferences (for recent labor applications, see Hyslop 1999; Keane and Sauer 2010). But in the consumer demand context, the Xs are not usually characteristics of people but rather of products, including marketing variables like price and advertising. Here, I think the CRE model is not very compelling.

In particular, I argued earlier that a standard random effects assumption on  $\alpha_i$  is plausible in the consumer demand context (see equation (25)). The most obvious time-varying attribute of a product is price. It is clearly implausible that price would be affected by individual brand preferences. But before ruling out correlation between  $\alpha_i$  and price we should also ask, "What is the source of price variation across consumers and over time?" Erdem, Imai, and Keane (2003) argue that almost all price variation in scanner data is exogenous from the point of view of consumers. Pesendorfer (2002) and Hong, McAfee, and Nayyar (2002) argue that a type of inter-temporal price discrimination strategy on the part of firms, where retailers play mixed strategies, most plausibly explains the frequent week-to-week price fluctuations for frequently purchased consumer goods that we see in scanner data.<sup>6</sup> Such price variation would appear random to consumers.

In light of these observations, I would place considerable confidence in results in the marketing and IO literatures that find substantial evidence of state dependence in consumer choice behavior (provided the studies in question include adequate controls for consumer heterogeneity and serial correlation in tastes). The existence of state dependence is important, as it implies that current marketing actions, such as price discounts, affect not only current but also future demand. But an even more important question is what mechanism generates state dependence. I turn to this question in the next section.

## 18.5 EMPIRICAL WORK ON STATE DEPENDENCE AND SOURCES OF DYNAMICS IN DEMAND

---

In this section, I discuss attempts to identify and quantify sources of state dependence and choice dynamics more generally. The field of marketing has reached rather broad consensus on many key issues related to the dynamics of consumer demand over the past 20 years, as I discuss below. The potential explanations for state dependence include learning, inventories and/or reference prices, habit persistence, variety seeking, and switching costs. All of these have been examined, but learning and inventories have received the most attention in the literature. I will start by discussing some of the more influential work on the functional form of state dependence.

After Guadagni and Little (1983), the main approach to modeling state dependence in the marketing literature was to let current utility depend on an exponentially smoothed weighted average of lagged purchase indicators, denoted  $GL_{ijt}$ . Specifically, replace  $d_{ij,t-1}$  in (1) with:

$$GL_{ijt} = \theta GL_{ij,t-1} + (1 - \theta)d_{ij,t-1} = (1 - \theta)\sum_{s=1}^{t-1} \theta^{s-1} d_{ij,t-s} + \theta^{t-1} GL_{ij1}. \quad (31)$$

Guadagni and Little (GL) famously called  $GL_{ijt}$  the “brand loyalty” variable. The smoothing parameter  $\theta \in [0, 1]$  determines how quickly the impact of lagged purchases on current utility decays. If  $\theta = 0$ , then only  $d_{ij,t-1}$  matters and we are back to a first order Markov process as in (1). As  $\theta \rightarrow 1$ , we get substantial inertia in brand preferences. For typical panel lengths and reasonable values of  $\theta$  the initial setting of  $GL_{ij1}$  is not very important.

GL estimated their model using scanner data on coffee purchases of 100 households in Kansas City for 32 weeks in 1979. They estimated a MNL model with eight alternatives. But they had no controls for heterogeneity or serial correlation in preferences (as this was not technically possible in 1983). Their complete model implied that “brand loyalty,” along with price and promotional activity, are strong predictors of brand choice.

Keane (1997) considered the impact of allowing for random effects and AR(1) errors in a model with the GL form of state dependence. The data cover 51 weeks of ketchup purchases by 1,150 consumers in Sioux Falls, South Dakota, in 1987–88. The choice set contained seven alternatives,<sup>7</sup> and up to 30 purchases per household. Thus, the required order of integration for the model with AR(1) errors is  $T \cdot (J - 1) = 180$ , and choice probabilities were evaluated using the GHK algorithm.

Keane assumed that  $\alpha_i \sim N(0, \Sigma)$  and  $\eta_{it} \sim N(0, \Omega)$ , giving a multinomial multi-period probit model. A major problem was that unrestricted  $\Sigma$  and  $\Omega$  would contain  $T \cdot (J - 1) \cdot J/2 - 1 = 631$  parameters. To deal with this, he assumed that both  $\Sigma$  and  $\Omega$  had a one factor structure. Then the covariance structure is characterized by (i) the

AR(1) parameter  $\rho$ ; (ii) the six factor loadings on the common factor that underlies  $\Sigma$ ; (iii) the uniquenesses of  $\Sigma$ , which are assumed equal for all brands and denoted by  $\kappa$ ; and (iv) the same seven parameters for  $\Omega$ . This gives only 15 parameters. Although this structure is very parsimonious, additional factors were not significant.

One goal of Keane (1997) was to give a taxonomy of types of heterogeneity. He argued that to rationalize the most general models in the literature on needs seven types: (i) observed and unobserved heterogeneity in tastes for observed attributes; (ii) observed heterogeneity in brand intercepts; (iii) unobserved heterogeneity in tastes for unobserved common and unique attributes for which consumers have fixed tastes; and (iv) the same for attributes where consumers have time-varying tastes. The basic strategy in Keane (1997) was to add more and more types of heterogeneity and see how estimates of state dependence were affected.

Keane's "Model 1" is very similar to Guadagni and Little (1983) but with normal errors (panel probit). He estimates  $\theta = .813$  and  $\gamma = 1.985$ . Note that  $\gamma(1 - \theta) = .37$  is the extra utility from buying brand  $j$  at  $t$  if you bought it at  $t - 1$ . The estimate of the price coefficient is  $-1.45$ , so this is equivalent to a 27 cent price cut. As mean price is roughly \$1.20, this is about a 22.5% price cut.

Keane's "Model 2" eliminates state dependence but includes heterogeneity in brand intercepts of the form  $\alpha_i \sim N(0, \kappa I_{J-1})$ . So we have unique factors but no common factors. The unique factors account for 48% of total error variance, implying substantial taste heterogeneity.

Keane's "Model 3" includes both the GL form of state dependence and " $\kappa$ -heterogeneity" (i.e., unique factors). When both are included, each becomes less important. The fraction of the error variance due to unique factors drops to 31%. We now get  $\theta = .833$ ,  $\gamma = 0.889$ , and a price coefficient of  $-1.66$ . So the effect of lagged purchase is equivalent to only a 9-cent price cut.

In the full model ("Model 16"), which includes all seven types of heterogeneity,  $\gamma = 1.346$  and  $\theta = .909$ . The price coefficient is heterogeneous, but for a typical family it is about  $-2.4$ . So lagged purchase has an effect on demand that is similar to roughly a 5-cent price cut (4%).

The effect of a purchase today on the probability of a purchase tomorrow is known as the "purchase carry-over effect" in marketing. The bottom line of Keane (1997) is that extensive controls for heterogeneity reduce the estimated carry-over effect from being the equivalent of a 22.5% price cut to being the equivalent of only a 4% price cut—thus reducing the carry-over effect by roughly 80%. So most of the observed persistence in brand choice does appear to be due to taste heterogeneity, but there is still a significant fraction that is due to state dependence.<sup>8</sup>

Of course, as we discussed in Section 18.4, inferences about the relative importance of heterogeneity and state dependence are always functional form dependent. Erdem and Keane (1996) showed that a Bayesian learning model implies a very different form of state dependence from that in GL. In their model, prior to receiving any information, consumers perceive that the true quality of brand  $j$ , denoted  $Q_j$ , is distributed

normally with mean  $Q_{j0}$  and variance  $\sigma_{j0}^2$ . Over time a consumer receives noisy information about a brand through use experience and ad signals. Let  $d_{jt}$  be an indicator for whether brand  $j$  is bought at time  $t$ , and let  $\sigma_\varepsilon^2$  denote the noise in experience signals. Let  $d_{jt}^A$  be an indicator for whether an ad for brand  $j$  is seen at time  $t$ , and let  $\sigma_A^2$  denote the noise in ad signals. Let  $N_j(t)$  and  $N_j^A(t)$  denote the total number of experience and ad signals received up through time  $t$ , respectively. Finally, let  $Q^E$  denote experience signals and  $A$  denote ad signals. Then the Bayesian learning model implies:

$$Q_{jt} = \frac{(1/\sigma_\varepsilon^2)}{(1/\sigma_{j0}^2)} \sum_{s=1}^{t-1} Q_{js}^E d_{js} + \frac{(1/\sigma_A^2)}{(1/\sigma_{jt}^2)} \sum_{s=1}^t A_{js} d_{js}^A + \frac{(1/\sigma_{j0}^2)}{(1/\sigma_{jt}^2)} Q_j \quad (32)$$

$$\sigma_{jt}^2 = \frac{1}{(1/\sigma_{j0}^2) + N_j(t)(1/\sigma_\varepsilon^2) + N_j^A(t)(1/\sigma_A^2)}. \quad (33)$$

Here,  $Q_{jt}$  is the perceived quality of brand  $j$  based on information received up through time  $t$ , and  $\sigma_{jt}^2$  is the perception error variance.

Note that the Bayesian learning model implies a very different form of state dependence than GL. First, note that more lagged purchases ( $N_j(t)$ ) reduce perceived uncertainty about the quality of a brand ( $\sigma_{jt}^2$ ). If consumers are risk-averse with respect to quality variation, this makes familiar brands more attractive, generating state dependence. The Bayesian framework in (33) implies that only the *total* number of lagged purchases of a brand,  $N_j(t)$ , matters for its current demand, while the GL framework in (31) implies that more recent experience is more important.

A more subtle difference between the models is that, in the learning model, heterogeneity and state dependence are not neatly separable phenomena. In (32), perceived quality of brand  $j$  at time  $t$ ,  $Q_{jt}$  is a function of all quality signals received up through  $t$ . This is heterogeneous across consumers—some will, by chance, receive better quality signals than others. Thus, heterogeneity in brand preferences evolves through time via the same process that generates state dependence.

Because the  $Q_{jt}$  are serially correlated random variables, which depend on lagged signals, we must use simulation to approximate the likelihood. What we have is a very complex mixture of logits model, with the mixing distribution given by the distribution of the  $Q_{jt}$ . The method used to simulate the likelihood is a smooth frequency simulator, like that presented in equation (21), with the  $\varepsilon_{dt}$  playing the role of the draws for the  $Q_{jt}$ .

Erdem and Keane (1996) compared a Guadagni and Little (1983) style model with a Bayesian learning model where state dependence is governed by (32)–(33).<sup>9</sup> They used Nielsen scanner data on liquid detergent purchases of 167 households in Sioux Falls, South Dakota, for 51 weeks in 1987–88. Telemeters were attached to panelists' TVs to measure ad exposures. The data include seven brands and a no purchase option. Three brands were introduced during the period, generating variability in brand familiarity.

Erdem and Keane (1996) augment the GL model by including a GL-type variable for ad exposures. Thus, both past use experience and ad exposures affect current utility.

When Erdem and Keane (1996) estimated the GL model they obtained  $\theta = .770$  and  $\gamma = 3.363$ , so  $\gamma(1 - \theta) = .773$ . The price coefficient was  $-1.077$ , implying that the impact of lagged purchase is equivalent to roughly a 72-cent price cut. Mean price is roughly \$3.50, so this is 21%. This is strikingly close to the effect Keane (1997) found in the GL model for ketchup. Surprisingly, the  $\gamma$  for advertising was only 0.14 with a standard error of .31 (not significant). Thus, the GL model implies the awkward result that advertising has no effect on demand!

However, Erdem and Keane (1996) found the Bayesian learning model gave a much better fit to the data than the GL model. The log likelihood (LL) and Bayes Information Criterion (BIC) for the GL model were  $-7463$  and  $7531$ . But for the learning model they obtained LL and BIC values of  $-7312$  and  $7384$ . Thus, the BIC improvement is 147 points. The key parameters that generate state dependence are  $\sigma_{j0}^2 = 0.053$ ,  $\sigma_\varepsilon = 0.374$ , and  $\sigma_A = 3.418$ .

The Erdem-Keane model is too complex to give simple calculations of the impact of lagged choices on current demand as we did with the GL and Keane (1997) models. The effects of price changes and changes in ad exposure frequency can only be evaluated by simulating the model. Unfortunately, Erdem-Keane only report advertising and not price simulations. But they do find clear evidence of state dependence in the advertising simulations. As they state, “although the short run effect of advertising is not large, advertising has a strong cumulative effect on choice over time as it gradually reduces the perceived riskiness of a brand.”<sup>10</sup>

Based on the evidence in Erdem and Keane (1996) and Keane (1997), as well as a large body of subsequent work, much of which is very well described by Neslin (2002a,b), there is now a broad consensus on three issues: (i) state dependence in demand does exist; (ii) as a result, both price promotion and advertising have long run effects; but (iii) consumer taste heterogeneity is a much stronger source of the observed persistence in choice behavior than is state dependence.

In contrast to the consensus on existence of state dependence, there is no clear consensus on its source. The Guadagni and Little (1983) and Keane (1997) types of model can be viewed as structural models where prior use experience literally increases the utility of current consumption of a brand through a habit persistence mechanism. Alternatively, these models can be viewed as flexible approximations to a broad (but unspecified) set of models that generate state dependence that is well described by the “brand loyalty” variable. The Erdem and Keane (1996) model and the large body of subsequent work derived from it (for a review, see Ching, Erdem, and Keane 2013) definitively take a stand that state dependence derives from the learning mechanism. Other work, especially Erdem, Imai, and Keane (2003) and Hendel and Nevo (2006), posits that inventories are an importance source of dynamics. Erdem, Keane and Sun (2008) show that the learning and inventory mechanisms are actually very hard to disentangle empirically, if one allows for a priori consumer taste heterogeneity. Thus,

there is little consensus on the relative importance of the different *mechanisms* that may generate state dependence.

The third key research objective that I mentioned in the introduction is to understand the dynamics of demand. Most important is to understand the sources of the observed increase in demand when a brand is on sale. Here, I think the literature has reached a high degree of consensus. Consider the demand for frequently purchased consumer goods. There is broad consensus that own price elasticities (given temporary price cuts) are about  $-3$  to  $-4.5$ . But it is also widely accepted by firms and academics that just knowing how much sales go up when you cut prices is not very interesting. What really matters is where the increase comes from.

Erdem, Imai, and Keane (2003) and Erdem, Keane, and Sun (2008) estimate that roughly 20–30% of the increase in sales due to a temporary price cut is cannibalization of future sales. Of the remaining incremental sales, 70–80% is due to category expansion and only about 20–30% is due to brand switching. It is hard to exaggerate the importance of this three-way decomposition of the price elasticity of demand, as it determines the profitability of price promotion. And a remarkable consensus has emerged on these figures in recent years. Some key papers on cannibalization rates are van Heerde, Leeflang, and Wittink (2000, 2004) and Ailawadi et al. (2006). And some important studies of brand switching are Pauwels, Hanssens, and Siddarth (2002), van Heerde, Gupta, and Wittink (2003), Sun, Neslin, and Srinivasan (2003), and Mace and Neslin (2004).

## 18.6 CONCLUSION

---

As we have seen, there is broad consensus that state dependence in consumer demand exists. There is also clear evidence that dynamic demand models fit the data much better than static models (see Ching, Erdem, and Keane 2009, 2013). And there is broad agreement that only about 20–25% of the incremental sales that accompany a price cut is due to brand switching, with the rest due to category expansion and cannibalization of own future sales. On the other hand, there is little agreement on the fundamental mechanism that generates dynamics in demand. The main competing theories are learning, inventories, and habit persistence. Progress in this area is severely hindered by the computational difficulty of nesting all these mechanisms in one model, but Ching et al. (2014) discuss some new advances in this area.

Much of demand modeling is done with the ultimate goal of merging the demand side with supply side models of industry competition. Such equilibrium models can be used for merger analysis, advertising regulation, anti-competitive pricing regulation, etc. But existing work in this area has typically used static demand models, due to the

computational difficulty of solving the problem of oligopolistic firms when demand is dynamic.

Unfortunately, static demand models greatly exaggerate cross-price elasticities, as they attribute far too much of incremental sales to brand switching (see Sun, Neslin, and Srinivasan 2003; Erdem, Imai, and Keane 2003; Erdem, Keane, and Sun 2008). As cross-price elasticities of demand summarize the degree of competition between products, this upward bias creates serious problems in attempting to predict effects of mergers. This example makes clear the importance of further work on developing dynamic models, particularly models sophisticated enough to capture observed dynamics, yet simple enough to merge with supply side models.

## ACKNOWLEDGMENTS

---

I thank the editor and an anonymous referee for helpful comments. This work was supported by Australian Research Council grant FL110100247.

## NOTES

---

- <sup>1</sup> Product attributes can be “vertical” or “horizontal.” A vertical attribute is something like quality that all consumers would like more of. A horizontal attribute is something like saltiness of crackers, which some people would like and others dislike. Thus, for horizontal attributes, even the sign of  $\beta$  may differ across consumers.
- <sup>2</sup> In the one-dimensional case, optimal quadrature points and weights can be found in references like Abramowitz and Stegun (1964). Stroud (1971) presents an extensive treatment of the two-dimensional case.
- <sup>3</sup> Of course, lagged purchases do affect current inventory, which is a state variable. And, if there are inventory carrying costs, consumers are less likely to buy, *ceteris paribus*, if current inventory is high. Furthermore, current inventory is more likely to be high in cases where recent lagged prices were low. But note that inventory is affected by past purchase of a *category*, not purchase of a particular *brand*.
- <sup>4</sup> Note that a random effect will generate a situation where all lags are equally informative about the current error term. In contrast, a process like a stationary AR(1) generates a situation where more recent choices are more informative, although the whole history still matters. Even if the errors are MA(1), the whole history of the process helps to predict the current choice.
- <sup>5</sup> Chamberlain (1985)’s negative results on nonparametric identification of state dependence do raise some interesting methodological questions. I will not attempt to address them here, but it is still worth raising them: (i) Chamberlain allows for extraordinarily

general patterns of heterogeneity. Does Occam's razor (or just common sense modeling practice) suggest limiting ourselves to much more parsimonious forms like (25) or (30)? (ii) It is not clear how a model where  $\alpha_i$  is allowed to depend in a very general way on time-varying  $X$ s can be used for forecasting. Should we limit ourselves to more parsimonious models in the interest of forecasting ability? (iii) In light of Chamberlain's negative results, and our own discussion surrounding equation (24), should we conclude that state dependence is not a useful construct in demand modeling? Would it be more fruitful to focus directly on modeling the dynamics of how lagged  $X$ s affect current and future choices, without the mediating concept of state dependence? (iv) Alternatively, is the state dependence construct useful because it enables us to develop more convenient and parsimonious functional forms compared to including many lagged covariates in a model?

6. There are sensible arguments for why consumer types may be correlated with brand prices, but I do not believe they are empirically relevant. Scanner data is typically collected from all the (large) stores in a particular area, like Sioux Falls, South Dakota, or Springfield, Missouri. So regional variation is not a potential source of price variation, but cross-store variation potentially is. However, while it is likely that stores differ in their average price level (e.g., some stores are more "up-scale," or are located in wealthier areas, and therefore charge higher prices in general), it is not clear why *relative* prices of brands would differ by store. Another idea is that consumers may actively seek out stores where their preferred brand is on sale. Or, even if they regularly visit only one store, to time visits for when that store is having a sale on their preferred brand. Such behavior *might* be relevant for expensive goods (e.g., meat, wine, diapers), but I doubt that anyone would decide when or what store to visit based on the price of Oreo cookies. Some years ago I attempted (in joint work with Tulin Erdem) to develop a model of store choice based on prices of various items. But we abandoned the project as we could not find any products that predicted store choice.
7. These were Hunt's (32 oz.), Del Monte (32 oz.), and five sizes of Heinz (40, 64, 14, 28, and 32 oz.). For Heinz the 32 and 14 oz. were glass and the other sizes were plastic. The Heinz 40-oz. size was introduced during the sample, creating a nice source of variation in the choice set. Heinz 32 oz. is the base alternative whose utility is normalized to zero.
8. Short run vs. long run price elasticities of demand are also of interest. In model 1 a 50% price cut leads to 257% sales increase in current period (elasticity of 5.1) but only about a 17% sales increase in subsequent periods (elasticity of roughly 0.34). In model 16 a 50% price cut leads to 313% sales increase in current period (elasticity of 6.3) but only about a 12% sales increase in subsequent periods (elasticity of roughly 0.24).
9. Erdem and Keane estimated two versions of their model where consumers are either myopic or forward-looking. Here I discuss only the myopic version, which is very similar to GL except for the different form of state dependence. The myopic model can be estimated using methods discussed in Section 18.3. The forward-looking version requires dynamic programming, which is beyond the scope of this chapter.
10. Unfortunately, their paper contains a major typo in a key figure (Figure 1) that shows this result. Figure 1 in their paper just duplicates Figure 3. Fortunately,

the basic result can also be seen in Figure 2 (for the model with forward-looking consumers).

## REFERENCES

---

- Abramowitz, M., and I. A. Stegun (1964), *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*, National Bureau of Standards Applied Mathematics Series, Vol. 55, US Government Printing Office, Washington, DC.
- Ackerberg, D. (2003), “Advertising, learning, and consumer choice in experience good markets: A structural empirical examination,” *International Economic Review*, 44(3), 1007–1040.
- Ailawadi, K., K. Gedenk, C. Lutzky, and S. Neslin (2007), “Decomposition of the sales impact of promotion-induced stockpiling,” *Journal of Marketing Research*, 44(3), 450–467.
- Allenby, G. M., and P. J. Lenk (1994), “Modeling household purchase behavior with logistic normal regression,” *Journal of American Statistical Association*, 89, 1218–1231.
- Andrews, R. L., and A. K. Manrai (1999), “MDS maps for product attributes and market response: An application to scanner panel data,” *Marketing Science*, 18(4), 584–604.
- Berkovec, James, and Steven Stern (1991), “Job exit behavior of older men,” *Econometrica*, 59(1), 189–210.
- Berry, Steven (1994), “Estimating discrete choice models of product differentiation,” *RAND Journal of Economics*, 25, 242–262.
- Berry, S., J. Levinsohn, and A. Pakes (1995), “Automobile Prices in Market Equilibrium,” *Econometrica*, 63, 841–890.
- Block, H., and J. Marschak (1960), “Random Orderings and Stochastic Theories of Response,” in I. Olkin, ed., *Contributions to Probability and Statistics*, pp. 97–132, Stanford University Press, Stanford, CA.
- Burda, M., M. Harding, and J. Hausman (2008), “A Bayesian mixed logit-probit model for multinomial choice,” *Journal of Econometrics*, 147, 232–246.
- Butler, J. S., and Robert Moffitt (1982), “A computationally efficient quadrature procedure for the one-factor multinomial probit model,” *Econometrica*, 50(3), 761–764.
- Chamberlain, Gary (1984), “Panel Data,” in Z. Griliches and M. Intriligator, eds., *Handbook of Econometrics*, Volume 2, pp. 1247–1318, North-Holland, Amsterdam.
- Chamberlain, Gary (1985), “Heterogeneity, Omitted Variable Bias, and Duration Dependence,” in J. Heckman and B. Singer, eds., *Longitudinal Analysis of Labor Market Data*, pp. 3–38, Cambridge University Press, Cambridge.
- Ching, A. T. (2010), “Consumer learning and heterogeneity: Dynamics of demand for prescription drugs after patent expiration,” *International Journal of Industrial Organization*, 28(6), 619–638.
- Ching, A., T. Erdem, and M. Keane (2009), “The price consideration model of brand choice,” *Journal of Applied Econometrics*, 24(3), 393–420.
- Ching, A., T. Erdem, and M. Keane (2013), “Learning models: An assessment of progress, challenges and new developments,” *Marketing Science*, 32(6), 913–938.
- Ching, A., T. Erdem, and M. Keane (2014), “A Simple Approach to Estimate the Roles of Learning and Inventories in Consumer Choice,” working paper, Nuffield College.

- Crawford, G., and M. Shum (2005), "Uncertainty and learning in pharmaceutical demand," *Econometrica*, 73(4), 1137–1173.
- Dubé, Jean-Pierre, Günter J. Hitsch, and Peter E. Rossi (2010), "State dependence and alternative explanations for consumer inertia," *RAND Journal of Economics*, 41(3), 417–445.
- Elrod, Terry (1988), "Choice map: Inferring a product map from observed choice behavior," *Marketing Science*, 7(Winter), 21–40.
- Elrod, T., and M. Keane (1995), "A factor-analytic probit model for representing the market structure in panel data," *Journal of Marketing Research*, 32, 1–16.
- Erdem, T., and M. Keane (1996), "Decision making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets," *Marketing Science*, 15(1), 1–20.
- Erdem, T., S. Imai, and M. Keane (2003), "Brand and quantity choice dynamics under price uncertainty," *Quantitative Marketing and Economics*, 1(1), 5–64.
- Erdem, T., M. Keane, and B. Sun (2008), "A dynamic model of brand choice when price and advertising signal product quality," *Marketing Science*, 27(6), 1111–1125.
- Geweke, J., and M. Keane (1999), "Mixture of Normals Probit Models," in C. Hsiao, K. Lahiri, L. F. Lee, and H. Pesaran, eds., *Analysis of Panels and Limited Dependent Variable Models*, pp. 49–78, Cambridge University Press, Cambridge.
- Geweke, J., and M. Keane (2001), "Computationally Intensive Methods for Integration in Econometrics," in J. J. Heckman and E. E. Leamer, eds., *Handbook of Econometrics*, Volume 5, pp. 3463–3568, Elsevier Science B.V., Amsterdam.
- Geweke, J., M. Keane, and D. Runkle (1994), "Alternative computational approaches to statistical inference in the multinomial probit model," *Review of Economics and Statistics*, 76(4), 609–632.
- Geweke, J., M. Keane, and D. Runkle (1997), "Statistical inference in the multinomial multiperiod probit model," *Journal of Econometrics*, 80, 125–165.
- Guadagni, Peter M., and John D. C. Little (1983), "A logit model of brand choice calibrated on scanner data," *Marketing Science*, 2(Summer), 203–238.
- Hajivassiliou, Vassilis, Daniel McFadden, and Paul Ruud (1996), "Simulation of multivariate normal rectangle probabilities and their derivatives theoretical and computational results," *Journal of Econometrics*, 72(1–2), 85–134.
- Harris, K., and M. Keane (1999), "A model of health plan choice: Inferring preferences and perceptions from a combination of revealed preference and attitudinal data," *Journal of Econometrics*, 89, 131–157.
- Heckman, J. J. (1981), "Heterogeneity and State Dependence," in S. Rosen, ed., *Studies in Labor Markets*, pp. 91–140, University of Chicago Press, Chicago.
- Hendel, I., and A. Nevo (2006), "Measuring the implications of sales and consumer inventory behavior," *Econometrica*, 74(6), 1637–1673.
- Hong, Pilky, R. Preston McAfee, and Ashish Nayyar (2002), "Equilibrium price dispersion with consumer inventories," *Journal of Economic Theory*, 105, 503–517.
- Hyslop, Dean R. (1999), "State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women," *Econometrica*, 67(6), 1255–1294.
- Kamakura, Wagner, and Gary Russell (1989), "A probabilistic choice model for market segmentation and elasticity structure," *Journal of Marketing Research*, 26, 379–390.

- Keane, M. (1992), "A note on identification in the multinomial probit model," *Journal of Business and Economic Statistics*, 10(2), 193–200.
- Keane, Michael 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 2: Econometrics*, pp. 545–571, Elsevier Science Publishers, Amsterdam.
- Keane, Michael P. (1994), "A computationally practical simulation estimator for panel data," *Econometrica*, 62(1), 95–116.
- Keane, Michael P. (1997), "Modeling heterogeneity and state dependence in consumer choice behavior," *Journal of Business and Economic Statistics*, 15(3), 310–327.
- Keane, Michael P. (2010a), "Structural vs. atheoretic approaches to econometrics," *Journal of Econometrics*, 156(1), 3–20.
- Keane, Michael P. (2010b), "A structural perspective on the Experimentalist School," *Journal of Economic Perspectives*, 24(2), 47–58.
- Keane, M. P., and R. Sauer (2010), "A computationally practical simulation estimation algorithm for dynamic panel data models with unobserved endogenous state variables," *International Economic Review*, 51(4), 925–958.
- Keane, M. P., and Tony Smith (2003), "Generalized Indirect Inference for Discrete Choice Models," Working paper, Yale University.
- Keane, M. P., and N. Wasi (2013), "Comparing alternative models of heterogeneity in consumer choice behavior," *Journal of Applied Econometrics* 28(6), 1018–1045.
- Keane, M. P., and K. Wolpin (2001), "The effect of parental transfers and borrowing constraints on educational attainment," *International Economic Review*, 42(4), 1051–1103.
- Keller, Kevin (2002), "Branding and Brand Equity," in B. Weitz and R. Wensley, eds., *Handbook of Marketing*, pp. 151–178, Sage Publications, London.
- Koopmans, T. C., H. Rubin, and R. B. Leipnik (1950), "Measuring the Equation Systems of Dynamic Economics," in T. C. Koopmans, ed., *Cowles Commission Monograph No. 10: Statistical Inference in Dynamic Economic Models*, pp. 53–237, John Wiley & Sons, New York.
- Lancaster, Kelvin J. (1966), "A new approach to consumer theory," *Journal of Political Economy*, 74, 132–157.
- 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. 305–319, MIT Press, Cambridge, MA.
- Mace, S., and S. Neslin (2004), "The determinants of pre- and postpromotion dips in sales of frequently purchased goods," *Journal of Marketing Research*, 41(3), 339–350.
- McFadden, D. (1974), "Conditional Logit Analysis of Qualitative Choice Behavior," in P. Zarembka, ed., *Frontiers in Econometrics*, pp. 105–142, Academic Press, New York.
- McFadden, D. (1989), "A method of simulated moments for the estimation of discrete response models without numerical integration," *Econometrica*, 57(5), 995–1026.
- McFadden, D., and K. Train (2000), "Mixed MNL models for discrete response," *Journal of Applied Econometrics*, 15, 447–470.
- Mousavi, S. E., H. Xiao, and N. Sukumar (2010), "Generalized Gaussian quadrature rules on arbitrary polygons," *International Journal for Numerical Methods in Engineering*, 82, 99–113.
- Neslin, Scott (2002a), *Sales Promotion*, Relevant Knowledge Series, Marketing Science Institute, Cambridge, MA.

- Neslin, Scott A. (2002b), "Sales Promotion," in Barton A. Weitz and Robin Wensley, eds., *Handbook of Marketing*, pp. 310–338, Sage Publications, London.
- Paap, Richard, and Philip Hans Franses (2000), "A dynamic multinomial probit model for brand choice with different long-run and short-run effects of marketing-mix variables," *Journal of Applied Econometrics*, 15(6), 717–744.
- Pauwels, K., D. Hanssens, and S. Siddarth (2002), "The long-term effects of price promotions on category incidence, brand choice, and purchase quantity," *Journal of Marketing Research*, 39(4), 421–439.
- Pesendorfer, Martin (2002), "Retail sales: A study of pricing behavior in supermarkets," *Journal of Business*, 75(1), 33–66.
- Rossi, P., G. Allenby, and R. McCulloch (2005), *Bayesian Statistics and Marketing*, John Wiley and Sons, Hoboken, NJ.
- Smith, Martin D. (2005), "State dependence and heterogeneity in fishing location choice," *Journal of Environmental Economics and Management*, 50(2), 319–340.
- Srinivasan, T. C., and Russell Winer (1994), "Using neoclassical consumer-choice theory to produce a market map from purchase data," *Journal of Business and Economic Statistics*, 12 (January), 1–9.
- Stern, Steven (1992), "A method for smoothing simulated moments of discrete probabilities in multinomial probit models," *Econometrica*, 60(4), 943–952.
- Stigler, George J. (1984), "Economics—the imperial science?" *Scandinavian Journal of Economics*, 86(3), 301–313.
- Stroud, A. H. (1971), *Approximate Calculation of Multiple Integrals*, Prentice-Hall, Englewood Cliffs, NJ.
- Sun, B., S. Neslin, and K. Srinivasan (2003), "Measuring the impact of promotions on brand switching under rational consumer behavior," *Journal of Marketing Research*, 40(4), 389–405.
- Train, K. (2003), *Discrete Choice Methods with Simulation*, Cambridge University Press, Cambridge.
- Van Heerde, H., S. Gupta, and D. Wittink (2003), "Is 75% of the sales promotion bump due to brand switching? No, only 33% is," *Journal of Marketing Research*, 40(4), 481–491.
- Van Heerde, H., P. Leeflang, and D. Wittink (2000), "The estimation of pre- and post-promotion dips with store-level scanner data," *Journal of Marketing Research*, 37(3), 383–395.
- Van Heerde, H., P. Leeflang, and D. Wittink (2004), "Decomposing the sales promotion bump with store data," *Marketing Science*, 23(3), 317–334.
- Winer, Russell S. (1986), "A reference price model of brand choice for frequently purchased products," *Journal of Consumer Research*, 13(September), 250–256.
- Wooldridge, J. (2003a), *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, MA.
- Wooldridge, J. (2003b), "Simple Solutions to the Initial Conditions Problem in Dynamic, Nonlinear Panel Data Models with Unobserved Heterogeneity," Working paper, Michigan State University.

## CHAPTER 19

---

# PANEL ECONOMETRICS OF LABOR MARKET OUTCOMES

---

THOMAS J. KNIESNER AND JAMES P. ZILIAK

### 19.1 INTRODUCTION

---

OVER the past four decades, advances in panel data econometrics have been intertwined with advances in labor economics. The tight link surely owes in part to the early availability of household panel data such as the Panel Study of Income Dynamics (PSID), the National Longitudinal Survey (NLS), and the negative income tax experiments. But it also stems from the fact that panel data offer numerous benefits for labor market research in terms of economically and econometrically richer models. This led to seminal research on life-cycle models of labor supply that developed new panel methods for separating state dependence from unobserved heterogeneity, allowing endogenous wages (Heckman 1978; MaCurdy 1981), considering earnings dynamics with growth rate heterogeneity and other autocorrelation processes (Lillard and Weiss 1979; MaCurdy 1982), modeling human capital investments that yielded new approaches to controlling for latent ability (Hausman and Taylor 1981), and to considering labor market interventions that in turn yielded new techniques for the evaluation of programs (Heckman and Robb 1985). The use of panel data also come with additional complications for labor market research that in turn led to advances in panel econometrics, such as how to control for attrition (Hausman and Wise 1979) and measurement error (Griliches and Hausman 1986).

In our chapter we cover some of the most commonly encountered issues in panel-data applications in labor economics.

Specifically, our interests here include whether and how to introduce heterogeneity in intercept and slope parameters; measurement errors in regressors; endogeneity bias and associated panel instrumental variables estimators; sample composition dynamics to control for selection on (un)observables; and model specification and selection

issues such as a static or dynamic framework. The formal econometric theory behind many of the topics we cover appears elsewhere (Baltagi 2008 and Chapters 3, 5, and 11 of the handbook). We also do not cover other important labor-related topics in our brief survey, such as program evaluation and limited-dependent variable models (Chapters 6 and 9). Our aim is instead to demonstrate how the panel-data techniques labor economists use manifest themselves in an applied setting, specifically the canonical hedonic labor-market equilibrium model of the wage-fatal risk trade-off (Thaler and Rosen 1975).

The pedagogical framework of labor-market wage-fatal risk setting is of interest for several reasons. First, the empirical framework builds on the standard Mincer wage equation used in scores of papers on inequality, returns to schooling and experience, and race and gender discrimination. In the case we consider, the Mincer model is augmented with a proxy for the risk of a fatal injury on the job, which is the central parameter that governs estimates of the so-called value of statistical life (VSL) or value of mortality risk reduction (VMRR) (Viscusi 2013). That is, the VSL captures a worker's willingness to trade off wages for a small probability of death on the job and, as such, is used in a variety of settings to assess the cost-effectiveness of health and safety programs. So, for example, a group of workers each requiring a \$700 annual wage premium to face an extra 1 in 10,000 probability of death on the job would have a VSL of \$7 million ( $700/0.0001$ ). A second reason for focusing on the compensating wage model is that obtaining reliable estimates of the VSL has long been challenging owing to the central roles of latent individual heterogeneity that is correlated with the regressors and state dependence which affects both the market offer curve and individual preferences. Perhaps surprising, with the lone exception of Brown (1980), the VSL literature has not until recently used panel data techniques to control for the variety of econometric problems in estimation (Kniesner, Viscusi, and Ziliak 2010; Kniesner et al. 2012). A third reason to focus on the wage-fatal risk model is that the wide variation of VSL estimates in the literature, which spanned from \$0 to \$20 million (Viscusi and Aldy 2003; Viscusi 2013), has generated concern that underlying econometric problems may jeopardize the validity of the estimates.

We begin by demonstrating the critical importance of controlling for latent intercept heterogeneity to estimates of the VSL by comparing results from pooled OLS and between-groups estimators to estimates obtained from standard random and fixed effects estimators. Here we find a real payoff to panel data in that the estimated VSL falls by 75% once allowing for intercept heterogeneity. But we also obtain the surprising result in our focal application that how one controls for intercept heterogeneity matters little. Typically random effects are statistically rejected in favor of fixed effects, and the same is true in our application, but economically there is little distinction in VSL estimates across fixed and random effects models, which is not necessarily a standard result in labor. We then examine how the VSL varies across the wage distribution by using some relatively new quantile estimators for panel data (Koenker 2004; Lamarche 2010), showing how there is nontrivial slope heterogeneity in the VSL.

Once we develop some basics of heterogeneity we next explore possible solutions to measurement error in the panel setting. Our emphasis here is on mismeasured covariates and, in particular, in our measure of fatality risk. Because on-the-job fatalities are relatively rare in many industries and occupations, there are possible concerns that measurement error may bias downward estimates of the VSL (Black and Kniesner 2003). Panel data offer several possible solutions, including averaging across several periods as is common in the intergenerational mobility literature (Solon 1992), taking wider differences to raise the signal to noise (Griliches and Hausman 1986), or instrumental variables (Baltagi 1981, 2008). We find that measurement error (attenuation) bias is pervasive in our application and tends to bias estimates of the VSL downward by around 25%. This is fairly consistent across techniques.

More generally, labor economists confront the challenge of endogenous regressors, and again panel data offer a wider array of possible solutions than is available in cross-sectional or repeated cross-sectional data. With few exceptions, most research in labor relies on exclusion restrictions to identify endogenous variables. For example, early work on cross-sectional labor supply would often use nonlinearities and interaction terms to identify the endogenous wage rate (Pencavel 1986); that is, if age and education affect hours of work, then age squared and age times education would be included in the reduced-form wage equation. In the last two decades labor economists have considered the so-called natural experiment approach to identifying endogenous regressors, exploiting some exogenous variation in policy that is external to the model (Card 1990; Card and Krueger 1994). Panel data admit both approaches and permit additional instruments by imposing structure on the time-series process (MacCurdy 1982) or the latent heterogeneity (Hausman and Taylor 1981; Cornwell, Schmidt, and Wyhowski 1992; Keane and Runkle 1992). Estimates of the VSL are fairly robust to several of the alternative approaches.

Missing data in the form of item nonresponse, wave nonresponse, and panel non-response also pervades panel data research in labor. This leads to unbalanced panels where the key issue is whether the unbalanced design imparts any bias on labor market parameters such as the risk of fatality. That is, determining whether the data are missing completely at random, missing conditionally at random, or missing non-randomly is key to consistency of model estimates (Baltagi and Chang 1994, 2000; Verbeek and Nijman 1992; Ziliak and Kniesner 1998; Baltagi and Song 2006; Wooldridge 2007). Few, if any, applications in labor satisfy the missing completely at random assumption, but the missing conditionally at random is more widely applicable with robust controls for selection on observables (Fitzgerald, Gottschalk, and Moffitt 1998). Ziliak and Kniesner (1998) find that modeling the missing non-randomly process as a person-specific fixed effect is adequate for males' life-cycle labor supply estimates, and the addition of a selection correction based on the idiosyncratic time-varying error adds little. This implies that the decision to attrite is time invariant, and a similar result holds in our VSL application here, suggesting that at least for some labor outcomes, controlling for latent heterogeneity is sufficient for nonrandom attrition.

Panel data offer opportunities to expand the economic framework from the static setting to the life cycle. Moreover, within the context of the life-cycle model, additional model richness can be explored such as time nonseparability in preferences owing to state dependence (Hotz, Kydland, and Sedlacek 1988), or nonseparability in budgets owing to endogenous human capital (Shaw 1989) or tax policy (Ziliak and Kniesner 1999). Related, there has been an explosion of research on the econometrics of dynamic panels (Arellano and Bond 1991; Chapters 4 and 14, this volume). We conclude our chapter with a summary of what research that panel data have facilitated in the important area of wage outcomes. Among other things, we note how estimates of the VSL differ in the short run versus the long run when we attempt to separate state dependence from latent heterogeneity. We find that the long-run parameter is about 10–15% higher than the short-run VSL, suggesting that the adjustment to labor-market equilibrium in wages is relatively quick.

## 19.2 ECONOMIC FRAMEWORK

---

The organizing economic framework for our chapter on panel-data labor econometrics is based on the compensating differences model of the wage-fatal injury risk trade-off. The most common motivation for this hedonic model is a desire to identify society's willingness to pay for alternative policy effects, such as the expected number of lives saved from a government policy or regulation. The standard measure used to capture the willingness-to-pay value is the trade-off rate between money and fatal injury risks—higher risk jobs are compensated with higher wages—or what is known as the value of statistical life (VSL). Because there is great heterogeneity in the risk of injury across jobs, the most common method of eliciting the VSL is via labor-market studies that estimate the effect of fatal injuries on hourly wage rates, or

$$\ln w_i = \alpha + \beta \pi_i + X_i \gamma + u_i, \quad (1)$$

where  $\ln w_i$  is the natural log of the hourly wage rate of worker  $i$ ,  $\pi_i$  is the measure of labor-market risk of fatal injury on the job,  $X_i$  is a vector of demographic controls found in the typical Mincer wage equation such as education, race, age, and marital status, and  $u_i$  is an idiosyncratic error term.

Hedonic equilibrium in the labor market means that equation (1) traces out the locus of labor market equilibria involving the offer curves of firms and the supply curves of workers. This implies that the focal parameter of interest is  $\beta$  as this summarizes the wage-fatal risk trade-off and is central to the VSL, defined as

$$VSL_i = \frac{\partial w_i}{\partial \pi_i} = 100,000 \times \beta \times w_i \times h_i, \quad (2)$$

where the expression in equation (2) accounts for the fact that the model in equation (1) is estimated in log wages, but the VSL is expressed in dollars and thus  $\beta$  is multiplied by the hourly wage. In addition the fatality rate  $\pi_i$  is generally measured in terms of the number of fatalities per 100,000 workers and thus we need to scale the VSL up by 100,000. Lastly, because the wage rate is measured as wages per hour of work, to arrive at an annual estimate of the VSL we multiply by annual hours of work,  $h_i$ . Even though the rate at which wages are traded for risk is constant across workers in equation (1) via a common  $\beta$ , the VSL can vary across workers as the expression in (2) can be evaluated at different points in the wages and hours distributions. In our examples below, we follow most studies and report only the effect at mean wages and a fixed hours point of 2000. Viscusi and Aldy (2003) provide a comprehensive review of estimates of the VSL based on cross-sectional data. Our primary objective in this chapter is to examine how following systematic econometric practices for panel data models refines the estimated range of VSL.

### 19.3 HETEROGENEOUS INTERCEPTS AND SLOPES

---

The most widely stated advantage of panel data in labor economics is that it permits the introduction of latent heterogeneity in intercepts; that is, with data on the same worker  $i$  ( $i = 1, \dots, N$ ) for multiple time periods  $t$  ( $t = 1, \dots, T$ ) we can rewrite the hedonic model as

$$\ln w_{it} = \alpha_i^+ + \alpha_i^- + \beta \pi_{it} + X_{it}\gamma + u_{it}, \quad (3)$$

where equation (3) contains two latent individual effects: one that is positively correlated with wages and the fatality rate ( $\alpha_i^+$ ) and one that is positively correlated with wages and negatively correlated with the fatality rate ( $\alpha_i^-$ ).

The first individual effect reflects unmeasured individual differences in personal safety productivity that leads higher wage workers to take what appears to be more dangerous jobs because the true danger level for such a worker is lower than the measured fatality rate; the second individual effect reflects unmeasured job productivity that leads more productive/higher wage workers to take safer jobs.<sup>1</sup> The economic interpretation of  $\alpha_i$  in the Mincer-type wage model for returns-to-schooling is that it captures pre-labor market factors that are fixed over time such as unobserved labor-market skill that reflects nature (genetic factors based down through families), nurture (environmental factors determined by familial and social forces), and the possible interaction of nature and nurture. One of the factors in the schooling model is the usual latent person-specific cognitive ability that makes the person have a higher wage rate and go to school more, the other is a trait such as beauty or likability that makes the person have a higher wage rate but go to school less because one does not need as much investment (Heckman et al. 2008). In a life-cycle model the latent heterogeneity will also embed the marginal utility of initial wealth (MacCurdy 1981).

For the remainder of the analysis, we suppress the  $+/-$  distinction on the latent heterogeneity.

If we impose the following zero conditional mean assumptions A.1 and A.2

$$\text{A.1 } E[u_{it}|\alpha_i, \pi_{it}, X_{it}] = 0$$

$$\text{A.2 } E[\alpha_i|\pi_{it}, X_{it}] = 0,$$

then OLS estimation of the hedonic equilibrium in equation (3) using pooled cross-section time-series data is consistent. The standard errors are serially correlated because of the presence of  $\alpha_i$  and need to be clustered to account for repeated observations of the same individual. An alternative to pooled OLS under the same assumptions is the between-groups estimator, which as we discuss below in the section on measurement error is potentially useful in mitigating attenuation bias as the averages smooth out idiosyncratic noise in any given time period.

In the first two columns of Table 19.1, we report the results of pooled OLS and between groups estimates of the fatality rate coefficient,  $\hat{\beta}$ , along with the implied VSL evaluated at the mean wage of \$21 and 2,000 hours of work (Kniesner et al. 2012). The data are of prime-age working men from the 1993–2002 Panel Study of Income Dynamics (PSID). The panel is unbalanced and in later sections we discuss the role of possible nonrandom nonresponse.<sup>2</sup> Both estimates reveal that there is economically and statistically strong evidence of a wage-fatal risk trade-off. The pooled OLS model shows that the implied VSL is just over \$15 million, while the between-groups estimate suggests it is closer to \$26 million, perhaps indicating significant attenuation bias. The estimates of the VSL are within the range of cross-sectional results summarized in Viscusi and Aldy (2003).

**Table 19.1 Linear cross-section and panel data estimates of value of statistical life**

	Pooled OLS Estimator	Between- Groups Estimator	Random- Effects Estimator	First- Difference Estimator	Difference-in- Differences Estimator
	(1)	(2)	(3)	(4)	(5)
Annual Fatality Rate	0.0037 (0.0013)	0.0063 (0.0021)	0.0015 (0.0007)	0.0013 (0.0006)	0.0015 (0.0006)
Implied VSL (\$Millions)	15.4 [5.3, 25.6]	25.9 [8.9, 42.9]	6.2 [0.3, 12.2]	5.8 [0.8, 10.8]	6.8 [1.1, 12.5]

Source: Calculations based on Kniesner et al. (2012). Standard errors are in parentheses, and 95% confidence intervals are in square brackets.

### 19.3.1 Intercept Heterogeneity Models

The pooled OLS and between groups estimators ignore the latent heterogeneity in equation (3), and thus under assumptions A.1 and A.2 a more efficient estimator is possible with the random-effects estimator. The random-effects model is a GLS estimator by explicitly including the latent heterogeneity term  $\alpha_i$  in the model's error structure to account for autocorrelation, but is similar to OLS in that the additional source of error is treated as exogenous to the regressors. The random effects structure is attractive for several reasons, including that it permits inference to the general population and it preserves identification of time-invariant regressors. For example, race is a typical covariate in Mincer wage models, and once there are controls for other observed factors of productivity, the race coefficient reflects the unexplained racial wage gap. Under random effects we can tighten the estimate of the racial gap by inclusion of (random) unobserved productivity. However, the implication of assumption A.2 for our application to the hedonic wage model is that selection into possibly risky occupations and industries on the basis of unobserved productivity and tastes is purely random across the population of workers. Such random sorting is unlikely to hold in the data, and in most labor-market applications the data reject the assumption of zero correlation with unobserved heterogeneity.

Instead, most work in labor economics relaxes A.2 to allow correlation between the latent heterogeneity and the regressors,

$$\text{A.2}' \quad E[\alpha_i | \pi_{it}, X_{it}] \neq 0,$$

yielding the fixed-effects estimator, whether in the form of the least-squares dummy variable model, the within estimator, the first-difference estimator, or the orthogonal deviations estimator.

The least-squares dummy variable model is convenient when there are only a limited number of cross-sectional units; for example, in US state panel-data models we need only include 50 dummy variables for each state, but the dummy variable method becomes impractical in most household panel surveys with large  $N$ . Most often the decision confronting the labor economist is whether to apply the within or first-difference transformation in equation (3). The two estimators yield identical results when there are two time periods and when the number of periods converges toward infinity. When there is a finite number of periods ( $T > 2$ ), estimates from the two different fixed-effects estimators can diverge due to possible non-stationarity, measurement errors, or model misspecification (Wooldridge 2010). Because wages, hours, and consumption from longitudinal data on individuals have been shown to be non-stationary in other contexts (Abowd and Card 1989; MaCurdy 2007), in our application to the hedonic wage model we adopt the first-difference model. However, first-differences comes at a cost of fewer observations than the within estimator because the first time period must be dropped for each cross-sectional unit, and as we discuss below, it may exacerbate measurement errors.

Lillard and Weiss (1979) demonstrated that earnings functions may not only have idiosyncratic differences in levels but also have idiosyncratic differences in growth. A straightforward approach to modeling growth rate heterogeneity is to put an idiosyncratic factor loading on a linear trend

$$\ln w_{it} = \alpha_i + \beta \pi_{it} + X_{it}\gamma + \delta_i t + u_{it}, \quad (4)$$

where  $\delta_i$  reflects person-specific growth in wages. Lillard and Weiss (1979) treated the growth heterogeneity as random in their error-components model, but a more robust approach is to modify assumption A.2' as

$$\text{A.2}'' \quad E[\alpha_i, \delta_i | \pi_{it}, X_{it}] \neq 0.$$

Notice that first differencing equation (4) still leaves  $\delta_i (= \delta_i t - \delta_i(t-1))$ . In models with limited cross-sections, it is possible to add dummy variables akin to the least squares dummy variable approach to control for latent growth heterogeneity (this is common among applications using 50-state panel data), but again in large panels it is impractical. Instead one can double difference the model as

$$\Delta^2 \ln w_{it} = \beta \Delta^2 \pi_{it} + \Delta^2 X_{it}\gamma + \Delta^2 u_{it}, \quad (5)$$

where  $\Delta^2 \equiv \Delta_t - \Delta_{t-1}$ , commonly known in the evaluation literature as the difference-in-difference operator.

In columns (3)–(5) of Table 19.1, we report estimates of the fatal risk parameter and the corresponding estimate of the VSL for the random-effects estimator, the first-difference estimator, and the double-difference estimator. A formal Breusch-Pagan test rejects the null hypothesis of common intercepts as assumed in the pooled OLS and between-groups estimators, and the rejection has significant economic implications in our application—the VSL from random effects is 60% lower than the pooled OLS estimate, and 75% lower than the between-groups estimate. In many labor-market applications there is not such a divergence between the OLS and random effects estimate as here, but it underscores the importance of testing for heterogeneous effects and focusing attention on the economic consequences of the homogeneity assumption.

Although a formal Hausman test rejects random effects in favor of fixed effects, comparing columns (3) and (4) shows that the estimated VSL is fairly robust to the random effects assumption. Again, formal testing can determine the statistical and economic consequences of seemingly innocuous assumptions. In the final column, we see that the estimated VSL is robust to growth heterogeneity as the difference-in-difference estimate of the VSL is comparable to the random effects and first-difference models.

The takeaway of our empirical example is that controlling for latent intercept heterogeneity is crucial to produce more accurate estimates of the VSL, but how one controls for the latent heterogeneity is less important. Although we believe the former result of the primacy of controlling for latent heterogeneity pervades most applications in labor economics, the latter result that the form of the heterogeneity does not matter economically is perhaps more unique to the VSL

application but underscores the importance of estimating alternative models in practice.

### 19.3.2 Intercept and Slope Heterogeneity Models

Models with marginal effect heterogeneity where regression coefficients are individual specific fixed or random effects are well known (Swamy 1971; Chapter 12 this volume), but until recently have been much less widely adopted in practice.<sup>3</sup> Here our exploration of the heterogeneity of regression parameters considers slope differences using a recently developed model of panel quantile regression (Koenker 2004; Lamarche 2010). We demonstrate the applicability of the estimator to an important policy question, which is how VSL might be indexed for income growth (Kniesner, Viscusi, and Ziliak 2010). Specifically, because the VSL from a quantile wage regression varies with the potential wage ( $\hat{w}$ ), it admits the possibility that the VSL varies positively with income levels, which captures distributional issues not evident in the mean regression models discussed above. Even a simple comparison of the mean versus median VSL is instructive for safety policy where the VSL is a benefit comparison point for evaluating life-saving programs with different cost levels. In particular, using the median program benefit instead of the mean benefit as a cutoff value ensures that a majority of the affected population will benefit from the program.

In a quantile regression model one no longer has a single parameter vector to estimate but rather a parameter vector for each  $\tau_j^{th}$  quantile, or

$$Q_{lnw_{it}}(\tau_j | \pi_{it}, X_{it}, \alpha_i) = \alpha_i(\tau_j) + \beta(\tau_j)\pi_{it} + X_{it}\gamma(\tau_j). \quad (6)$$

With a finite number of time periods it is general practice to assume that the latent heterogeneity is common to all the conditional quantiles of the regression outcomes,  $\alpha_i(\tau_j) = \alpha_i \forall \tau_j$ . However, even with such quantile invariant intercept heterogeneity the panel quantile model is computationally intensive if there are a large number of cross-sectional units.

Koenker (2004) and Lamarche (2010) offer an innovative solution to the panel quantile version of the incidental parameters problem by proposing a shrinkage estimator wherein a tuning parameter controls the degree of inter-person intercept differences. The tuning parameter, which if fixed ex ante, allows one to vary the degree of heterogeneity from no individual heterogeneity to individual heterogeneity modeled as fixed effects. As an alternative one can solve for the tuning parameter optimally by minimizing the trace of the covariance matrix when estimating the tuning parameter and thus implicitly permit the data to determine the extent of latent heterogeneity.

Let us write the minimization problem as

$$\arg \min(\alpha, \beta, \gamma) \sum_{j=1}^J \sum_{t=1}^T \sum_{i=1}^N \omega_{\tau_j} \rho_{\tau_j} (\ln w_{it} - \alpha_i - \beta(\tau_j) \pi_{it} - X_{it} \gamma(\tau_j)) + \lambda \sum_{i=1}^N |\alpha_i|, \quad (7)$$

where  $\omega_{\tau_j}$  is the relative weight of the  $j^{\text{th}}$  quantile,  $\rho_{\tau_j}(u) = u(\tau_j - I(u \leq 0))$  is the quantile loss function, and  $J$  is the number of quantiles that are estimated simultaneously. The tuning parameter regulates the influence of the latent heterogeneity on the quantile functional. In the case where  $\lambda = 0$  the fixed effects estimator emerges and for the case where  $\lambda > 0$  a penalized (shrinkage) fixed-effects estimator appears (Lamarche 2010). Note that in solving for the unknown parameters in equation (7) there are two kinds of heterogeneity: intercept heterogeneity, whereby the wage equation intercepts vary with the person indicator ( $i$ ); and slope heterogeneity, whereby the curvature of the hedonic locus varies with  $\tau$  to reflect both latent worker and firm differences in risk tolerance and cost functions.

Table 19.2 presents quantile regression estimates of the fatal injury risk for no latent intercept heterogeneity and the regression for  $\lambda = 1$ , which minimized the trace of the variance-covariance matrix, where at the median quantile of 0.5 the estimated VSL is about \$7.6 million, which is less than half that obtained at the same point of the wage distribution but without intercept heterogeneity.<sup>4</sup> Note that there is a sharp increase in the implied VSL at the 75th and 90th quantiles such that the VSL jumps to \$14.6 million and \$22 million, respectively. Again, though, the statistically preferred panel

**Table 19.2 Panel quantile estimates of value of statistical life**

Quantile	No Unobserved Heterogeneity	Implied VSL (in \$Million)	Unobserved Heterogeneity ( $\lambda = \text{Tuning Parameter}$ )	
			$\lambda = 1.0$	Implied VSL (in \$Million)
0.10	0.0025 (0.0010)	4.33	0.0020 (0.0012)	3.46
0.25	0.0022 (0.0009)	5.15	0.0021 (0.0011)	4.92
0.50	0.0049 (0.0010)	16.82	0.0022 (0.0011)	7.55
0.75	0.0062 (0.0010)	31.11	0.0029 (0.0010)	14.55
0.90	0.0075 (0.0015)	55.11	0.0030 (0.0011)	22.04

Source: Calculations based on Kniesner, Viscusi, and Ziliak (2010). Standard errors are in parentheses, calculated from 500 bootstrap replications. The panel quantile bootstrap standard errors are obtained by sampling (with replacement) the dependent variable and regressors for each cross-sectional unit.

quantile estimates are less than half those obtained if we were to incorrectly ignore intercept heterogeneity, highlighting the crucial importance of panel data with models permitting greater heterogeneity.

## 19.4 MEASUREMENT ERROR

---

Labor-market data reported in household surveys are noisy (Bound and Krueger 1991; Bound, Brown, and Mathiowetz 2001). Problems of measurement error generally arise owing to misreporting by the respondent or interviewer, or possibly the respondent refusing to report information and the statistical agency using imputation methods to allocate values. Traditionally measurement error is addressed by using proxy variables or instrumental variables for continuous variables, but in the case of discrete regressors bounding techniques may be necessary (Bollinger 1996).

As detailed in Meijer, Spierdijk, and Wansbeek (Chapter 11), the standard panel data estimator generally exacerbates measurement error compared to the cross-sectional counterpart, and within the class of fixed effect estimators, the first-difference estimator tends to have a lower signal to noise ratio compared to the within estimator (Griliches and Hausman 1986). For example, in Table 19.1 we reported both cross-sectional and first-difference estimates of the wage-fatal risk trade-off, and likely part of the difference in the estimated VSLs across estimators is due to attenuation effects in the first-difference estimates.

At the same time that panel data create additional problems of measurement error, they also open up new avenues for correcting attenuation bias. For example, in Table 19.1, we see that the between-groups estimates of the VSL are nearly 70% higher than the corresponding pooled OLS estimates. Under assumptions A.1 and A.2, the advantage of constructing time means to mitigate measurement error is clear. However, as we demonstrated earlier, the assumption of A.2 is rejected, leading us to panel estimators that admit latent intercept heterogeneity. But measurement error in the fatality risk variable also leads to a violation of assumption A.1 (Black and Kniesner 2003; Ashenfelter and Greenstone 2004b; Ashenfelter 2006).

Griliches and Hausman (1986) provide the seminal treatment of measurement error in panel data, and one transparent solution they suggest is to take wider differences. That is, instead of subtracting the  $(t - 1)$  value of a variable, one instead could subtract the  $(t - 2)$  or further lag of the variables. Indeed, Hahn, Hausman, and Kuersteiner (2007) suggest that it is optimal to take the widest possible difference available in the data. For example, with five years of data on wages the corresponding regression model would be

$$\ln w_{it} - \ln w_{it-4} = \beta(\pi_{it} - \pi_{it-4}) + (X_{it} - X_{it-4})\gamma + (u_{it} - u_{it-4}), \quad (8)$$

**Table 19.3 Instrumental variables estimates of value of statistical life**

	First-Difference IV Estimator, $t - 1$ and $t - 3$ Fatality as Instruments	First-Difference IV Estimator, Lag Differenced Fatality as Instrument	First-Difference IV Estimator, $t - 2$ and $t - 3$ Fatality as Instruments	First-Difference IV Estimator, Lag Differenced Fatality as Instrument	First-Difference IV Estimator, $t - 2$ and $t - 4$ Fatality as Instruments	First-Difference IV Estimator, Lag Differenced Fatality as Instrument
	(1)	(2)	(3)	(4)	(5)	(6)
Annual Fatality Rate	0.0018 (0.0009)	0.0018 (0.0009)	0.0020 (0.0011)	0.0020 (0.0011)	0.0014 (0.0011)	0.0013 (0.0012)
Implied VSL (\$Millions)	7.6 [ -0.1, 15.2]	7.8 [ 0.1, 15.4]	8.7 [ -0.9, 18.4]	8.5 [ -1.1, 18.1]	6.4 [ -3.6, 16.3]	5.7 [ -4.3, 15.7]
Number of Observations	4,338	4,338	4,338	4,338	3,235	3,235

Source: Calculations based on Kniesner et al. (2012). Standard errors are recorded in parentheses and 95% confidence intervals in square brackets. Standard errors are robust to heteroscedasticity and within industry-by-occupation autocorrelation.

which amounts to a simple cross-sectional regression. In results not tabulated, in our application we find that the estimated VSL is about 20% higher in the wide difference model compared to the first-difference estimate reported in Table 19.1.<sup>5</sup>

Another solution to the measurement error problem proposed by Griliches and Hausman (1986) is to exploit the *iid* assumption of the model's error term  $u_{it}$  and to use various combinations of lags of the endogenous regressors as instrumental variables. For example, in the first-difference model, it is possible to use interchangeably the  $(t - 1)$  and  $(t - 3)$  levels of the fatality risk, the  $(t - 1) - (t - 3)$  difference, the  $(t - 2)$  and  $(t - 3)$  levels and difference, and the  $(t - 2)$  and  $(t - 4)$  levels and difference. In Table 19.3, we report the results of various differenced instrumental variable models; the mode result is that the first-difference OLS estimates in Table 19.1 are attenuated by measurement error by about 20% compared to the first-difference IV estimates in Table 19.3.

## 19.5 ENDOGENEITY

One of the most commonly encountered challenges in empirical labor economics is identification of model parameters in the presence of endogenous regressors. Access to

panel data at once introduces additional opportunities for and complications to identification. For example, one opportunity with panel data as described in the last section on measurement error is access to lags of the endogenous regressor ( $\pi_{t-1}, \pi_{t-2}, \dots$ ) in the case of our wage equation model with potentially endogenous fatality rate. A complication, though, is that to use lags of endogenous regressors one has to impose structure on the time-series process. For example, a common approach is to assume the time-varying error  $u_{it}$  is *iid* or MA(1); the *iid* case permits use of instruments dated ( $t - 2$ ) and earlier if the regressor is endogenous, and the MA(1) case permits use of instruments dated ( $t - 3$ ) and earlier under endogeneity. If  $u_{it}$  is an AR(1) process, then lags of the endogenous regressor are ruled out as instruments.

A second complication arises from the fact that the choice of transformation to sweep out the latent heterogeneity is not innocuous. If the within transformation is used to eliminate the fixed effect, then lags of endogenous (or predetermined for that matter) regressors are not valid instruments because they are correlated with the time mean of the error term. Instead, one must use the first-difference or orthogonal-deviations transformation, along with assumptions on the structural error term, or must have access to strictly exogenous instrumental variables such as policy reforms used in the natural experiment literature.

To formalize ideas we begin by rewriting the wage equation in matrix form as

$$W_i = D_i \Gamma + \varepsilon_i, \quad (9)$$

where  $W_i$  is the  $T \times 1$  vector of log wages for person  $i$ ,  $D_i = [\iota_T F'_i, X_i]$  is the  $T \times (G + P)$  matrix of regressors for person  $i$  in which  $\iota_T$  is a  $T \times 1$  vector of ones,  $F_i$  is a  $G \times 1$  vector of time invariant regressors such as race, gender, and education attainment (assuming no students in the sample),  $X_i$  is a  $T \times P$  vector of time-varying regressors including the fatality rate,  $\Gamma$  is a  $(G + P) \times 1$  vector of unknown parameters to estimate, and  $\varepsilon_i = \iota_T \alpha_i + u_i$  is the  $T \times 1$  error component.

The most general treatment of the problem is by Arellano and Bover (1995), who proposed GMM estimation of equation (7) within the context of the correlated random effects setup of Hausman and Taylor (1981). The idea is to find a nonsingular transformation,  $C$ , and a matrix of instruments,  $M_i$ , such that the population moment conditions  $E(M'_i C \varepsilon_i) = 0$  are satisfied. Arellano and Bover suggest the transformation

$$C = \begin{pmatrix} H \\ \iota'_T / T \end{pmatrix}, \quad (10)$$

where  $H$  is a  $(T - 1) \times T$  matrix containing the first-difference, within, or orthogonal deviations operator, and  $\iota'_T / T$  converts a variable into its time mean. Notice that  $H$  eliminates  $\alpha_i$  from the first ( $T - 1$ ) rows, thus allowing the identification of the coefficients on time-varying regressors, while  $\iota'_T / T$  creates an equation in levels (between-groups), and permits identification of the coefficients on time-invariant regressors. For the instruments, define the block-diagonal matrix  $M_i = I_T \otimes [m'_i, m'_i, \dots, m'_i, m'']$ ,

where  $I_T$  is a  $T \times T$  identity matrix,  $m_i = (F_i, x_i)$  is a typical row from  $D_i$ , and  $\tilde{m}'$  is a subset of  $m_i$  that is assumed to be uncorrelated in levels with  $\alpha_i$ .

Stacking the observations across all  $i$ , the GMM estimator then is given as

$$\hat{\Gamma} = [D' \bar{C}' M (M' \bar{C} \hat{\Omega} \bar{C}' M)^{-1} M' \bar{C} D]^{-1} D' \bar{C}' M (M' \bar{C} \hat{\Omega} \bar{C}' M)^{-1} M' \bar{C} W, \quad (11)$$

where  $\bar{C} = I_N \otimes C$ , with  $I_N$  an  $N \times N$  identity matrix, and  $\hat{\Omega}$  is a conformable matrix  $\hat{u}\hat{u}'$  with estimated residuals from a first-stage 2SLS regression.

### 19.5.1 Identification

The key to identification for correlated random effects is the choice of instruments that comprise  $\tilde{m}_i$ . It is important to emphasize that, unlike standard instrumental variables, identification of the correlated random effects estimator generally does not come from exclusion restrictions outside of the system, but instead from inside the system via assumptions about correlation with  $\alpha_i$  and  $u_i$ . Cornwell, Schmidt, and Wyhowski (1992) proposed a classification scheme where the time-varying  $X_i$  are decomposed as  $[X_{1i}, X_{2i}, X_{3i}]$  and the time-invariant  $F_i$  as  $[F_{1i}, F_{2i}, F_{3i}]$ .  $X_{1i}$  and  $F_{1i}$  are called endogenous because they are correlated with both  $\alpha_i$  and  $u_i$  (assumptions A.1 and A.2 do not hold),  $X_{2i}$  and  $F_{2i}$  are called singly exogenous because they are assumed to be correlated with  $\alpha_i$  but not  $u_i$  (assumption A.1 holds but not A.2), and  $X_{3i}$  and  $F_{3i}$  are called doubly exogenous as they are assumed to be uncorrelated with both  $\alpha_i$  and  $u_i$  (both assumptions A.1 and A.2 hold). It is the doubly exogenous  $X_3$  that are critical for identification of the endogenous  $F_1$ ; that is, identification requires the number of time-varying doubly exogenous variables ( $X_3$ ) to be at least as large as the number of time-invariant endogenous variables ( $F_1$ ).

Hausman and Taylor (1981) suggest one possibility for  $m_i = [\bar{x}_{3i}, F_{3i}]$ , where  $\bar{x}_{3i}$  is the individual time-mean of the doubly exogenous  $X$ s. In addition, leads and lags of the singly and doubly exogenous time-varying variables ( $X_2$  and  $X_3$ ) can be used to identify the endogenous time-varying  $X_1$ , as can lags of the endogenous  $X_1$  depending on the time series properties of  $u_i$ . Amemiya and MacCurdy (1986) and Breusch, Mizon, and Schmidt (1989) suggest additional instruments for the correlated random effects model.

A number of important cases obtain from the general endogeneity framework we have been discussing. First, the typical panel application in labor economics eschews identification of the coefficients on time-invariant variables and thus  $F_i$  is dropped from the analysis. This in turn implies that the transformation in equation (10) is now reduced to  $H$  because the time-mean transform  $\iota'_T/T$  is no longer needed, and demands for identification are relaxed as there is no longer a need for doubly exogenous  $X_3$ . Second, it is also typical to assume that all time-varying regressors are singly exogenous, implying that even if there was an interest in including time invariant variables the  $F_1$  and  $F_2$  cannot be identified unless rather stringent assumptions are imposed or there are instruments available outside the system (Breusch et al. 1989).

An example of the outside instruments case is Ziliak (2003), who used state-by-year variation in the generosity of a state's welfare program to identify the effect endogenous (time-invariant) welfare income on asset holdings. Third, it is common that the structural model of interest contains at least one endogenous  $X_1$ . Identification in the one or more endogenous components of  $X_1$  case rests on time-series properties of the structural error. If  $u_i \sim iid(0, \sigma^2)$  and  $H$  is a first-difference transformation, then  $x_{1t-2}, x_{1t-3}, \dots$  are valid instrumental variables. If  $u_i$  is MA(1), then  $x_{1t-3}$  and earlier are valid instruments. In the *iid* case the panel length must be at least three periods, in the MA(1) case the panel length must be at least four periods, so that the time dimension increases in importance as the researcher introduces more flexibility into the error process.

For example, Kniesner et al. (2012) assumed that all regressors were singly exogenous, but they were also concerned that the fatality rate in equation (3) might be endogenous if there are idiosyncratic preferences for on-the-job risk correlated with wages beyond the fixed effect. They assumed that the time-varying error  $u_i$  was *iid*, and thus after first-differencing the fatality rate at  $(t - 2)$  was a valid instrument for identification. Indeed, as they were concerned about endogeneity emanating from measurement error, as well as the more general simultaneity bias, they used the results of Griliches and Hausman (1986) and considered a number of alternative instrument sets such as the  $(t - 1)$  and  $(t - 3)$  levels of the fatality risk, the  $(t - 1) - (t - 3)$  difference, the  $(t - 2)$  and  $(t - 3)$  levels and difference, and the  $(t - 2)$  and  $(t - 4)$  levels and difference. As noted earlier in our discussion of IV as a treatment for measurement error bias, the main result of such regressions is a fairly narrow range for the estimated VSL, approximately \$6 million to \$8.5 million, though the confidence intervals widen as is typical in IV models compared to OLS.

An obvious question arises. Which set of IV results are most preferred? That is, it is fairly standard to estimate a variety of specifications and examine how sensitive the key parameters of interest are to alternative identification assumptions. Deciding which models to emphasize are guided first and foremost by the theory underlying the economic model. Theory is then supplemented with model specification tests such as the first-stage  $R^2$  (or partial  $R^2$  with multiple endogenous regressors) of instrument correlation (Shea 1997; Stock et al. 2002), and the Sargan-Hansen test of overidentifying restrictions (Hansen 1982). Moreover, in the GMM framework of Arellano and Bover (1995), one can employ pseudo likelihood ratio tests comparing unrestricted to restricted instrument sets to test the validity of instruments (Newey and West 1987), and test the assumptions of singly and doubly exogenous variables.

## 19.5.2 State Dependence and Dynamic Panels

Over the past couple of decades the panel econometrics literature has devoted significant attention to a particular form of endogeneity in the form of lagged dependent variables (Anderson and Hsiao 1982; Arellano and Bond 1991; Kiviet 1995; Blundell

and Bond 1998). In labor economics dynamic models arise in a variety of situations, such as habit persistence in preferences over consumption and hours of work choices, in explicit and implicit labor contracts that cause wages to change only sluggishly to changing economic conditions, or in employers tagging workers as potentially risky if they have a prior history of unemployment. In the unemployment tagging case, the very fact that a worker has been unemployed in the past has a direct effect on the chances of a future spell of unemployment, which is behaviorally distinct from a worker that is predisposed to being unemployed based on some unobserved factor (to the econometrician at least). Thus, separating state dependence from unobserved heterogeneity is of economic and econometric interest.

In the context of our wage equation example we modify the specification to admit state dependence as:

$$\ln w_{it} = \alpha_i + \rho \ln w_{it-1} + \beta \pi_{it} + X_{it}\gamma + u_{it}, \quad (12)$$

where in this case  $\beta$  is the short-run effect of the fatality rate on wages and  $\beta/(1 - \rho)$  is the long-run effect. In terms of estimation, it is important to recognize that  $E[\ln w_{t-1}\alpha_i] \neq 0$ , which follows by definition because  $\alpha_i$  is a determinant of current wages it must also be a determinant of lagged wages. Moreover, under the standard case where  $u_{it}$  is *iid*, then  $E[\ln w_{t-1}u_{it}] = 0$ . In the Cornwall et al. terminology the lagged dependent variable is singly exogenous because it is correlated with the fixed effect but not the overall model error, and thus its inclusion is no different than any other typical regressor in the model (assuming known initial conditions). Indeed, the usual time series stationarity requirement for dynamic models that  $|\rho| < 1$  is not necessary provided that the number of cross-sectional units ( $N$ ) is large relative to the number of time units ( $T$ ), though economic interpretation gets muddled (Holtz-Eakin, Newey, and Rosen 1988).

The problem with the dynamic model occurs once we first-difference to eliminate the latent heterogeneity  $\alpha_i$

$$\Delta \ln w_{it} = \rho \Delta \ln w_{it-1} + \beta \Delta \pi_{it} + \Delta X_{it}\gamma + \Delta u_{it}, \quad (13)$$

where we now have an induced endogeneity because the MA(1) error term  $\Delta u_{it}$  is correlated with the change in the lagged dependent variable  $\Delta \ln w_{it-1}$ . Identification, however, is no more difficult than the case of the endogenous fatality rate above wherein in this case we can use lag levels of the dependent variable dated at time  $(t - 2)$  and earlier as instruments for the change in the lagged dependent variable. Arellano and Bond (1991) discuss how it is possible to overidentify the model by using different lags of the dependent variable for each period with a block-diagonal instrument matrix; that is, if there is a total of five time periods then  $(t - 4)$  is an instrument in time  $(t - 2)$ ;  $(t - 3)$  and  $(t - 4)$  are valid instruments in time  $(t - 1)$ ; and  $(t - 2)$ ,  $(t - 3)$ , and  $(t - 4)$  are valid instruments in time  $t$ . If there are also endogenous regressors beyond the lagged dependent variable such as the fatality rate, then the same rules apply as discussed in the prior section.

We have also estimated dynamic first-difference regressions based on both the simple Anderson-Hsiao just-identified IV estimator and the heavily over-identified Arellano-Bond dynamic GMM estimator, which permit separating short-run versus long-run steady state estimates.<sup>6</sup> Remember that our first-differences estimator focuses on changes in wages in response to changes in risk. The mechanism by which the changes will become reflected in the labor market hinges on how shifts in the risk level will affect the tangencies of the constant expected utility loci with the market offer curve. To the extent that the updating of risk beliefs occurs gradually over time, which is not unreasonable because even release of the government risk data is not contemporaneous, one would expect the long-run effects on wages of changes in job risk to exceed the short-run effects. Limitations on mobility will reinforce a lagged influence (state dependence).

As one would then expect, the steady state estimates of VSL after the estimated three-year adjustment period are larger than the short-run estimates. The difference between the short-run and long-run VSL is about \$1 million, ranging from \$6 million to \$7 million versus \$7 million to \$8 million using a standard work year. Again, the central tendency of VSL estimates is not greatly affected when panel data are used with estimators that accommodate generic endogeneity, weak instruments, measurement error, latent heterogeneity, and possible state dependence.

## 19.6 SAMPLE COMPOSITION DYNAMICS

---

Fundamental to identification of panel data models is the relative importance of the between-group versus within-group variation. If too much of the variation is between groups, then the within or first-difference estimator is not identified. On the other hand, if the bulk of variation is within groups, then the random effects estimator, which is a weighted average of within and between groups variation, converges to the within estimator as the ratio of between to within variance approaches zero. Concern over sample composition dynamics is exemplified in our wage-fatal risk example, where the fatality rate is not person-specific but instead is at the industry-occupation level, implying that between-group variation may be a significant fraction of the total variation. Indeed, Kniesner et al. (2012) demonstrated that most of the variation in aggregate fatality risk is between groups (across occupations or industries at a point in time) and not within groups (within either occupations or industries over time). However, the within-group variation in the fatality measure accounts for about one-third of the total and thus it is feasible to identify the risk parameter in the first-difference wage model.

What is less well appreciated in many labor-market studies is the underlying source of the within variation that permits identification. For example, in labor supply applications that have a focus on identifying the wage elasticity of labor supply, it is typical

to take the within-group variation in wages and hours as given, although Altonji and Paxson (1992) showed that for most workers it was necessary to change jobs if they wanted to change hours or wages in a significant way. In our wage-fatal risk example, is the within variation largely coming from workers who stay in the same job but face varying on-the-job fatal risk over time, or is it from workers who change jobs either to a safer work environment at a lower wage or to a riskier work environment at a higher wage? Kniesner et al. (2012) show that the main source of variation identifying compensating differentials for fatal injury risk comes from workers who switch industry-occupation cells over time. That is, the within-group variation is eight times higher for job changers than job stayers, and thus job changers are key to identifying compensating differentials.

If job changers are essential to identification for this, and potentially many labor models, then an important consideration is whether the dynamics of changing sample composition is endogenous, and if so, what form the endogeneity takes. For example, if the decision to change jobs is idiosyncratic and time invariant, then any potential bias from job switchers is swept out with first-differencing. Or, if the decision to change jobs is trending over time, then including a person-specific trend and double differencing as in equation (5) will eliminate the endogeneity of job changers. However, suppose that workers who switch to more dangerous jobs require a large wage increase to accept a new job that is more dangerous but workers who seek a safer job do not accept a safer job if it is accompanied by much of a wage cut. The result will be an estimated hedonic locus in a panel data set that is driven by idiosyncratic worker selection effects that change over time and thus may not be well captured by a fixed effect or trend.

The time-varying selection effects can be introduced either as selection on observables or selection on unobservables (Barnow, Cain, and Goldberger 1980; Heckman 1979). In the observables case a control function of the form  $g_{it} = f(Z_{it}\eta)$  can be appended to the model to control for the probability of switching jobs, where the  $Z_{it}$  are observed covariates and the  $\eta$  are unknown parameters. This could be implemented in a variety of ways such as directly including  $g_{it}$  in the regression model and estimating the parameters of the selection model jointly with the fatality risk parameter, or as a two-step propensity score estimator where in the first step estimates a flexible model of job change and then append  $\hat{g}_{it}$  to the wage equation.

The canonical selection on unobservables model of Heckman (1979) likewise involves a function similar to the selection on observables, but in the unobservables case  $g_{it}$  is a generalized residual. So, for example, if the first stage model is a probit then  $g_{it} = \frac{\varphi(Z_{it}\eta)}{\Phi(Z_{it}\eta)}$ , which is the inverse Mill's ratio (for alternative specifications in the panel data context, see Wooldridge 2010, chapter 19). Kniesner, Viscusi, and Ziliak (2012) develop an interactive factor model as described in Bai (2009), which for wage levels here is

$$\ln w_{it} = \alpha_i + \beta\pi_{it} + X_{it}\gamma + \delta_t + u_{it} \quad (14)$$

with

$$E[\alpha_i u_{it}] = 0, \quad (15)$$

$$E[\delta_t u_{it}] = 0, \quad (16)$$

and

$$E[u_{it} | \pi_{it}, X_{it}] \neq 0 = \lambda_i(\theta_0 + \theta_1 t) + e_{it}, \quad (17)$$

where  $\lambda_i$  is the inverse Mills ratio of the probability of ever changing jobs. Given our specification of the conditional mean function in equation (17), where  $\lambda_i$  is a time-invariant factor loading, the first-differenced model for the selection bias corrected panel data regression estimated on the subsample job changers is

$$\Delta \ln w_{it} = \beta \Delta \pi_{it} + \Delta X_{it} \gamma + \Delta \delta_t + \theta_1 \lambda_i + \Delta e_{it}. \quad (18)$$

The first-stage probit model for constructing the inverse Mills ratio regresses whether the worker ever changes a job on the time-means of the variables used in the wage equation. Because the regression in equation (18) does not include time means, exclusion restrictions as well as nonlinearities identify the effect of the inverse Mills ratio.

Kniesner, Viscusi, and Ziliak (2012) demonstrate that among job changers the estimated VSL is robust to the inclusion of selection effects that change over time. In other words, the correlation between preferences for risk and the risk of the job are well captured by controlling for simple latent intercept heterogeneity via first-differences. The robustness obtains even when they estimate a difference-in-difference with selection correction. Although it is not possible to extrapolate the result to the wider scope of models in labor economics, we note that controlling for fixed intercept heterogeneity is useful in mitigating bias from missing data in many other labor applications (Wooldridge 2010).

## 19.7 SUMMARY AND POLICY IMPLICATIONS

---

We have offered a brief summary of some key problems confronting empirical applications in labor economics and how panel data can be utilized to robustly estimate parameters of economic interest such as the value of statistical life. Our pedagogical framework of the hedonic equilibrium model was motivated in part because obtaining reliable estimates of compensating wage differentials has long been challenging because of the central roles of individual heterogeneity and state dependence in affecting both the market offer curve and individual preferences, and the lack of disaggregated, longitudinal data on fatal job risks.

The wide variation of VSL estimates in the literature also has generated concern that underlying econometric problems may jeopardize the validity of the estimates. The

range for VSL in the cross-section empirical literature is extremely wide, from about \$0 million to \$20 million, which is also the case in our own cross-section based estimates with the PSID. Earlier research did not control for the host of econometric problems we address here. A most important finding in our empirical example is that controlling for latent time-invariant intercept heterogeneity is crucial—much more so than how one does it econometrically—so much so that it reduces the estimated VSL by as much as two-thirds to about \$6 million to \$10 million depending on the time-frame (short-run versus long-run) and whether or not measurement error is addressed. In short, the models that yield the \$6 million to \$10 million range are preferred because they control comprehensively for selection on unobservables (via fixed effects, state effects, and industry occupation effects), selection on observables, and measurement error.

Using panel data econometric methods to narrow the VSL as we do here has substantial benefits for policy evaluation. In its Budget Circular A4 (Sept. 17, 2003), the U.S. Office of Management and Budget requires that agencies indicate the range of uncertainty around key parameter values used in benefit-cost assessments. Because of the wide range of estimates from the earlier cross-sectional research, agencies often have failed to provide any boundaries at all to the key VSL parameter in their benefit assessments. For example, in comparing the cost estimates of health and safety regulations found in Breyer (1993), 23 of the 53 policies are in the indeterminate zone (neither pass nor fail the benefit-cost test) based on the cross-section range of VSL, but this is reduced to just two policies with our refined range of VSL. Panel data methods confer to labor economics the dual benefits of statistical robustness with economic relevance.<sup>7</sup>

---

## NOTES

---

1. We note that as an alternative to individual-level panel data we may instead follow cohorts over time. For a discussion on the merits and restrictions of pseudo panel (repeated cross-sections) data, see, e.g., Deaton 1985.
2. In the wage-hedonic literature the individual's risk exposure is proxied with the fatal risk on the job. As risk information is not collected at the person level, the literature has historically used the publically released fatality rates by one or two-digit industry. Kniesner et al. (2012) obtained proprietary data from the Census of Fatal Occupational Industries to construct a measure of fatality risk that varies by 72 two-digit industries and 10 one-digit occupations, yielding a potential of 720 different fatal risk outcomes. Each regression model controls for a quadratic in age, years of schooling, indicators for marital status, union status, race, one-digit occupation, two-digit industry, region, state, and year. Standard errors are clustered by industry and occupation and are also robust to the relevant heteroscedasticity.
3. Slope heterogeneity is distinct from more ubiquitous parametric forms of heterogeneity obtained by interacting a parameter of interest with other variables, often demographic factors. For example, Aldy and Viscusi (2008) interact the fatal risk variable with the worker's age to estimate how the VSL varies over the life cycle, while Kniesner et al. (2006) show how the VSL varies with the age profile of life-cycle consumption and emphasize the need for panel data.

4. The implied VSL at the median is evaluated at the quantile-specific wage rate in conjunction with the quantile-specific estimate of the fatal risk parameter,  $\hat{\beta}(0.5) = 0.0022$ .
5. Another solution to measurement error problems that is somewhat idiosyncratic to the hedonic model in Kniesner et al. (2012) is to combine the ideas of the between-groups estimator with first (or wider) differences. Specifically, the fatality risk is measured at the industry and occupation level, and because this is exogenous to the individual, Kniesner et al. compute three-year moving averages of the fatality risk to include in the regression model. In the first-difference model reported in Table 19.1, the use of three-year averages results in an estimate of the VSL of \$7.7 million, or about one-third higher than reported in Table 19.1.
6. The Arellano-Bond model has also proved useful in studying job injury risk as the outcome of interest. See Kniesner and Leeth 2004.
7. For more discussion, see Kniesner and Leeth 2009.

## REFERENCES

---

- Abowd, J., and D. Card. (1989). "On the Covariance Structure of Earnings and Hours Changes." *Econometrica* 57(2): 411–445.
- Aldy, J. E., and W. K. Viscusi. (2008). "Adjusting the Value of a Statistical Life for Age and Cohort Effects." *Review of Economics and Statistics* 90(3): 573–581.
- Altonji, J. G., and C. H. Paxson. (1992). "Labor-Supply, Hours Constraints, and Job Mobility." *Journal of Human Resources* 27(2): 256–278.
- Amemiya, T., and T. MaCurdy. (1986). "Instrumental Variable Estimation of an Error Components Model." *Econometrica* 54(4): 869–881.
- Anderson, T. W., and C. Hsiao. (1982). "Formulation and Estimation of Dynamic Models Using Panel Data." *Journal of Econometrics* 18: 67–82.
- Arellano, M., and S. Bond. (1991). "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *Review of Economic Studies* 58: 277–297.
- Arellano, M., and O. Bover. (1995). "Another Look at the Instrumental Variable Estimation of Error-Components Models," *Journal of Econometrics* 68(1): 29–51.
- Ashenfelter, O. (2006). "Measuring the Value of a Statistical Life: Problems and Prospects." *Economic Journal* 116(510): C10–C23.
- Ashenfelter, O., and M. Greenstone. (2004a). "Using Mandated Speed Limits to Measure the Value of a Statistical Life." *Journal of Political Economy* 112(1, pt. 2): S226–S267.
- Ashenfelter, O., and M. Greenstone. (2004b). "Estimating the Value of a Statistical Life: The Importance of Omitted Variables and Publication Bias." *American Economic Association Papers and Proceedings* 94(2): 454–460.
- Baltagi, B. H. (1981). "Simultaneous Equations with Error Components." *Journal of Econometrics* 17: 189–200.
- Baltagi, B. H. (2008). *Econometrics of Panel Data*, 4th Edition, New York: John H. Wiley and Sons, Inc.
- Baltagi, B. H., and Y. J. Chang. (1994). "Incomplete Panels: A Comparative Study of Alternative Estimators for the Unbalanced One-Way Error Component Regression Model." *Journal of Econometrics* 62: 67–89.

- Baltagi, B. H., and Y. J. Chang. (2000). "Simultaneous Equations with Incomplete Panels." *Econometric Theory* 16: 269–279.
- Baltagi, B. H., and S. H. Song. (2006). "Unbalanced Panel Data: A Survey." *Statistical Papers* 47: 493–523.
- Barnow, B., G. Cain, and A. Goldberger. (1980). "Issues in the Analysis of Selectivity Bias." In E. Stromsdorfer and G. Farkas, eds., *Evaluation Studies Review Annual*, Volume 5, pp. 43–59. Beverly Hills, CA: Sage Publications.
- Black, D. A., and T. J. Kriesner. (2003). "On the Measurement of Job Risk in Hedonic Wage Models." *Journal of Risk and Uncertainty* 27(3): 205–220.
- Blundell, R., and S. Bond. (1998). "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models." *Journal of Econometrics* 87: 115–144.
- Bollinger, C. R. (1996). "Bounding Mean Regressions When a Binary Regressor is Mismeasured." *Journal of Econometrics* 73(2): 387–399.
- Bound, J., and A. Krueger. (1991). "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?" *Journal of Labor Economics* 9: 1–24.
- Bound, J., C. Brown, and N. Mathiowetz. (2001). "Measurement Error in Survey Data." In E. E. Leamer and J. J. Heckman, eds., *Handbook of Econometrics*, Vol. 5, pp. 3705–3843. Amsterdam: North Holland.
- Breusch, T. S., G. E. Mizon, and P. Schmidt. (1989). "Efficient Estimation Using Panel Data." *Econometrica* 57(3): 695–700.
- Breyer, S. G. (1993). *Breaking the Vicious Cycle: Toward Effective Risk Regulation*. Cambridge, MA: Harvard University Press.
- Brown, C. (1980). "Equalizing Differences in the Labor Market." *Quarterly Journal of Economics* 94(1): 113–134.
- Card, D. (1990). "The Impact of the Mariel Boatlift on the Miami Labor Market." *Industrial and Labor Relations Review* 43(2): 245–257.
- Card, D., and A. Krueger. (1994). "Minimum Wages and Employment: A Case Study of the Fast Food Industry in New Jersey and Pennsylvania." *American Economic Review* 84(4): 772–793.
- Cornwell, C., P. Schmidt, and D. Wyhowski. (1992). "Simultaneous Equations and Panel Data." *Journal of Econometrics* 51(1/2): 151–181.
- Fitzgerald, J., P. Gottschalk, and R. Moffitt. (1998). "An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics." *Journal of Human Resources* 33(2): 251–299.
- Griliches, Z., and J. A. Hausman. (1986). "Errors in Variables in Panel Data." *Journal of Econometrics* 31(1): 93–118.
- Hahn, J., J. Hausman, and G. Kuersteiner. (2007). "Long Difference Instrumental Variables Estimation for Dynamic Panel Models with Fixed Effects." *Journal of Econometrics* 140(2): 574–617.
- Hausman, J., and W. Taylor. (1981). "Panel Data and Unobservable Individual Effects." *Econometrica* 49(5): 1377–1398.
- Hausman, J., and D. Wise. (1979). "Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment." *Econometrica* 47(2): 455–73.
- Heckman, J. J. (1978). "Simple Statistical Models for Discrete Panel Data Developed and Applied to Test the Hypothesis of True State Dependence against the Hypothesis of Spurious State Dependence." *Annales de l'INSEE*, no. 30–31, 227–269.

- Heckman, J. J. (1979). "Sample Selection Bias as Specification Error." *Econometrica* 47(1): 153–161.
- Heckman, J.J., L. Lochner, and P. Todd. (2008). "Earnings Functions and Rates of Return." *Journal of Human Capital* 2(1): 1–31.
- Heckman, J. J., and R. Robb. (1985). "Alternative Methods for Evaluating the Impact of Interventions: An Overview." *Journal of Econometrics* 30(1–2): 239–267.
- Holtz-Eakin, D., W. Newey, and H. S. Rosen. (1988). "Estimating Vector Autoregressions with Panel Data." *Econometrica* 56: 1371–1395.
- Hotz, V. J., F. Kydland, and G. Sedlacek. (1988) "Intertemporal Preferences and Labor Supply." *Econometrica* 56: 335–360.
- Keane, M. P. (1993). "Individual Heterogeneity and Interindustry Wage Differentials." *Journal of Human Resources* 28(1): 134–161.
- Keane, M., and D. Runkle. (1992). "On the Estimation of Panel-Data Models with Serial Correlation When Instruments Are Not Strictly Exogenous." *Journal of Business and Economic Statistics* 10(1): 1–9.
- Kiviet, J. (1995). "On Bias, Inconsistency, and Efficiency of Various Estimators in Dynamic Panel Data Models." *Journal of Econometrics* 68(1): 53–78.
- Kniesner, T. J., and J. D. Leeth. (2004). "Data Mining Mining Data: MSHA Enforcement Efforts, Underground Coal Mine Safety, and New Health Policy Implications." *Journal of Risk and Uncertainty* 29: 83–111.
- Kniesner, T. J., and J. D. Leeth. (2009). "Hedonic Wage Equilibrium: Theory, Evidence, and Policy." *Foundations and Trends® in Microeconomics* 5(4): 229–299.
- Kniesner, T. J., and J. P. Ziliak. (2002). "Tax Reform and Automatic Stabilization." *American Economic Review* 92(3): 590–612.
- Kniesner, T. J., W. K. Viscusi, and J. P. Ziliak. (2006). "Life-Cycle Consumption and the Age-Adjusted Value of Life." *Contributions to Economic Analysis & Policy* 5(1): Article 4, <http://www.bepress.com/bejap/contributions/vol5/iss1/art4>.
- Kniesner, T. J., W. K. Viscusi, and J. P. Ziliak. (2010). "Policy Relevant Heterogeneity in the Value of a Statistical Life: Evidence from Panel Quantile Regressions." *Journal of Risk and Uncertainty* 40(1): 15–31.
- Kniesner, T. J., W. K. Viscusi, and J. P. Ziliak. (2012). "Willingness to Accept Equals Willingness to Pay for Labor Market Estimates of the Value of Statistical Life." IZA Working Paper 6816.
- Kniesner, T. J., W. K. Viscusi, C. Woock, and J. P. Ziliak. (2012). "The Value of a Statistical Life: Evidence from Panel Data." *The Review of Economics and Statistics* 94(1): 74–87.
- Koenker, R. (2004). "Quantile Regression for Longitudinal Data." *Journal of Multivariate Analysis* 91(1): 74–89.
- Lamarche, C. (2010). "Robust Penalized Quantile Regression Estimation for Panel Data." *Journal of Econometrics* 157: 396–408.
- Lillard, L., and Y. Weiss. (1979). "Components of Variation in Panel Earnings Data: American Scientists." *Econometrica* 47(2): 437–454.
- MacCurdy, T. (1981). "An Empirical Model of Labor Supply in a Life Cycle Setting." *Journal of Political Economy* 89: 1059–1085.
- MacCurdy, T. (1982). "The Use of Time Series Processes to Model the Error Structure of Earnings in a Longitudinal Data Analysis." *Journal of Econometrics* 18(1): 83–114.
- MacCurdy, T. (2007). "A Practitioner's Approach to Estimating Intertemporal Relationships Using Longitudinal Data: Lessons from Applications in Wage Dynamics." In J. Heckman

- and E. Leamer, eds., *Handbook of Econometrics*, Volume 6A, Chapter 62, pp. 4057–4167. Amsterdam: Elsevier B.V.
- Newey, W. K., and K. D. West. (1987) “Hypothesis Testing with Efficient Method of Moments Estimation.” *International Economic Review* 28: 777–787.
- Pencavel, J. (1986). “Labor Supply of Men.” In O. Ashenfelter and R. Layard, eds., *Handbook of Labor Economics*, Volume 1, no. 5, pp. 3–102. Amsterdam: North Holland.
- Shaw, K. (1989). “Life-Cycle Labor Supply with Human Capital Accumulation.” *International Economic Review* 30(1): 431–56.
- Shea, J. (1997). “Instrument Relevance in Multivariate Linear Models: A Simple Measure.” *Review of Economics and Statistics* 79(2): 348–352.
- Solon, G. (1986). “Bias in Longitudinal Estimation of Wage Gaps.” NBER Working Paper 58.
- Solon, G. (1989). “The Value of Panel Data in Economic Research.” In D. Kasprzyk, D. Gregory, K. Graham, and M. P. Singh, eds., *Panel Surveys*, pp. 486–496. Hoboken, NJ: Wiley.
- Solon, G. (1992). “Intergenerational Income Mobility in the United States.” *The American Economic Review* 82(3): 393–408.
- Stock, J., J. Wright, and M. Yogo. (2002). “A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments.” *Journal of Business and Economic Statistics* 20(4): 518–529.
- Swamy, P. A. V. B. (1971). *Statistical Inference in Random Coefficient Regression Models*. New York: Springer-Verlag.
- Thaler, R., and S. Rosen. (1975). “The Value of Saving a Life: Evidence from the Labor Market.” In N. E. Terleckyj, ed., *Household Production and Consumption*, pp. 265–300. New York: Columbia University Press.
- U.S. Department of Transportation, Office of the Assistant Secretary for Transportation Policy. (2005). *Revised Departmental Guidance: Treatment of the Value of Preventing Fatalities and Injuries in Preparing Economic Analyses*. Washington, DC.
- U.S. Office of Management and Budget. (2003). *OMB CIRCULAR A-4, Regulatory Analysis* (Rep. No. A-4). Washington, DC.
- Verbeek, M., and T. Nijman. (1992). “Testing for Selectivity Bias in Panel Data Models.” *International Economic Review* 33(3): 681–703.
- Viscusi, W. K. (2009). “Valuing Risks of Death from Terrorism and Natural Disasters.” *Journal of Risk and Uncertainty* 38(3): 191–213.
- Viscusi, W. K. (2013). “The Value of Individual and Societal Risks to Life and Health.” In M. Machina and W. K. Viscusi, eds., *Handbook of the Economics of Risk and Uncertainty*, Chapter 8. Amsterdam: Elsevier Publishing.
- Viscusi, W. K., and J. Aldy. (2003). “The Value of a Statistical Life: A Critical Review of Market Estimates throughout the World.” *Journal of Risk and Uncertainty* 27(1): 5–76.
- Viscusi, W. K., and W. N. Evans. (1990). “Utility Functions That Depend on Health Status: Estimates and Economic Implications.” *American Economic Review* 80(3): 353–374.
- Weiss, Y., and L. A. Lillard. (1978). “Experience, Vintage, and Time Effects in the Growth of Earnings: American Scientists, 1960–1970.” *Journal of Political Economy* 86(3): 427–447.
- Wooldridge, J. M. (2007). “Inverse Probability Weighted M-Estimation for General Missing Data Problems.” *Journal of Econometrics* 141: 1281–1301.
- Wooldridge, J. M. (2010). *Econometric Analysis of Cross-Section and Panel Data*, Second Edition. Cambridge, MA: MIT Press.

- Ziliak, J. P. (2003). "Income Transfers and Assets of the Poor." *Review of Economics and Statistics* 85(1): 63–76.
- Ziliak, J. P., and T. J. Kniesner. (1998). "The Importance of Sample Attrition in Life-Cycle Labor Supply Estimation." *Journal of Human Resources* 33(2): 507–530.
- Ziliak, J. P., and T. J. Kniesner. (1999). "Estimating Life-Cycle Labor Supply Tax Effects." *Journal of Political Economy* 107(2): 326–359.

## CHAPTER 20

---

# PANEL DATA GRAVITY MODELS OF INTERNATIONAL TRADE

---

BADI H. BALTAGI, PETER EGGER, AND MICHAEL  
PFAFFERMAYR

## 20.1 INTRODUCTION

---

THIS chapter focuses on the estimation of gravity models of bilateral trade of goods (or services) and other bilateral international outcomes such as foreign direct investment or migration stocks or flows. Gravity models assume that these bilateral relationships can be modelled as a multiplicative function of the economic masses of two economies (incomes, expenditures, or endowments), the inverse of economic distance (trade costs, investment costs, or migration costs), and some constant, akin to Isaac Newton's law of gravity. Stochastic versions of this model have become the empirical workhorse to study gravity models of trade and migration since the nineteenth century. The estimates obtained (especially for bilateral geographical distance) reflect some of the most robust relationships not only in international economics but in economics at large (see Leamer and Levinsohn 1995).

This chapter discusses the application of panel econometric methods to gravity estimation. It also discusses single cross-section, as well as repeated cross-sections of country pairs over time, estimation. The nature of such data calls for panel econometric methods due to their inherent double and even triple indexation, respectively.

## 20.2 THEORETICAL BACKGROUND

---

### 20.2.1 Fundamentals

A large class of new trade models generate aggregate bilateral demand equations of consumers in country  $j = 1, \dots, F$  from producers in country  $i = 1, \dots, E$  at time  $s = 1, \dots, S$  of the following isomorphic form:

$$X_{ijs} = l_{is} m_{js} t_{ijs}^b, \quad (1)$$

where  $X_{ijs}$  are aggregate, nominal bilateral exports,  $l_{is}$  are exporter-time-specific factors,  $m_{js}$  are importer-time-specific factors,  $t_{ijs}$  is a compact measure of all bilateral (and potentially time-specific) ad-valorem trade costs, and  $b$  is referred to as the partial elasticity of trade (flows) with respect to variable trade costs.

The interpretation of  $l_{is}$  and  $m_{js}$  depends upon the underlying guiding theoretical model. Leading examples in the literature are the following. Krugman (1980) presents a model which has been outlined for multiple countries in Bergstrand, Egger, and Larch (2013), where  $l_{is}$  is the product of country-specific numbers of identical (single-product) monopolistically competitive firms and producer (mill) prices of identical prices per product, where the latter is exponentiated by  $b$ . Krugman's (1980) model is one of a love of variety, and  $b$  in that context reflects both (one minus) the elasticity of substitution among varieties (i.e., firms' products) and (one minus) the elasticity of demand. Eaton and Kortum (2002) formulate a Ricardian model where firms in each country draw their productivity from a Fréchet distribution. In this model,  $l_{is}$  represents the product of the average level of productivity in a country and marginal production costs. This product is exponentiated by  $b$ , which captures the dispersion of productivity draws. Anderson and van Wincoop (2003) formulate an endowment-economy model along the lines of Anderson (1979), where  $l_{is}$  represents the product of a preference mass parameter and the unique producer price relevant to country-time  $is$ , both exponentiated by  $b$ . In their context,  $b$  measures (one minus) the elasticity of substitution between products from different countries of origin. In any one of those models,  $Y_{is} = \sum_{j=1}^F X_{ijs}$  measures total sales of country  $i$  in period  $s$ . This corresponds to gross domestic product, if there is a single sector or activity. Moreover, in any one of the aforementioned models in case of  $E = F$  it holds that  $Y_{js} = \sum_{i=1}^E X_{ijs}$ ,  $m_{js} \equiv \frac{Y_{js}}{\sum_{i=1}^E l_{is} t_{ijs}^b}$  and  $Y_{is} = \sum_{j=1}^F X_{ijs} = l_{is} \sum_{j=1}^F m_{js} t_{ijs}^b$ .

While the frameworks of Eaton and Kortum (2002), Anderson and van Wincoop (2003), or Bergstrand, Egger, and Larch (2013) are based on consumer preferences and firms' supply behavior which ensures that all country-pairs trade with each other, zero bilateral trade can be generated by other frameworks. For instance, in the probabilistic model of Eaton and Kortum (2002), zero trade does not occur in expectation but

very well may take place in realization. The models of Anderson (1979) and Krugman (1980) can generate zero trade flows if fixed trade costs are considered beyond variable trade costs, no matter whether firms are heterogeneous in terms of productivity (see Helpman, Melitz, and Rubinstein 2008) or not (see Egger and Larch 2011; and Egger et al. 2011). The latter study considers endowment-economy versions with zero trade flows.

### 20.2.2 Multiple Sectors

An extension of the model in Section 20.2.1 to the case of multiple sectors is straightforward. For instance, if consumers spend a fixed share of their income per product, the model extension is particularly easy. Bilateral demand per product or sector can be represented by an equation such as (1) and all subscripts have to be augmented by a sector index, say,  $h = 1, \dots, H$ . Moreover,  $Y_{ihs}$  would be sales per sector  $h$  of country  $i$  in year  $s$ , and GDP would be  $Y_{is} = \sum_{h=1}^H Y_{ihs} = \sum_{h=1}^H \sum_{j=1}^F X_{ijhs}$ . Sector-level model versions have been formulated by Anderson (1979) as well as Anderson and van Wincoop (2003) for endowment-economy models, by Levchenko and Zhang (2012) for a Ricardian model, and by Egger, Larch, and Staub (2012) for a Krugman-type model.

### 20.2.3 Outcomes beyond Goods Trade

While the aforementioned models have been introduced originally to specify bilateral demand for goods, models as in Section 20.2.1 have recently been derived for services trade (see Egger, Larch, and Staub 2012), for portfolio capital flows (see Okawa and van Wincoop 2013), for foreign direct investment (see Baier, Bergstrand, and Gainer 2012),<sup>1</sup> and for migration (see Anderson 2011).

## 20.3 CROSS-SECTION OF COUNTRY PAIRS

---

Notice that the model in (1) is triple-indexed. Hence, even an analysis of a cross-section of country pairs (based on a single year  $s$  or an average across several years) in essence involves panel data, although the format tends to be quadratic (with about as many exporting countries as there are importing ones) resulting in a two-way panel. Let us specify the deterministic part of bilateral exports (or imports)  $\ln(l_i m_j t_{ij}^b)$  in (1) as  $h_{ij}\gamma + \alpha$ , where  $h_{ij}$  is a  $1 \times k$  vector of observable determinants,  $\gamma$  is a conformable

(unknown) parameter vector, and  $\alpha$  is a constant. Moreover, let us introduce a log-additive stochastic term  $e_{ij}$  and write the stochastic counterpart to equation (1) in a cross-section as

$$X_{ij} = \exp(h_{ij}\gamma + \alpha + e_{ij}), \quad i = 1, \dots, E, j = 1, \dots, F. \quad (2)$$

Using  $x_{ij} \equiv \ln X_{ij}$ , the log-linear form of the model is thus given as

$$x_{ij} = h_{ij}\gamma + \alpha + e_{ij}. \quad (3)$$

Since the model is double-indexed, it is reasonable to assume that  $e_{ij}$  exhibits a two-way error component structure of the form

$$e_{ij} = u_i + v_j + \varepsilon_{ij}, \quad (4)$$

where  $\varepsilon_{ij}$  are *i.i.d.* random disturbances and  $u_i$  and  $v_j$  capture exporter- and importer-specific-effects. In matrix form, the model can be written as

$$x = h\gamma + \alpha \iota_{EF} + e \quad (5)$$

$$e = \Delta_u u + \Delta_v v + \varepsilon, \quad (6)$$

where  $x$  is an  $EF \times 1$  vector and the dimensions of  $h$ ,  $u$ , and  $v$  are, respectively,  $EF \times k$ ,  $E \times 1$ , and  $F \times 1$ . In general, we denote by  $\iota_{EF}$  a column vector of ones whose dimension is  $EF$ . Assume that the data are sorted first by exporter and then by importer country ( $i$  is the slow index and  $j$  the fast one). Then, the (indicator variables) design matrices  $\Delta_u \equiv (I_E \otimes \iota_F)$  and  $\Delta_v \equiv (\iota_E \otimes I_F)$ , where  $I_E$  and  $I_F$  are identity matrices of dimensions  $E$  and  $F$ , respectively. In case of an unbalanced panel one obtains the design matrices  $\Delta_u$  and  $\Delta_v$  by skipping the rows of  $(I_E \otimes \iota_F)$  and  $(\iota_E \otimes I_F)$  that correspond to missing values. Depending on the assumptions on the exporter and importer-specific terms ( $u$  and  $v$ ) several econometric models can be specified.<sup>2</sup>

### 20.3.1 The Two-Way Fixed Effects Model

Treating  $u_i$  and  $v_j$  as fixed parameters, i.e., subsuming observed and unobserved exporter-specific factors (such as  $\ln l_i$ ) and importer factors (such as  $\ln m_j$ ) of the generic gravity model into the effects, estimation of the parameters  $\gamma$  is straightforward, even if the panel is unbalanced.

Following Davis (2002), define the projection matrices  $P_{[A]} = A(A'A)^{-}A'$ , where  $^{-}$  denotes the pseudo inverse, and  $Q_{[A]} = I - P_{[A]}$ . His Lemma 1 states that for conformable matrices  $\Delta = (\Delta_u, \Delta_v)$ , it follows that  $P_{[\Delta]} = P_{[\Delta_u]} + P_{[Q_{[\Delta_u]}\Delta_u]}$ . The within transformation that eliminates fixed exporter and importer effects is therefore defined as

$$\begin{aligned} Q_{[\Delta]} &= I - P_{[\Delta]} = Q_{[\Delta_u]} - P_{[Q_{[\Delta_u]}\Delta_v]} \\ &= Q_{[\Delta_u]} - Q_{[\Delta_u]}\Delta_v(\Delta_v Q_{[\Delta_u]}\Delta_v)^{-1}Q_{[\Delta_u]}\Delta_v. \end{aligned} \quad (7)$$

Estimation is thus straightforward as one can apply OLS to the within transformed model of the form

$$Q_{[\Delta]}x = Q_{[\Delta]}h\gamma + Q_{[\Delta]}\varepsilon \quad (8)$$

to obtain consistent estimates of  $\gamma$ . To guard against equicorrelation in the data due to the presence of common shocks to exporters and importers, it may be advisable to use two-way clustering when estimating the variance of the estimated parameters (see Wooldridge 2003).

### 20.3.2 The Two-Way Random Effects Model

Under the random effects assumption exporter and importer effects are assumed to be random with  $u_i|h \sim iid(0, \sigma_u^2)$ ,  $v_j|h \sim iid(0, \sigma_v^2)$ ,  $E[u_i v_j|h] = 0$  as well as  $E[u|h] = 0$  and  $E[v|h] = 0$ . The variance-covariance matrix of the disturbances in the two-way gravity model with random effects is then given by (see Baltagi 2008, p. 37)

$$\Omega_e := E[ee'] = \sigma_u^2 \Delta_u \Delta'_u + \sigma_v^2 \Delta_v \Delta'_v + \sigma_\varepsilon^2 I_{EF}. \quad (9)$$

For the unbalanced panel case we define

$$\begin{aligned} \tilde{\Delta}_E &= \Delta'_u \Delta_u + \frac{\sigma_\varepsilon^2}{\sigma_u^2} I_E \\ \tilde{\Delta}_F &= \Delta'_v \Delta_v + \frac{\sigma_\varepsilon^2}{\sigma_u^2} I_F \\ \tilde{P} &= \tilde{\Delta}_F - J_{FE} \tilde{\Delta}_E^{-1} J'_{EF} \\ V &= I_n - \Delta_u \tilde{\Delta}_E^{-1} \Delta'_u \end{aligned}$$

where  $n$  is the overall number of observations. Wansbeek and Kapteyn (1989) obtain the inverse variance-covariance matrix as

$$\Omega_e^{-1} = V - V \Delta_v \tilde{P}^{-1} \Delta'_v V. \quad (10)$$

One can use GLS-estimation, i.e., applying OLS to the GLS-transformed data using  $\sigma_\varepsilon \Omega_e^{-1/2} x$  and  $\sigma_\varepsilon \Omega_e^{-1/2} h$ . An important advantage of the random effects specification is that the explanatory variables,  $h$ , may include variables that vary only in the exporter or importer dimension, but not in both. However, the main drawback of the random effects specification lies in its restrictive exogeneity assumptions that require zero

correlation of the explanatory variables with both the random exporter and importer effects.

## 20.4 THREE-WAY PANELS OF COUNTRY PAIRS—REPEATED OBSERVATION OF CROSS-SECTION DATA OVER TIME

---

Gravity models with time variation typically involve a large number of country-pairs but a short time span ( $EF \gg S$ ). In this context, the workhorse model is usually a two-way model with country-pair and time effects of the form

$$x_{ijs} = h_{ijs}\gamma + \alpha + u_{ij} + v_s + \varepsilon_{ijs}, \quad (11)$$

(see Egger and Pfaffermayr 2003), using an obvious amendment of the notation for cross-section data in equation (2).

However, unless  $h_{ijs}$  includes structural or approximated measures of  $l_{is}$  and  $m_{js}$ , this specification is not able to account for time varying country-time-specific factors such as producer and consumer goods price indices, etc. In the absence of such terms,  $h_{ijs}$  would be endogenous and the standard estimates of  $\gamma$  ignoring this endogeneity will be inconsistent. The latter could be avoided by including fixed exporter-time and importer-time effects in addition to (fixed or random) country-pair effects. Baltagi, Egger, and Pfaffermayr (2003) proposed and analyzed such a specification with fixed exporter-time, importer-time, and exporter-importer effects of the form:

$$\begin{aligned} x_{ijs} &= h_{ijs}\gamma + \alpha + u_{ij} + v_{is} + w_{js} + \varepsilon_{ijs} \\ &\text{or in vector form} \\ x &= h\gamma + \alpha\iota_n + \Delta_u u + \Delta_v v + \Delta_w w + \tau\beta + \varepsilon, \end{aligned} \quad (12)$$

where again  $\Delta_u$ ,  $\Delta_v$ , and  $\Delta_w$  denote the corresponding dummy design matrices, and  $\iota_n$  is a vector of ones of length  $n$ , where  $n$  is the number of observations. Clearly, only the coefficients  $\gamma$  of the  $ijs$ -indexed explanatory variables may be identified with such an approach.

Applying Davis's (2002) Lemma twice (as in his Corollary 1), yields the projection on  $\Delta = [\Delta_u, \Delta_v, \Delta_w]$

$$P_{[\Delta]} = P_{[\Delta_u]} + P_{[Q_{[\Delta_u]}\Delta_v]} + P_{[Q_{[\Delta_u]}\Delta_v]Q_{[\Delta_v]}\Delta_w]}. \quad (13)$$

The within transformation is therefore given by

$$\begin{aligned} Q[\mathbf{z}] &= I - P_{[\Delta]} = Q_{[\Delta_u]} - P_{[Q_{[\Delta_u]}\Delta_v]} - P_{[Q_{[Q_{[\Delta_u]}\Delta_v]}\mathbf{Q}_{[\Delta_v]}\Delta_w]} \\ P_{[Q_{[\Delta_u]}\Delta_v]} &= Q_{[\Delta_u]}\Delta_v (\Delta'_v Q_{[\Delta_u]}\Delta_v)^{-1} \Delta_v Q_{[\Delta_u]} \\ P_{[Q_{[Q_{[\Delta_u]}\Delta_v]}\mathbf{Q}_{[\Delta_v]}\Delta_w]} &= Q_{[\Delta_u]} Q_{[Q_{[\Delta_u]}\Delta_v]}\Delta_w (\Delta'_w Q_{[\Delta_u]} Q_{[Q_{[\Delta_u]}\Delta_v]}\Delta_w)^{-1} \\ &\quad \times \Delta'_w Q_{[\Delta_u]} Q_{[Q_{[\Delta_u]}\Delta_v]}. \end{aligned}$$

In the *balanced case* with  $n \equiv EFS$  observations, this generates a block-diagonal structure, if one sorts the data first by exporter, then importer, and lastly by time to obtain

exporter-importer indicators :  $\Delta_u = (I_E \otimes I_F \otimes I_S)$

exporter year indicators :  $\Delta_v = (I_E \otimes I_F \otimes I_S)$

importer year indicators :  $\Delta_w = (I_E \otimes I_F \otimes I_S)$ .

Davis (2002) also discusses the random effects and mixed random and fixed specifications of this model. In practice, many researchers exploit the fact that *ES* and *FS* are much smaller than *EF*. They use the *ES* exporter-time, *FS* importer-time indicator variables, in conjunction with a within-transformation to wipe out the *EF* country-pair fixed effects. This obtains identical estimates of  $\gamma$  as the three-way within transformation following Davis (2002).

## 20.5 A SMORGASBORD OF EMPIRICAL TOPICS

---

### 20.5.1 Econometric Issues

#### 20.5.1.1 Heteroskedasticity

##### The case of cross-section data

If one estimates gravity models in levels rather than in logs, heteroskedasticity typically arises due to the large variation in country size. To account for heteroskedasticity, Santos Silva and Trenreyno (2006) suggest estimating the gravity model in levels rather than in logs. A by-product of this strategy is that information on zero bilateral trade flows may be used when identifying parameters. They propose a nonlinear, exponential-family gravity model with an additive error term of the form:

$$X_{ij} = \exp(h_{ij}\gamma + \alpha + u_i + v_j) + \varepsilon_{ij}, \quad i = 1, \dots, E, j = 1, \dots, F. \quad (14)$$

This model may be easily estimated by either assuming that  $u_i$  and  $v_j$  are zero (as Santos Silva and Tenreyro did) or by assuming that  $u_i$  and  $v_j$  are fixed and part of  $h_{ij}\gamma$ . If

$X_{ij} = 0$ ,  $\varepsilon_{ij} = -\exp(h_{ij}\gamma + \alpha + u_i + v_j)$ . Following McCullagh and Nelder (1989), Santos Silva and Tenreyro (2006) assume that the conditional variance is proportional to the conditional mean:  $V[X_{ij}|h_{ij}] \propto E[X_{ij}|h_{ij}] = \exp(h_{ij}\gamma + \alpha + u_i + v_j)$ . The first order conditions for weighted nonlinear least squares are given by:

$$\sum_{i=1}^E \sum_{j=1}^F (X_{ij} - \exp(h_{ij}\gamma + \alpha + u_i + v_j)) h_{ij} = 0 \quad (15)$$

$$\sum_{i=1}^E \sum_{j=1}^F (X_{ij} - \exp(h_{ij}\gamma + \alpha + u_i + v_j)) 1 = 0 \quad (16)$$

$$\sum_{j=1}^F (X_{ij} - \exp(h_{ij}\gamma + \alpha + u_i + v_j)) 1 = 0 \quad (17)$$

$$\sum_{i=1}^E (X_{ij} - \exp(h_{ij}\gamma + \alpha + u_i + v_j)) 1 = 0. \quad (18)$$

These are numerically identical to the Poisson pseudo-maximum-likelihood (PPML) estimator that is often used for count data. Note that under pseudo-maximum-likelihood estimation, only the conditional mean has to be specified correctly to obtain consistent parameter estimates. Since, both  $E$  and  $F$  are large but (much) smaller than  $EF$ , one can add exporter and importer dummies to the model to account for unobserved variables that vary either in the exporter or the importer dimension. Let us define  $\widehat{\mu}_{ij} \equiv \exp(h_{ij}\widehat{\gamma} + \widehat{\alpha} + \widehat{u}_i + \widehat{v}_j)$  and  $z_{ij} = (h_{ij}, \Delta_u, \Delta_v)$  with  $\Delta_u$  denoting exporter indicators and  $\Delta_v$  importer indicators. Following, White (1982), one can show that the PPML estimator is asymptotically normal and the corresponding variance-covariance matrix can be estimated as

$$V(\widehat{\gamma}, \widehat{\alpha}) = \left( \sum_{i=1}^E \sum_{j=1}^F \widehat{\mu}_{ij} z_{ij} z'_{ij} \right)^{-1} \left( \sum_{i=1}^E \sum_{j=1}^F V(X_{ij}|z_{ij}) z_{ij} z'_{ij} \right) \\ \times \left( \sum_{i=1}^E \sum_{j=1}^F \widehat{\mu}_{ij} z_{ij} z'_{ij} \right)^{-1}. \quad (19)$$

Clearly, if the data are generated in full (!) compliance with the model in Section 20.2.1 (invoking multilateral trade balance),  $\exp(\widehat{u}_i)$  and  $\exp(\widehat{v}_j)$  are consistent estimates of the cross-section counterpart  $l_i$  and  $m_j$  in equation (1) (see Fally 2012). To see this, use  $\lambda_{ij} = \exp(h_{ij}\gamma + \alpha)$ ,  $l_i = \exp(u_i)$ , and  $m_j = \exp(v_j)$  to write the first-order conditions of the Poisson likelihood (15) with respect to the  $k$ th variable in  $h_{ij}$  and to  $u_i$  and  $v_j$  as

$$\begin{aligned}
\frac{\partial \ln L}{\partial \lambda_{ij}} \frac{\partial \lambda_{ij}}{\partial \gamma_k} &= \sum_{i=1}^E \sum_{j=1}^F \left( -l_i m_j + \frac{X_{ij}}{\lambda_{ij}} \right) h_{ij,k} = 0 \\
\frac{\partial \ln L}{\partial l_i} &= \sum_{j=1}^F \left( -\lambda_{ij} m_j + \frac{X_{ij}}{l_i} \right) = 0 \rightarrow \hat{l}_i \sum_{j=1}^F \lambda_{ij} m_j = \sum_{j=1}^F X_{ij} \\
\frac{\partial \ln L}{\partial m_j} &= \sum_{i=1}^E \left( -\lambda_{ij} l_i + \frac{x_{ij}}{m_j} \right) = 0 \rightarrow \hat{m}_j \sum_{i=1}^E \lambda_{ij} l_i = \sum_{i=1}^E X_{ij}.
\end{aligned} \tag{20}$$

The latter two conditions imply

$$\begin{aligned}
\sum_{i=1}^E X_{ij} &= \sum_{i=1}^E \lambda_{ij} \hat{l}_i \hat{m}_j \\
\sum_{j=1}^F X_{ij} &= \sum_{j=1}^F \lambda_{ij} \hat{l}_i \hat{m}_j \\
\sum_{i=1}^E \sum_{j=1}^F X_{ij} &= \sum_{i=1}^E \sum_{j=1}^F \lambda_{ij} \hat{l}_i \hat{m}_j.
\end{aligned} \tag{21}$$

In a structural gravity model, as in Section 20.2.1, setting  $Y \lambda_{ij} l_i m_j = t_{ij}^b l_i m_j$  with  $Y = \sum_{j=1}^F Y_j$  denoting world expenditures on goods, bilateral goods exports (or imports) are defined as

$$X_{ij} = Y \lambda_{ij} l_i m_j. \tag{22}$$

Denoting income and expenditure shares as  $\theta_i = Y_i / Y$  and  $\theta_j = Y_j / Y$ ,

$$\theta_i = l_i \sum_{h=1}^F \lambda_{ih} m_h, \quad \theta_j = m_j \sum_{h=1}^F \lambda_{hj} l_h. \tag{23}$$

It is now readily seen that the fixed effects estimates  $\exp(\hat{v}_i)$  and  $\exp(\hat{w}_j)$  are consistent estimates of  $l_i$  and  $m_j$  in (1) as they solve

$$\begin{aligned}
\theta_i &= \frac{1}{Y} \sum_{j=1}^F X_{ij} = \frac{1}{Y} \sum_{j=1}^F \lambda_{ij} \hat{l}_i \hat{m}_j \\
\theta_j &= \frac{1}{Y} \sum_{i=1}^E X_{ij} = \frac{1}{Y} \sum_{i=1}^E \lambda_{ij} \hat{l}_i \hat{v}_j \\
1 &= \frac{1}{Y} \sum_{i=1}^E \sum_{j=1}^F X_{ij} = \frac{1}{Y} \sum_{i=1}^E \sum_{j=1}^F \lambda_{ij} \hat{l}_i \hat{m}_j.
\end{aligned} \tag{24}$$

With this data-generating process, the exporter and importer fixed effects in the PPML-model would be structural estimates of the terms  $l_i$  and  $m_j$ . Yet, evidence suggests that this is not the case, and the data-generating process of bilateral export and import flows appears to violate some of the fundamental assumptions (see Egger, Larch, and Staub 2012). Among others, the PPML estimator has been applied to cross-sectional gravity models by Santos Silva and Silvana Tenreyro (2009) and Egger et al. (2011).

### 20.5.1.2 The Case of Data with Repeated Cross-Sections (Three-Way Panels)

The PPML-estimator discussed for the (two-way) cross-sectional model can also be applied to the two-way (or even three-way) panel models with a large number of country pairs and a small number of time periods ( $N = EF \gg S$ ). Subsuming  $h_{ijs}$ , time indicators, and the constant into  $z_{ijs}$  with conformable parameter  $\phi$ , the model becomes

$$X_{ijs} = \exp(z_{ijs}\phi + u_{ij}) + \varepsilon_{ijs}, \quad i = 1, \dots, E, j = 1, \dots, F, s = 1, \dots, S \quad (25)$$

The underlying likelihood may be based on  $X_{ijs} \sim iid \mathcal{P}(u_{ij}e^{z_{ijs}\phi})$ . Defining  $\lambda_{ijs} = \exp(z_{ijs}\phi)$ , Cameron and Trivedi (2005) demonstrate that the first order condition for  $u_{ij}$  yields  $\widehat{u}_{ij} = \sum_{s=1}^S x_{ijs} / \sum_{s=1}^S \ln \lambda_{ijs}$  so that one can eliminate (concentrate out)  $u_{ij}$  and there is no incidental parameters problem (akin to the one-way within-transformation with a large number of cross-sections and a short time period in a linear panel model).

For any generic variable  $z_{k,ijs}$ , the score function of the concentrated likelihood uses the transformed values  $z_{k,ijs} - \frac{\lambda_{ijs}}{\lambda_{ij}} \bar{z}_{k,ij}$ , where a bar indicates averages of a variable across all years in the data, to obtain:

$$\frac{\partial \ln L(\phi, u_{ij})}{\partial \phi} = \sum_{i=1}^E \sum_{j=1}^F \sum_{s=1}^S \left[ \lambda_{ijs} \left( z_{ijs} - \frac{\lambda_{ijs}}{\bar{\lambda}_{ij}} \bar{z}_{ij} \right) \right]. \quad (26)$$

The fixed effects Poisson estimator has strong robustness properties as it is consistent under the conditional mean assumption, whereby  $E[X_{ijs}|z_{ijs}, u_{ij}] = \exp(u_{ij})e^{\lambda_{ijs}\phi}$ . Also the distribution of  $z_{ijs}$  does not need to be discrete, there is no restriction on the dependence of  $z_{ijs}$  and  $z_{ijs'}$ ,  $s \neq s'$ , and one can apply PPML (see Wooldridge 2010). Uniqueness holds under general identification assumptions. However, a similar transformation is not available for the two-way or three-way fixed effects problems. The fixed country-pair effects PPML estimator has been applied with panel data, among others, by Egger and Nelson (2011).

### 20.5.1.3 Zeros and Missing Data

#### The case of cross-section (two-way) data

In general, trade data often include a large number of reported zeros or missing trade flows as, e.g., small countries may not have trade relations with all possible

trading partners or because statistical offices do not report trade flows below certain thresholds. Non-randomly missing zero trade flows require the estimation of sample selection models or two-part models. For example, the model discussed in Helpman, Melitz, and Rubinstein (2008) naturally leads to a sample selection approach if the disturbances in the outcome equation (log bilateral exports or imports,  $x_{ij}$ ) are correlated with those of the selection equation. A latent variable for the propensity of exports from  $i$  to  $j$  may be defined as

$$V_{ij}^* = f_{ij}\delta + k_i + r_j + \eta_{ij}, \quad (27)$$

where  $f_{ij}$  includes, among other variables, measures of log bilateral trade barriers and log bilateral fixed costs of exporting (or importing). Since variable and fixed costs to trade depend, at least partly, on the same determinants (such as log bilateral distance),  $\delta$  only identifies their joint impact.

$$\begin{aligned} \ln X_{ij} &= \begin{cases} h_{ij}\gamma + u_i + v_j + \varepsilon_{ij} & \text{if } V_{ij} = 1 \\ \text{unobserved} & \text{if } V_{ij} = 0 \end{cases} \\ V_{ij} &= 1[\ln V_{ij}^* > 0] \\ (\eta_{ij}, \varepsilon_{ij}) &\sim N\left[0, \begin{pmatrix} 1 & \rho\sigma_\varepsilon \\ \rho\sigma_\varepsilon & \sigma_\varepsilon^2 \end{pmatrix}\right]. \end{aligned} \quad (28)$$

The standard sample selection model assumes normality and homoskedasticity. The latter means that  $k_i$  and  $r_j$ , as well as  $u_i$  and  $v_j$ , must be assumed as fixed, and they can be identified, since  $EF \gg E + F$ . The parameters can then be estimated by maximizing the likelihood

$$\begin{aligned} \ln L &= \sum_{i=1}^E \sum_{j=1}^F V_{ij} \left( \ln \Phi\left(\frac{f_{ij}\delta + k_i + r_j + \frac{\rho}{\sigma_e}(\ln X_{ij} - h_{ij}\gamma - u_i - v_j)}{(1-\rho^2)^{1/2}}\right) \right. \\ &\quad \left. - \sigma_e \ln \phi\left(\frac{\ln X_{ij} - h_{ij}\gamma - u_i - v_j}{\sigma_e^2}\right) \right) \\ &\quad + (1 - V_{ij}) \ln \Phi(f_{ij}\delta + k_i + r_j). \end{aligned} \quad (29)$$

Formally, the first order conditions for the score of the likelihood can be solved without exclusion restrictions, although it may be poorly identified if large values of  $V_{ij}^*$  are not in the data (see Cameron and Trivedi 2005). Also, two-step estimators that include the estimated mills ratio from a first step may be used.<sup>3</sup> The main advantage of the sample-selection approach lies in its ability to predict potential unobserved outcomes. This allows intuitive comparative static analysis consistent in broad terms with the model structures in Section 20.2.1. Using  $x_{ij} = \ln X_{ij}$ , this model can be written as

$$E[x_{ij} | V_{ij}^* > 0] = h_{ij}\gamma + u_i + v_j + \rho\sigma_e \frac{\phi(f_{ij}\delta + k_i + r_j + \eta_{ij})}{\Phi(f_{ij}\delta + k_i + r_j + \eta_{ij})}. \quad (30)$$

Sample selection gravity models of this kind have been estimated, among others, by Helpman, Melitz, and Rubinstein (2008), Egger et al. (2011), and Head and Mayer et al. (2011).

Alternatively, one can rely on a two-part model (see, e.g., Egger et al. 2011). This model specifies the conditional mean for positive models separately as

$$E[x_{ij}|X_{ij} > 0] = h_{ij}\gamma + u_i + v_j + \varepsilon_{ij} \quad (31)$$

and estimates a probit or logit model for the probability to export (or import) as

$$P(V_{ij}^* > 0) = P(f_{ij}\delta + k_i + r_j + \eta_{ij} > 0). \quad (32)$$

In both cases the unconditional expectation is then given as

$$P(V_{ij}^* > 0)E[x_{ij}|X_{ij} > 0]. \quad (33)$$

#### 20.5.1.4 The Case of Repeated Cross-Section (Three-Way) Data

In case of non-randomly missing bilateral trade flows with time-series-cross-section data, one has to rely on panel data sample selection models as discussed in Wooldridge (1995), Dustmann and Rochina-Barrachina (2007), and Semykina and Wooldridge (2010). For ease of notation, and without loss of generality, let us subsume the constant and fixed time effects into  $h_{ijs}\gamma$ . Furthermore, in line with typical data, let us assume that  $EF \gg S$ . The sample selection model with (fixed or random) country-pair effects ( $u_{ij}$ ) can be written as:

$$x_{ijs} = \begin{cases} h_{ijs}\gamma + u_{ij} + \varepsilon_{ijs} & \text{if } V_{ijs} = 1 \\ \text{unobserved} & \text{if } V_{ijs} = 0 \end{cases} \quad (34)$$

$$V_{ijs} = 1[\ln V_{ijs}^* > 0]$$

$$V_{ijs}^* = f_{ijs}\delta_s + k_{ij} + \eta_{ijs},$$

Wooldridge (1995) shows that one can correct for selection bias arising from the correlation of the disturbances  $\varepsilon_{ijs}$  and  $\eta_{ijs}$  using Mundlak's (1978) and Chamberlain's (1982) ideas. Note the coefficient of  $f_{ijs}$  in the selection equation may vary over time.

The conditional expectation  $E[\varepsilon_{ijs}|f_{ijs}, V_{ijs}]$  has a complicated nonlinear form. But if  $E[u_{ij} + \varepsilon_{ijs}|f_{ijs}, V_{ijs}] = 0$ , a pooled OLS control function approach on a Mundlak-type model will be consistent, which leads to the specification

$$x_{ijs} = h_{ijs}\gamma + \underbrace{\bar{h}_{ij}\pi_{x,s} + \varsigma_{ij}}_{u_{ij}} + \underbrace{E[\varepsilon_{ijs}|f_{ijs}, V_{ijs} = 1]}_{=\gamma E[\eta_{ijs}|f_{ijs}, V_{ijs}=1]} + \nu_{ijs},$$

$$= h_{ijs}\gamma + \bar{h}_{ij}\pi_{x,s} + \varsigma_{ij} + \gamma\lambda_{ijs} + \nu_{ijs} \quad \text{for } V_{ijs} = 1, \quad (35)$$

where  $\nu_{ijs}$  denote the remainder random disturbances and  $\varsigma_{ij}$  is a random country-pair effect that comes from parameterizing  $u_{ij}$  by a systematic term  $\bar{h}_{ij}\pi_{x,s}$  and a random country-pair effect  $\varsigma_{ij}$ . In addition,  $E[\nu_{ijs}|f_{ijs}, V_{ijs} = 1] = 0$  is assumed (see Semykina and Wooldridge 2010, Assumption 4.1.1, p. 387). For implementation, one estimates for each year  $s = 1, \dots, S$  a probit model  $P(V_{ijs} = 1|f_{ijs}) = \Phi(f_{ijs}\delta_s)$  to obtain

an estimator of the inverse Mills' ratio  $\widehat{\lambda}_{ijs} = \lambda(f_{ijs}\delta_s)$ . Then one can use pooled OLS to estimate the outcome model given above. However to calculate the standard errors one has to use a panel bootstrap or to estimate the asymptotic variance as described in the Appendix of Wooldridge (1995) and Semykina and Wooldridge (2010), since the model includes estimated right-hand-side variables. Such a model has been estimated by Egger et al. (2009) and Wamser (2011) for bilateral foreign direct investment.

Raymond et al. (2010) propose a maximum-likelihood estimation procedure for a panel random effects sample selection model of the form:

$$\begin{aligned} x_{ijs} &= h_{ijs}\gamma + u_{ij} + \varepsilon_{ijs} = A_{ijs} + u_{ij} + \varepsilon_{ijs} \\ V_{ijt}^* &= f_{ijs}\delta + k_{ij} + \eta_{ijs} = B_{ijs} + k_{ij} + \eta_{ijs}. \end{aligned} \quad (36)$$

For the disturbances, they assume

$$\begin{aligned} \begin{bmatrix} u_{ij} \\ k_{ij} \end{bmatrix} &\sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & \rho_{uk}\sigma_u\sigma_\eta \\ \rho_{uk}\sigma_u\sigma_\eta & \sigma_\eta^2 \end{bmatrix}\right) \\ \begin{bmatrix} \varepsilon_{is} \\ \eta_{is} \end{bmatrix} &\sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_\varepsilon^2 & \rho_{\varepsilon\eta}\sigma_\varepsilon \\ \rho_{\varepsilon\eta}\sigma_\varepsilon & 1 \end{bmatrix}\right). \end{aligned} \quad (37)$$

Note that  $f_{ijs}$  may include  $V_{ij0}$  and  $V_{ijs-1}$  to account for initial values and state dependence (see Wooldridge 2005). The bilateral country-pair effects  $u_{ij}$  and  $k_{ij}$  may be modeled along the lines of Mundlak (1978) as discussed above.

Clearly, the presence of equi-correlation through  $u_{ij}$  and  $k_{ij}$  would render maximum likelihood estimates inconsistent. Therefore, these effects have to be integrated out. Using the change of variables  $v_{ij} = u_{ij}/\sigma_u(2(1-\rho_{uk}^2))^{1/2}$  and  $z_{ij} = k_{ij}/\sigma_k(2(1-\rho_{uk}^2))^{1/2}$ , so that  $du_{ij} = \sigma_u(2(1-\rho_{uk}^2))^{1/2}dv_{ij}$  and  $dk_{ij} = \sigma_k(2(1-\rho_{uk}^2))^{1/2}dz_{ij}$ , the likelihood may be written as

$$\begin{aligned} \mathcal{L}_{ij} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \prod_{s=1}^S L_{ijs} \phi(v_{ij}, z_{ij}) dv_{ij} dz_{ij}, \\ L_{ijs} &= \Phi(-B_{ijs} - z_{ij})^{1-V_{ijs}} \left[ \frac{1}{\sigma_\varepsilon} \phi\left(\frac{x_{ijs}-A_{ijs}-v_{ij}}{\sigma_\varepsilon}\right) c \right. \\ &\quad \times \left. \Phi\left(\frac{B_{ijs}+z_{ij}+\frac{\rho\eta}{\rho\varepsilon\eta}(x_{ijs}-A_{ijs}-v_{ij})}{(1-\rho\varepsilon\eta)^{1/2}}\right)\right]^{V_{ijs}}, \end{aligned} \quad (38)$$

where  $\phi(u_{ij}, k_{ij})$  denotes the density of the bivariate normal of the random effects. The likelihood can be maximized using numerical optimization procedures. For gravity models, this procedure has been applied by Egger and Pfaffermayr (2011) to estimate effects of dynamic export market entry in general equilibrium.

### 20.5.1.5 Systems of Equations

Many gravity models of bilateral trade lump all goods (and even services) transactions into one sector. This has the advantage that only few parameters have to be estimated, but it involves the problem that those parameters may be inadequate for use at the sectoral level. As a consequence, there could be aggregation error and inadequate out-of-sample predictions. For instance, the same types of goods could be used as final products versus intermediate goods. This might require fundamentally different modeling to describe economic behavior. Similarly, there might be different types of foreign direct investment such as mergers and acquisitions and greenfield investment that might require different modeling. The same arguments might apply to flows of different types of migrants (such as skilled and unskilled ones).

Hence, it might be useful to allow the outcome to be determined by type-specific variables which carry type-specific parameters. Moreover, it would be desirable to model outcome types to depend on each other. For instance, expenditures on one type of good should not be fully independent of expenditures on another type of good for the same reasons as imports from one country and imports from another country are not independent. Consumers are bound by resource constraints and so are whole countries. Finally, if outcome types are not fully independent of each other, it is logical to allow their shocks to be correlated. All of this suggests that bilateral outcomes could be viewed as systems of equations or, in a structural form, at least as seemingly unrelated regressions.

For instance, Egger and Pfaffermayr (2004a) consider bilateral sector-level goods exports and stocks of foreign direct investment in 1989–1999 as seemingly unrelated regressions. The disturbances are allowed to be correlated between exports and foreign direct investment within a sector and year in a reduced-form model, where the two outcomes depend on endogenous bilateral geographical distance in a log-linear way. The stochastic model is a seemingly unrelated regression-type model for two equations, one for (log) exports and one for (log) outward foreign direct investment. Egger and Pfaffermayr (2004a) find that, indeed, the shocks between exports and foreign direct investment were correlated, in particular, with regard to the time invariant error component. Egger, Larch, and Staub (2012) employ a structural approach for bilateral goods and services trade. In their case, the two outcomes are related to each other for three reasons: they utilize the same factor of production (labor), they are bound by total consumer income, and they face correlated shocks on disturbances. Their model involves structurally restricted exporter-time and importer-time fixed effects and an error component structure regarding the disturbances with a random country-pair term and a random idiosyncratic term. This model represents a generalization of the pseudo-maximum likelihood estimation procedure for exponential models with heteroskedastic disturbances as proposed by Santos Silva and Tenreyro (2006) for cross-sectional, single-equation gravity models.

### 20.5.1.6 Dynamics

Dynamic patterns in bilateral outcome may occur for various reasons. A key factor for rationalizing dynamics are adjustment costs in a broad sense. For instance, such adjustment costs may occur for trade flows due to staggered contracts which will prevent trade flows from responding immediately to technology shocks or trade costs. For stocks or flows of bilateral foreign direct investment, adjustment costs can easily be rationalized from a host of investment models. Finally, adjustment costs might matter for migration flows due to information lags about economic circumstances in possible countries of residence. For instance, Eichengreen and Irwin (1998) proposed considering dynamic adjustment when estimating gravity equations.

A log-linear version of a model as in equation (1) with a time lag would be<sup>4</sup>

$$\ln X_{ijs} = \delta \ln X_{ij,s-1} + \ln(l_{is}m_{js}t_{ijs}^b) + u_{ijs}. \quad (39)$$

Such a model could be estimated by employing the generalized method of moments procedure for differenced data as proposed by Arellano and Bond (1991) or the systems estimator (which involves both the differenced and the levels equations) proposed by Blundell and Bond (1998). Differenced gravity equations à la Arellano and Bond (1991) are almost never motivated on the basis of a structural model and have been estimated for bilateral trade (see Egger 2001; Bun and Klaassen 2002; Millimet and Osang 2007; Olivero and Yotov 2012), for bilateral foreign direct investment (see Egger 2001; Egger and Pfaffermayr 2004b; Egger and Merlo 2011, 2012; or Egger et al. 2009), and for bilateral migration (see Mayda 2010). Systems estimator versions à la Blundell and Bond (1998) have been estimated for bilateral trade flows (see Martínez-Zarzoso, Nowak-Lehmann, and Horsewood 2009; or Martínez-Zarzoso et al. 2009), bilateral foreign direct investment (Abbott and De Vita 2011), and bilateral migration (Etzo 2009).

### 20.5.1.7 Endogenous Regressors

#### Standard instrumental variables and control function procedures

Suppose elements of trade costs  $t_{ij}$  (with cross-section data) or  $t_{ijs}$  (with repeated cross-section data over time) were endogenous. Using the notation given above, this would lead to the endogeneity of  $h_{ij}$  with  $E[e_{ij}|h_{ij}] \neq 0$  and the endogeneity of  $h_{ijs}$  with  $E[e_{ijs}|h_{ijs}] \neq 0$ . Suppose we had instruments  $q_{ij}$  with cross-sections or  $q_{ijs}$  with repeated cross-sections over time. With column rank  $k$  for  $h_{ij}$  and  $h_{ijs}$  and column rank  $k' \geq k$  for  $q_{ij}$  and  $q_{ijs}$ , one could apply standard instrumental-variable estimation or employ a linear or nonlinear (e.g., polynomial) control function of  $h_{ij} - \hat{h}_{ij}$  or  $h_{ijs} - \hat{h}_{ijs}$  in the outcome equation (see Wooldridge 2010; Semykina and Wooldridge 2010). For instance, such a procedure had been applied to gravity models for foreign direct investment by Egger and Merlo (2011, 2012). Egger et al. (2011) estimated a cross-sectional gravity model on bilateral exports with a mass point at zero and

used an instrumental variables procedure in the PPML context to guard against the endogeneity of preferential trade agreement membership.

However, no matter whether trade costs are endogenous or not, the variables  $l_i$  and  $m_j$  (and the logs thereof) might be endogenous in cross-sectional data. Clearly,  $\ln l_i$  and  $\ln m_j$  cannot be estimated if fixed effects  $u_i$  and  $v_j$  are included in the model as in Section 20.3.1. Moreover,  $t_{ij}$  and, hence,  $h_{ij}$  may include time-invariant variables,<sup>5</sup> so that the parameters of such trade costs (endogenous or not) cannot be identified when fixed effects  $u_{ij}$  are included in the model. The subsequent parts of this section address this problem.

#### Endogenous variables with cross-section data in Hausman and Taylor type models

Following Mundlak (1978), Kang (1985), Hausman and Taylor (1981), and Wyhowski (1994), one may include exporter- and importer-specific means of  $h_{ij}$  in a model as in equation (3), if the explanatory variables are supposed to be either correlated with  $u_i$  or  $v_j$  or both to obtain the within estimator of  $\gamma$ . In matrix form, the generalized Mundlak-model for two-way data reads

$$\begin{aligned} x &= h\gamma + P_{[\Delta_u]} h\pi_u + P_{[\Delta_v]} h\pi_v + e \\ e &= \Delta_u u + \Delta_v v + \varepsilon. \end{aligned} \tag{40}$$

A Hausman-Taylor (1981) type estimation procedure can guard against possible correlation of a subset of the explanatory variables in  $h$  with either  $u$  or  $v$ , but not with  $\varepsilon$ . For the balanced case, this model has been analyzed by Wyhowski (1994), who specifies

$$\begin{aligned} x_{ij} &= \alpha + h_{ij}\gamma + o_i\delta + r_j\theta + e_{ij} = z_{ij}\phi + e_{ij} \\ e_{ij} &= u_i + v_j + \varepsilon_{ij}, \end{aligned} \tag{41}$$

where  $o_i$  contains variables that vary only in the exporter dimension, while those in  $r_j$  vary only in the importer dimension. For instance, gross domestic product or per capita income would be variables of that kind in a gravity context. Wyhowski (1994) introduces the following partition of the set of explanatory variables:  $o_j = [o_i^e, o_i^u]$ ,  $r_j = [r_j^e, r_j^v]$  and  $h_{ij} = [h_{ij}^e, h_{ij}^u, h_{ij}^v, h_{ij}^{uv}]$ . Superscript  $e$  indicates that the corresponding variables are uncorrelated with the error components and are called doubly exogenous, while superscript  $u$  indicates correlation with  $u_i$  and superscript  $v$  correlation with  $v_j$ , respectively. The dimensions of  $o_i^e$  and  $o_i^u$  are  $(1 \times k_{oe})$  and  $(1 \times k_{ou})$ , respectively. Those of  $r_j^e$  and  $r_j^v$  are  $(1 \times k_{re})$  and  $(1 \times k_{rv})$ . Lastly,  $h_{ij}^e$ ,  $h_{ij}^u$ ,  $h_{ij}^v$ , and  $h_{ij}^{uv}$  are of dimension  $(1 \times k_{hp})$ , where  $p \in \{e, u, v, uv\}$ . Theorem 1 of Wyhowski (1994) shows that the parameter vector  $\phi = (\alpha, \gamma', \delta', \theta')'$  is identified if (i)  $k_{he} + k_{hu} \geq k_{ou}$  and (ii)  $k_{he} + k_{hv} \geq k_{rv}$ .

Wyhowski (1994) demonstrates that one obtains consistent parameter estimates if one uses a comprehensive set of instruments with a Hausman and Taylor (1981) type instrumental variables estimator. In the context of gravity models, this procedure has

been applied by Egger (2005) to identify the parameters of the potentially endogenous gross domestic product and per-capita income variables in a cross-section setting.

### Endogenous variables with repeated cross-section (panel) data in Hausman and Taylor type models

The lion's share of the variance in log bilateral trade flows  $x_{ijs}$  is contributed by the exporter-importer (time-invariant) component (see Baltagi, Egger, and Pfaffermayr 2003). Hence, there is a particularly big chance for endogeneity of time-variant or time-invariant variables in gravity models to be correlated with the time-invariant part of the error term. Accordingly, if cross-sections of country-pairs are observed repeatedly over time, it may be desirable to include country-pair fixed effects to guard against this endogeneity. Clearly, the parameters on most observable measures of trade costs—such as log bilateral distance, common official bilateral language, contiguity, etc.—cannot be identified anymore. The model proposed by Hausman and Taylor (1981) offers a suitable strategy to avoid this problem. It allows a subset of the explanatory variables to be correlated with the bilateral (country-pair) effects, and at the same time provides consistent parameter estimates of the time-invariant variables such as log bilateral distance. Note this model can also be estimated with unbalanced data when data are missing at random. The model can be written as

$$\begin{aligned} x_{ijs} &= \alpha + h_{ijs}\gamma + r_{ij}\delta + e_{ijs} = z_{ijs}\phi + e_{ijs} \\ e_{ijs} &= u_{ij} + \varepsilon_{ijs}. \end{aligned} \tag{42}$$

Similar to the case with double-indexed cross-sectional data, the set of explanatory variables is partitioned in doubly and singly exogenous variables:  $r_{ij} = [r_{ij}^e, r_{ij}^u]$  and  $h_{ijs} = [h_{ijs}^e, h_{ijs}^u]$ . Superscript  $e$  indicates that the corresponding variables are uncorrelated with either type of error component and are called doubly exogenous, while superscript  $u$  indicates correlation of a variable with  $u_{ij}$ . Both types of variables are assumed to be uncorrelated with the remainder disturbances such that  $E[r_{ij}e_{ijs}] = 0$  and  $E[h_{ijs}e_{ijs}] = 0$ . Hence, the fixed effects (or within-type) estimator  $\tilde{\gamma}_W$  which wipes out  $u_{ij}$  provides a consistent estimator of  $\gamma$  but not  $\delta$ . The dimensions of  $r_{ij}^e$  and  $r_{ij}^u$  are  $(1 \times k_{re})$  and  $(1 \times k_{ru})$ , respectively, and those of  $h_{ijs}^e$  and  $h_{ijs}^u$  are dimension  $(1 \times k_{he})$  and  $(1 \times k_{hu})$ , respectively. The parameter vector  $\phi = (\alpha, \gamma', \delta')'$  is identified if  $k_{he} \geq k_{ru}$  holds. The Hausman and Taylor (1981) procedure provides a more efficient estimator  $\hat{\gamma}_{HT}$  provided the choice of exogenous variables is correct and the model is over-identified, i.e.,  $k_{he} - k_{zu} > 0$ . This over-identification condition can be tested against the within estimator using a Hausman-type test. In the context of gravity models, this procedure has been first applied by Egger (2004b), illustrating that log bilateral distance is likely endogenous in gravity models of bilateral trade. The estimator has been generalized in Egger (2002) to account for serial correlation in  $e_{ijs}$ , and to illustrate properties of so-called export potentials. Moreover, it has been used by Egger

and Pfaffermayr (2004b) to quantify the log distance coefficient for exports and foreign direct investment in a system with bilateral exports and foreign direct investment of the United States.

#### 20.5.1.8 Cross-Sectional Interdependence

Notice that the models discussed in Section 20.2 generally involve contagious effects of explanatory variables on outcome through general equilibrium. Suppose bilateral trade costs or—for exogenous reasons—the real production or the consumer base of a country change. This will have repercussions on the producer as well as consumer prices of goods or services. This can be seen from the definition of outcome in equation (1),  $X_{ijs} = l_{is} m_{js} t_{ijs}^b$  and the fact that  $l_{is}$  and  $m_{js}$  are endogenous, since

$$m_{js} \equiv \frac{Y_{js}}{\sum_{k=1}^F l_{ks} t_{kjs}^b} \quad (43)$$

and, through  $\sum_{j=1}^F X_{ijs} = Y_{is} = l_{is} \sum_{j=1}^F m_{js} t_{ijs}^b$ . Hence, given (estimates of)  $t_{kjs}^b$  and data on gross domestic product,  $Y_{is}$ ,  $l_{is}$ , and  $m_{js}$  are determined as multilateral, triangular functions of those variables.

Of course, accounting for  $l_{is}$  and  $m_{js}$  by fixed  $i$ s-specific and  $j$ s-specific effects, respectively, allows taking this cross-sectional correlation into account. In this context, a standard framework with random  $i$ s-specific and  $j$ s-specific effects will be inappropriate not only for reasons of endogeneity (since these effects are functions of  $t_{ijs}^b$ ), but also since the country-year-specific effects are not independent of each other (i.e., the one for  $i$ s is correlated with the one for  $i'$ s for  $i' \neq i$ , and the one for  $j$ s is correlated with the one for  $j'$ s for  $j' \neq j$ ). Clearly, this makes  $X_{ijs}$  dependent on  $X_{i'js}$ , on  $X_{ij's}$ , and on  $X_{i'j's}$  for all  $i' \neq i$  and  $j' \neq j$ . The source of this cross-sectional interdependence are general equilibrium or resource constraints.

Two other forms of cross-sectional dependence that may possibly play a role are strategic interaction and cross-sectional dependence through unobservable determinants of outcome captured by the disturbances. The former would emerge, for example, if the entry of exporters into a market, that of multinational firms into a market, or that of migrants into a country, would strategically by way of information diffusion depend upon other such units' decisions. A log-linear version of the model in equation (1) with strategic interaction and an additive disturbance  $e_{ij(s)}$  would then read

$$\ln X_{ij(s)} = \lambda \left( \sum_{h \neq i} \sum_{k \neq j} w_{ij,hk(s)} \ln X_{hk(s)} \right) + \ln (l_{i(s)} m_{j(s)} t_{ij(s)}^b) + e_{ij(s)}, \quad (44)$$

where  $\lambda$  could be referred to as a spatial lag parameter and  $w_{ij,hk}$  as a spatial weight (see Anselin 1988; Kelejian and Prucha 1999; for the necessary properties of such weights and parameters). Such models have been estimated for bilateral trade flows (see Lebreton and Roia 2009), for bilateral foreign direct investment (see Blonigen et al. 2007; Blonigen et al. 2008), and for bilateral migration (see LeSage and Pace 2008).

If contagion surfaces by way of correlation of the disturbances instead, one could reformulate the disturbance term of a cross-sectional gravity model as a spatial autoregressive model of the form

$$\varepsilon = R\varepsilon + \xi, \quad (45)$$

where  $\xi$  is the independently distributed counterpart to  $\varepsilon$  and  $R$  is an  $EF \times EF$  matrix which exhibits zero diagonal elements and has finitely summable row and column elements (see Chapter 5 on spatial panels in this Handbook). A leading example in the literature is

$$\varepsilon_{ij} = \rho \left( \sum_{h \neq i} \sum_{k \neq j} w_{ij,hk} \varepsilon_{hk} \right) + \xi_{ij}, \quad (46)$$

which is referred to as a first-order spatially auto-regressive model, involving known weights  $w_{hk}$  and one unknown (spatial auto-regressive) parameter  $\rho$  (for the underlying assumptions see Kelejian and Prucha 1999, 2007). A similar structure could be assumed for  $e_{ij} = u_i + v_j + \varepsilon_{ij}$ . Similarly, with repeated cross-section data of bilateral trade over time and stacking the data in an  $(EFS \times EFS)$  block-diagonal matrix  $R_S = \text{diag}(R, \dots, R)$  the same notation might be used.

Models allowing for an error structure akin to the one in (45) have been estimated for trade flows (see Porojan 2001; and Behrens, Ertur, and Koch 2012),<sup>6</sup> for foreign direct investment (see Coghlin and Segev 2000; Baltagi, Egger, and Pfaffermayr 2007, 2008), and for migration flows (see Bertolia and Fernández-Huertas Moraga 2012).

From a theoretical perspective, many of the aforementioned applications can build on research which motivates cross-sectional interdependence in flow models between two spatial units, but they commonly lack the rigorous treatment of interdependence as in structural models underlying the generic framework in Section 20.2. For this reason, spatial lag models as the one in (44) have not found much recognition in empirical international economics. Considering interdependence structures in disturbances may be a fruitful econometric strategy though, since it can be combined with structural modeling of the deterministic part, and permit improvements in inference and comparative statics compared to procedures that consider the disturbances to be cross-sectionally independent. Cross-sectional dependence of the disturbances leads to biased standard errors of the parameters  $\gamma$  and the comparative static effects based on a structural model as the one in Section 20.2.1. The latter may be particularly relevant when quantifying trade or welfare effects of changes in trade costs. However, consistent estimates of the standard errors could be obtained by procedures outlined in Kelejian and Prucha (2007). Such procedures have been applied in the context of bilateral trade flow models by Behrens, Ertur, and Koch (2012) and for foreign direct investment by Baltagi, Egger, and Pfaffermayr (2007, 2008).

### 20.5.1.9 Normalized Outcomes

Head and Ries (2001) proposed to transform the deterministic part of equation (1) in the following way:

$$\sqrt{\frac{X_{ijs} X_{iis}}{X_{jis} X_{jjs}}} = \sqrt{\frac{t_{ijs}^b t_{jis}^b}{t_{jis}^b t_{iis}^b}} = \sqrt{t_{ijs}^b t_{jis}^b} \text{ if } t_{iis}^b, t_{jjs}^b = 1 \quad (47)$$

for all  $i, j, s$ , and analogously for cross-sections. The appeal of this normalization lies in the elimination of all exporter-time and importer-time-specific factors from equation (12). However, this advantage comes at a cost: with cross-sections or panels involving a large number of countries, the above normalization aggravates the problem of zeros in the bilateral trade matrix dramatically;<sup>7</sup> and exporter-time or importer-time-specific trade costs cancel out so that their parameters cannot be estimated, akin to models with fixed exporter-time and importer-time effects.<sup>8</sup>

Adding a log-additive error term and transforming the model in logs yields

$$\ln \frac{X_{ijs} X_{iis}}{X_{jis} X_{jjs}} = 0.5(h_{ijs} + h_{jis})\gamma + (\varepsilon_{ijs} + \varepsilon_{jis} - \varepsilon_{iis} - \varepsilon_{jjs}). \quad (48)$$

In particular, such a stochastic model is difficult to estimate if the stochastic terms are not independent and there is cross-sectional correlation.

The gravity model could alternatively be transformed to yield a logistic specification in terms of log-odds ratios (see, e.g., Head, Mayer, and Ries 2010). Similar to the two-way within model, the logistic transformation normalizes nominal trade flows so that exporter- and importer-specific determinants are eliminated but the impact of the bilateral ones still remain identified. The gravity model can be written in terms of relative log-odds or tetradiac terms

$$\ln \frac{X_{ij}}{X_{ik}} - \ln \frac{X_{lj}}{X_{lk}} = (h_{ij} - h_{ik} - h_{lj} + h_{lk})\gamma + (\varepsilon_{ij} - \varepsilon_{ik} - \varepsilon_{lj} + \varepsilon_{lk}) \quad (49)$$

where exporter country  $l$ , importer country  $k$ , and the bilateral trade flow between them serves as the basis. An advantage of this approach is that the base categories  $X_{ik}$  and  $X_{lk}$  are readily observed if picking country  $k$  properly, while this is not the case with intranational sales  $X_{iis}$  and  $X_{jjs}$  employed in the previous approach. With either type of normalization, it may be advisable to use multi-way clustering as proposed by Cameron, Gelbach, and Miller (2011); see Head, Mayer, and Ries (2010).

### 20.5.1.10 Linearly Approximated Models

Baier and Bergstrand (2009) proposed to approximate the structural model in Section 20.2.1 for a cross-section of bilateral trade flows as follows. Parameterize  $b \ln t_{ij}$  as  $h_{ij}\gamma$  and denote the  $k$ th elements of  $h_{ij}$  and  $\gamma$  by  $h_{k,ij}$  and  $\gamma_k$ , respectively. Moreover, define

$\check{h}_{k,i} \equiv \sum_{j=1}^F \theta_j h_{k,ij}$ ,  $\check{h}_{k,j} \equiv \sum_{i=1}^E \theta_i h_{k,ij}$ , and  $\check{h}_{k,..} \equiv \sum_{i=1}^E \sum_{j=1}^F \theta_i \theta_j h_{k,ij}$  to obtain

$$\tilde{h}_{k,ij} \equiv h_{k,ij} - \check{h}_{k,i} - \check{h}_{k,j} + \check{h}_{k,..}. \quad (50)$$

Then, the linearly approximated model à la Baier and Bergstrand (2009) involves

$$x_{ij} - \ln(Y_i Y_j) \approx \alpha + \tilde{h}_{ij} \gamma, \quad (51)$$

where  $\alpha$  is a constant. To the panel econometrician, the analogy of this approximation to a two-way (quasi-)within transformation of  $h_{ij}$  with balanced panels is evident: if  $E = F$  and all countries were of identical size,  $\theta_i, \theta_j = 1/F$ . Then,  $\check{h}_{k,ij} \equiv h_{k,ij} - h_{k,i} - h_{k,j} + h_{k,..}$ ,  $h_{k,j} = \frac{1}{F} \sum_{i=1}^F h_{k,ij}$ , and  $h_{k,..} = \frac{1}{F^2} \sum_{i=1}^F \sum_{j=1}^F h_{k,ij}$ . However, the analogy is not perfect, since countries are not symmetric. The weighting by expenditure shares ( $\theta_i$  and  $\theta_j$ ) in the transformation for  $\tilde{h}_{k,ij}$  does not fully eliminate country-specific trade costs (which is an advantage).

As long as the income constraints matter for each time period separately, the model can be readily adapted with repeated observation of cross-sections over time, involving

$$\tilde{h}_{k,ij,s} \equiv h_{k,ij,s} - \check{h}_{k,i,s} - \check{h}_{k,j,s} + \check{h}_{k,..s}, \quad (52)$$

which are just period-specific transformations, so that one obtains

$$x_{ij,s} - \ln(Y_{is} Y_{js}) \approx \alpha_s + \tilde{h}_{ij,s} \gamma, \quad (53)$$

where  $\alpha_s$  is a time-specific constant. Baier and Bergstrand (2009) applied this approximation with cross-section data, and Carrère (2006) and Egger and Nelson (2011) applied it with panel data.

However, the replacement of  $\ln(l_i m_j t_{ij}^b)$  by  $\widetilde{\ln t_{ij}^b}$  has further consequences. For instance, binary and other discrete variables in  $h_{ij}$  (such as regional trade agreement membership indicators or trade freedom or political freedom indices) are transformed into continuous (bounded) variables. In case of endogeneity, proper methods have to be used to avoid the endogeneity bias. For instance, with a control function approach, the residuals from a first-stage model have to be transformed analogously to  $h_{k,ij}$  in  $\tilde{h}_{k,ij}$ . Moreover, notice that the approximation error is a (nonlinear) function of the variables in the model (all countries'  $\theta_i$  and  $h_{k,ij}$  for bilateral trade of a given pair  $ij$ ). The latter involves both heteroskedasticity as well as cross-sectional dependence in the disturbances.

### 20.5.1.11 Interpretation of Disturbances

There are three main strands of interpretation of the disturbances  $e_{ij}$  in the literature. Some authors (implicitly) assume that  $e_{ij}$  contains an unmeasured part of true trade costs (see the definition of trade costs in equation (29) in Eaton and Kortum (2002), and the discussion between equations (29) and (30), there). In that case, true trade costs are  $t_{ij}^b \exp u_{ij}$  rather than  $t_{ij}^b$ . Others assume that true trade costs are

$t_{ij}^b$  but what is measured is  $t_{ij}^b \exp e_{ij}$ , implying the existence of measurement error. Finally, some authors assume that  $e_{ij}$  is simply a measurement error for  $X_{ij}$ . Obviously, the last interpretation is unproblematic. The second interpretation requires methods for modeling measurement error to avoid an associated (endogeneity) bias. The first interpretation—which is also one of measurement error—is most difficult to deal with, since the constraint about the importer-time effect mentioned in Section 20.2 now reads  $m_{js} \equiv \frac{Y_{js}}{\sum_{k=1}^E l_{ks} t_{kjs}^b \exp(e_{kjs})}$ . Such a measurement error can be addressed using the fact that aggregate bilateral demand adds up to gross domestic product,  $Y_i = \sum_{j=1}^F X_{ij}$ , and export shares as well as import shares add up to unity, i.e.,  $\sum_{j=1}^F (X_{ij}/Y_i) = 1$  (see Eaton, Kortum, and Sotelo 2012).

## 20.5.2 Specific Topics

### 20.5.2.1 Fixed Effects Versus Random Effects

The use of fixed effects by way of binary indicator variables is quite established in gravity modeling. Pöyhönen (1963) was probably the first to control for country-specific fixed effects in cross-sectional data. The use of fixed country effects is now viewed as an acceptable procedure in structural modeling of trade flows (see Feenstra 2002; Eaton and Kortum 2002; Anderson and van Wincoop 2003; Egger and Larch 2012; Bergstrand, Egger, and Larch 2013). In general, an appeal for using fixed effects is that the parameters on the regressors which are not fully collinear with the fixed effects can be estimated with less danger of an endogeneity bias. However, this advantage comes at a potentially high cost of efficiency loss. In cross-sectional models, the use of country-specific fixed effects is largely unproblematic, since the number of observations tends to be not much short of  $(E - 1)(F - 1)$  (typically, intra-national sales are not included in the data; and missing data may lead to a further loss of observations) while the number of fixed effects estimated is only  $E + F$ . With triple-indexed models and repeated observations of country-pairs over time, one gets up to  $(E - 1)(F - 1)S$  observations (with the number of time periods  $S$  being relatively small) and  $(E - 1)(F - 1)$  fixed country-pair effects as well as  $(E - 1)S$  and  $(F - 1)S$  exporter-time and importer-time effects, respectively.

Clearly, with repeated observations of country-pairs' bilateral trade flows over time and a triple-indexed aggregate gravity equation, there are many options for modeling fixed effects. One version would be to include main effects only, namely, exporting country, importing country, and time effects. Such a model has been proposed by Mátyás (1997, 1998). The most general version is one which includes exporter-time, importer-time, and country-pair (exporter-importer) effects.<sup>9</sup> This general version has been proposed by Baltagi, Egger, and Pfaffermayr (2003). Any triple-indexed model with fewer effects than the one of Baltagi, Egger, and Pfaffermayr (2003) can be thought of as a restricted version of this general framework, and it can be tested against it. Models with separate country-pair and common time effects are quite prominent in the

literature. For instance, Baldwin (1994) used such a framework to estimate effects of economic integration, assuming that the country-pair effects were random. Usually, gravity models with random country-pair effects are only estimated for comparison with fixed effects models (see Egger 2000, 2002, 2004a, b; Egger and Pfaffermayr 2003). Egger and Pfaffermayr (2003) discussed models with fixed (exporter, importer, and year) main effects and fixed versus random exporter-importer interaction (i.e., country-pair) effects to test the pair effects model against the country effects model. They find that pair effects should not be ignored. Models with fixed country-pair effects have been estimated by Egger and Pfaffermayr (2004b), Baltagi, Egger, and Pfaffermayr (2008), Egger et al. (2009). Egger and Merlo (2011, 2012) used models with fixed country-pair effects for foreign direct investment, and Orefice (2013) estimated models involving country-pair effects on migration. Other articles ignore the country-pair equi-correlation in the disturbances altogether and condition on country-time effects. For the data-generating process outlined in Section 20.2.1, country-time effects are implicit functions of bilateral trade or transaction costs.

Empirically, country-pair effects explain much more of the variation in bilateral exports or imports, foreign direct investment, or migration than country-time effects. This implies a relatively big chance for omitted country-pair-specific effects to induce endogeneity of pair-specific time-invariant variables (such as bilateral distance, adjacency, common language, etc.) or even of pair-time-specific covariates (such as trade agreement membership, bilateral tariffs, or other measures of preferentialism). The endogeneity of time-invariant trade cost measures such as (log) bilateral distance has been documented in Egger (2004a) and Egger and Pfaffermayr (2004a) for both bilateral trade and foreign direct investment. Common language and other historical, institutional, and legal time-invariant characteristics are endogenous. The reason is that common culture is a multifaceted driver of trade and the measured characteristics typically included in gravity models of bilateral goods trade or other outcomes measure only a small fraction of the universe of time-invariant determinants of bilateral trade (see Egger and Lassmann 2012). Moreover, the endogeneity of preferentialism has been documented for goods trade, services trade, and foreign direct investment (see Egger and Pfaffermayr 2004b; Egger, Egger, and Greenaway 2008; Baier and Bergstrand 2007; Egger et al. 2011; Egger and Wamser 2013a, b). Clearly, with endogenous triple-indexed variables in triple-indexed models one has to rely on nonparametric identification by relying on selection on observables or on *outside instruments* (through control function or instrumental variables estimation).

However, double-indexed (or, generally, single-indexed) endogenous variables (such as log bilateral distance, common language, etc.) can be instrumented from *within the triple-indexed (or, generally, higher-indexed) model* by splitting all higher-indexed variables into two (or, if necessary, more) dimensions of the data and eventually using them as separate variables for instrumentation. Related work follows the idea of Hausman and Taylor (1981), Amemiya and MacCurdy (1986), Breusch, Mizon, and Schmidt (1989), Cornwell, Schmidt, and Wyhowski (1992), and Wyhowski (1994) to discern exogenous single-indexed and endogenous single-indexed variables, on the one hand,

and fully exogenous all-indexed and partly exogenous all-indexed variables, on the other hand. For instance, Egger (2004b) considers log bilateral distance as an endogenous single-indexed variable (because it is time-invariant and the country-pair index is used as a single cross-sectional index) and time-plus-country-pair-indexed variables as covariates in a non-structural gravity equation of bilateral exports. All of the time-variant variables are assumed to be uncorrelated with the time-variant error component. But some of them are fully exogenous (also uncorrelated with the time-invariant error component) and some of them are partly exogenous (correlated with the time-invariant error component). For identification, there must be at least as many doubly exogenous all-indexed variables as there are endogenous single-indexed ones in the model. Then, the respective single-indexed component of all-indexed doubly exogenous variables can be used as an instrument for endogenous single-indexed variables in the model (which log distance is the example of in Egger 2004b, and Egger and Pfaffermayr 2004b). The papers by Egger (2004) and Egger and Pfaffermayr (2004b) show that the coefficient of the instrumented log distance is largely different from the one of a random (country-pair) effects model or a pooled OLS model. Serlenga and Shin (2007) apply this strategy to gravity models of bilateral goods trade in the context of European economic integration. A general conclusion from that literature is that the distance coefficient in cross-sectional gravity models of bilateral goods trade or bilateral foreign direct investment is likely biased, and the bias is probably large. Similar conclusions might hold for other coefficients on time-invariant geographical or institutional variables.

#### 20.5.2.2 *Effects of Preferential Agreements on Outcome*

A large literature in trade analyzes the effects of a country's membership in trade agreements on trade flows. Similar literature assess the impact of services trade agreements on services flows, and of investment (or tax) agreements on foreign direct investment. Since the 1950s, hundreds of goods trade agreements have been notified and most of them are now governed by the World Trade Organization. Similarly, a huge number of bilateral investment treaties, tax treaties, and services trade agreements have been signed by now. Egger and Wamser (2013a, b) provide a broad overview of the theoretical and empirical literature on preferentialism. Most of the studies on the effects of preferential agreements assess the impact of bilateral goods trade agreement membership on bilateral goods trade, but more recent work considers agreements and outcomes beyond goods trade. While earlier research focused on cross-sectional data, more recent studies exploit the time variation in the data. Studies of the latter kind typically employ country-pair fixed effects by way of a within transformation in static models or by estimating dynamic differenced models.

An overwhelming number of studies analyzed preferential agreement effects (in particular, those of goods trade agreements) in cross-sections of country-pairs. A key problem of such a strategy is the potential endogeneity of preferential agreement membership. In cross-sectional data, this endogeneity is typically not overcome in

a two-way fixed effects framework with exporter and importer country fixed effects. Then, instrumental variable strategies, switching regression, or other strategies based on selection on observables should be employed. For instance, Baier and Bergstrand (2007) propose matching on the propensity score which is estimated from a bilateral trade agreement membership model in cross-sectional bilateral trade data. Egger et al. (2011) develop a model where, in a cross-section country-specific effects are controlled for by exporter- and importer-specific fixed effects. Trade agreements are allowed to be endogenous in a two-part model with control function to account for zero trade flows.

More recent work tends to favor the identification of preferential agreements from time series of country-pairs. For instance, using pooled data points on country-pairs' trade for every fifth year between 1960 and 2000, Baier and Bergstrand (2007) find that the bias from self-selection of country pairs into trade agreements may be substantially reduced when employing pair-specific fixed effects. Egger, Larch, and Staub (2012) estimate the impact of goods and services trade memberships on trade flows (and welfare), employing fixed country-time effects using annual panel data covering the period 1996–2005. Egger (2001) and Martínez-Zarzoso, Nowak-Lehmann, and Horsewood (2009), among others, difference out fixed country-pair effects (and allow for partial adjustment). However, the finding that the selection bias of preferential (trade) agreement membership can be avoided by conditioning on country-pair effects seems to depend on the data at hand. For instance, Egger and Wamser (2013b) find that a nonparametric framework of selection on observables still results in estimates of the effect of preferential agreements (goods or services trade agreements, bilateral investment treaties, or double taxation treaties) on bilateral trade flows which differs from the simple fixed effects model. The latter suggests that the parameter bias on preferential agreement indicators accruing to self-selection is not fully overcome by removing the within variation in the data. Egger, Egger, and Greenaway (2008) proposed to combine a selection-on-observables identification strategy with a difference-in-difference model for switching into preferential agreements. In this model, the selection bias is not fully removed from differencing the data.

Baltagi, Egger, and Pfaffermayr (2003) proposed a more general framework for estimating panel data gravity models which employs country-pair, exporter-year, and importer-year fixed effects. Such a model explains almost all of the variation in data on bilateral exports or imports. As long as the fixed effects do not wipe out all of the variation of interest, it will likely remove most sources of inconsistency. Cross-sectional fixed (exporter and importer) country effects models which involve the traditional pair-specific trade cost variables (log distance, common border, common language, preferential trade agreement membership, etc.) can explain about 60–80% of the variation in the data on bilateral trade (see Bergstrand, Egger, and Larch 2013). Panel data models with country-pair, exporter-time, and importer-time fixed effects explain 95–98% of the variation in the data (see Baltagi, Egger, and Pfaffermayr 2003). When imposing proper constraints, such a model even has a structural interpretation and can be used for general equilibrium comparative static analysis (see Egger and Nigai 2013).

A variant of a model with fixed pair and fixed country-time effects has been proposed by Egger and Pfaffermayr (2013) to assess the effect of the formation of the

European Union on bilateral trade flows. Rather than imposing country-year fixed effects, they introduce country blocs. These are associated with groups of countries that entered the European Union until 2001 at a time (six founding countries; three entrants in 1973, one entrant in 1981, two in 1986, three in 1995—and one outside group) and the phases of a particular size of the Union (1960–1964; 1965–1972; 1973–1980; 1981–1985; 1986–1994; 1995–2001). They follow intra-bloc and inter-bloc trade for all six blocs over the six phases since 1960 which permits them to focus on trade creation and trade diversion effects of European integration. In a somewhat more restrictive framework regarding the functional form assumptions of integration effects, but with the appeal of structural (general equilibrium) interpretation, this approach was applied in Carrère (2006), Egger and Larch (2011), and Egger and Nelson (2011).<sup>10</sup>

### 20.5.2.3 Potential Outcomes

An old interest in estimating gravity models relates to the issue of so-called trade (or foreign direct investment) potentials (see Baldwin 1993, 1994; Gros and Gonciarz 1996; Brenton and DiMauro 1998; Nilsson 2000; De Benedictis and Vicarelli 2005). In essence, this literature is about comparing model predictions about bilateral outcome,  $\hat{X}_{ij(s)}$ , to the data on outcome,  $X_{ij(s)}$ . Depending on whether  $\hat{X}_{ij(s)} > X_{ij(s)}$  or  $\hat{X}_{ij(s)} < X_{ij(s)}$ , country-pairs are said to under- or over-exhaust their (trade or foreign direct investment) potentials (at time  $s$ , if the data carry a time index). One root of that literature is policy advice surrounding the question which funding lines and country initiatives should be prioritized to stimulate trade or foreign direct investment.

Clearly, this line of research faces a number of problems. First, if models are estimated on log-transformed data but potentials are calculated on bilateral outcome in levels, the predictions tend to under-predict the data because the log-transformation does not work well if the variance in the data is large, as is typically the case (Jensen's inequality; see Egger 2010). Second, even in the absence of this problem, it is the case that, e.g.,  $E[F^{-1} \sum_{j=1}^F (X_{ij(s)} - \hat{X}_{ij(s)})] \neq 0$  if  $[F^{-1} \sum_{j=1}^F (\ln X_{ij(s)} - \ln \hat{X}_{ij(s)})] \neq 0$  is an indication of model mis-specification (see Egger 2002). Clearly, with fixed country(-time) effects what is referred to as the country-specific difference between potential and actual outcome is absorbed in the fixed effects. Similarly, for the average year, that difference would be absorbed in a model with country-pair fixed effects. Overall, this leaves little meaningful scope for the literature on unexhausted potentials.

### 20.5.2.4 Non-structural Versus Structural Estimates

With non-structural estimates, we have to distinguish between two types. First, suppose the researcher estimates a log-linear model of the form

$$x_{ij(s)} = \alpha_{(s)} + \zeta_1 \ln Y_{i(s)} + \zeta_2 Y_{j(s)} + \zeta_3 \ln L_{i(s)} + \zeta_4 L_{j(s)} + h_{ij(s)}\gamma + e_{ij(s)}, \quad (54)$$

where  $\ln Y_{i(s)}$  and  $\ln L_{i(s)}$  measure the log gross domestic income and log population size of country  $i$  (in year  $s$ ), respectively. Clearly, this is a simple re-parametrization of a model which employs per-capita incomes ( $\ln(Y_{i(s)}/L_{i(s)})$ ) instead of  $\ln L_{i(s)}$ . As

long as no country(-time) fixed effects are included, this model will generate biased parameters if the structural model in Section 20.2 applies. The reason is simply that  $\ln l_{i(s)}$  and  $\ln m_{j(s)}$  are non(log-)linear functions of  $Y_{i'(s)}$  and  $h_{i'j'(s)}$  for all  $i'j'$  in the model. The latter generates both endogeneity as well as heteroskedasticity in the above adhoc reduced-form model versions that have been estimated for decades to date.<sup>11</sup>

A second type of non-structural model

$$x_{ij(s)} = h_{ij(s)}\gamma + u_{ij} + v_{i(s)} + w_{j(s)} + \varepsilon_{ij(s)} \quad (55)$$

is estimated with  $v_{i(s)}$  and  $w_{j(s)}$  denoting fixed country(-time) effects for exporters and importers, respectively. This model obtains consistent estimates of  $\gamma$ . Given  $\hat{\gamma}$  and, hence,  $t_{ij(s)}^b$  and expenditures/income  $Y_{is}$  for all countries  $i$  and  $j$ , one may solve for  $\hat{l}_{i(s)}$  and  $\hat{m}_{j(s)}$  as introduced in Section 20.2.1. In empirical data sets,  $\hat{l}_{i(s)}$  and  $\hat{m}_{j(s)}$  will not be identical to  $v_{i(s)}$  and  $w_{j(s)}$  for reasons discussed in Egger, Larch, and Staub (2012). The latter means that the center of the predictions of the structural model

$$\hat{X}_{ij(s)} = \exp(h_{ij(s)}\hat{\gamma} + \hat{l}_{i(s)} + \hat{m}_{j(s)}) \quad (56)$$

will not necessarily lie in the center of the data, unlike those of a PPML model of the form

$$\hat{X}_{ij(s)} = \exp(h_{ij(s)}\hat{\gamma} + \hat{u}_{ij} + \hat{v}_{i(s)} + \hat{w}_{j(s)}). \quad (57)$$

One consequence of the latter is that the structural model predictions based on trade cost estimates in fixed country effects models will predict structural model solutions that may be largely different from the actual data (see Bergstrand, Egger, and Larch 2013). Only an iterative structural model or a constrained fixed effects model as in Egger, Larch, and Staub (2012) will close the gap between fixed effects estimation and structural model estimation with cross-sections or repeated cross-sections of bilateral trade flows in gravity models.

## ACKNOWLEDGMENTS

---

Egger acknowledges funding from GA ČR through grant no. P402/12/0982.

## NOTES

---

1. Earlier studies integrating goods trade and foreign direct investment were conducted by Eaton and Tamura (1994) and Egger and Pfaffermayr (2004b).
2. Gravity equations have also been estimated for remittances of migrants (see Docquier, Rapoport, and Salomone 2010).

3. In general, one may use  $\ln \frac{X_{ij}}{l_i m_j}$  as a dependent variable instead of  $x_{ij}$ . Then,  $h_{ij}\gamma$  would measure  $b \ln t_{ij}$ . In particular, this may be desirable with structural (iterative) model estimation. Of course, this is only relevant with random effects estimates, since in a fixed effects  $\ln l_i m_j$  are fully captured by the exporter and importer fixed effects.
4. Alternatively, one may use a nonparametric control function approach to correct the estimated model for the conditional mean of the selection model given positive trade flows (see Cameron and Trivedi 2005).
5. An exponential model with a time lag could be estimated, e.g., by a linear backfitting procedure as in Blundell, Griffith, and Windmeijer (2002), or by the control function procedure proposed by Cameron and Trivedi (2005).
6. Notice that the majority of trade cost variables employed in empirical research on gravity models are time invariant. Most variables except for trade agreement membership and bilateral tariffs are of that kind (e.g., geographical variables such as log bilateral distance, contiguity, land accessibility, or cultural variables such as common official language, historical colonial relationships, etc.).
7. Behrens, Ertur, and Koch (2012), use more theoretical rigor than others to derive a spatial econometric model which involves weighted bilateral exports of other pairs as a determinant of bilateral exports of a given pair of trading partners.
8. In large data sets  $P[X_{ij} > 0]$  often takes values of 0.3 (especially, in the 1960s and earlier) to 0.7 (more recently). Hence, such-transformed data contain about  $P[X_{ij} > 0] - P[\frac{X_{ij}}{X_{jj}} \frac{X_{is}}{X_{ks}} > 0] \approx (P[X_{ij} > 0])^2$ —or one-to-two thirds—more zeros than the untransformed data.
9. In essence, this means that the overall level of fixed costs  $t_{ij}^b$  cannot be estimated consistently. This is a severe problem, since exporter-time and importer-time-specific fixed costs can be shown to be important, and comparative static effects of changes in observable trade costs are extremely sensitive to large measurement errors about total trade cost levels (see Egger, Larch, and Staub 2012).
10. In general, one could think of dyadic pair effects where the fixed effect for pair  $ij$  would be forced to be identical to the one for pair  $ji$ . However, unless the data themselves are symmetric (e.g., because bilateral exports plus imports are used as a dependent variable, which is not advisable, see Baldwin and Taglioni 2006), this restriction is unlikely to be justified.
11. Models with time effects that vary in the cross-section but not quite across countries are often used for a different reason than in Egger and Pfaffermayr (2013), namely with non-linear models or in large samples to reduce the amount of fixed effects to be estimated. For that reason, e.g., Egger and Wamser (2013b) employ fixed country-pair effects along with fixed exporter-continent-time and importer-continent-time effects.
12. This argument holds true whether a log-linear model, a PPML, or a nonlinear least-squares exponential model is estimated.

## REFERENCES

- Abbott, Andrew J. and Glauco De Vita, 2011. Evidence on the impact of exchange rate regimes on bilateral FDI flows. *Journal of Economic Studies* 38(3–4), 253–274.
- Amemiya, Takeshi and Thomas E. MacCurdy, 1986. Instrumental-variable estimation of an error-components Model. *Econometrica* 54(4), 869–880.

- Anderson, James E., 1979. A theoretical foundation for the gravity equation. *American Economic Review* 69(1), 106–116.
- Anderson, James E., 2011. The gravity model. *Annual Review of Economics* 3(1), 133–160.
- Anderson, James E. and Yoto V. Yotov, 2012. Gold standard gravity. NBER Working Papers 17835, Cambridge, MA: National Bureau of Economic Research, Inc.
- Anderson, James E. and Eric van Wincoop, 2003. Gravity with gravitas: A solution to the border puzzle. *American Economic Review* 93(1), 170–192.
- Anselin, Luc, 1988. *Spatial econometrics: Methods and models*. London and Dordrecht: Kluwer Academic Publishers.
- Arellano, Manuel and Eric Bond, 1991. Some tests of specification for panel data: Monte carlo evidence and an application to employment. *Review of Economic Studies* 58(2), 277–297.
- Baier, Scott L. and Jeffrey H. Bergstrand, 2007. Do free trade agreements actually increase members' international trade? *Journal of International Economics* 71(1), 72–95.
- Baier, Scott L. and Jeffrey H. Bergstrand, 2009. Bonus vetus OLS: A simple method for approximating international trade-cost effects using the gravity equation. *Journal of International Economics* 77(1), 77–85.
- Baier, Scott L., Jeffrey H. Bergstrand, and Mitch Gainer, 2012. Structural gravity for foreign direct investment. Unpublished manuscript. University of Notre Dame.
- Baldwin, Richard E., 1993. The potential for trade between the countries of EFTA and Central and Eastern Europe. CEPR Discussion Papers no. 853, Centre for Economic Policy Research (CEPR), London.
- Baldwin, Richard E., 1994. Towards an Integrated Europe. Centre for Economic Policy Research (CEPR), London.
- Baldwin, Richard; and Daria Taglioni, 2006. Gravity for dummies and dummies for gravity equations, CEPR Discussion Papers no. 5850, Centre for Economic Policy Research (CEPR), London.
- Baltagi, Badi H., 2008. *Econometric Analysis of Panel Data*. 4th edition. Chichester: Wiley & Sons.
- Baltagi, Badi H., Peter H. Egger, and Michael Pfaffermayr, 2003. A generalized design for bilateral trade flow models. *Economics Letters* 80(3), 391–397.
- Baltagi, Badi H., Peter H. Egger, and Michael Pfaffermayr, 2007. Estimating models of complex FDI: Are there third-country effects? *Journal of Econometrics* 140(1) 260–281.
- Baltagi, Badi H., Peter H. Egger, and Michael Pfaffermayr, 2008. Estimating regional trade agreement effects on FDI in an interdependent world. *Journal of Econometrics* 145(1–2), 194–208.
- Behrens, Kristian, Cem Ertur, and Wilfried Koch, 2012. 'Dual' gravity: Using spatial econometrics to control for multilateral resistance. *Journal of Applied Econometrics* 27(5), 773–794.
- Bergstrand, Jeffrey H., 1989. The generalized gravity equation, monopolistic competition, and the factor-proportions theory in international trade. *Review of Economics and Statistics* 71(1), 143–153.

- Bergstrand, Jeffrey H., Peter H. Egger, and Mario Larch, 2013. Gravity redux: Estimation of gravity-equation coefficients, elasticities of substitution, and general equilibrium comparative statics under asymmetric bilateral trade costs. *Journal of International Economics* 89(1), 110–121.
- Bertolia, Simone and Jesús Fernández-Huertas Moraga, 2012. Multilateral resistance to migration. Unpublished manuscript, Fundación de Estudios de Economía Aplicada (FEDEA).
- Blonigen, Bruce A., Ronald B. Davies, Helen T. Naughton, and Glen R. Waddell, 2008. Spacey parents: Spatial auto-regressive patterns in inbound FDI, in S. Brakman and H. Garretsen (eds.), *Foreign Direct Investment and the Multinational Enterprise*. Cambridge, MA: MIT Press, 173–198.
- Blonigen, Bruce A., Ronald B. Davies, Glen R. Waddell, and Helen T. Naughton, 2007. FDI in space: Spatial auto-regressive relationships in foreign direct investment. *European Economic Review* 51(5), 1303–1325.
- Blundell, Richard and Stephen Bond, 1998. Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* 87(1), 115–143.
- Blundell, Richard, Rachel Griffith, and Frank Windmeijer, 2002. Individual effects and dynamics in count data models. *Journal of Econometrics* 108(1), 113–131.
- Brenton, Paul and Francesca DiMauro, 1998. Is there any potential in trade in sensitive industrial products between the CEECs and the EU? *World Economy* 21(3), 285–304.
- Breusch, Trevor S., Grayham E. Mizon and Peter Schmidt, 1989. Efficient estimation using panel data. *Econometrica* 57(3), 695–700.
- Bun, Maurice J.G., and Franc J.G.M. Klaasen, 2002. The importance of dynamics in panel gravity models of trade, mimeo.
- Cameron, Colin A. and Pravin K. Trivedi, 2005. *Microeometrics, Methods and Applications*. Cambridge: Cambridge University Press.
- Cameron, Colin A., J.B. Gelbach, and D.L. Miller, 2011. Robust inference with multiway clustering. *Journal of Business & Economic Statistics* 29(2), 238–249.
- Carrère, Céline. 2006. Revisiting the effects of regional trade agreements on trade flows with proper specification of the gravity model. *European Economic Review* 50(2), 223–247.
- Chamberlain, Gary 1982. Multivariate regression models for panel data. *Journal of Econometrics* 18(1), 5–46.
- Coghlin, Cletus C. and Eran Segev, 2000. Foreign direct investment in China: A spatial econometric study. *World Economy* 23(1), 1–23.
- Cornwell, Christopher, Peter Schmidt and Donald Wyhowski, 1992. Simultaneous equations and panel data. *Journal of Econometrics*, January–February 51(1–2), 151–181.
- Davis, Peter., 2002. Estimating multi-way error components models with unbalanced data structures. *Journal of Econometrics* 106(1), 67–95.
- de Arce, Rafael and Ramon Mahia, 2009. Determinants of bilateral immigration flows between the European Union and some mediterranean partner countries: Algeria, Egypt, Morocco, Tunisia and Turkey. MPRA Paper No. 14547, Munich Personal REPEC Archive.
- De Benedictis, Luca and Claudio Vicarelli, 2005. Trade potentials in gravity panel data models. *Topics in Economic Analysis and Policy* 5(1), 1386–1386.

- De la Mata, Tamara and Carlos Llano, 2011. Social networks and trade of services: modeling interregional tourism flows with spatial and network autocorrelation effects. Unpublished manuscript, Universidad Autónoma de Madrid.
- Not in text: Dekle, Robert, Jonathan Eaton, and Samuel S. Kortum, 2007. Unbalanced trade. *American Economic Review* 97(2), 351–355.
- Docquier, Frédéric, Hillel Rapoport, and Sara Salomone, 2010, Remittances and skills: Evidence from bilateral data. Unpublished manuscript. Bar Ilan-University.
- Dustmann, Christian and M.E. Rochina-Barrachina, 2007. Selection correction in panel data models: An application to the estimation of females' wage. *Econometrics Journal* 10(2), 263–293.
- Eaton, Jonathan and Akiko Tamura, 1994. Bilateralism and regionalism in Japanese and U.S. trade and direct foreign investment patterns. *Journal of the Japanese and International Economies* 8(4), 478–510.
- Eaton, Jonathan and Samuel S. Kortum, 2002. Technology, geography, and trade. *Econometrica* 70(5), 1741–1779.
- Eaton, Jonathan, Samuel S. Kortum, and Sebastian Sotelo, 2012. International trade: Linking micro and macro. NBER Working Paper no. 17864.
- Egger, Hartmut, Egger, Peter H. and David Greenaway, 2008. The trade structure effects of endogenous regional trade Agreements. *Journal of International Economics* 74(2), 278–298.
- Egger, Peter H., 2000. A note on the proper econometric specification of the gravity equation. *Economics Letters* 66(1), 25–31.
- Egger, Peter H., 2001. European exports and outward foreign direct investment: A dynamic panel data approach. *Review of World Economics (Weltwirtschaftliches Archiv)* 137(3), 427–449.
- Egger, Peter H., 2002. An econometric view on the estimation of gravity models and the calculation of trade potentials. *World Economy* 25(2), 297–312.
- Egger, Peter H., 2004a. Estimating trading bloc effects with panel data. *Review of World Economics/Weltwirtschaftliches Archiv* 140(1), 151–166.
- Egger, Peter H., 2004b. On the problem of endogenous unobserved effects in the estimation of gravity models. *Journal of Economic Integration* 19(2), 182–191.
- Egger, Peter H., 2005. Alternative techniques for estimation of cross-section gravity models. *Review of International Economics* 13(5), 881–891.
- Egger, Peter H., 2010. Bilateral FDI potentials for Austria. *Empirica* 37(1), 5–17.
- Egger, Peter H. and Mario Larch, 2011. An assessment of the Europe agreements' effects on bilateral trade, GDP, and welfare. *European Economic Review* 55(2), 263–279.
- Egger, Peter H. and Mario Larch, 2012. Estimating consistent border effects in gravity models with multilateral resistance. *World Economy* 35(9), 1121–1125.
- Egger, Peter H. and Andrea Lassmann, 2012. The language effect in international trade: A meta-analysis. *Economics Letters* 116(2), 221–224.
- Egger, Peter H. and Valeria Merlo, 2011. Statutory corporate tax rates and double-taxation treaties as determinants of multinational firm activity, *Finanzarchiv* 67(2), 145–170.
- Egger, Peter H. and Valeria Merlo, 2012. BITs bite: An anatomy of the impact of bilateral investment treaties on multinational firms. *Scandinavian Journal of Economics* 114(4), 1240–1266.
- Egger, Peter H. and Douglas R. Nelson, 2011. How bad is antidumping? Evidence from panel data. *Review of Economics and Statistics* 93(4), 1374–1390.

- Egger, Peter H. and Sergey Nigai, 2013. Structural constrained ANOVA-type estimation of gravity panel data models. Unpublished manuscript, ETH Zurich.
- Egger, Peter H. and Michael Pfaffermayr, 2003. The proper panel econometric specification of the gravity equation: A three-way model with bilateral interaction effects. *Empirical Economics* 28(3), 571–580.
- Egger, Peter H. and Michael Pfaffermayr, 2004a. Distance, trade and FDI: A Hausman-aylor SUR approach. *Journal of Applied Econometrics* 19(2), 227–246.
- Egger, Peter H. and Michael Pfaffermayr, 2004b. Foreign direct investment and European integration in the 90s. *World Economy* 27(1), 99–110.
- Egger, Peter H. and Michael Pfaffermayr, 2011. Structural estimation of gravity models with path-dependent market entry. CEPR Discussion Papers 8458, London: Centre for Economic Policy Research.
- Egger, Peter H. and Michael Pfaffermayr, 2013. The pure effects of European integration on intra-EU trade. *World Economy* 36(6), 701–712.
- Egger, Peter H. and Georg Wamser, 2013a. Multiple faces of preferential market access: Their causes and consequences. *Economic Policy* 28(73), 145–187.
- Egger, Peter H. and Georg Wamser, 2013b. Effects of the endogenous scope of preferentialism on international goods trade. CESifo Working Paper no. 4208.
- Egger, Peter H., Mario Larch, and Kevin E. Staub, 2012. Trade preferences and bilateral trade in goods and services: A structural approach. CEPR Discussion Papers no. 9051, London: C.E.P.R.
- Egger, Peter H., Mario Larch, Kevin Staub, and Rainer Winkelmann, 2011. The trade effects of endogenous preferential trade agreements. *American Economic Journal: Economic Policy* 3(3), 113–143.
- Egger, Peter H., Simon Loretz, Michael Pfaffermayr, and Hannes Winner, 2009. Bilateral effective tax rates and foreign direct investment. *International Tax and Public Finance* 16(6), 2009, 822–849.
- Eichengreen, Barry, and Douglas A. Irwin, 1998. The role of history in bilateral trade flows, in J.A. Frankel (ed.) *The Regionalization of the World Economy*, (Chicago: University of Chicago Press), 33–57.
- Etzo, Ivan, 2009. The end of the “Empirical Puzzle” and the determinants of interregional migration in Italy: A panel data analysis. Unpublished manuscript, Università degli Studi di Cagliari.
- Fally, Thibault, 2012. Structural gravity and fixed effects. Unpublished manuscript, University of Colorado.
- Feenstra, Robert C., 2002. Border effects and the gravity equation. Consistent methods for estimation. *Scottish Journal of Political Economy* 49(5), 491–506.
- Gros, Daniel and Andrzej Gonciarz, 1996. A note on the trade potential of Central and Eastern Europe. *European Journal of Political Economy* 12(4), 709–721.
- Hausman, Jerry A. and William E. Taylor, 1981. Panel data and unobservable individual effects. *Econometrica* 49(6), 1377–1398.
- Head, Keith and Thierry Mayer, 2011. Gravity, market potential and economic development. *Journal of Economic Geography* 11(2), 281–294.
- Head, Keith and John Ries, 2001. Increasing returns versus national product differentiation as an explanation for the pattern of US–Canada trade. *American Economic Review* 91(4), 858–876.

- Head, Keith, Thierry Mayer, and John Ries, 2010. The erosion of colonial trade linkages after independence. *Journal of International Economics* 81(1), 1–14.
- Helpman, Elhanan, Marc Melitz, and Yona Rubinstein, 2008. Estimating trade flows: Trading partners and trading volumes. *Quarterly Journal of Economics* 123(2), 441–487.
- Kang, S., 1985. A note on the equivalence of specification tests in the two-factor multivariate variance components model. *Journal of Econometrics* 28(2), 193–203.
- Kelejian, Harry H. and Ingmar R. Prucha, 1999. A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review* 40(2), 509–533.
- Krugman, Paul 1980. Scale economies, product differentiation, and the pattern of trade *American Economic Review*, December 70(5), 950–959.
- Leamer, Edward E. and James Levinsohn, 1995. International trade theory: The evidence, in G. M. Grossman and K. Rogoff (eds.), *Handbook of International Economics*, 1st edition. volume 3, chapter 26, Elsevier, North-Holland, Amsterdam, New York and Oxford, 1339–1394.
- Lebreton, Marie and Laïsa Roia, 2009. A spatial interaction model with spatial dependence for trade flows in Oceania: A preliminary analysis. Unpublished manuscript, Université Montesquieu Bordeaux IV.
- LeSage, James P. and R. Kelley Pace, 2008. Spatial econometric modeling of origin-destination flows. *Journal of Regional Science* 48(5), 941–967.
- Levchenko, Andrei and Jing Zhang, 2012. Comparative advantage and the welfare impact of European integration. *Economic Policy* 27(72), 567–602.
- Martínez-Zarzoso, Immaculada, Felicitas Nowak-Lehmann, and Nicolas Horsewood, 2009. Are regional trading agreements beneficial? Static and dynamic gravity models, *North American Journal of Economics and Finance* 20(1), 46–65.
- Martínez-Zarzoso, Immaculada, Felicitas Nowak-Lehmann, Stephan Klasen, and Mario Larch, 2009. Does German development aid promote German exports? *German Economic Review* 10(3), 317–338.
- Mátyás, László, 1997. Proper econometric specification of the gravity model. *World Economy* 20(3), 363–368.
- Mátyás, László, 1998. The gravity model: Some econometric considerations. *World Economy* 21(3), 397–401.
- Mayda, Anna Maria, 2010. International migration: A panel data analysis of the determinants of bilateral flows, *Journal of Population Economics* 23(4), 1249–1274.
- McCullagh, P. and J.A. Nelder, 1989. *Generalized Linear Models*. 2nd edition. CRC Monographs on Statistics & Applied Probability. London and New York: Chapman & Hall.
- Millimet, Daniel L. and Thomas Osang, 2007. Do state borders matter for U.S. intranational trade? The role of history and internal migration. *Canadian Journal of Economics* 40(1), 93–126.
- Mundlak, Yair, 1978. On the pooling of time series and cross-section data. *Econometrica* 46(1), 69–85.
- Nilsson, Lars, 2000. Trade integration and the EU economic membership criteria. *European Journal of Political Economy* 16(4), 807–827.
- Okawa, Yohei and Eric van Wincoop, 2013. Gravity in international finance. *Journal of International Economics* 87(2), 205–215.

- Olivero, Maria Pia and Yoto V. Yotov, 2012. Dynamic gravity: Endogenous country size and asset accumulation. *Canadian Economic Journal* 45(1), 64–91.
- Orefice, Gianluca, 2013. International Migration and Trade Agreements: the new role of PTAs. FIW Working Paper no. 111. Centre d'Études Prospectives et d'Informations Internationales, Paris.
- Porofjan, A., 2001. Trade flows and spatial effects: The gravity model revisited. *Open Economies Review* 12(3), 265–280.
- Pöyhönen, Pentti, 1963. A tentative model for the volume of trade between countries. *Weltwirtschaftliches Archiv* 90(1), 93–99.
- Raymond, W., P. Mohnen, F. Palm, and S. Schim van der Loeff, 2010. Persistence of innovation in Dutch manufacturing: Is it spurious? *Review of Economics and Statistics* 92(3), 495–504.
- Santos Silva, João M.C. and Silvana Tenreyro, 2006. The log of gravity. *Review of Economics and Statistics* 88(4), 641–658.
- Santos Silva, João M.C. and Silvana Tenreyro, 2009. Trading partners and trading volumes: Implementing the Helpman-Melitz-Rubinstein model empirically. CEP Discussion Papers dp0935, London: Centre for Economic Performance, LSE.
- Santos Silva, João M.C. and Silvana Tenreyro, 2011. Further simulation evidence on the performance of the Poisson pseudo-maximum likelihood estimator. *Economics Letters* 112(2), 220–222.
- Semykina, A. and J.M. Wooldridge, 2010. Estimating panel data models in the presence of endogeneity and selection. *Journal of Econometrics* 157(2), 375–380.
- Serlenga, Laura and Yongcheol Shin, 2007. Gravity models of intra-EU trade: Application of the CCEP-HT estimation in heterogeneous panels with unobserved common time-specific factors. *Journal of Applied Econometrics* 22(2), 361–381.
- Stock, James H. and M.W. Watson, 2008. Heteroskedasticity-robust standard errors for fixed effects panel data regression. *Econometrica* 76(1), 155–174.
- Wamser, Georg, 2011. Foreign (in)direct investment and corporate taxation. *Canadian Journal of Economics* 44(4), 1497–1524.
- Wansbeek, Tom and Arye Kapteyn, 1989. Estimation of the error-components model with incomplete panels. *Journal of Econometrics* 41(3), 341–61.
- White, Halbert, 1982. Maximum likelihood estimation of misspecified models. *Econometrica* 50(1), 1–25.
- Wooldridge, Jeffrey M., 1995. Selection corrections for panel data models under conditional mean independence assumptions. *Journal of Econometrics* 68(1), 115–132.
- Wooldridge, Jeffrey M., 2003. Cluster-sample methods in applied econometrics. *American Economic Review* 93(2), 133–138.
- Wooldridge, Jeffrey M., 2005. Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics* 20(1), 39–54.
- Wooldridge, Jeffery M. 2010. *Econometric Analysis of Cross-Section and Panel Data*. 2nd edition. Cambridge, MA: MIT-Press.
- Wyhowski, Donald J., 1994. Estimation of a panel data model in the presence of correlation between regressors and a two-way error component. *Econometric Theory* 10(1), 130–139.



## AUTHOR INDEX

Page numbers followed by *f* or *t* indicate figures or tables, respectively. Numbers followed by n indicate endnotes.

- Aasness, J., 342, 359  
Abadie, A., 268, 281  
Abbott, Andrew J., 622, 635  
Abbring, J.H., 258, 270, 281  
Abowd, J., 589, 603  
Abramowitz, M., 577n2, 579  
Abrevaya, J., 188, 196  
Ackerberg, D., 568, 579  
Adams, J., 515  
Adamson, A., 515  
Adda, J., 351, 359  
Afifi, A.A., 150, 168  
Afriat, S.N., 519, 540  
Ahn, S., 151, 168  
Ahn, S.C., 16, 41, 78–79, 86, 90, 97, 107, 118, 126, 145, 439, 446, 458, 484–485, 487–488, 531–533, 540–541  
Ai, C., 285, 319  
Aigner, D., 519, 541  
Aihie Sayer, A., 515  
Ailawadi, K., 576, 579  
Akaike, H., 413, 416, 540n4, 541  
Alam, I.M.S., 519, 541  
Alan, S., 352–353, 359  
Alberman, E., 513  
Aldy, J., 584, 587–588, 602n3, 603, 606  
Allenby, G., 173, 196, 560–562, 579, 582  
Allers, M., 395–396  
Allison P.D., 242, 254  
Almanidis, P., 534, 541  
Altonji, J., 183, 196, 310, 319  
Altonji, J.G., 600, 603  
Alvarez, J., 81, 86, 107, 122, 145, 364, 391, 396, 455, 460, 466, 488  
Amemiya, T., 596, 603, 630, 635  
Amengual, D., 16, 41  
Anders, I., 541  
Andersen, R., 422, 425, 446  
Andersen, S.L., 444n1, 447  
Anderson, D.R., 534, 540n4, 542  
Anderson, E.B., 208, 231  
Anderson, J.E., 609–610, 629, 636  
Anderson, T.W., 78, 86, 107, 118, 145, 297, 308, 319, 344, 359, 407, 409, 416, 597, 603  
Andrews, D., 3, 41  
Andrews, R.L., 556, 579  
Andreyeva, T., 350, 356, 361  
Angrist, J.D., 257, 260, 265–268, 280n8, 281, 283  
Anselin, L., 4, 28, 42, 363–366, 367t, 368, 373, 393, 397, 625, 636  
Antman, F., 343, 359  
Antweiler, W., 151, 169  
Appelbe, T.W., 414, 416, 416n1  
Aquaro, M., 430–432, 446  
Arbia, G., 195–196  
Arellano, M., xi, xv, 76–79, 81, 83, 86–87, 107–108, 118, 120, 122, 124, 138, 141–142, 145–146, 151, 160–161, 169, 189, 193, 196, 203, 231, 246, 254, 258, 282, 285, 308, 319, 327, 344, 359, 364, 391, 396–397, 438–439, 447, 455, 458–460, 466, 469, 478, 488, 595, 597–598, 603, 622, 636  
Arrow, K.J., 519, 541  
Ashenfelter, O., 593, 603  
Åslund, A., 546  
Aten, B., 547  
Athey, S., 282

- Attanasio, O., 352, 359  
 Augustin, N.H., 542
- Badunenko, O., 537, 541  
 Bago d'Uva, T., 175, 196, 250, 254  
 Bai, J., xii, 15–16, 22–23, 42, 51–53, 61–62,  
     64, 66, 71–72, 126, 129, 146, 151–152,  
     169, 461, 466–467, 474, 483–484,  
     487n7, 488, 523, 531, 541  
 Bai, Z.D., 5, 42, 45  
 Baicker, K., 375, 397  
 Baier, S.L., 610, 627–628, 630, 632, 636  
 Bailey, N., 6–7, 14, 32, 42  
 Baker, R.M., 82, 108  
 Baldwin, R.E., 630, 633, 635n10, 636  
 Balestra, P., 196  
 Balk, B.M., 521, 541  
 Baltagi, B.H., xi, xii, xv, 28, 32–33, 42, 50, 72,  
     77, 107, 118, 120, 146, 150–151, 161,  
     169, 203, 231, 269, 282, 285, 314, 319,  
     327, 359, 363–367, 367t, 392–395, 397,  
     419, 433–440, 444, 447, 455, 471, 477,  
     488, 496, 512, 523, 541, 584–585,  
     603–604, 612–613, 624, 626, 629–630,  
     632, 636  
 Bañbara, M., 154, 169  
 Banerjee, A., 59, 70, 71n3, 72, 318–319  
 Banerjee, M., 442, 447  
 Banks, J., 354, 359  
 Bargmann, R.E., 150, 170  
 Bari, W., 441, 446n14, 447, 450  
 Barker, D., 514–515  
 Barnett, V., 420, 447  
 Barnow, B., 600, 604  
 Barro, R.J., 545  
 Bartlett, M.S., 476, 488  
 Bartolucci, F., 212, 217–220, 231  
 Bassett, G., 303, 321  
 Bates, J.M., 540n5, 541  
 Battese, G.E., 519–520, 526–527, 530, 532,  
     541–542  
 Beale, E.M.L., 150, 169  
 Beck, N., 474, 478, 488  
 Behrens, Kristian, 626, 635n7, 636  
 Bekker, P.A., 80, 107, 334, 359, 460, 488  
 Bell, K., 195–196  
 Belsley, D.A., 418, 446n16, 447  
 Bera, A., 28, 42, 184, 196  
 Berger, J.O., 112, 143, 146, 444, 447  
 Bergstrand, J.H., 609–610, 627–630, 632,  
     634, 636–637  
 Berk, K.N., 41n13, 42  
 Berkovec, J., 563, 579  
 Berliner, L.M., 444, 447  
 Berndt, E.R., 540n3, 547  
 Beron, K., 195–196  
 Berry, S., 551, 560, 566, 579  
 Bertolia, Simone, 637  
 Bertschuk, I., 182, 196  
 Besley, T., 76, 107  
 Bester, A., 480, 488  
 Bester, C., 191, 196  
 Bester, C.A., 395, 397  
 Bhargava, A., 86, 107, 388, 398, 466, 488  
 Bhat, C., 172, 180, 195–197  
 Bhattacharya, P.K., 303, 319  
 Bhopal, R., 516  
 Bickel, P.J., 291, 319  
 Biewen, M., 216, 231  
 Bilias, Y., 309, 321  
 Binder, M., 81, 85, 107, 468, 488  
 Biorn, E., 150, 169, 269, 282  
 Biørn, E., 337, 342–343, 359  
 Black, D.A., 585, 593, 604  
 Blanden, J., 496, 502, 510, 512  
 Block, H., 552, 579  
 Blonigen, B.A., 625, 637  
 Bloom, N., 520, 540n3, 541  
 Blume, L., 198  
 Blundell, R., 76–77, 79–80, 83, 86, 88, 93,  
     96–98, 107–108, 119, 146, 169, 242,  
     244, 246–247, 254, 265, 282, 354, 359,  
     440, 447, 458, 460, 489, 493, 509, 512,  
     597–598, 604, 635n5, 637  
 Böckenholdt, U., 254  
 Bockstaal, N., 195–196  
 Boente, G., 150, 169  
 Bollinger, C.R., 593, 604  
 Bond, E., 622, 636  
 Bond, S., 76–80, 83, 86, 88, 90, 93, 96–98,  
     107–108, 118–119, 145–146, 169, 246,  
     254, 258, 282, 308, 319, 344, 359,  
     438–440, 447, 458–460, 488–489,  
     597–598, 603–604, 622, 637

- Bondell, H.D., 310, 319  
 Bonhomme, S., 138, 141–142, 145, 493, 510,  
   512  
 Bontempi, M.E., 461, 489  
 Börsch-Supan, A., 188, 229, 231  
 Bound, J., 82, 108, 348, 357, 359–360, 593,  
   604  
 Bover, O., 76, 79, 83, 86, 107, 120, 145, 169,  
   488, 595, 597, 603  
 Bowsher, C.G., 87, 93, 108  
 Box, G.E.P., 444n1, 447  
 Boyd, A., 505, 512–513  
 Boyle, P., 515  
 Bramati, M.C., 419, 425, 427, 429–430, 434,  
   436, 444n3, 447  
 Brame, R., 231  
 Brännäs, K., 239, 244, 246, 254  
 Brant, R., 191, 197  
 Breitung, J., xii, xiv, xv, 16, 42, 46, 50, 59, 67,  
   72, 134, 146, 160, 169, 461, 464–465,  
   469, 471–472, 483–484, 487n6, 489  
 Brenton, Paul, 633, 637  
 Breslow, N.E., 441, 447  
 Bresson, G., 393–395, 397, 419, 433–434,  
   436–438, 447  
 Brett, C., 513–514  
 Breusch, T.S., 28, 30–31, 33, 42, 183, 186,  
   197, 447, 596, 604, 630, 637  
 Breyer, S.G., 602, 604  
 Brock, W., 178, 198  
 Brown, C., 348, 357, 360, 584, 593, 604  
 Browning, M., 352–353, 359  
 Brownstone, D., 358, 360  
 Bruno, G.S.F., 487n1, 489  
 Bryk, A., 191, 197  
 Buchanan, J.M., 541  
 Buchinsky, M., 303, 319  
 Buckland, S.T., 540n4, 542  
 Budge, D., 514  
 Bühlmann, P., 66, 72  
 Bukodi, E., 499, 512  
 Bun, M.J.G., xii, 81–86, 93, 95–97, 108, 464,  
   487n4, 489, 622, 637  
 Buonaccorsi, J., 335, 360  
 Burda, M., 560, 579  
 Burgess, S., 502, 512, 514–515  
 Burnham, K.P., 534, 540n4, 542  
 Burridge, P., 392, 398  
 Bushway, S., 213, 231  
 Busso, M., 264, 282  
 Butler, J., 180, 184, 197  
 Butler, J.S., 554–555, 562, 579  
 Butler, N., 500, 512  
 Bynner, J., 494, 496, 513, 515  
 Cai, Z., 303, 318–319  
 Cain, G., 600, 604  
 Calderwood, L., 514–515  
 Cameron, A.C., 243, 253–254  
 Cameron, C., 171, 197, 336, 360  
 Cameron, C.A., xiii, 618, 627, 635nn4–5, 637  
 Cameron, N., 516  
 Campbell, H., 513  
 Canay, I.A., 309, 320  
 Canova, F., 415–416, 467–468, 489  
 Cantoni, E., 441–442, 447  
 Card, D., 279n2, 282, 585, 589, 603–604  
 Cardot, H., 150, 169  
 Carlin, B.P., 446n12, 447  
 Carree, M.A., 85, 108, 464, 487n4, 489  
 Carrère, Céline, 71, 628, 633, 637  
 Carrion-i-Silvestre, J.L., 64, 71  
 Carro J., 189, 191, 197  
 Carroll, C., 352, 360  
 Carroll, R.J., 288–290, 293–294, 309, 314,  
   321–324, 540n4, 542  
 Case, A., 76, 107, 375, 398, 500, 508, 513  
 Casella, G., 112, 146–147  
 Caves, D.W., 521, 542  
 Chabé-Ferret, S., 282  
 Chakir, R., 174, 195, 197  
 Chamberlain, G., xi, xv, 42, 119, 146, 186,  
   188, 190–194, 197, 203, 208–209, 212,  
   214, 224, 231, 241–242, 246, 254, 565,  
   567–571, 577–578n5, 579, 619, 637  
 Chang, Y.J., 151, 169, 585, 603–604  
 Charnes, A., 519, 542  
 Chatterjee, S., 420, 446n16, 447  
 Chaturvedi, A., 450  
 Chaudhuri, P., 303, 320  
 Chen, E.K., 538, 542  
 Chen, J., 294, 300–302, 316–318, 320, 322  
 Chen, S., 178, 197  
 Chen, X., 354, 358, 360, 375, 398

- Chernozhukov, V., 145n9, 146, 176, 197, 308, 310–311, 313, 320
- Chesher, A., 176, 183–184, 197–198
- Chiang, M.H., 47–49, 51, 73
- Chib, S., 240, 254, 446n12, 447
- Ching, A., 568, 575–576, 579
- Chintagunta, P., 216, 230–231
- Cho, H., 442–443, 450
- Choi, I., xii, xv, 16, 42, 46, 49–51, 58, 62, 72, 160, 169
- Christensen, B.J., 351, 360
- Christensen, L.R., 521, 542
- Chudik, A., xi, 5–7, 12, 14, 20–22, 24–27, 34, 41n13, 42–43, 374–375, 393, 398, 474, 485, 489
- Chue, T., 51, 72
- Ciccarelli, M., 467–468, 489
- Čížek, P., 430–432, 441, 446–447
- Claeskens, G., 540n4, 542
- Clark, P., 514
- Clayton, D.G., 441, 447
- Clemen, R.T., 540n5, 542
- Cliff, A., 4, 43
- Cliff, A.D., 392, 398
- Coakley, J., 15, 20–21, 43
- Coe, D.T., 519, 542
- Coelli, T., 520–522, 526–527, 530, 532, 541–542
- Coghlin, C.C., 637
- Conley, T., 480, 488
- Conley, T.G., 375, 395, 397–398
- Conniffe, D., 150, 170
- Connor, G., 72
- Conti, G., 510, 513
- Contoyannis, C., 172–173, 192–194, 198
- Cook, R.D., 442, 446n16, 447
- Cooper, C., 515
- Cooper, R., 351, 359
- Cooper, W.W., 519, 542
- Corcoran, M., 214, 231
- Corley, J., 513
- Cornwell, C., 438, 447, 520, 523, 527–528, 532, 542, 585, 596, 604, 630, 637
- Corra, G.S., 519, 541
- Costa Dias, M., 265, 282
- Couteur, A.le, 515
- Cox, D., 183, 198
- Cox, D.R., 208, 232, 377, 398
- Cox, L., 514
- Crawford, G., 568, 580
- Crepon, B., 244, 246, 254
- Croux, C., 419, 425, 427, 429–430, 434, 436, 444n3, 445n7, 447–448
- Cuesta, R.A., 527, 542
- Currie, J., 500, 513
- Cutler, D., 500, 509, 513
- Dagenais, D.L., 339, 360
- Dagenais, M.G., 339, 360
- Dale, A., 515
- Das, M., 198
- Davey Smith, G., 512–514
- Davies, A., 540n5, 542
- Davies, R.B., 637
- Davis, P., 151, 170, 611, 613–614, 637
- Dawid, A.P., 280n11, 282
- Deanfield, J., 515
- de Arce, Rafael, 637
- Dearden, L., 502, 504, 512–513
- Deary, I., 513–514
- Deaton, A., 258, 282, 496, 513, 602n1
- Deb, P., 249, 251–252, 254
- Debarsy, N., 373–374, 381, 393, 398
- De Benedictis, Luca, 633, 637
- Debreu, G., 519, 537, 542
- Dees, S., 486, 489
- De Finetti, B., 443, 448
- Dehon, C., 418, 445n6, 448
- de Jong, R., 364, 368, 382, 386–387, 401
- Dekle, Robert, 638
- De la Mata, T., 638
- de Leeuw, J., 339, 362
- Demetrescu, M., 477, 489, 534, 542
- Demidenko, E., 335, 360
- Deng, Y., 395, 397
- Dennison, E., 515
- Der, G., 443, 448
- Dette, H., 310, 315, 320
- De Vita, Glauco, 622, 635
- Dhaene, G., 26, 43, 85, 108, 124, 146, 249, 254, 440, 448, 464, 489
- Diao, X., 519, 542
- Dickey, D.A., 41n13, 45
- Diebold, F.X., 540n5, 542

- Diewert, W.E., 521, 542  
 Diggle, P.J., 244, 255  
 Diggle, R., 182, 198  
 diMauro, F., 486, 489, 633, 637  
 DiNardo, J., 264, 282  
 Dineen, C.R., 414, 416, 416n1  
 Ding, W., 281n22, 282  
 Docquier, Frédéric, 634n2, 638  
 Dolado, J.J., 59, 72  
 Donoho, D.L., 420, 448  
 Doornik, J.A., 108  
 Doran, H.E., 461, 489  
 Drewett, R., 515  
 Driscoll, J.C., 480, 489  
 Druska, V., 364, 398, 534, 543  
 Dubé, Jean-Pierre, 570, 580  
 Dufour, J.M., 31, 43, 480, 490  
 Duguet, E., 244, 246, 254  
 Durlauf, S., 178, 198  
 Durlauf, S.N., 415–416  
 Dustmann, C., 619, 638  
 Dynan, K.E., 343, 360
- Eaton, Jonathan, 609, 628–629, 634n1, 638  
 Eckstein, 568  
 Ecob, R., 443, 448  
 Edgerton, D.L., 66, 75  
 Egger, H., 630, 632, 638  
 Egger, P.H., xv, 230, 232, 364–365, 367, 367t, 392–393, 397, 455, 490, 609–610, 613, 617–622, 624–626, 628–634, 634n1, 635n9, 635n11, 636–639  
 Eichengreen, Barry, 622, 639  
 Eickmeier, S., 486, 490  
 Ekelund, U., 515  
 Elashoff, R.M., 150, 168  
 Elhorst, J., 363–364, 366, 367t, 371, 373, 382, 385–389, 395–396, 398  
 Ellerman, A.D., 547  
 Elliott, G., 135, 146, 198  
 Elliott, J., 500–501, 513, 515  
 Elrod, Terry, 551, 553, 556, 561, 580  
 Emanuel, I., 495, 513  
*Empirical Economics*, 303  
 Engle, R.F., 46, 55, 72, 490  
 Entorf, H., 47, 73  
 Entur, C., 534, 543
- Erdem, T., 568, 573–577, 578n6, 578n9, 579–580  
 Erickson, T., 339, 360  
 Ertur, C., 364, 373–374, 393, 398, 626, 635n7, 636  
 Etzo, Ivan, 622, 639  
 Eubank, R.L., 543  
 Evans, S., 513  
 Evans, W.N., 606  
 Evdokimov, K., 310, 320  
 Everaert, G., 41n10, 43, 77, 96, 108
- Fairley, L., 516  
 Fall, C., 514–515  
 Fally, T., 615, 639  
 Fan, J., 290–291, 303, 305, 320  
 Fan, Z., 335, 360  
 Färe, R., 521, 543  
 Farrell, M.J., 519, 537, 543  
 Feenstra, Robert C., 629, 639  
 Fei, L., 195, 201  
 Feng, Q., 477, 488, 534, 543  
 Fernández-Huertas Moraga, Jesúis, 637  
 Fernández-Val, I., 145n9, 146, 177, 187–189, 193, 197–198, 243, 255, 310, 320  
 Fertig, A., 500, 502, 511, 513  
 Fethi, M., 541  
 Filakti, H., 513  
 Fingleton, B., 363, 366, 367t, 394, 397–398  
 Firpo, S., 258, 282  
 Fischer, S., 546  
 Fisher, I., 543  
 Fitzenberger, B., 281n21, 282  
 Fitzgerald, J., 585, 604  
 Flores-Lagunes, A., 195, 198  
 Fok, D., 174, 201  
 Forni, M., 4, 43  
 Førsund, F.R., 521, 543  
 Fraiman, R., 150, 169  
 Franses, P.H., 570, 582  
 Franzese, R.J., 364, 398  
 Fraser, A., 505, 512–513  
 Frazier, C., 364, 398  
 Freedman, L.S., 542  
 Frees, E.W., 31, 43, 442, 447, 490  
 Freyberger, J., 302, 320  
 Friberg, P., 515

- Fried, H.O., 520–521, 543  
Friedman, M., 414, 416  
Frijters, P., 504, 508, 513  
Frölich, M., 264, 268–269, 280n19, 282  
Fu, B., 305, 321  
Fu, L., 304–305, 320  
Fuertes, A.M., 43  
Fujiki, H., 410, 410*t*, 411, 411*t*, 412, 416  
Fung, W., 306, 321  
Fung, W.K., 305, 321  
  
Gainer, Mitch, 610, 636  
Galichoni, A., 310, 320  
Galindo-Rueda, F., 500, 510, 514  
Galvao, A.F., 307–308, 320–321  
Gangopadhyay, A.K., 303, 319  
Gao, J., 294, 300–302, 316–318, 320, 322  
Garrett, J., 173, 196  
Garvin, H., 150, 169  
Gassner, M., 418, 445n6, 448  
Gayle, V., 515  
Geary, R.C., 339, 360, 448  
Gedenk, K., 579  
Gelbach, J.B., 627, 637  
Gengenbach, C., 60–61, 63–64, 71, 73  
Gervini, D., 431, 448  
Geweke, J., 4, 43, 554, 557, 560, 580  
Ghosh, A., 515  
Ghosh, J.K., 112, 143, 146  
Giavazzi, F., 429, 448  
Gill, R.D., 270, 282  
Girón, F.J., 112, 146–147  
Glass, A., 534, 543  
Goldberger, A., 600, 604  
Goldberger, A.S., 394, 399  
Golding, J., 512–513  
Goldstein, H., 500, 512–513, 515  
Goldthorpe, J., 499, 512, 512n2, 514  
Gonciarz, Andrzej, 633, 639  
Good, D.H., 518, 521, 527, 540, 543  
Goodman, A., 512  
Goolsbee, A., 360  
Gottschalk, P., 585, 604  
Gourieroux, C., 33, 43, 120, 146  
Gow, A., 513  
Granger, C.W., 540n5, 541  
Granger, C.W.J., 24–25, 43, 46, 55, 72  
Grassetti, L., 466, 490  
Gravelle, H., 192, 198  
Greaves, E., 512  
Greenaway, David, 630, 632, 638  
Greenaway-McGrevy, R., 484, 490  
Greenberg, E., 240, 254, 446n12, 447  
Greene, W.H., xii, 171–172, 177, 180,  
    183–186, 188–190, 198–199, 242–243,  
    251, 255, 528–530, 534, 543, 547  
Greenland, S., 270, 284  
Greenstone, M., 593, 603  
Gregg, P., 505, 507, 512, 514  
Grifell-Tatjé, E., 521, 543  
Griffin, J.H., 455, 488  
Griffin, J.M., 523, 541  
Griffith, R., 242, 244, 246–247, 254, 635n5,  
    637  
Griliches, Z., 184, 187, 199, 240, 242, 255,  
    325, 331, 360, 470, 491, 517–519, 528,  
    540nn2–3, 543–544, 583, 585, 593–594,  
    597, 604  
Groen, J.J.J., 69–71, 73, 468, 490  
Groote, T.D., 41n10, 43  
Gros, D., 633, 639  
Grosskopf, S., 543  
Guadagni, P.M., 548, 552, 570, 572–575, 580  
Guimarães, P., 255  
Gupta, S., 576, 582  
Gutierrez, L., 57–58, 73  
  
Hachem, W., 5, 43  
Hadi, A.S., 420, 446n16, 447–448  
Hahn, J., 82, 108, 120–121, 124–125, 129,  
    131, 138–140, 145n4, 145n9, 146–147,  
    176, 186, 189–190, 192–193, 196–199,  
    320, 385, 399, 455, 463, 490, 593, 604  
Haining, R.P., 4, 28, 43  
Hajivassiliou, V., 557, 580  
Hales, C., 503, 514  
Hall, A.R., 487n3, 490  
Hall, B., 184, 187, 199  
Hall, B.H., 240, 242, 255  
Hall, J., 513  
Hall, R.E., 351, 360  
Halliday, T.J., 209, 232  
Hallin, M., 16, 43  
Halunga, A., 487n10, 490

- Ham, J., 192, 199  
 Hampel, E.K., 420, 448  
 Hampel, F.R., 420, 444n1, 448  
 Han, C., 83, 85, 108, 112, 119, 143, 145n3,  
     145n6, 147, 465, 484, 490, 492  
 Hansen, 597  
 Hansen, B.E., 52, 65, 74, 540n4, 544  
 Hansen, C., 191, 196, 308, 320, 480, 488  
 Hansen, C.B., 395, 397  
 Hansen, K., 502, 514  
 Hansen, L.P., 93, 108, 351, 360  
 Hanssens, D., 576, 582  
 Hao, J., xiv, 540n6, 541, 544, 547  
 Harding, M., 16, 43, 308, 320, 560, 579  
 Härdle, W., 314–315, 320–321  
 Hardy, R., 514–516  
 Harrington, D.P., 442, 450  
 Harris, K., 560, 580  
 Harris, M., 175, 184, 199  
 Harvey, D.I., 540n5, 546  
 Harwood, N., 515  
 Hastie, T.J., 150, 170  
 Hatton, T., 513  
 Haupt, H., 319n1, 322  
 Hause, J.C., 87, 109  
 Hausman, J., xiv, 82, 108, 184, 187, 192, 199,  
     240, 242, 255, 325, 331, 354, 360, 412,  
     419, 434–436, 448, 519, 528, 544, 560,  
     579, 583, 585, 593–597, 604, 623–625,  
     630, 639  
 Hawkes, D., 515  
 Hayakawa, K., 77, 81, 85, 92–93, 96–97, 109  
 He, X., 305–306, 310, 321  
 Head, J., 514  
 Head, K., 618, 627, 639–640  
 Heagerty, P., 255  
 Heckman, J., 173, 183, 193–194, 199, 245,  
     255, 257–258, 262, 268, 280n5,  
     281–283, 510, 513, 549, 552, 569, 580,  
     583, 587, 600, 604–605  
 Helpman, E., 519, 542, 610, 618, 640  
 Hendel, I., 575, 580  
 Henderson, D.J., 290, 293–294, 309, 314,  
     321, 537, 541  
 Henderson, J., 512–513  
 Hendry, D.F., 490  
 Hensher, D., 172, 183–185, 190, 198–199  
 Hensher, D.A., 251, 255  
 Hertfordshire Cohort Study Group, 515  
 Hess, W., 55, 75  
 Heston, A., 547  
 Hidalgo, J., 314, 319  
 Hill, M.S., 214, 231  
 Hines, J.R., 375, 398  
 Hinkley, D., 183, 198  
 Hinloopen, J., 429, 448  
 Hirano, K., 263, 283  
 Hitsch, G.J., 570, 580  
 Hjalmarsson, L., 521, 543  
 Hjort, N.L., 540n4, 542  
 Hjorth, J.U., 534, 544  
 Hlouskova, J., 60, 70, 75  
 Ho, Y.-Y., 443, 450  
 Hocking, R.R., 150, 170  
 Hoefller, A., 90, 108  
 Hoeting, J.A., 540n4, 544, 546  
 Hoffmaister, A., 519, 542  
 Holderlein, S., 209, 232  
 Hollingsworth, B., 184, 199  
 Holly, S., 14, 43, 364, 374–375, 399  
 Holtz-Eakin, D., 78, 109, 126, 147, 467,  
     487n8, 490, 598, 605  
 Homm, U., 477, 489, 534, 542  
 Honda, T., 318, 321  
 Hong, H., 354, 358, 360  
 Hong, P., 571, 580  
 Honoré, B., xi, xv, 107, 118, 138, 146, 160,  
     169, 178, 193, 196, 199, 203, 212, 217,  
     230–232, 285, 319, 430, 448  
 Horenstein, 16  
 Horowitz, J., 176, 184, 199  
 Horowitz, J.L., 339, 360  
 Horrace, W., 534, 543  
 Horrace, W.C., 364, 398  
 Horsewood, N., 622, 632, 640  
 Horsky, 568  
 Hotz, V.J., 262, 282, 605  
 Howard, S., 208, 232  
 Hsiao, C., xi, xiii, xv, 33, 43, 77–78, 81,  
     85–86, 107, 109, 118, 120, 138, 145, 147,  
     151, 160, 170, 188, 193, 199, 203, 232,  
     285, 297, 308, 319, 321, 327, 334, 344,  
     359–360, 364, 378, 399, 402, 407,  
     409–410, 410*t*, 411, 411*t*, 412–416,

- 416n1, 440, 449, 465, 468, 487n6, 488, 490, 597, 603
- Hu, F., 270, 284
- Hu, T., 303, 305, 320
- Hu, Y., 358, 361
- Huang, C.J., 538, 544
- Huang, T., 290, 298–299, 305, 320
- Huber, M., 264, 281n26, 283
- Huber, P.J., xv, 418, 421, 444n1, 448–449
- Huggins, R.M., 305, 321
- Hughes, A., 515
- Hughes, G., 515
- Hultberg, P.T., 520, 544
- Hurwicz, L., 41n11, 43
- Hwang, H.S., 150, 170
- Hypolite, J., 250, 255
- Hyslop, D.R., 350, 361, 580
- Hyson, R., 500, 513
- Iacovou, M., 505, 514
- Ibrahim, J.G., 442–443, 450
- Ichimura, H., 360
- Im, K.S., 58, 64, 67, 70, 73, 470, 473, 492
- Imai, S., 575–577, 580
- Imbens, G.W., 257–258, 260, 263–265, 267–268, 275, 280n8, 280n12, 282–283, 350, 361
- Ioannides, Y., 198
- Irish, M., 184, 197
- Irwin, D.A., 622, 639
- Irwin, E., 395, 399
- Isaksson, A., 544
- Jackson, M., 512n2, 514
- Jacobs, R., 198
- Jacobson, T., 69, 73
- Jaeger, D.A., 82, 108
- James, G.M., 150, 170
- Janz, N., 438, 440, 449
- Jappelli, T., 429, 448
- Jarque, C., 184, 196
- Jayet, H., 363–364, 366, 367t, 368, 373, 397
- Jeanty, P., 395, 399
- Jensen, P.S., 43
- Jeon, B.M., 521, 544
- Jeong, H., 16, 42
- Jeong, K., 315, 321
- Jin, S., 298–299, 315–317, 321, 323, 396, 400
- Jochmans, K., 26, 43, 85, 108, 124, 146, 249, 254, 464, 489
- Johansen, S., 55, 67–68, 73
- Johansson, P., 239, 246, 254
- John, S., 32, 44, 477, 490
- Johnson, L., 514
- Johnson, P., 415–416
- Johnson, W., 495, 504, 514
- Johnston, D., 514
- Johnstone, I., 51, 73
- Jolliffe, I.T., 155, 170
- Jones, A., 172–173, 192–194, 198, 509, 514
- Jones, A.J., xiv
- Jones, E., 513
- Jones, M.C., 303, 324
- Jöreskog, K.G., 341–342, 361
- Jorgenson, D.W., 517–518, 540n1, 544
- Joshi, H., 494, 496, 513–515
- Juarez, M.A., 440, 449
- Judge, 405
- Judson, R.A., 460–462, 490
- Juhl, T., 471–472, 490
- Jun, S.J., 305, 321
- Jung, B.C., 151, 169, 397
- Jung, S., 305, 321
- Juodis, A., 77, 109, 463, 468, 487n4, 490
- Kai, B., 306, 321
- Kalton, G., 515
- Kamakura, Wagner, 561, 580
- Kang, S., 623, 640
- Kao, C., xii, xv, 47–54, 56–57, 59–61, 64–66, 71–74, 356, 361, 477, 488, 523, 541
- Kapetanios, G., 16, 20–21, 42, 44, 53–54, 64, 73, 461, 484, 490–491
- Kapoor, N.M., 363–366, 367t, 377, 381, 389, 392, 394, 399
- Kapteyn, A., 151, 170, 343, 350, 356, 361–362, 641
- Karlsson, A., 304, 321
- Kato, K., 307, 321
- Katz, E., 189, 199
- Katz, J.N., 474, 478, 488
- Kauppi, H., 49, 73
- Keane, M.P., xiv, 550–554, 556–557, 560–561, 564, 568, 570–577, 578n9, 579–581, 585, 605

- Kelejian, H.H., 364–366, 367*t*, 377, 381, 389, 392, 394–395, 399, 625–626, 640
- Keller, K., 568, 581
- Keller, W., 364, 399
- Kelly, E., 497, 500, 514
- Kendrick, J.W., 518, 544
- Kenjegalieva, K., 534, 543
- Kerman, S., 150, 169
- Ketende, S., 502, 514–515
- Khalaf, L., 31, 43, 480, 490
- Khan, S., 178, 197
- Kho, 366
- Kiefer, N., 126, 147
- Kiefer, N.M., 351, 360
- Kim, B., 348, 361
- Kim, J.H., 203, 232
- Kim, J.I., 544
- Kim, M.O., 318, 321
- Kim, M.S., 395, 399
- Kim, S., 520, 544
- Kimhi, A., 184, 200
- Kipnis, V., 542
- Kiviet, J.F., 77, 79, 81, 85–86, 93, 95–97, 108–109, 409, 416, 462–463, 487*n1*, 490, 597, 605
- Klaasen, F.J.G.M., 622, 637
- Klasen, S., 640
- Kleibergen, F., 69–71, 73, 82, 84, 86, 108–109, 468, 490
- Klein, L., 522, 544
- Klein, L.R., 414, 417
- Klein, R., 176, 199
- Klein, T.J., 258, 283
- Klette, J.T., 335, 337, 359, 361
- Klier, T., 195, 200
- Kline, B., 501, 514
- Kloek, T., 419, 438–440, 449, 461, 491
- Kluve, J., 279*n2*, 282
- Knaap, T., 82, 110
- Knapp, M., 195, 200
- Knapp, T., 130, 148
- Kneip, A., 151, 170, 530–532, 534, 545
- Kniesner, T.J., xiv, 584–585, 588, 588*t*, 591, 592*t*, 593, 594*t*, 597, 599–601, 602*nn2*–3, 603*nn5*–7, 604–605, 607
- Koch, W., 364, 398, 626, 635*n7*, 636
- Kockelman, K., 195, 200, 364, 398
- Koenker, R., 80, 109, 303, 306–309, 319*n1*, 321–322, 584, 591
- Koh, W., 42, 363, 365, 392, 397
- Kompas, T., 197
- Komunjer, I., 319*n1*, 322
- Koop, G., 186, 200, 540*n4*, 545, 547
- Koopmans, T.C., 570, 581
- Korajczyk, R., 72
- Korniotis, G.M., 385, 399
- Kortum, S.S., 609, 628–629, 638
- Kraay, A.C., 480, 489
- Krailo, M., 188, 200
- Krailo, M.D., 208, 232
- Krasker, W.S., 420, 428, 449
- Krätsig, M., 468, 491
- Krishnainiah, P.R., 45
- Krishnakumar, J., 269, 282
- Krishnan, T., 170
- Kristian, B., 636
- Krueger, A., 585, 593, 604
- Krueger, A.B., 267, 281, 348, 359
- Krugman, P., 545, 609–610, 640
- Kruiniger, H., 82–84, 109, 465, 491
- Kuersteiner, G., 82, 108, 189, 193, 199, 385, 399, 455, 463, 490, 593, 604
- Kuersteiner, G.M., 120–121, 129, 131, 138–140, 145*n4*, 146
- Kuh, D., 498*t*, 499, 514–516
- Kuh, E., 418, 420, 446*n16*, 447, 449
- Kumbhakar, S.C., 520, 524–525, 528–530, 534, 545–547
- Kutlu, L., 519, 545
- Kwiatkowski, 65
- Kydland, F., 605
- Kyriazidou, E., 178, 193, 199, 212, 217, 230–232
- Lahiri, K., 540*n5*, 542, 545
- Lai, H.P., 538, 544
- Lai, T.L., 441, 449
- Laisney, F., 186, 200
- LaLonde, R.J., 257, 283
- Lamarche, C., 307–308, 320, 322, 584, 591–592, 605
- Lambros, L.A., 540*n5*, 547
- Lancaster, K.J., 556, 581

- Lancaster, T., 85, 109, 145nn1–2, 147–148, 186, 188–189, 200, 242, 255, 462, 491
- Langenberg, C., 515
- Larch, M., 609–610, 617, 621, 629, 632–634, 635n9, 637–640
- Larsson, R., 66–71, 73
- Lassmann, A., 630, 638
- Lau, L.J., 544
- Lawlor, D., 512–514
- Lawlor, F., 516
- Leamer, E.E., 608, 640
- Lebreton, Marie, 625, 640
- Lechner, M., xiii, 182, 186, 196, 200, 262, 264–266, 268, 270, 274–275, 279nn2–3, 280n15, 280n19, 281n26, 281nn22–23, 282–284
- Lee, J., 300, 322
- Lee, J.W., 545
- Lee, L., 177, 183, 195–197, 200
- Lee, L.F., xiii, 4, 40, 44, 363–364, 366–367, 367t, 368, 373, 375–379, 381–383, 385–387, 389–391, 393–394, 399–401, 520, 524, 546
- Lee, M., xii, 41n9, 184, 200
- Lee, M.J., 202–203, 216–218, 232
- Lee, N., 44
- Lee, S., 303, 318, 322, 534, 545
- Lee, Y., 151, 168, 295, 322, 534, 545
- Lee, Y.H., 90, 107, 126, 144–145, 147, 484–485, 488, 531–533, 540, 544–545
- Leeb, H., 534, 540n4, 545
- Leeflang, P., 576, 582
- Leeth, J.D., 603nn6–7, 605
- Le Gallo, J., 363–364, 366, 367t, 368, 373, 397
- Lehrer, S.F., 281n22, 282
- Leili, R., 198
- Leipnik, R.B., 581
- Lenk, P.J., 561–562, 579
- Lerman, S., 553, 562, 581
- Lerner, J., 540n3
- Leroy, A.M., 418, 422, 444n1, 444n3, 450
- LeSage, J.P., 364, 366, 368, 373–374, 387, 398, 400, 625, 640
- Levchenko, A., 610, 640
- Levina, E., 291, 319
- Levinsohn, J., 560, 566, 579, 608, 640
- Lewbel, A., 339–340, 354, 359, 361, 411, 417
- Lewis, T., 420, 447
- Li, B., 290, 305, 323
- Li, D., 294, 300–302, 316–318, 320, 322–323, 364, 394, 397
- Li, K., 151, 169
- Li, Q., xiii, 285–286, 290, 293–295, 298, 305, 309, 314, 319, 321–322
- Li, R., 290–294, 306, 320–321, 323–324
- Liang, K.-Y., 239, 255–256
- Liang, P., 182, 198
- Liao, Y., xii
- Lillard, L., 583, 590, 605–606
- Lin, X., 288–290, 294, 322, 324
- Lin, Z., 294–295, 298, 309, 314, 322
- Lindeboom, M., 497, 507, 515
- Lindley, D.V., 417
- Lindsay, B.G., 290, 305, 323
- Linton, O.B., 294, 322
- Lippi, M., 4, 43
- Lipsitz, S.R., 442, 450
- Liska, R., 16, 43
- Little, J.D.C., 548, 552, 570, 572–575, 580
- Little, R.J.L., 150, 169
- Liu, J., 533–534, 545
- Liu, L., 50, 72
- Llano, C., 638
- Llena-Nozal, A., 515
- Lleras-Muney, A., 500, 509, 513
- Llewellyn, C., 514
- Lochner, L., 605
- Lopez, J.A., 540n5, 542
- Loretan, M., 49, 74
- Loretz, Simon, 639
- Löthgren, M., 66–67, 73
- Loubaton, P., 43
- Lovell, C., 519, 541
- Lovell, C.A.K., 520–521, 543, 545
- Lovell, C.K., 520–521, 525, 543, 546
- Lu, X., 295–297, 323
- Lu, Z., 303, 324
- Lucas, A., 419–420, 438–440, 449
- Lucas, R.E., Jr., 351, 362, 519, 545
- Lugovskyy, O., 471–472, 490
- Lütkepohl, H., 468, 491
- Lutzky, C., 579
- Lyhagen, J., 66–71, 73

- Macdonald-Wallis, C., 513  
 Mace, S., 576, 581  
 Macfarlane, P., 515  
 Machado, J.A.F., 80, 109  
 Machin, S., 502, 512  
 Macleod, J., 512–513  
 Macmillan, L., 512  
 MaCurdy, T.E., 487n5, 491, 583, 585, 587, 589, 596, 603, 605, 630, 635  
 Maddala, G.S., 58, 73, 188, 200, 364, 379–380, 400, 453, 491  
 Madigan, D., 544, 546  
 Madsen, E., 49, 74  
 Magnac, T., 208–209, 215, 232  
 Magnus, J.R., 337, 361, 379, 400  
 Mahia, R., 637  
 Mairesse, J., 470, 491, 519, 540n2, 544–545  
 Mallows, C.L., 428, 449, 540n4, 545  
 Mammen, E., 232, 294, 314, 320, 322  
 Mammi, I., 461, 489  
 Manning, A., 509, 515  
 Manrai, A.K., 556, 579  
 Mansfield, E., 518, 546  
 Manski, C., 176, 200, 553, 562, 581  
 Marcellino, M., 70, 71n3, 72, 461, 490  
 Mark, N.C., 47, 49, 51, 74  
 Maronna, R.A., 427, 436–437, 445n7, 449  
 Marschak, J., 552, 579  
 Marshall, E.C., 443, 449  
 Martin, H., 515  
 Martin, J., 495, 515  
 Martin, R., 513  
 Martin, R.D., 449  
 Martinez, M.L., 146  
 Martínez-Zarzoso, I., 622, 632, 640  
 Mastromarco, C., 534, 546  
 Mathiowetz, N., 348, 357, 360, 593, 604  
 Mátyás, L., xi, xv, 77, 109, 257, 284, 629, 640  
 Matzkin, R., 176, 183, 196, 200, 310, 319  
 Maughan, B., 513  
 Mayda, A., 622, 640  
 Mayer, T., 618, 627, 639–640  
 McAfee, R.P., 571, 580  
 McArdle, P., 515  
 McCoskey, S., 65–66, 74  
 McCrary, J., 264, 282  
 McCullagh, P., 615, 640  
 McCulloch, R., 560, 582  
 McDonald, J., 514  
 McFadden, D., 171, 174, 176, 185, 200, 550, 552–553, 557, 560, 563, 580–581  
 McKenzie, C., 183–184, 198  
 McKenzie, D., 343, 359, 496, 515  
 McLachlan, G., 170  
 McMillen, D., 195, 200  
 MCS Team, 514  
 Meeusen, W., 519, 546  
 Meijer, E., xiii, 327, 334–340, 343–347, 350, 354, 356, 361–362  
 Meinecke, J., 186, 199  
 Melitz, Marc, 610, 618, 640  
 Mennes, L.B.M., 461, 491  
 Merckens, A., 334, 359  
 Merlo, V., 622, 630, 638  
 Mesnard, A., 513  
 Mestre, R., 59, 72  
 Michael, R., 513  
 Midthune, D., 542  
 Miller, D.L., 627, 637  
 Millimet, D.L., 622, 640  
 Million, A., 172, 184, 201  
 Min, C.K., 414, 417  
 Miquel, R., 262, 270, 274–275, 279n2, 281n21–281n22, 283–284  
 Mishra, G., 515  
 Mizon, G.E., 447, 596, 604, 630, 637  
 Modugno, M., 154, 169  
 Moffitt, R., 180, 184, 197, 359, 361, 554–555, 562, 579, 585, 604  
 Mohnen, P., 641  
 Molloy, L., 512–513  
 Monfort, A., 43  
 Montes-Rojas, G.V., 307–308, 320–321  
 Mooijaart, A., 339, 362  
 Moon, H.R., xii, 21–22, 44, 47–48, 51, 74, 120, 125–129, 131–137, 145n7, 146–148, 151–152, 170, 192–193, 199, 318, 322, 483, 491  
 Moran, P.A.P., 4, 28, 44, 391–392, 400  
 Moreira, M.J., 109  
 Moreno, E., 112, 146–147  
 Moscone, F., 41n16, 44, 195, 200, 476, 487n10, 491

- Moulin, H., 535, 546  
Mousavi, S.E., 581  
Mu, Y., 309, 322  
Mukherjee, A., 319n1, 322  
Mukhopadhyay, N., 112, 143, 146  
Mullahy, J., 175, 200  
Müller, H., 150, 170  
Mundlak, Y., 187, 191–192, 200, 241, 255,  
    473, 491, 528, 546, 619–620, 623, 640  
Musolesi, A, 534, 543  
Mutl, J., 364, 381, 389, 400
- Na, S., 50, 72  
Nadiri, M.I., 518, 520–521, 527, 543–544  
Nagata, S., 77, 93, 96–97, 109  
Najim, J., 43  
Nathan, G., 515  
Nauges, C., 83, 108  
Naughton, H.T., 637  
Navarro-Lozano, S., 258, 283  
Nayyar, Ashish, 571, 580  
Nekipelov, D., 354, 360  
Nelder, J.A., 615, 640  
Nelson, D.R., 617, 628, 633, 638  
Nelson, S., 513  
Neocleous, T., 310, 322  
Nerlove, M., 196, 466, 491  
Neslin, S., 575–577, 579, 581–582  
Ness, A., 512–513  
Nessen, M., 73  
Neuburger, J., 499, 515  
Neudecker, H., 337, 361  
Nevo, A., 575, 580  
Newbold, P., 540n5, 546  
Newey, W., 56, 74, 78, 84, 109, 126, 138,  
    145n6, 145n9, 146–147, 189–190, 197,  
    199, 247, 255, 263, 283, 320, 354, 360,  
    467, 480, 487n8, 490–491, 597–598,  
    605–606  
Neyman, J., 85, 109, 111, 147, 188, 200, 376,  
    400  
Ng, P., 303, 322  
Ng, S., 16, 32, 42, 44, 53, 61–62, 64, 72, 169,  
    461, 474, 477, 488, 491, 523, 531, 541  
Ng, T., 486, 490  
Nguyen, H., 197  
Nickell, S., 85, 109, 112, 114–115, 147
- Nielsen, J.P., 294, 322  
Nigai, S., 632, 639  
Nigro, V., 212, 217–220, 231  
Nijman, T., 192, 201, 585, 606  
Nilsson, L., 633, 640  
Norris, M., 543  
Nowak-Lehmann, F., 622, 632, 640
- Oberhofer, W., 319n1, 322  
O'Connell, P.G.J., 4, 44  
O'Donnell, C.J., 542  
Ogaki, M., 51, 74  
O'Hagan, A., 443, 449  
Okawa, Y., 610, 640  
Olive, D.J., 444n1, 445n8, 450  
Olivero, M.P., 622, 641  
Olley, G.S., 519, 546  
Olsen, M.K., 248, 255  
Onatski, A., 16, 44  
Ord, J.K., 4, 43, 392, 398  
Orea, L., 528, 534, 546  
Orefice, G., 630, 641  
Orme, C.D., 487n10, 490  
Osang, T., 622, 640  
Osbat, C., 70, 71n3, 72  
Osiewalski, J., 186, 200, 547  
Osikuminu, A., 281n21, 282  
Osmond, C., 503, 514–515  
Ouliaris, S., 55, 58, 61, 74  
Owen, A.L., 460–462, 490
- Paap, R., 174, 201, 570, 582  
Pace, K., 366, 373, 400  
Pace, R.K., 625, 640  
Pagan, A., 183, 197  
Pagan, A.R., 28, 30–31, 33, 42  
Pagan, M., 429, 448  
Pakes, A., 519, 544, 546, 560, 566, 579  
Pal, M., 339, 361  
Palm, F.C., 60–61, 73, 641  
Palmgren, J., 242, 255  
Palta, M., 337, 362  
Parent, O., 174, 195, 197, 364, 368, 387, 400  
Park, B., 319n1, 321  
Park, B.U., 527–528, 546  
Parkinson, K., 506, 515  
Parks, R., 487n12, 491  
Parmeter, C., 534, 545

- Parslow, R., 516  
 Parsons, S., 495, 515  
 Partridge, M., 395, 399  
 Paternoster, R., 231  
 Pattie, A., 513  
 Paul, D., 150, 154, 170  
 Paul, M., 281n21, 282  
 Pauwels, K., 576, 582  
 Paxson, C., 500, 508, 513  
 Paxson, C.H., 600, 603  
 Pearce, M., 515  
 Pedroni, P., 47, 49, 57–61, 63, 71n2, 74  
 Peixe, F.P.M., 487n3, 490  
 Pellerin, D., 515  
 Pencavel, J., 585, 606  
 Peng, H., 540n5, 545  
 Peng, J., 150, 154, 170  
 Perelman, S., 521, 542  
 Perktold, J., 231  
 Perron, B., xii, 51, 74, 134–137, 147  
 Perron, P., 58, 66, 72, 74  
 Peruggia, M., 443, 450  
 Pesaran, M.H., xi, xii, xv, 7–8, 14–15, 17–22,  
     24–27, 30–32, 34, 41n13, 41n15,  
     41nn5–7, 42–44, 46, 53–54, 58, 64, 67,  
     70–74, 81, 85, 107, 109, 134, 145n5,  
     146, 148, 151, 170, 298–299, 317, 322,  
     364, 374–375, 378, 393, 398–400, 407,  
     409, 411, 413, 416–417, 440, 449, 465,  
     468, 470–471, 473–474, 476–477,  
     483–486, 487n6, 487n9, 488–492, 534,  
     546  
 Pesendorfer, M., 571, 582  
 Petherick, E., 516  
 Pfaffermayr, M., xv, 232, 364–365, 367, 367t,  
     381, 392–393, 397, 400, 455, 490, 613,  
     620–621, 624–626, 629–632, 634n1,  
     635n11, 636, 639  
 Phillips, G.D.A., 409, 416  
 Phillips, P.C.B., xii, 3, 45, 47–49, 52, 55, 58,  
     61, 64–65, 74, 83, 85, 108, 112,  
     119–120, 126, 130–137, 143, 145n3,  
     145n6, 145n8, 146–148, 308, 318, 322,  
     465, 490, 492  
 Pick, A., 43  
 Pickett, K., 516  
 Pickles, A., 180, 183, 201  
 Pierce, M., 515  
 Piesse, A., 515  
 Pigorsch, U., 16, 42  
 Pike, M., 188, 200  
 Pike, M.C., 208, 232  
 Pinske, J., 195, 200, 305, 321  
 Piorier, D.J., 545  
 Pirotte, A., 393–395, 397, 455, 492  
 Pischke, J.-S., 257–258, 265–266, 281  
 Pischke, S., 509, 515  
 Pitt, M., 520, 524, 546  
 Pleus, M., 79, 109  
 Plewis, I., 500–502, 515  
 Ploberger, W., 64, 136, 148  
 Plümper, T., 186, 191, 201  
 Poldermans, R., 79, 109  
 Porajan, A., 626, 641  
 Porteous, D., 513  
 Portnoy, S., 303, 310, 322  
 Pötscher, B.M., 534, 540n4, 545  
 Powell, J.L., 309, 323, 354, 360, 430, 448  
 Power, C., 500–501, 515  
 Pöyhönen, P., 629, 641  
 Prabhakar Rao, R., 335, 360  
 Preisser, J.S., 442, 450  
 Propper, C., 505, 507, 514–515  
 Prucha, I.R., 364–366, 367t, 377, 381, 389,  
     392, 394–395, 399, 625–626, 640  
 Pudney, S., 184, 201  
 Pulugurta, V., 172, 196  
 Purdon, S., 513  
 Pyke, 32, 45  
 Qaqish, B., 239, 255  
 Qaquish, B.F., 442, 450  
 Qian, J., 294, 323, 534, 541  
 Qu, A., 290–291, 305, 323  
 Quah, D., 415–416  
 Raban, 568  
 Rabe-Hesketh, S., 180, 183, 201, 255  
 Racine, J., 286, 322  
 Raftery, A.E., 540n4, 544, 546  
 Rao, D.P., 542  
 Rapoport, H., 634n2, 638  
 Rasch, G., 190, 201, 208, 232  
 Rathbun, S., 195, 201  
 Rattsø, J., 519, 542

- Raudenbush, S., 191, 197  
Raymond, W., 620, 641  
Raynor, P., 516  
Reed, H., 512  
Reed, W.R., 492  
Reich, B.J., 310, 319  
Reichlin, L., 43  
Reiersøl, O., 334, 361  
Reifschneider D., 527, 546  
Reikard, G., 546  
Reilly, J., 515  
Relton, C., 515  
Renault, E., 43  
Rhodes, E., 519, 542  
Rice, J.A., 150, 170  
Rice, N., 172–173, 192–194, 198, 514  
Richard, J.F., 490  
Richards, M., 515–516  
Ridder, G., 263, 283, 358–359, 361  
Ries, John, 627, 639–640  
Rigg, J., 515  
Rincke, J., 375, 400  
Ring, S., 512–513  
Riphahn, R., 172, 184, 201  
Robb, R., 583, 605  
Roberts, 568  
Robertson, D., 3, 34, 45, 90, 110, 318, 323, 393, 400, 478, 492  
Robins, J.M., 270, 281n21, 282, 284  
Robinson, P., 299–300, 302, 322–323  
Robinson, P.M., 395–396, 400  
Rochina-Barrachina, M.E., 619, 638  
Roeller, L.H., 527, 543  
Roia, Laïsa, 625, 640  
Roling, C., 471–472, 489  
Romer, P.M., 518–519, 546  
Ronchetti, E., 432, 441–442, 444n1, 447–450  
Ronning, G., 353, 361  
Roodman, D., 77, 81, 85, 87, 110, 461, 492  
Rosa Dias, P., 514  
Rose, J., 185, 199  
Rosen, A., 176, 197, 307, 323  
Rosen, H., 126, 147  
Rosen, H.S., 78, 109, 375, 398, 467, 487n8, 490, 598, 605  
Rosen, S., 584, 606  
Rosenbaum, P.R., 264, 284  
Ross, E., 500, 512  
Rossi, P., 173, 196, 560, 570, 580, 582  
Rothenberg, T., 135, 146  
Rousseeuw, P.J., 418, 422, 425–426, 426t, 427, 444n1, 444n3, 445n8, 448, 450  
Roy, N., 288, 290, 323  
Rubin, D.B., 259, 263–264, 284, 358, 361  
Rubin, H., 581  
Rubinstein, Y., 610, 618, 640  
Ruckstuhl, A.F., 289, 323  
Runkle, D., 557, 580, 585, 605  
Rupert, P., 438, 447  
Russell, G., 561, 580  
Russell, R.R., 537, 541  
Rust, J., 351, 361  
Ruud, P.A., 400, 557, 580  
  
Sacerdote, B., 500, 515  
Sachs, J.D., 546  
Said, E., 41n13, 45  
Saikkonen, P., 49, 51, 67, 74  
Salibian-Barrera, M., 427, 445n7, 450  
Salish, N., 471–472, 489  
Salomone, S., 634n2, 638  
Santner, T., 443, 450  
Santos Silva, J.M.C., 614–615, 617, 621, 641  
Sarafidis, V.T., xii, 3, 34, 45, 90, 110, 161, 170, 316, 318, 323, 393, 400, 478, 492  
Sargan, J., 34, 45  
Sargan, J.D., 86, 93, 107, 110, 388, 398, 458, 466, 488, 492  
Sargent, T.J., 4, 45  
Sauder, U., 493, 510, 512  
Sauer, R., 564, 581  
Schafer, J.L., 170, 248, 255  
Schennach, S.M., 547  
Scherer, F.M., 518, 546  
Schim van er Loeff, S., 641  
Schmidt, P., 65, 78–79, 86, 90, 97, 107, 118, 126, 145, 150–151, 168, 170, 439, 446–447, 458, 461, 484–485, 487–489, 519–521, 523, 525, 527–528, 531–533, 540–542, 545–547, 585, 596, 604, 630, 637  
Schmidt, S.S., 520, 543  
Schmidt, T.D., 43  
Schneeweiss, H., 353, 361

- Schnell, J.F., 356, 361  
 Schnier, K., 195, 198  
 Schoon, I., 495, 515  
 Schott, J.R., 31, 33, 45  
 Schuermann, T., 485, 491  
 Schulman, C., 150, 170  
 Schurer, S., 514  
 Schwarz, G., 413, 417, 540n4, 547  
 Scott, E., 111, 147, 188, 200  
 Scott, E.L., 85, 109, 376, 400  
 Sedlacek, G., 605  
 Segev, Eran, 637  
 Semykina, A., 173, 193, 201, 619–620, 622, 641  
 Sener, I., 195, 197  
 Serlenga, L., 641  
 Sevestre, P., xi, xv, 77, 109, 257, 284  
 Sevilla-Sanz, A., 505, 514  
 Shang, C., xiv, 547  
 Shao, J., 337, 362  
 Shaw, J., 513  
 Shaw, K., 606  
 Shea, J., 597, 606  
 Shen, Y., 410, 410t, 411, 411t, 412, 416  
 Sheng, X., 540nn5–6, 545  
 Shepherd, P., 501, 513  
 Shepherd, W.G., 519, 547  
 Shields, M., 184, 201, 513–514  
 Shih, M.C., 441, 449  
 Shin, D.W., 45, 58, 64–65, 67, 70  
 Shin, Y., 73, 486, 491–492, 534, 546, 641  
 Shiue, C.H., 364, 399  
 Shum, M., 568, 580  
 Sickles, R.C., xiv, 151, 170, 518–521, 523, 527–528, 530–534, 540, 540n6, 541–547  
 Siddarth, S., 576, 582  
 Silverstein, J.W., 5, 42  
 Simar, L., 521, 527–528, 546–547  
 Simmonds, S., 515  
 Simonoff, J.S., 448  
 Sims, C.A., 4, 45  
 Singer, B., 183, 199  
 Singh, A., 450  
 Singleton, K.J., 351, 360  
 Sinha, S.K., 441–442, 450  
 Skjerpen, T., 342, 359  
 Skrondal, A., 180, 183, 201, 255  
 Slade, M., 195, 200  
 Small, N., 516  
 Smirnov, A., 195, 201  
 Smith, A.F.M., 417  
 Smith, J.A., 257, 283  
 Smith, L.V., 489  
 Smith, M.D., 570, 582  
 Smith, R., 18, 43–44, 74, 409, 417, 470, 473, 487n9, 492  
 Smith, T., 564, 581  
 Smith, W.B., 150, 170  
 Smolinsky, K., 176, 198  
 Snow, J., 264–265, 284  
 So, B.S., 45  
 Solon, G., 348, 361, 606  
 Solow, R.M., 518, 547  
 Song, M., 23, 45  
 Song, S., 42, 315, 321  
 Song, S.H., 151, 169, 363, 365, 392, 397, 585, 604  
 Song, W., 151, 170, 530–532, 545  
 Sotelo, Sebastian, 638  
 Spady, R., 176, 199  
 Spiegelhalter, D.J., 443, 449  
 Spierdijk, L., xiii, 335, 337–340, 343–347, 361  
 Srinivasan, K., 576–577, 582  
 Srinivasan, T.C., 582  
 Staiger, D., 82, 110  
 Stalel, W.A., 448  
 Stander, J., 303, 324  
 Star, J., 513  
 Starr, J., 513  
 Staub, K.E., 610, 617, 621, 632, 634, 635n9, 639  
 Steel, M., 186, 200  
 Steel, M.F., 547  
 Steel, M.K.F., 440, 449  
 Stegun, I.A., 577n2, 579  
 Steinbuks, J., 534, 546  
 Stephen, A., 515  
 Stern, Steven, 563, 579, 582  
 Stevenson R., 527, 546  
 Stewart, M., 255  
 Stigler, G.J., 548, 582  
 Stiroh, K., 547  
 Stock, J., 135, 146, 597, 606

- Stock, J.H., 16, 45, 49, 64, 75, 82, 84, 110, 150, 154, 159, 170, 474, 492, 641
- Stoker, T.M., 412, 417, 519, 547
- Stokey, N.L., 351, 362
- Stokke, H.E., 519, 542
- Stone, M., 112, 143, 148
- Støve, B., 294, 322
- Street, A., 198
- Stroud, A.H., 577n2, 582
- Sturgis, P., 501, 515
- Su, L., 285, 289, 294–299, 302, 310–311, 315–318, 321, 323, 364, 385–389, 396, 400
- Sugar, C.A., 150, 170
- Sukumar, N., 581
- Sul, D., 3, 45, 47, 49, 51, 74, 85, 108, 112, 119, 126, 130–131, 143, 145n3, 145n8, 147–148, 308, 322, 465, 484, 490, 492
- Sullivan, L., 501, 515
- Summers, R., 547
- Sun, B., 575–577, 580, 582
- Sun, B.H., 413–414, 416
- Sun, Y., xiii, 294–295, 298, 303, 305–306, 309, 314–315, 318, 322–323, 395, 399
- Sutradhar, B.C., 335, 360, 441, 446n14, 447, 450
- Swamy, P.A.V.B., 405, 413, 417, 470, 492, 591, 606
- Syddall, H., 503, 515
- Sylva, K., 513
- Tacmisioglu, A.K., 85, 109
- Tae, Y.H., 216, 232
- Taglioni, D., 635n10, 636
- Tahmisioglu, A., 378, 399, 407, 409, 416, 440, 449, 465, 487n6, 490
- Tamer, E., 358, 360
- Tamura, A., 634n1, 638
- Taylor, G., 334, 360
- Taylor, M., 513
- Taylor, W., 583, 585, 595–596, 604
- Taylor, W.E., xiv, 419, 434–436, 448, 528, 544, 623–625, 630, 639
- Teigland, C., 540n5, 545
- Temple, J., 90, 108
- Tenhofen, J., 160, 169, 483–484, 489
- Tenreyro, S., 614–615, 617, 621, 641
- Thaler, R., 584, 606
- Theil, H., 412, 417
- Thomas, A., 208–209, 232
- Tilling, K., 513
- Timmermann, A., 540nn4–5, 547
- Tjøstheim, D., 294, 322
- Tobias, J.L., 501, 514, 545
- Todd, P., 605
- Tollison, R.D., 541
- Tosetti, E., 8, 18, 20, 41n16, 43–44, 195, 200, 374–375, 393, 398, 400, 474, 476, 487n10, 489, 491
- Tosteson, T., 335, 360
- Traferri, A., 189, 191, 197
- Train, K., 176, 180, 183, 185, 200–201, 560, 581–582
- Trapani, L., 54, 73
- Trawinski, I.M., 150, 170
- Trivedi, P.K., xiii, 171, 197, 243, 249–255, 336, 360, 412, 417, 617–618, 635nn4–5, 637
- Troeger, V., 186, 191, 201
- Trognon, A., 43
- Trojani, F., 432, 450
- Troxel, A.B., 442, 450
- Truong, Y.K., 303, 305, 320
- Tsionas, E.G., 528–530, 533–534, 545, 547
- Tuffnell, D., 516
- Tukey, J.W., 420, 444n1, 450
- Tullock, G., 535, 541, 547
- Ullah, A., 44, 285, 288–290, 294, 310–311, 321, 323, 393, 400, 477, 492
- Urbain, J.P., 21, 45, 47, 60–61, 63–64, 71, 73, 75, 483–484, 492
- Urban, 568
- Urga, G., 54, 73
- Urzua, S., 513
- U.S. Department of Transportation, 606
- U.S. Office of Management and Budget, 602, 606
- Valletta, R.G., 358, 360
- van den Berg, G.J., 258, 281
- van den Broeck, J., 519, 534, 546–547
- van der Klaauw, B., 515
- van Dijk, R., 174, 201, 419, 438–440, 449
- van Driesssen, K., 445n8, 450

- van Heerde, H., 576, 582  
 van Jaarsveld, C., 506, 514  
 van Montfort, K., 339, 362  
 van Reeden, J., 520, 541  
 van Soest, A., 198  
 van Wincoop, Eric, 609–610, 629, 636, 640  
 Verardi, V., 418, 425, 429, 445n6–445n7, 448, 450  
 Verbeek, M., 192, 201, 496, 516, 585, 606  
 Vicarelli, Claudio, 633, 637  
 Vignoles, A., 500, 510, 512, 514  
 Vijverberg, W., 195–196  
 Viscusi, W.K., 584, 587–588, 591, 592<sub>t</sub>, 600–601, 602n3, 603, 605–606  
 Visscher, P., 513  
 Vogelsang, T.J., 395, 400, 480, 492  
 Volgushev, S., 310, 315, 320  
 Volinsky, C.T., 544  
 von Hinke Kessler Scholder, S., xiv, 505, 514  
 Vuong, Q., 319n1, 322  
 Vytlacil, E., 268, 280n5, 283–284
- Waddell, Glen R., 637  
 Wadsworth, M., 499, 513–516  
 Wagener, J., 315, 320  
 Wagenvoort, J.L.M., 429, 448  
 Wagenvoort, R., 419, 429, 432–434, 436, 438, 450  
 Wagner, J., 429, 450  
 Wagner, M., 60, 70, 75  
 Waiblinger, D., 516  
 Waldfogel, J., 502, 516  
 Waldmann, R., 419, 429, 432–434, 436, 438, 450  
 Wambach, A., 172, 184, 201  
 Wamser, G., 620, 630–632, 635n11, 639, 641  
 Wang, C., 195, 200  
 Wang, H., 306, 310, 319, 323  
 Wang, J., 150, 170  
 Wang, L., 294, 323  
 Wang, N., 288, 323  
 Wang, W., 393, 401  
 Wang, Y., 304–305, 320  
 Wansbeek, T.J., xiii, 45, 82, 90, 110, 130, 148, 151, 170, 316, 323, 327, 334–340, 343–347, 350, 354, 356, 359, 361–362, 492, 641  
 Ward, M., 197  
 Wardle, J., 514  
 Warner, A., 546  
 Washbrook, E., 502, 514, 516  
 Wasi, N., 556, 561, 581  
 Waterman, R.P., 242, 254  
 Watson, M.W., 16, 45, 49, 64, 75, 150, 154, 159, 170, 474, 492, 641  
 Webb, R., 492  
 Weber, A., 232, 279n2, 282  
 Wei, Y., 309, 322  
 Weidner, M., 21–22, 44, 126–129, 145n7, 147, 151–152, 170, 483, 491  
 Weiner, S.M., 485, 491  
 Weiss, A.A., 319n1, 323  
 Weiss, Y., 583, 590, 605–606  
 Welsch, R.E., 418, 420, 428, 446n16, 447, 449  
 Welsh, A., 288–289, 323–324  
 West, J., 516  
 West, K.D., 56, 74, 84, 109, 480, 491, 597, 606  
 Westerlund, J., 21, 45, 47, 51–53, 55, 59–60, 62–66, 71n1, 75, 483–484, 492  
 Whalley, L., 513  
 White, H., 615, 641  
 Whited, T.M., 339, 360  
 Whittle, P., 4, 45  
 Wiehler, S., 281n22, 283  
 Wildasin., D., 364, 401  
 Wilson, P.W., 521, 547  
 Wilson, D., 512  
 Wilson, V., 513  
 Windmeijer, F., 77, 80–83, 86, 93, 96–98, 108, 110, 145n6, 147, 242, 244, 246–247, 254–255, 460, 489, 514, 635n5, 637  
 Winer, R.S., 582  
 Winkelmann, R., 240, 254, 446n12, 447, 639  
 Winner, Hannes, 639  
 Winter, P., 514–515  
 Wise, D., 583, 604  
 Wittink, D., 576, 582  
 Wolpin, K., 564, 581  
 Wong, A., 515  
 Wong, W.H., 377, 401  
 Woock, C., 605  
 Wooldridge, J.M., xi, xv, 173, 175, 177–178, 190–194, 201, 203, 232, 242, 245–246, 256, 264–265, 267, 283, 327, 362, 442,

- 450, 565, 582, 585, 589, 600–601, 606, 612, 617, 619–620, 622, 641  
World Health Organization, 192, 201  
Wright, C., 515  
Wright, J., 506, 516, 606  
Wright, J.H., 82, 84, 110  
Wu, C., 150, 170  
Wu, D., 192, 201  
Wu, G.L., 534, 543  
Wu, H., 318, 324  
Wu, P., 496, 512  
Wu, S., 58, 73  
Wunsch, C., 262, 264, 279n2, 281n26, 283–284  
Wyhowski, D., 585, 596, 604, 623, 630, 637, 641  
  
Xiao, H., 581  
Xiao, Z., 318–319, 337, 362  
Xu, R., 337, 362  
Xu, X., 318–319  
  
Yamagata, T., 34, 41n5, 43–45, 53–54, 64, 73, 161, 170, 318, 323, 364, 374–375, 387, 393, 399–400, 413, 417, 471, 477–478, 487n10, 490, 492  
Yang, J., xii  
Yang, Z., 364, 385–386, 388–389, 393, 397, 400  
Yao, F., 150, 170  
Yao, Q., 290, 320  
Yao, W., 291–293, 324  
Yin, Y.Q., 5, 45  
  
Yogo, M., 82, 110, 606  
Yohai, V.J., 425–426, 426*t*, 427, 431, 436–437, 445n7, 448–449  
Yotov, Y.V., 622, 636, 641  
Young, A., 547  
Yu, J., xiii, 4, 40, 44, 120, 146, 177, 195, 200, 363–364, 366–367, 367*t*, 368, 373, 375–379, 381–383, 385, 387, 389–391, 393–394, 399–401  
Yu, K., 232, 303, 324  
Yuan, M., 306, 324  
  
Zaporowski, M., 545  
Zaprovska, 540n5  
Zarnowitz, V., 540n5, 547  
Zeger, S., 182, 198  
Zeger, S.L., 239, 255–256  
Zellner, A., 4, 45, 414, 417  
Zhang, J., 318, 324, 610, 640  
Zhang, Y., 302, 317–318, 323  
Zhang, Y.Y., xiii  
Zhang, Z., 543  
Zhao, Q., 319n1, 324  
Zhao, Y., 175, 199, 545  
Zhao, Z., 409, 417  
Zheng, J., 303, 314–315, 324  
Zhou, J., 306, 323  
Zhu, H., 442–443, 450  
Zhu, Y., 440, 448  
Zhu, Z., 306, 321, 323  
Ziliak, J.P., xiv, 81, 110, 584–585, 591, 592*t*, 597, 600–601, 605, 607  
Zou, H., 291, 306, 321, 324

## S U B J E C T I N D E X

Page numbers followed by *f* or *t* indicate figures or tables, respectively. Numbers followed by n indicate endnotes.

- additive individual effects and time effects, 125–126
- ADF* (augmented Dickey-Fuller statistic), 57
- advertising, 575
- affine equivariance, 444n3
- aggregate analysis, 410–412
- aggregation, 411
- agricultural economics, 364
- Ahn, Lee, and Schmidt model, 532–533
- Akaike Information Criterion (AIC), 143
- $\alpha$ -trimmed mean, 444n2
- ALSPAC (Avon Longitudinal Study of Parents and Children) (“Children of the 90s”), 498t, 505, 507
- alternative asymptotics, 122–124
- alternative bias correction method, 124
- alternative procedures, 84–85
- alternatives, irrelevant, 553
- Anderson-Hsiao estimator, 344–345, 345f
- applications, 453–641
- approximate factor model, 9
- area birth cohorts, 495–497
- Arellano and Bond IV matrix, 439
- Arellano-Bond estimator, 460
- AR(1)* errors, 562
- arithmetic weighted average, 518
- ARMA process, 487n5
- AR(1)* panel models, 446n13
- AR(1)* specifications, 552, 561
- ASF (average structural function), 311
- AS (nonlinear) GMM estimators, 98–99, 105
- Asian countries, 536, 537, 538–539
- asymptotics
  - alternative, 122–124
  - joint, 127–129
  - asymptotic standard errors, 80
  - attrition, 173, 192
  - augmented Dickey-Fuller statistic (*ADF*), 57
  - augmented regression model, 113
  - autocorrelation, 392, 393, 394, 395
  - autoregressive models, 113
  - integer-valued autoregressive model of order 1 (INAR(1)), 244
  - linear models with spatial autoregression, 195
  - average marginal effect (average treatment effect), 145n9
  - average regression coefficient, long-run, 47
  - average structural function (ASF), 311
  - average treatment effect (ATE), 145n9, 260, 311
  - dynamic (DATE), 271, 276
  - local (LATE), 260
  - for non-treated (ATENT), 260
  - for treated (ATET), 260
- Avon Longitudinal Study of Parents and Children (ALSPAC) (“Children of the 90s”), 498t, 505, 507
- bad leverage points, 418, 419f
- Battese and Coelli model, 526–527
- Bayes estimator, 409–410
- Bayesian approach, 405–407
- Bayesian Information Criterion (BIC), 112, 143, 538–539
- Bayesian Stochastic Frontier Model, 533–534
- BCS (Birth Cohort Study) (UK), 498t, 501, 508, 510–511
- Berkson model, 350

- best linear unbiased predictor (BLUP), 394  
 $\beta$   
 fully modified OLS (FM-OLS) estimator of, 48  
 Within-OLS estimator of, 48–49  
 between equation, 379–380  
 BHPS (British Household Panel Survey), 172, 173, 192  
 bias, 332, 332f  
 of coefficient estimators, 98–99  
 within estimator, 332, 332f  
 of first differences and longest differences, 332, 332f  
 of fixed effect estimator, 124  
 Nickell bias, 85, 115–118, 455  
 bias-corrected estimators, 121, 409, 462–465  
 bias correction, 385  
 alternative method, 124  
 direct, 189  
 bias reduction  
 for binary choice models, 189–190  
 definition of, 142  
 bias stability assumptions, 264  
 BIC (Bayesian Information Criterion), 112, 143, 538–539  
 BI (bounded influence) estimators, 428–429  
 binary choice models, 175–192  
 binomial AR(1) panel model, 446n13  
 birth cohorts, 494, 495–497, 497–502, 498t, 503–506  
 Birth Cohort Study (BCS) (UK), 498t, 501, 508, 510–511  
 biweight function (Tukey bisquare weight function), 422, 423t, 424f  
 block missing, 152, 153f  
 BLUP (best linear unbiased predictor), 394  
 body mass index (BMI), 501  
 bootstrap method, 66, 366  
 Born-in-Bradford cohort, 498t, 505–506  
 Bounded Inefficiency Model, 534  
 bounded influence (BI) estimators, 428–429  
 Box-Cox transformed dependent variables, 309  
 Boyd Orr Cohort, 498t, 504, 508–509  
 Bradford Royal Infirmary, 505–506  
 brand choice models, 550–553  
 brand loyalty, 549, 572, 575  
 brand loyalty variable ( $GL_{ijt}$ ), 572, 575  
 brand switching, 576  
 breakdown points (BDPs), 420–425  
 British Household Panel Survey (BHPS), 172, 173, 192  
 Burnside, Ethel Margaret, 503  
 Carnegie United Kingdom Trust, 504  
 CCE (Common Correlated Effects)  
 estimation, 17  
 advantages of, 20  
 extension to non-/semi-parametric panel data models with large n and large t, 298  
 CCE (Common Correlated Effects) estimators, 17–21  
 application to unbalanced panels, 38–40  
 dynamic, 24–28  
 CCE Mean Group (CCEMG) estimators, 18–19, 23, 27–28, 29t  
 CCE Pooled (CCEP) estimators, 18–19, 27–28, 29t  
 CCR (conditionally correlated random) effects model, 241, 245  
 $CD_{LM}$  test, 31, 33  
 $CD_P$  test, 32, 34, 35, 36t, 38, 38t  
 CD tests, 38–40  
 Census of Fatal Occupational Industries, 602n2  
 Chamberlain estimator, 190  
 “Children of the 90s” (Avon Longitudinal Study of Parents and Children, ALSPAC), 498t, 505, 507  
 China, 537  
 choice, 552. *see also* T (number of choice situations or time series observations)  
 binary, 175–192  
 discrete choice models, 171–201, 548–582  
 lagged, 552, 569  
 multinomial, 174  
 ordered, 174, 190–191  
 choice probability, 554, 556  
 Cholesky decomposition, 287, 288  
 CIA (conditional independence assumption), 261–264, 280n16, 280n18  
 weak dynamic (W-DCIA), 273–274  
 cluster correction, 175–176

- clustering, 176
- Cobb-Douglas distance function, 522
- coefficient estimators, 98–99. *see also specific estimators*
- cohort data, 493–516
- cohort studies, 493–494
- area birth cohorts, 495–497
  - birth cohorts, 494, 495–497
  - econometric methods applied to, 506–511
  - limitations to, 496
- cointegration, 46–75
- combining forecasts, 534–535
- common cointegrating vectors, 50
- Common Correlated Effects (CCE)
- estimation, 17
  - advantages of, 20
  - extension to non-/semi-parametric panel data models with large n and large t, 298
- Common Correlated Effects (CCE) estimators, 17–21
- application to unbalanced panels, 38–40
- CCE Mean Group (CCEMG) estimators, 18–19, 23, 27–28, 29 $t$
- CCE Pooled (CCEP) estimators, 18–19, 27–28, 29 $t$
- dynamic, 24–28
- common culture, 630
- common factor models, 8–14
- dynamic model with smooth factors, 163–168, 166 $t$
  - dynamic model with stochastic factors, 163–168, 167 $t$
  - static model with smooth factors, 163–168, 164 $t$
  - static model with stochastic factors, 163–168, 165 $t$
- common factors, 482–485
- assumptions typically made regarding, 8–9
  - deterministic, smooth factors, 162, 163 $f$
  - generation procedures, 162, 163 $f$
  - non-smooth factors, 162, 163 $f$
  - observed ( $d_t$ ), 41n8
  - semi-strong, 12, 13–14
  - smooth factors, 162, 163–168, 163 $f$ , 164 $t$
  - stochastic, 163–168, 165 $t$
  - stochastic, non-smooth factors, 162, 163 $f$
- strong, 10–11, 12–13, 14
  - unobserved, 21
  - weak, 10–11, 12–13, 13–14
- common-trend assumptions, 264, 265
- composite quantile regression (CQR) method, 306
- computational issues, 553–557
- concentrated likelihood approach, 138–141
- conditional independence, 177, 178
- conditional independence assumption (CIA), 261–264, 280n16, 280n18
- weak dynamic (W-DCIA), 273–274
- conditional log likelihood function, 139
- conditionally correlated random (CCR) effects model, 241, 245
- conditional mean regression models, 286–302
- conditional quantile regression models, 303–310
- consistent estimation, 335–341
- AH method, 344–345, 345 $f$
  - constant-correlated effects, 86–90, 93
  - constant correlation, 79
  - consumer demand, 548–582
  - dynamics of, 549, 576
  - price elasticity of, 576, 578n8
  - sources of dynamics in, 572–576
  - state dependence in, 575
  - consumer demand models, 576–577
  - static, 577
  - typical structure of, 550–553
- consumer learning models, 568
- consumer taste, 575
- contest function, 534–535
- contiguity matrix, 195
- continuously updated fully modified (CUP-FM) estimators, 52
- control function procedures, 622–623
- Cook's distance, 446n16
- Cornwell, Schmidt, and Sickles (CSS) panel stochastic frontier model, 523–524, 527
- correlated ("fixed") effects, 326
- correlated random effects, 187, 191–192, 595–596
- correlated random effects (CRE) model, 571
- count-dependent variables, 233–256

- count models  
*AR(1)* panel model, 446n13  
 individual effects in, 237  
 static, 237–243
- count panel data, 233–256
- country effects  
 cross-sectional fixed (exporter and importer), 632, 635n9  
 fixed, 629–631, 632–633  
 random, 629–631
- country-pair fixed effects, 632, 635n11
- country pairs  
 cross-section of, 610–613  
 three-way panels, 613–614
- covariances, 477
- CQR (composite quantile regression)  
 method, 306
- CRE (correlated random effects) model, 571
- criminology, 214
- cross-correlations, 477
- cross-sectional fixed (exporter and importer)  
 country effects, 632, 635n9
- cross-sectional gravity models, 614–617, 622–623
- cross-sectional independence, 30, 47–50, 316–318
- cross-sectional interdependence, 625–626
- cross-sectionally correlated panels, 51–55
- cross-section data, 614–617  
 count data, 234–236  
 endogenous variables with, 623–624  
 grouped, 209  
 repeated, 624–625  
 repeated observation over time of, 613–614  
 two-way, 617–619
- cross-section dependence, 374–376, 473–486  
 error, 28–38  
 large models with, 3–45  
 modeling, 474–476  
 panel data models with, 297–302  
 strong, 6  
 testing for, 33  
 tests for, 34, 393, 476–478  
 types of, 5–8  
 weak, 6, 7–8
- cross-section regression models, 446n16
- cross-sections  
 of country pairs, 610–613  
 repeated, 266–268, 617, 619–620
- CSS (Cornwell, Schmidt, and Sickles) panel  
 stochastic frontier model, 523–524, 527
- culture, common, 630
- CUP-FM (continuously updated fully modified) estimators, 52, 53
- data. *see also* Panel data  
 cohort, 493–516  
 cross-section, 209  
 missing, 150, 152, 159, 168, 585, 617–619
- degrees of freedom correction, 188–189
- $\delta_i$ : fixed effects for, 208
- demand  
 consumer demand, 548–582  
 good, 364
- Denmark, 494
- dependent data, 395
- dependent variables  
 Box-Cox transformed dependent variables, 309  
 lagged, 21
- deterministic, smooth factors, 162, 163f
- DGP, 59
- Dickey-Fuller coefficient  
 augmented (*ADF*), 57  
 modified, 56–57
- difference-in-difference (DiD) approach, 264–268, 276–277
- DIF (first-differenced) GMM estimators, 86, 98, 99, 105
- directed tests, 477
- disaggregate analysis, 410–412
- disclosure avoidance, 353–354
- discrete choice models, 171–201  
 computational issues, 553–557  
 of consumer demand, 548–582  
 estimation of, 553–564  
 extension to serially correlated taste shocks, 561–563  
 extension to state dependence, 563–564  
 general overview of, 553–557  
 other models, 184–185  
 over two alternatives, 174

- spatial panels and, 195–196  
typical structure of, 550–553  
discrete outcome models, 172–175  
disturbances, 628–629  
DOLS (dynamic OLS), 48–49  
double fixed effects, 208–209  
double-indexed variables, 630–631  
double index process, 5  
dummy variables, 178  
Durbin–Hausman test, 62–63  
Durbin regressors, 366  
dynamic average treatment effect (DATE),  
    271, 276  
dynamic average treatment effect (DATE) on  
    the treated (DATET), 271  
dynamic CCE estimators, 24–28  
dynamic models, 76–110, 160–161, 455–469  
    additional parameters, 211–212  
    alternative procedures, 84–85  
    of discrete choice, 192–193  
    existing results, 96–97  
    with factor error structure, 21–28  
    fixed effects, 245–247  
    incidental parameters, 111–148  
    initial conditions for, 85–93  
    in labor economics, 597–599  
    latent class models, 252–253  
    linear, 438–441  
    literature review, 77–85  
    logit models, 214  
    measurement error models, 343–347  
    Monte Carlo design, 93–96  
    new results, 97–105  
    panel count models, 243–247  
    pooled, 244  
    prototype model, 113  
    random coefficient models, 407–410  
    random effects, 245  
    second-order, 214  
    simulation results, 93–105, 100*t*, 101*t*,  
        102*t*, 103*t*, 104*t*  
    with smooth factors, 163–168, 166*t*  
    with spatial errors, 385–389  
    with spatial lag, 382–385  
    spatial models, 364, 368–374, 382–391,  
        387*t*  
    specifications for, 243–244  
static representation of, 456–458, 459*t*  
with stochastic factors, 163–168, 167*t*  
treatment models, 270–276  
unbalanced, with interactive effects,  
    160–161  
dynamic OLS (DOLS), 48–49  
dynamic panel conditional logit estimators,  
    211–220  
four periods or more with no regressor,  
    212–215  
four periods or more with  $y_T$  conditioned  
    on, 218–219  
four periods with the same last two-period  
    regressors, 215–217  
three periods or more using an estimator  
    for  $\delta_i$ , 219–220  
three periods or more without  $y_T$   
    conditioned on, 217–218  
three periods or more with regressors,  
    217–220  
dynamic panels, 586  
nonlinear, 138–143  
order selection in, 143–144  
PC estimators for, 23–24  
dynamics  
    in demand, 576  
    in labor economics, 599–601  
    in panel logit models, 211  
    in trade flows, 622  
EC3SLS estimation, 395  
ECHP (European Community Household  
    Panel Survey), 172  
econometrics, 583–607  
economic growth. *see also* Productivity  
    growth  
    decomposition of, 521–534  
    world, 535–539  
economics, 548–549  
efficiency  
    predictors of, 526–527  
    technical, 537  
efficiency change  
    identified by regression, 521–534  
    index number approaches to, 520–521  
Efficiency IV estimator, 524  
efficient instrumental variables, 524

- ELFE (Étude Longitudinale Française depuis d'Enfance) (French Longitudinal Study of Children), 494
- EM (Expectation-Maximization) algorithm, 158, 530  
     for factor models with missing data, 159  
     LS-EM-PCA estimation, 158–160, 168  
     for spatial panel data, 394
- empirical Bayes estimator, 409
- empirical Monte Carlo studies, 264
- endogeneity, 594–599
- endogeneous regressors, 622–625
- endogenous attrition, 173
- endogenous human capital, 586
- endogenous spatial weights matrices, 393
- endogenous variables  
     with cross-section data, 623–624  
     double-indexed, 630–631  
     with repeated cross-section (panel) data, 624–625  
     triple-indexed, 630
- England. *see* United Kingdom
- EQS software, 341
- equilibrium, hedonic, 586–587
- equivariance  
     affine, 444n3  
     crucial properties of, 444n3  
     of estimators, 421  
     to monotone transformations, 309  
     regression, 444n3  
     scale, 421, 425, 444n3
- error correction model (ECMs), 371  
     vector error correction model (VECM), 485–486
- error cross-sectional dependence tests, 28–38
- errors  
     asymptotic standard errors, 80  
     factor error structure, 14–21, 21–28  
     homoskedastic, 33  
     idiosyncratic, 8–9  
     measurement, 325–362  
     one-way components, 151  
     panel corrected standard errors (PCSE), 474, 478–480  
     spatial, 385–389  
     weak or semi-strong factors in, 13–14
- estimates  
     combining, 534–535  
     non-structural versus structural, 633–634  
     pooling of, 380
- estimation  
     consistent, 335–341, 344–345, 345*f*  
     iterative, 154  
     nonparametric, 287  
     optimal, 345–347  
     parametric, 395–396
- estimators. *see also specific estimators*  
     alternative, 472–473  
     coefficient, 98–99  
     general principles of, 472–473  
     logit, 202–232
- Étude Longitudinale Française depuis d'Enfance (ELFE) (French Longitudinal Study of Children), 494
- Euler equations, 351–353
- European Community Household Panel Survey (ECHP), 172
- European Union (EU), 632–633
- exchangeable covariance matrices, 304
- exogenous regressors, 339–341
- Expectation-Maximization (EM) algorithm, 158, 530  
     for factor models with missing data, 159  
     LS-EM-PCA estimation, 158–160, 168  
     for spatial panel data, 394
- explosive case, 372–373
- exponential feedback, 243–244
- exporter and importer (cross-sectional fixed)  
     country effects, 632, 635n9
- exporter-continent-time effects, fixed, 635n11
- exporter-time fixed effects, 632
- external information, 357–359
- factor error structure  
     dynamic panel data models with, 21–28  
     large panels with strictly exogenous regressors and, 14–21
- factor loadings, 8, 10–11, 533
- factor models, 534. *see also* Common factor models
- factors, 533. *see also* Common factors
- fatal injury effects, 586–587

- feedback, exponential, 243–244  
 finance economics, 374  
 finite mixture models (FMMs), 249–252  
 first difference, 118–120, 430–432, 594  
   bias of, 332, 332f  
 first-differenced (DIF) GMM estimators, 86, 98, 99, 105  
 first-order spatial auto-regressive models, 9, 626  
 Fisher effects, 62  
 Fisher's test statistic, 58  
 fixed effects (FE), 41n8, 85, 122, 123, 243, 355  
   additive, 125–126  
 bias B of, 124  
 country effects, 629–631  
 country-pair, 632, 635n11  
 double, 208–209  
 elimination of, 330–332  
 exporter-continent-time effects, 635n11  
 exporter-time, 632  
 formulations and problems of interest, 113–114  
 in gravity models, 629–631  
 importer-continent-time effects, 635n11  
 importer-time, 632  
 individual effects, 115–124, 377–378  
 individual-specific effects, 114  
 instrumental variables quantile regression  
   with fixed effects (IVQRFE) estimator, 308  
 interactive, 126–129  
 limit distribution of, 121  
 versus random effects, 469, 629–631  
 in static models, 185–191  
 in static spatial models with serially  
   correlated disturbances, 377–378  
 time-specific effects, 114  
 fixed effects models, 241–243, 293–297, 329–330  
   advantages of, 364  
   dynamic models, 245–247  
   logit models, 185, 187–188  
   probit models, 185, 188, 190  
   quantile regression models, 306–309  
   standard, 402  
   two-way, 611–612  
 fixed specification, 413–415  
 FMMs (finite mixture models), 249–252  
 FM-OLS estimator, 49, 52  
 forecasting, 394, 534–535  
 foreign direct investment models, 622–623, 626  
 foreign direct investment potentials, 633  
 forward orthogonal difference (FOD)  
   transformation, 120, 390  
 FPCA (functional principal component analysis), 150, 155–156, 157–158  
 French Longitudinal Study of Children  
   (Étude Longitudinale Française depuis d'Enfance, ELFE), 494  
 FSMQL (fully standardized Mallow's type quasilikelihood approach), 441  
 F-test, 470–471  
 fully nonseparable models, 310  
 fully standardized Mallow's type  
   quasilikelihood approach (FSMQL), 441  
 functional principal component analysis  
   (FPCA), 150, 155–156, 157–158  
 future research directions, 539–540
- Gateshead Millennium Study (GMS), 498*t*, 506  
 Gauss-Chebyshev quadrature, 554  
 Gauss-Hermite quadrature, 554  
 Gauss-Legendre quadrature, 554  
 G-computation algorithm, 270  
 Gemini (birth cohort study), 498*t*, 506  
 generalized estimating equation (GEE), 182, 239  
 generalized least squares (GLS), 404, 405  
   breakdown point, 432  
   SUR-GLS estimator, 473–474, 480–482  
 generalized linear longitudinal mixed model  
   (GLLMM), 441, 442  
 generalized linear models (GLMs), 235  
 generalized *M*-estimators (*GM*-estimators), 428–429, 432–433  
 generalized method of moments (GMM), 76–77, 89, 105–106, 318, 364, 366, 456  
   bias and precision of, 98–99  
   breakdown point, 432

- first-differenced (DIF) estimators, 86, 98, 99, 105  
GMM-D estimator, 353  
GMM-LN estimator, 353  
LEV estimator, 86  
merits of, 390–391  
nonlinear (AS) estimators, 98–99, 105  
robust (*RGMM*), 432, 433–434, 438–439, 439–440  
in SDPD models, 389–391  
for spatial panel data models, 393–394  
system (SYS) estimators, 86, 98–99, 105  
when  $T$  is large, 458–460  
generalized PCSE estimator, 478  
geometric weighted average, 518  
German Socioeconomic Panel (GSOEP), 172, 173, 216  
GHK algorithm, 550, 557–560, 564  
Gibbs samplers, 407  
GLLMM (generalized linear longitudinal mixed model), 441, 442  
GLMs (generalized linear models), 235  
global VAR (GVAR), 468, 485–486, 486–487  
GLS (generalized least squares), 404, 405  
breakdown point, 432  
SUR-GLS estimator, 473–474, 480–482  
*GM*-estimators (generalized-*M* estimators), 428–429, 432–433  
GMM. *see* Generalized method of moments  
GMS (Gateshead Millennium Study), 498t, 506  
good demand, 364  
good leverage points, 418, 419f  
Granger causality, 468  
gravity equations, 634n2  
differenced, 622  
dynamic adjustment, 622  
systems of, 621  
gravity models  
balanced case, 614  
cross-sectional, 614–617, 622–623  
empirical topics, 614–634  
fixed effects in, 629–631  
for foreign direct investment, 622–623  
fundamentals of, 609–610  
of international trade, 608–641  
random effects in, 629–631  
repeated cross-sectional (three-way panel), 617, 619–620  
in terms of relative log-odds or tetrads terms, 627  
theoretical background, 609–610  
with time variation, 613  
two-way fixed effects model, 611–612  
two-way random effects model, 612–613  
Greene, Kumbhakar, and Tsionas latent class models, 528–530  
grouped cross-section data, 209  
group mean estimator, 409  
growth convergence, 364  
growth theory, 415, 518–519  
GSOEP (German Socioeconomic Panel), 172, 173, 216  
GVAR (global VAR), 468, 485–486, 486–487  
HAC (heteroskedasticity and autocorrelation consistent) estimation, 395  
Hampel M-estimator, 422, 423t, 424f  
Hannan-Quinn (HQ) procedure, 143  
HAS (Hertfordshire Ageing Study), 503  
Hausman and Taylor type models  
endogenous variables with cross-section data in, 623–624  
endogenous variables with repeated cross-section (panel) data in, 624–625  
Hausman specification test, 380–381, 412–413, 445n6  
Hausman-Taylor estimator, 434–438  
Hausman test statistic, 243  
HCS (Hertfordshire Cohort Study), 503  
health economics, 493–516  
Health Survey for England, 503  
hedonic equilibrium, 586–587  
Helmert transformation, 390  
Hertfordshire Ageing Study (HAS), 503  
Hertfordshire Cohort Study (HCS), 503  
Hertfordshire Studies, 498t, 503  
heterogeneity, 412, 566  
individual, 175–192  
intercept, 589–591  
intercept and slope, 587–593  
types of, 573  
unobservable, 174  
heterogeneity tests, 412–413, 470–472

- heterogeneous panels, 472–473  
 heteroskedasticity, 394, 396  
     in gravity models of international trade, 614–617  
     spatial HAC estimation, 395  
 heteroskedasticity and autocorrelation  
     consistent (HAC) estimation, 395  
 hidden Markov models, 252–253  
 hierarchical Bayes estimator, 409, 410  
 hierarchical models, 191–192  
 HILDA (Household Income and Labor  
     Dynamics in Australia), 172  
 homogeneity, 411  
 homoskedastic errors, 33  
 homoskedastic slopes, 33  
 Horenstein, A. R., 41  
 hourly wage rates, 586–587  
     interactive factor model for, 600–601  
     Mincer wage models for, 589  
 Household Income and Labor Dynamics in  
     Australia (HILDA), 172  
 housing prices, 395  
 HQ (Hannan-Quinn) procedure, 143  
 Huber M-estimator, 422, 423*t*, 424*f*  
 human capital, endogenous, 586  
 hurdle models, 175, 247–249
- identification  
     nonparametric, 261–263, 268–269  
     semiparametric, 264–266  
 idiosyncratic errors, 8–9  
 idiosyncratic shocks, serially correlated, 563  
 idiosyncratic taste shocks, 552, 566  
 IIA (independence of irrelevant alternatives), 553  
 importer-continent-time effects, fixed,  
     635*n*11  
 importer-time fixed effects, 632  
 incidental parameters, 111–148  
     in discrete choice cases, 188–189  
     estimation of  $\rho$  with, 115–134  
     in micro panel data models, 202  
     testing for unit roots with, 134–138  
 incidental trends, 130–134  
 inconsistency, 328  
 incremental Sargan tests, 97  
 independence  
     conditional, 177, 178  
     cross-sectional, 316–318  
     of irrelevant alternatives (IIA), 553  
     of random terms in utility functions, 178  
 index *i*, 46  
 index number approaches, 520–521  
 indicator functions, 563  
 individual effects, 151  
     in count models, 237  
     fixed effects, 115–124, 125–126, 377–378  
     random effects, 378–379  
 individual heterogeneity, 175–192  
 individual-specific effects, 114, 202  
 infeasible Bayes estimator, 409  
 inference, valid, 478  
 inference approach, 395  
 influence functions, 420, 421, 422, 424*f*  
 influential observations, 418–450  
 innovation  
     decomposition of, 521–534  
     index number approaches to,  
         520–521  
     measurement of, 537  
     technical change due to, 526–527  
 instrumental variables (IVs), 76, 118–120,  
     268–270, 326  
     Arellano and Bond IV matrix, 439  
     efficient, 524  
     estimates of VSL, 594, 594*t*  
     ideal IV matrix, 391  
     robust IV estimators, 432–441  
     standard, 622–623  
 instrumental variables quantile regression  
     with fixed effects (IVQRFE) estimator,  
         308  
 instruments  
     approaches for limiting the number of, 81  
     many, 80–81  
     proliferation of, 460–462  
     weak, 82–84  
 integer-valued autoregressive models of  
     order 1 (INAR(1)), 244  
 integrated likelihood approach, 141–143  
 interactive effects, 149–170  
     fixed effects, 126–129  
 intercept and slope heterogeneity models,  
     591–593

- intercept heterogeneity models, 589–591  
 international trade models, 608–641  
 inventory, current, 577n3  
 inventory effects, 567, 568  
 inverse Mill's ratio, 600–601  
 inverse normal test statistic, 58  
 inverse probability weighting, 192  
 Ireland, 494  
 IRI, 548  
*IRLS* (iterated reweighted least squares), 430–431  
 irrelevant alternatives, 553  
 iterated reweighted least squares (*IRLS*), 430–431  
 iterative estimation, 154  
 IVQRFE (instrumental variables quantile regression with fixed effects) estimator, 308  
 IVs. *see* Instrumental variables
- $J_{BFK}$  test, 34, 35, 37t  
 Jensen's inequality, 633  
 $J_n$ : transformation approach with, 383–384  
 joint asymptotics, 127–129
- KLIPS (Korean panel data), 216  
 Kneip, Sickles, and Song model, 530–532  
 Korean panel data (KLIPS), 216  
 Krugman-type models, 610  
 Kullback-Leibler approach, 112, 144  
 Kumbhakar model, 525
- labor economics  
   dynamic models of, 597–599  
   endogeneity in, 594–599  
   framework for, 586–587  
   hedonic equilibrium in, 586–587  
   heterogeneous intercepts and slopes in, 587–593  
   intercept and slope heterogeneity models of, 591–593  
   intercept heterogeneity models of, 589–591  
   panel econometrics, 583–607  
   sample composition dynamics, 599–601  
   state dependence in, 597–599  
   summary, 601–602  
 labor-market data, 593–594
- labor policy, 601–602  
 lagged choice, 569  
 lagged choice variables, 552  
 lagged dependent variables, 21  
 lagged prices, 567, 569  
 lagged purchases, 551, 568, 577n3  
 Lagrange multiplier (LM) test, 30–31, 363  
*LM<sub>Adj</sub>* test, 31–32, 34, 35, 36t  
*LM<sub>S</sub>* test, 33, 34, 35, 36t  
 robust modified versions, 471–472  
 for serial correlation and spatial autocorrelation, 392  
 for spatial effects, 392–393  
 standardized, 393  
 for unbalanced panels, 39–40
- large N, small T panel data sets, 453–454  
 large panel data models, 3–45  
 large panel data sets, 444  
 large panels, 14–21  
 latent class models, 183, 249–252, 561, 562  
   dynamic, 252–253  
   general expression for, 249  
   of Greene, Kumbhakar, and Tsionas, 528–530
- Latin American countries, 536, 537, 538–539  
 least squares. *see also* Ordinary least squares (OLS) estimators  
   of aggregate money demand function, 410–411, 410t  
   FPCA via (LS-FPCA), 157–158, 168  
   generalized (GLS), 404, 405, 432, 473–474, 480–482  
   iterated reweighted (*IRLS*), 430–431  
   LS-EM-PCA estimation, 158–160, 168  
   nonlinear (NLS), 393–394  
   robust and efficient weighted (*REWLS*) estimator, 431–432, 445n11  
   2SLS estimation, 393–394
- least squares functional principal component analysis (LS-FPCA), 157–158, 168
- Least Trimmed Squares (*LTS*) estimator, 425  
 reweighted LTS (*RLTS*), 431–432
- LEV GMM estimator, 86  
 Life Study (UK), 498t, 502, 512n1  
 likelihood, 205  
   concentrated, 138–141  
   integrated, 141–143

- 
- maximum likelihood (ML) estimator, 85  
modified estimation equations, 189  
quasi-maximum likelihood estimator (QMLE), 21–23  
limited-dependent variables, 355–356  
linearly approximated models, 627–628  
linear models  
  dynamic models, 438–441  
  generalized (GLMs), 235  
  with spatial autoregression, 195  
  static models, 403–407, 419–432  
LISCOMP model, 356  
Lisrel model, 341–342  
LISREL software, 341–342  
literature reviews  
  for dynamic panel data models, 77–85  
  empirical findings on *TFP* growth, 536–539  
  empirical Monte Carlo studies, 264  
  empirical work on state dependence, 572–576  
  for missing data problem, 150  
  for panel stochastic frontier models, 533–534  
  for PCLEs, 208–209  
LM (Lagrange multiplier) test, 30–31, 363  
   $LM_{Adj}$  test, 31–32, 34, 35, 36t  
   $LM_S$  test, 33, 34, 35, 36t  
  robust modified versions, 471–472  
  for serial correlation and spatial autocorrelation, 392  
  for spatial effects, 392–393  
  standardized, 393  
  for unbalanced panels, 39–40  
local average treatment effect (LATE), 260  
lock-in effects, 262–263  
logit estimators, 202–232  
longest differences, 332, 332f  
longitudinal data sets, unbalanced, 173  
longitudinal survey data sets, 172  
long-run average regression coefficient, 47  
loss functions, 422  
Lothian Birth Cohorts, 498t, 503–504  
LR statistic, 476  
LSDV estimator, 98  
LS-EM-PCA estimation, 158–160, 168  
LS-FPCA (least squares functional principal component analysis), 157–158, 168  
LTS (Least Trimmed Squares) estimator, 425  
  reweighted LTS (*RLTS*), 431–432  
macroeconomic data, 453–492  
macro shocks, 533  
MAD (mean absolute deviation), 421  
majority voting, 534–535  
Mahalanobis distance (or Rao's distance), 429  
Mallows's estimators, 428–429  
Mallow's type quasilielihood (MQL)  
  approach, 441  
many instruments problem, 460–462  
marginal analysis, 238  
marginal effects, 145n9, 209–210  
marketing, 573  
marketing mix variables, 565–566  
*Marketing Science*, 548  
market mapping, 553, 556  
Markov chain Monte Carlo (MCMC)  
  methods, 240–241  
masking effects, 420, 425  
matching, 267–268  
MatLab software, 171, 487n1, 487n14  
matrices  
  Arellano and Bond IV matrix, 439  
  endogenous spatial weights matrices, 393  
  exchangeable covariance, 304  
  spatial weight or contiguity matrix, 195  
  wage equation in matrix form, 595  
maximum likelihood estimator (ML)  
  estimator or MLE), 85, 122, 465–467  
  Gaussian, 114  
  issues with, 390  
  partial, 208  
  Poisson quasi-MLE, 234–235  
  quasi-maximum likelihood estimator (QMLE), 21–23, 115, 118, 234–235  
MCD (Minimum Covariance Determinant) estimator, 445n8  
MCS (Millennium Cohort Study) (UK), 494, 498t, 501–502, 511, 512n1  
MDE (minimum distance estimator)  
mean absolute deviation (MAD), 421

- mean group estimator, 409, 470  
measurement error, 325–362  
    basic results, 327–335  
    classical, 347–348  
    dynamic models of, 343–347  
    identification of, 334–335  
    in labor-market data, 593–594  
    multiplicative, 350–354  
    neglecting, 327–328  
    nonclassical, 347–350  
    nonlinear models of, 354–356  
    structural equation model (SEM) of, 342–343  
Medical Expenditure Panel Survey (MEPS) (US), 172  
M-estimators, 420–425  
    commonly used, 422, 423*t*  
    generalized (GM), 432–433  
    Hampel, 422, 423*t*, 424*f*  
    Huber, 422, 423*t*, 424*f*  
    influence functions of, 422, 424*f*  
    Tukey, 422, 423*t*, 424*f*  
    two-stage generalized (2SGM), 432–433  
    weight functions of, 422, 424*f*  
method of moments (MOM) estimation, 363, 377–381  
micro panel data models, 202  
migration flows, 626, 634n2  
Millennium Cohort Study (MCS) (UK), 494, 498*t*, 501–502, 511, 512n1  
Mill's ratio, inverse, 600–601  
Mincer models, 584, 589  
Minimum Covariance Determinant (MCD)  
    estimator, 445n8  
minimum distance estimator (MDE)  
minimum distance theory, standard, 356  
missing  
    block, 152, 153*f*  
    random, 152, 153*f*  
    regular, 152, 153*f*  
missing data, 152, 168, 585, 617–619  
    EM algorithm for factor models with, 159  
    literature review, 150  
missing observations, 168  
    in panel data, 152–154  
    simulation patterns, 162  
    for spatial panel data sets, 393–394  
types of patterns, 152, 153*f*  
mixed models, 183  
mixture of logits model (MIXL), 560  
    normal (N-MIXL), 560, 561, 563  
mixture-of-normals model, 561  
ML or MLE. *see* Maximum likelihood estimator  
MM-estimation, 426–427  
model averaging, 534–535  
models  
    binary choice models, 175–192  
    common factor models, 8–14  
    of consumer demand, 548–582  
    of cross-section count data, 234–236  
    of cross-section dependence, 474–476  
    discrete choice models, 171–201, 548–582  
    discrete outcome models, 172–175  
    dynamic, 76–110, 160–161, 211–212, 455–469  
    factor models with missing data, 159  
    gravity models of international trade, 608–641  
    nonlinear regression models, 172–175  
    panel data models and methods, 3–450  
    of world economic growth, 535–539  
modified estimation (likelihood) equations, 189  
moment conditions testing, 93  
MOM (method of moments) estimation, 363, 377–381  
money demand  
    least squares estimation of, 410–411, 410*t*  
    random coefficient estimation of, 411, 411*t*  
monotone transformations, 309  
Monte Carlo simulations  
    of dynamic panel data models, 93–96  
    empirical, 264  
    of unbalanced panel data models with interactive effects, 161–168  
Moran I test, 392  
mortality risk reduction: value of (VMRR), 584

- Mplus software, 341, 356  
 MQL (Mallow's type quasilikelihood)  
     approach, 441  
 MRC National Survey of Health and  
     Development (NSHD) (UK), 498t, 499  
 MS-estimators, 425–428  
 multinomial choice, 174  
 multinomial logit estimators, 202–232  
 multiplicative measurement error, 350–354  
 Mundlak-Chamberlain model, 467  
 Mundlak-type models, 619–620  
 Mx software, 341
- National Bureau of Economic Research, 517  
 National Child Development Study (NCDS)  
     (UK), 498t, 499–500, 507–511  
 National Children's Study (NCS) (US), 494  
 National Health Service (UK), 499  
 National Longitudinal Survey (NLS) (US),  
     172, 173, 583  
 National Survey of Health and Development  
     (NSHD) (UK), 494, 498t, 499  
 natural experiments, 500, 585  
 NCDS (National Child Development Study)  
     (UK), 498t, 499–500, 507–511  
 negative binomial (NB) models, 235,  
     236  
 neoclassical model, 518–519  
 new growth theory, 518–519  
 Nickell bias, 85, 115–118, 455  
 NLOGIT software, 171, 180  
 NLS (National Longitudinal Survey) (US),  
     172, 173, 583  
 N-MIXL (normal mixture of logits model),  
     560, 561, 563  
 noise, 83  
     adding, 563  
     signal-to-noise ratio (SNR), 95–96  
 noncointegration  
     null of, 56–64, 65–66  
     sequential approach to testing for, 61–62  
 nonlinear dynamic panels, 138–143  
 nonlinear (AS) GMM estimators, 98–99,  
     105  
 nonlinear least squares (NLS) method,  
     393–394
- nonlinear measurement error models,  
     354–356  
 nonlinear panel data models, 176, 441–442  
 nonlinear regression models, 172–175  
 nonparametric identification, 261–263,  
     268–269  
 nonparametric regression models, 285–324  
 nonparametric tests, 313–318  
 nonseparable models, 310–313  
 non-structural estimates, 633–634  
 non-structural models, 633–634  
 normalized outcomes, 627  
 normal mixture of logits model (N-MIXL),  
     560, 561, 563  
 Northern Ireland. *see* United Kingdom  
 Norway, 494  
 notation, 46, 259–261, 280n6, 281n20,  
     281n25, 454  
 NSHD (National Survey of Health and  
     Development) (MRC) (UK), 498t, 499  
 nuisance parameters, 81–82  
 null of noncointegration tests,  
     residual-based, 56–64, 65–66  
 number of choice situations ( $T$ ). *see T*  
     (number of choice situations or time  
         series observations)
- observables: selection on, 261–264, 600  
 observations  
     bad leverage points, 418, 419f  
     of cross-section data over time, 613–614  
     good leverage points, 418, 419f  
     influential, 418–450  
     missing, 152–154, 168  
     outliers, 442–444  
     random sampling of units, 178  
     repeated, 613–614  
     vertical outliers, 418, 419f  
 observed common factors ( $d_t$ ), 41n8  
 Occam's razor, 578n5  
 odds ratio, 209–210  
 OECD (Organization for Economic  
     Cooperation and Development)  
     countries, 536, 537, 538–539  
 OIR (overidentifying restrictions) tests, 97,  
     99–105  
 OLS (ordinary least squares) estimators, 364

- asymptotic properties, 54  
breakdown point, 432  
dynamic OLS (DOLS), 48–49  
FM-OLS estimator, 49, 52  
pooled, 114  
reduced form, 327–328  
Within-OLS estimator, 48–49
- OpenMx software, 341  
optimal estimation, 345–347  
ordered choice, 190–191  
ordered logit, panel conditional, 220–223  
ordered multinomial choice, 174  
order selection, 143–144  
ordinary least squares (OLS) estimators, 364  
    asymptotic properties, 54  
    breakdown point, 432  
    dynamic OLS (DOLS), 48–49  
    FM-OLS estimator, 49, 52  
    pooled, 114  
    reduced form, 327–328  
    Within-OLS estimator, 48–49
- Organization for Economic Cooperation and Development (OECD) countries, 536, 537, 538–539
- outliers, 442–444
- overidentifying restrictions (OIR) tests, 97, 99–105
- pairwise differences, 430–432, 445n9
- panel cointegration, 46–75
- panel conditional logit estimator (PCLE), 203, 230  
    dynamic, 211–220  
    with more than enough waves, 223  
    static, 204–211
- panel conditional multinomial logit (PCML), 224–230
- panel conditional ordered logit estimators, 220–223
- panel corrected standard error (PCSE), 474, 478–480
- panel count models  
    dynamic, 243–247  
    static, 237–243
- panel data  
    applications, 453–641
- compared to repeated cross-sections, 266–268  
count, 233–256  
detection of influential observations and outliers in, 442–444  
large N, small T sets, 453–454  
large sets, 444  
macroeconomic, 453–492  
measurement error in, 325–362  
missing observations in, 152–154  
repeated, 624–625  
treatment effects and, 257–284  
value of, 266–268
- panel data models and methods, 3–450  
    applied to cohorts, 506–511  
    approximate factor model, 9  
    augmented regression model, 113  
    autoregressive model, 113  
    binary choice models, 175–192  
    common factor models, 8–14  
    for count-dependent variables, 233–256  
    with country-pair, exporter-time, and importer-time fixed effects, 632  
    with cross-sectional dependence, 297–302  
    discrete choice models, 171–201, 548–582  
    dynamic models, 21–28, 76–110, 111–148, 163–168, 166t, 167t, 364, 382–385, 385–389, 438–441  
    fixed effects models, 293–297  
    with fixed pair and fixed country-time effects, 632–633  
    gravity models of international trade, 608–641  
    heterogeneous, 14–15  
    with interactive effects, 151–152  
    with lagged dependent variables and unobserved common factors, 21
- large models with cross-sectional dependence, 3–45
- linear dynamic models, 438–441
- linear static models, 419–432
- micro models, 202
- nonlinear, 176, 441–442
- nonparametric regression models, 285–324
- panel ordered logit models, 220–223
- parametric models, 235–236

- prototype model, 113  
 random coefficient models, 402–417  
 random effects models, 286–293  
 regression models, 48–49, 57  
 robust methods, 418–450  
 simplest models, 233  
 spatial models, 363–401  
 static models, 364, 365–368, 367*t*  
 stochastic frontier models, 520, 523–524,  
   525, 526–527, 527–528, 528–530,  
   533–534  
 unbalanced models with interactive  
   effects, 149–170  
 panel fully aggregated estimator (PFAE), 119  
 panel logit estimators, 202–232  
 panel regressions, 47–55  
 panels  
   cross-sectionally correlated, 51–55  
   cross-sectionally independent, 47–50  
   dynamic, 23–24, 138–143, 143–144  
   with fixed T, 118–120  
   heterogeneous, 472–473  
   with large T, 120–124  
   nonlinear dynamic, 138–143  
   pseudo-balanced, 482  
   random coefficient models in, 402–417  
   spatial, 195–196  
   three-way, 613–614, 617  
   unbalanced, 38–40, 192  
 Panel Study of Income Dynamics (PSID),  
   357, 583, 588  
 panel VAR cointegration tests, 66–70  
 parameters. *see also specific parameters*  
   nuisance parameters, 81–82  
   pooled versus individual specific, 469–473  
   testing for heterogeneity of, 470–472  
 parameter vector  $\beta$   
   fully modified OLS (FM-OLS) estimator  
     of, 48  
   Within-OLS estimator of, 48–49  
 parametric estimates, 395–396  
 parametric models, 235–236  
 Paris, France, 395  
 Park, Sickles, and Simar (PSS) models,  
   527–528  
 partial effects, 181–182  
 partially separable models, 310  
 partial MLE, 208  
 PCA (principal component analysis), 15–16,  
   150, 151, 483–484  
   for dynamic panels, 23–24  
   functional (FPCA), 150, 155–156, 157–158  
   with joint asymptotics, 127–129  
   LS-EM-PCA, 158–160, 168  
   LS-FPCA, 157–158, 168  
 PCLE (panel conditional logit estimator),  
   203, 230  
   dynamic, 211–220  
   with more than enough waves, 223  
   static, 204–211  
 PCML (panel conditional multinomial  
   logit), 224–230  
 PCSE (panel corrected standard error), 474,  
   478–480  
 penalized criterion, 189  
 penalized IVQRFE estimator, 308  
 penalized QRFE (PQRFE) estimator, 307  
 Penn World Tables, 174  
 Perinatal Mortality Survey (PMS) (UK),  
   499–500  
 PFAE (panel fully aggregated estimator), 119  
 Phillips–Perron coefficient, 58  
 placebo tests, 262–263  
 PMS (Perinatal Mortality Survey) (UK),  
   499–500  
 Poisson generalized estimating equations  
   (GEE) estimator, 239  
 Poisson models  
   fixed effects versions, 233  
   marginal density in, 187  
   random effects (RE) models, 239, 240  
   standard generalization of, 235–236  
 Poisson pseudo-maximum-likelihood  
   (PPML) estimator, 615, 617  
 Poisson quasi-MLE estimator, 234–235  
 polynomial models, 354–355  
 poolability tests, 54, 314–316, 470–471  
 pooled dynamic models, 244  
 pooled least square (OLS) estimator, 114  
 pooled models, 238–239, 243  
 pooled panel data quantile regression  
   models, 304–306  
 pooled parameters, 469–473  
 pooling of estimates, 380

- population-averaged models, 182, 190, 238–239
- potential outcomes, 259, 277–279
- PPML (Poisson pseudo-maximum-likelihood) estimator, 615, 617
- PQRFE (penalized quantile regression fixed effects) estimator, 307
- precision, 98–99
- preferentialism, 630
- preferential trade agreements, 631–633
- pre-program tests, 262
- price, 567
- inventory effects on, 567, 568
  - lagged price, 567–568, 569
  - reference price effects, 567, 568
  - relative price, 578n6
  - as signal of quality, 567, 568
- price coefficients, 551
- price elasticity, 576, 578n8
- price promotion, 575
- principal component analysis (PCA), 15–16, 150, 151, 483–484
- for dynamic panels, 23–24
  - functional (FPCA), 150, 155–156, 157–158
  - with joint asymptotics, 127–129
  - LS-EM-PCA, 158–160, 168
  - LS-FPCA, 157–158, 168
- probit models
- alternatives to, 560–561
  - correlated random effects (CRE) model, 571
  - discrete choice models, 554
  - fixed effects models, 185, 188, 190
  - random effects models, 179, 183–184, 554, 562
- problems of interest, 113–114
- product attributes, 577n1
- production function, 519
- productivity change, 537
- productivity growth, 518–521, 537
- productivity measurement, 517–547
- propensity score, 264
- prototype model, 113
- pseudo-balanced panels, 482
- pseudo-likelihood approaches, 442, 446n15
- pseudo-panels, 496
- PSID (Panel Study of Income Dynamics), 357, 583, 588
- p*-spacings, 477
- public economics, 364
- purchase carry-over effect, 573
- purchases, lagged, 551, 568, 577n3
- purchasing power parity (PPP) relations, 69
- pure space recursive models, 364
- $Q_j$  (true quality), 573
- QML estimation, 393
- QSF (quantile structural function), 311, 313
- QTE (quantile treatment effect), 311
- quadrature, 554–555
- quality
- price as signal of, 567, 568
  - quantile crossing, 309–310
- quantile regression (QR) models
- composite quantile regression (CQR) method, 306
  - conditional quantile regression models, 303–310
  - fixed effects panel data quantile regression models, 306–309
  - instrumental variables quantile regression with fixed effects (IVQRFE) estimator, 308
  - penalized quantile regression fixed effects (PQRFE) estimator, 307
  - pooled panel data quantile regression models, 304–306
- quantile structural function (QSF), 311, 313
- quantile treatment effect (QTE), 311
- quasi-differencing approach, 126
- quasilikelihood approach, 441
- quasi-maximum likelihood (QML)
- estimation, 364
- quasi-maximum likelihood estimator (QMLE), 21–23
- Poisson, 234–235
- random coefficient models, 402–417, 411*t*
- random coefficients, 533
- random effects (RE), 230, 243, 577n4
- correlated, 187, 191–192, 595–596
  - in count data models, 184
  - country effects, 629–631
  - versus fixed effects, 469, 629–631

- in gravity models, 629–631  
 individual effects, 378–379  
 specification with fixed T, 387–388  
 in static models, 179–185  
 in static spatial models with serially correlated disturbances, 378–379  
 random effects assumption, 566  
 random effects models, 183, 239–241, 243, 286–293, 333–334  
 alternatives, 182–183  
 definition of, 179  
 dynamic models, 245  
 estimators for, 50  
 fully specified logit model, 184  
 generalizations of, 240–241  
 logit models, 179, 184  
 probit models, 179, 183–184, 554, 562  
 standard, 402  
 two-way random effects model, 612–613  
 random missing, 152, 153f  
 random sampling, 178  
 random specification, 413–415  
 random terms, 178  
 random utility, 171–172  
 random utility models, 172–173, 552  
 rank ordered logit models, 174  
 Rao's distance (Mahalanobis distance), 429  
 Rasch/Chamberlain estimator, 191  
 real estate economics, 364, 374, 395  
 reduced form (RF) parameters, 221  
 reference price effects, 567, 568  
 regional markets, 364  
 regression  
   for cross-sectionally correlated panels, 51–55  
   for cross-sectionally independent panels, 47–50  
   efficiency change identification by, 521–534  
   heterogeneous coefficients, 57  
   long-run average coefficients, 47  
   modified Dickey-Fuller coefficient, 56–57  
   Phillips–Perron coefficient, 58  
   seemingly unrelated (SUR), 150, 395  
 regression equivariance, 444n3  
 regression models
- composite quantile regression (CQR)  
   method, 306  
 cross-section, 446n16  
 for dynamic OLS (DOLS), 48–49  
 fixed effects panel data quantile regression models, 306–309  
 global VAR (GVAR), 485–486  
 with heterogeneous coefficients, 57  
 nonlinear, 172–175  
 nonparametric, 285–324  
 pooled panel data quantile regression models, 304–306  
 vector auto-regressive (VAR) models, 467–468  
 regressors  
   endogenous, 622–625  
   strictly exogenous, 14–21  
 regularization parameters, 307  
 regular missing, 152, 153f  
 rejection frequency, 99  
 relative efficiency, 421  
 repeated cross-sections  
   compared to panel data, 266–268  
   data with, 617  
   three-way, 617, 619–620  
 repeated observations, 613–614  
 research directions, 539–540  
 residual-based tests, 56–64, 65–66  
 restrictions  
   benefits of, 336–337  
 reweighted LTS (*RLTS*) estimator, 431–432  
 REWLS (robust and efficient weighted least squares) estimator, 431–432, 445n11  
 RGMM (robust generalized method of moments estimator)  
   for linear dynamic models, 438–439, 439–440  
   for linear static models, 432, 433–434  
 $\rho$   
   estimation of, with incidental parameters, 115–134  
   limit distribution, 121  
   (concentrated) profile likelihood of, 123  
   QMLE of, 115  
*RLTS* (reweighted LTS) estimator, 431–432  
 RND1 model, 536

- robust (term), 444n1  
 robust and efficient weighted least squares (*REWLS*) estimator, 431–432, 445n11  
 robust dispersion, 425–428  
 robust estimators  
   for linear dynamic models, 438–441  
   for linear static models, 419–432  
   for nonlinear models, 441–442  
 robust generalized method of moments estimator (*RGMM*)  
   for linear dynamic models, 438–439, 439–440  
   for linear static models, 432, 433–434  
 robust Hausman-Taylor estimator, 434–438  
 robust IV estimators, 432–441  
 robust LTS (*RLTS*) estimator, 430–432  
 robust methods, 418–450  
 robust pseudo-likelihood (RPL) estimator, 442  
 RPL (robust pseudo-likelihood) estimator, 442  
 R software, 171  
 R's sem package, 341  
 Rubin rules, 358
- SAH (self-assessed health), 192  
 sample selection, 173  
 sampling  
   GHK algorithm for, 564  
   for linear static models, 403–405  
   sequential importance sampling, 550  
 SAR (spatial auto-regressive) disturbances, 363  
 Sargan-Bhargava statistic, modified, 64  
 Sargan tests, 97  
 SAR (spatial auto-regressive) models, 395–396  
   first-order, 9, 626  
   linear, 195  
   semiparametric, 396  
 SAS software, 171, 180, 341  
 scale equivariance, 421, 425, 444n3  
 scanner data, 548, 549, 565–566  
 scanner data panels, 565  
 Scottish Mental Surveys, 503, 504  
 SDPD models. *see* Spatial dynamic panel data (SDPD) models  
   second-order dynamic models, 214  
   seemingly unrelated regression (SUR), 395  
     SUR-GLS estimator, 480–482  
   self-assessed health (SAH), 192  
   semiparametric and parametric estimates, 395–396  
   semiparametric identification, 264–266  
   semi-strong factors, 12, 13–14  
   SEMs (simultaneous equations models), 395  
   SEMs (structural equation models), 326, 341–343  
   sequential importance sampling, 550  
   serial correlation  
     strongly serially correlated case, 131–134  
     tests for, 392  
     weakly serially correlated case, 130–131  
   serially correlated disturbances, 377–381  
   serially correlated idiosyncratic shocks, 563  
   serially correlated taste shocks, 561–563  
   serially uncorrelated disturbances, 376–377  
   S-estimators, 425–428, 426t, 445n6  
   sieve bootstrap method, 66  
   signal, 83  
   signal-to-noise ratio (SNR), 95–96  
   simulations, 555–556  
     dynamic panel data model, 93–105  
     unbalanced panel data model with interactive effects, 161–168  
   simultaneous equations models (SEMs), 395  
   single index models, 185  
   SIPP (Survey of Income and Program Participation) (US), 172  
   skewness problem, 534  
   slope, homoskedastic, 33  
   SMA process, 366  
   smooth factors  
     deterministic, 162, 163f  
     dynamic models with, 163–168, 166t  
     static model with, 163–168, 164t  
   SNR (signal-to-noise ratio), 95–96  
   software, 341  
   Solow model, 89–90  
   Solow Residual (SR), 538  
   space-time filters, 368  
   spatial autocorrelation tests, 392, 393  
   spatial auto-regressive (SAR) disturbances, 363

- spatial auto-regressive (SAR) models, 395–396  
 first-order, 9, 626  
 linear, 195  
 semiparametric, 396
- spatial cointegration, 371–372, 383, 387<sub>t</sub>
- spatial Durbin regressors, 366
- spatial dynamic panel data (SDPD) models  
 applications, 364  
 bias correction, 385  
 cases, 370  
 categories of, 364  
 error correction model (ECM)  
   representation, 371  
 estimation and inference, 382–391, 387<sub>t</sub>  
 explosive case, 372–373, 384–385  
 fixed effects specification with fixed T, 388–389  
 general specifications, 368–369  
 GMM estimation, 389–391  
 with individual and time effects, 390  
 parameter spaces, 370, 371  
 pure unit root case, 386–387  
 QML estimation, 393  
 random effects specification with fixed T, 387–388  
 reduced form, 369  
 situations, 370  
 spatial cointegration case, 383, 387<sub>t</sub>  
 specifications, 368–374  
 stable case, 382–383, 387<sub>t</sub>  
 with time dummies, 383–384  
 unit root case, 373–374, 387<sub>t</sub>
- spatial econometrics, 394
- spatial effects, 392–393
- spatial errors, 385–389
- spatial heteroskedasticity and  
 autocorrelation consistent (spatial HAC) estimation, 395
- spatial lag (SL), 363, 382–385
- spatial moving average (SMA) structures, 363–364
- spatial panel data models, 363–401  
 dynamic, 364, 368–374, 382–391  
 static, 364, 365–368, 367<sub>t</sub>, 376–381
- spatial panels, 195–196
- Spatial Stochastic Frontier model, 534
- spatial weight matrix, 195, 393
- specification tests, 183–184
- sphericity test, 477
- SPSS Amos software module, 341
- spurious state dependence, 569
- SR (Solow Residual), 538
- stability theory, 444n1
- standard errors, asymptotic, 80
- Stata software, 171, 180, 208  
 external XTCSD procedure, 487n11  
 external XTLDVC procedure, 487n1
- SEM module, 341
- state dependence, 552, 563–564, 586  
 in demand, 575  
 empirical work on, 572–576  
 in labor economics, 597–599  
 spurious, 569  
 testing for, 565–571
- static models  
 case example, 151–152  
 of consumer demand, 577  
 count models, 237–243  
 fixed effects in, 185–191  
 linear, 403–407, 419–432  
 random effects in, 179–185  
 with smooth factors, 163–168, 164<sub>t</sub>  
 spatial models, 364, 365–368, 367<sub>t</sub>, 376–381  
 with stochastic factors, 163–168, 165<sub>t</sub>  
 treatment models, 259–270
- static panel conditional logit estimators, 204–211
- stationarity, 371
- statistical life value (VSL), 584, 585, 586, 588, 588<sub>t</sub>, 592–593, 592<sub>t</sub>, 594, 594<sub>t</sub>, 599, 601–602, 603nn4–5
- steady state behavior: deviations from, 91–92
- stochastic factors  
 dynamic model with, 163–168, 167<sub>t</sub>  
 non-smooth factors, 162, 163<sub>f</sub>  
 static model with, 163–168, 165<sub>t</sub>
- stochastic frontier models, 520  
 additional models, 533–534
- Battese and Coelli model, 526–527
- Cornwell, Schmidt, and Sickles (CSS) model, 523–524, 527

- of Greene, Kumbhakar, and Tsionas, 528–530  
Kumbhakar model, 525  
in literature, 533–534  
Park, Sickles, and Simar (PSS) models, 527–528  
store choice, 578n6  
strong cross-sectional dependence, 6  
strong factors, 14  
definition of, 10  
example, 10–11  
theorem 2, 12–13  
strongly serially correlated case, 131–134  
structural equation models (SEMs), 326, 341–343  
structural estimates, 633–634  
structural form (RF) parameters, 221  
structural functions  
average structural function (ASF), 311  
quantile structural function (QSF), 311, 313  
structural learning models, 568  
structural models, 634  
structural parameters, 202  
sub-sampling algorithms, 445n7  
SUR (seemingly unrelated regression), 395  
SUR-GLS estimator, 473–474, 480–482  
Survey of Income and Program Participation (SIPP, US), 172  
swamping effects, 420  
system estimator, 459–460  
system (SYS) GMM estimators, 86, 98–99, 105  
systems of equations, 621
- T (number of choice situations or time series observations), 173  
balanced, 173  
fixed, 118–120, 387–388, 388–389  
large, 120–124, 409, 455, 458–460  
large N, small T panel data sets, 453–454  
sample log-likelihood function for general T, 218  
sufficiently large, 455  
unbalanced, 173  
taste shocks  
idiosyncratic, 552, 566
- serially correlated, 561–563  
technical efficiency, 537  
technical innovation change, 526–527, 537  
tests and testing. *see also specific tests*  
for constant-correlated effects, 93  
for cross-section dependence, 28–38, 33, 34, 476–478  
directed tests, 477  
for error cross-sectional dependence, 28–38  
with incidental parameters, 134–138  
for moment conditions, 93  
for noncointegration, 61–62  
overidentifying restrictions (OIR) tests, 97, 99–105  
for panel cointegration, 55–70  
residual-based, 56–64, 65–66  
for sphericity, 477  
for state dependence, 565–571  
for unit roots, 134–138  
TFP (total factor productivity) change, 537  
TFP (total factor productivity) growth, 537  
decomposition of, 522–523, 526–527, 530, 537  
empirical findings, 536–539  
measurement of, 518, 520–521, 528–529, 533–534  
world growth findings, 536–539  
third moments, 339  
three-way panels, 613–614, 617  
time dummies, 178, 383–384  
time effects, 125–126, 474–476  
time invariant individual variables (TIVs), 186  
time series models, 243–244  
time series observations (*T*). *see T* (number of choice situations or time series observations)  
time-space dynamics, 364, 368  
time-space recursive models, 364  
time-space simultaneous models, 364  
time-specific effects, 114  
time-varying parameters, 210–211  
time-varying regressors, 215  
TIVs (time invariant individual variables), 186  
total factor productivity (TFP) change, 537

- total factor productivity (*TFP*) growth, 537  
decomposition of, 522–523, 526–527, 530,  
537  
empirical findings, 536–539  
measurement of, 518, 520–521, 528–529,  
536  
world growth findings, 536–539
- trade cost variables, 635n6
- trade data  
cross-section (two-way), 617–619  
missing, 617–619  
repeated cross-sectional (three-way),  
619–620
- trade flows  
effects of preferential agreements on,  
631–633  
normalized outcomes, 627  
outcomes beyond goods trade, 610  
potential, 633
- trade models, 609  
with country-pair, exporter-time, and  
importer-time fixed effects, 632  
cross-sectional (three-way), 617, 619–620  
cross-sectional fixed (exporter and  
importer) country effects models, 632  
cross-sectional gravity models, 622–623  
cross-sectional of country pairs, 610–613  
double-indexed, 611  
empirical topics, 614–634  
with fixed pair and fixed country-time  
effects, 632–633  
gravity models, 608–641  
Hausman and Taylor type models,  
623–624, 624–625  
interpretation of disturbances, 628–629  
Krugman-type, 610  
linearly approximated models, 627–628  
with multiple sectors, 610  
Mundlak-type, 619–620  
nonlinear, 635n11  
outcomes beyond goods trade, 610  
repeated cross-sectional (three-way), 617,  
619–620  
with three-way panels of country pairs,  
613–614  
triple-indexed, 610–611, 629, 630–631
- two-part, 618, 619  
two-way fixed effects model, 611–612  
two-way random effects model, 612–613
- transformation  
based on X-differences, 465  
forward orthogonal difference (FOD),  
120, 390  
Helmert, 390
- transportation research, 364
- t-ratio, 50
- treatment effects, 257–284  
average (ATE), 145n9, 260, 311  
average for non-treated (ATENT), 260  
average treatment effect for treated  
(ATET), 260  
dynamic average (DATE), 271, 276  
dynamic average on the treated (DATET),  
271  
dynamic models of, 270–276  
identification of, 273–275  
local average (LATE), 260  
nonparametric identification of, 261–263,  
268–269  
quantile (QTE), 311  
semiparametric identification of, 264–266  
static models of, 259–270  
trends, incidental, 130–134  
triple-indexed trade models, 610–611, 629,  
630–631  
triple-indexed variables, 630  
“True” Fixed Effects Model, 534  
true quality ( $Q_j$ ), 573  
t-test statistic, 56–57, 63–64  
Tukey bisquare weight function (biweight  
function), 422, 423t, 424f  
Tukey M-estimator, 422, 423t  
tuning parameters, 307  
2SGM (two-stage generalized *M*-estimator),  
432–433, 438  
2SLS estimation, 393–394  
two-part models, 247–249  
two-stage generalized *M*-estimator (2SGM),  
432–433, 438  
two-way (cross-section) data, 617–619  
two-way fixed effects model, 611–612  
two-way random effects model, 612–613

- unbalanced longitudinal data sets, 173  
 unbalanced panel data models with  
   interactive effects, 149–170  
   data generation and implementation, 161–162  
   dynamic case, 160–161, 163–168, 166*t*, 167*t*  
   main findings, 163–168, 164*t*, 165*t*, 166*t*, 167*t*  
   Monte Carlo simulations, 161–168  
   static case, 151–152, 163–168, 164*t*, 165*t*  
 unbalanced panels, 192  
 UNIDO (United Nations Industrial Development Organization), 535–539  
 United Kingdom: cohort studies, 494, 497–502, 498*t*, 503–506, 511, 512n1  
 United Nations Industrial Development Organization (UNIDO), 535–539  
 United States, 494  
 unit roots  
   dynamic models with, 373–374, 386–387, 387*t*  
   testing for, 134–138  
 universal product code (UPC), 548  
 unobservable heterogeneity, 174  
 unobservables, 264–268, 268–270, 600  
 unordered multinomial choice, 174  
 UPC (universal product code), 548  
 U.S. Office of Management and Budget, 602  
 utility, 172–173  
 utility functions, 178  
  
 validation studies, 357–359  
 valid inference, 478  
 value of mortality risk reduction (VMRR), 584  
 value of statistical life (VSL), 584, 585, 586, 601–602, 603nn4–5  
 instrumental variables estimates of, 594, 594*t*  
 linear cross-section and panel data  
   estimates of, 588, 588*t*, 603n5  
 panel quantile estimates of, 592–593, 592*t*  
 short-run estimates of, 599  
 steady state estimates of, 599  
 variable intercept models, 402, 415  
 variance–covariance (VC) matrices, 395  
 variance ratio (VR), 95  
 vector auto-regressive (VAR) models, 467–468  
   global (GVAR), 485–486, 486–487  
 vector error correction model (VECM), 50, 485–486  
 vertical outliers, 418, 419*f*  
 video games, 174  
 VMRR (value of mortality risk reduction), 584  
 voting, majority, 534–535  
 VSL (value of statistical life), 584, 585, 586–587, 601–602, 603nn4–5  
 instrumental variables estimates of, 594, 594*t*  
 linear cross-section and panel data  
   estimates of, 588, 588*t*  
 panel quantile estimates of, 592–593, 592*t*  
 short-run estimates of, 599  
 steady state estimates of, 599  
  
 wage equation, 595  
 wage rates  
   effects of fatal injuries on, 586–587  
   interactive factor model for, 600–601  
   Mincer wage models for, 589  
 Wald tests, 63–64, 97, 99, 192, 392  
 Wales. *see* United Kingdom  
 Watson, M. W., 41  
 waves, more than enough, 223  
 weak cross-sectional dependence, 6, 7–8  
 weak dynamic conditional independence  
   assumption (W-DCIA), 273–274  
 weak factors  
   definition of, 10  
   in errors, 13–14  
   example, 10–11, 12  
   theorem 2, 12–13  
 weak instruments, 82–84  
 weakly serially correlated case, 130–131  
 WGM-estimator (Within GM-estimator), 429  
 Wiener process, 56  
 Wishart distribution, 407  
 Within GM-estimator (WGM-estimator), 429  
 within-group estimator, 458, 459*t*

- Within *MS*-estimator (*WMS*-estimator),  
427–428, 429
- Within-OLS estimator, 48–49
- World Bank, 517
- world economic growth models, 535–539
- World Productivity Database (WPD)  
(UNIDO), 535–539
- World Saving Data Base, 429
- World Trade Organization (WTO), 631
- WPD (World Productivity Database)  
(UNIDO), 535–539
- X-differences, 119, 465
- Young index, 521–522
- zero-inflated models, 248
- zeros, 617–619









