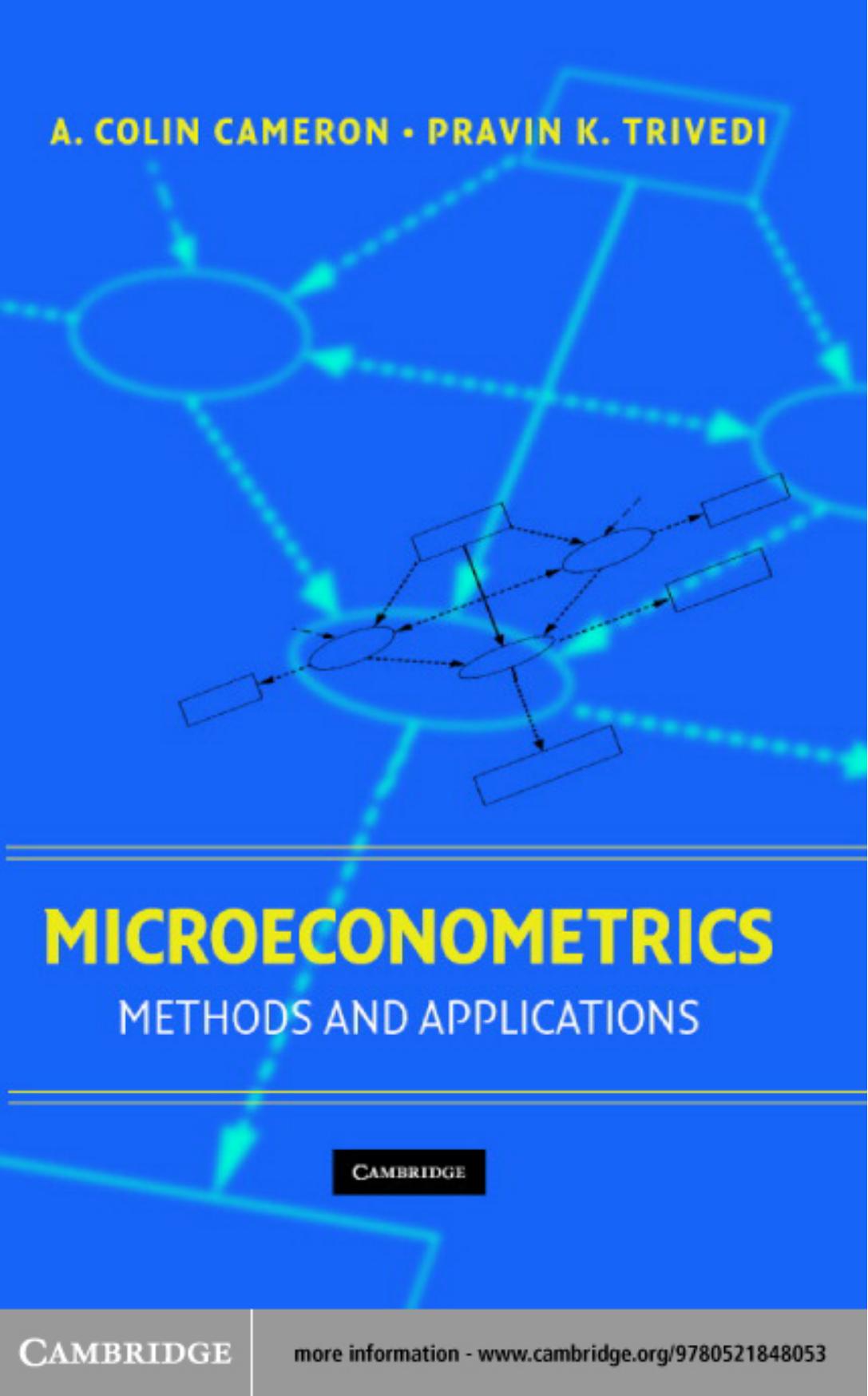


A. COLIN CAMERON • PRAVIN K. TRIVEDI



MICROECONOMETRICS

METHODS AND APPLICATIONS

CAMBRIDGE

CAMBRIDGE

more information - www.cambridge.org/9780521848053

This page intentionally left blank

Microeometrics

This book provides a comprehensive treatment of microeometrics, the analysis of individual-level data on the economic behavior of individuals or firms using regression methods applied to cross-section and panel data. The book is oriented to the practitioner. A good understanding of the linear regression model with matrix algebra is assumed. The text can be used for Ph.D. courses in microeometrics, in applied econometrics, or in data-oriented microeconomics sub-disciplines; and as a reference work for graduate students and applied researchers who wish to fill in gaps in their tool kit. Distinguishing features include emphasis on nonlinear models and robust inference, as well as chapter-length treatments of GMM estimation, nonparametric regression, simulation-based estimation, bootstrap methods, Bayesian methods, stratified and clustered samples, treatment evaluation, measurement error, and missing data. The book makes frequent use of empirical illustrations, many based on seven large and rich data sets.

A. Colin Cameron is Professor of Economics at the University of California, Davis. He currently serves as Director of that university's Center on Quantitative Social Science Research. He has also taught at The Ohio State University and has held short-term visiting positions at Indiana University at Bloomington and at a number of Australian and European universities. His research in microeometrics has appeared in leading econometrics and economics journals. He is coauthor with Pravin Trivedi of *Regression Analysis of Count Data*.

Pravin K. Trivedi is John H. Rudy Professor of Economics at Indiana University at Bloomington. He has also taught at The Australian National University and University of Southampton and has held short-term visiting positions at a number of European universities. His research in microeometrics has appeared in most leading econometrics and health economics journals. He coauthored *Regression Analysis of Count Data* with A. Colin Cameron and is on the editorial boards of the *Econometrics Journal* and the *Journal of Applied Econometrics*.

Microeconometrics

Methods and Applications

A. Colin Cameron
*University of California,
Davis*

Pravin K. Trivedi
Indiana University



CAMBRIDGE UNIVERSITY PRESS

Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press

The Edinburgh Building, Cambridge CB2 2RU, UK

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

www.cambridge.org

Information on this title: www.cambridge.org/9780521848053

© A. Colin Cameron and Pravin K. Trivedi 2005

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

First published in print format 2005

ISBN-13 978-0-521-12581-2 eBook (NetLibrary)

ISBN-10 0-521-12581-X eBook (NetLibrary)

ISBN-13 978-0-521-84805-3 hardback

ISBN-10 0-521-84805-9 hardback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication, and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

*To
my mother and the memory of my father
the memory of my parents*

Contents

List of Figures	<i>page</i>	xv
List of Tables		xvii
Preface		xxi

I Preliminaries

1 Overview	3
1.1 Introduction	3
1.2 Distinctive Aspects of Microeometrics	5
1.3 Book Outline	10
1.4 How to Use This Book	14
1.5 Software	15
1.6 Notation and Conventions	16
2 Causal and Noncausal Models	18
2.1 Introduction	18
2.2 Structural Models	20
2.3 Exogeneity	22
2.4 Linear Simultaneous Equations Model	23
2.5 Identification Concepts	29
2.6 Single-Equation Models	31
2.7 Potential Outcome Model	31
2.8 Causal Modeling and Estimation Strategies	35
2.9 Bibliographic Notes	38
3 Microeconomic Data Structures	39
3.1 Introduction	39
3.2 Observational Data	40
3.3 Data from Social Experiments	48
3.4 Data from Natural Experiments	54

CONTENTS

3.5	Practical Considerations	58
3.6	Bibliographic Notes	61
 II Core Methods		
4	Linear Models	65
4.1	Introduction	65
4.2	Regressions and Loss Functions	66
4.3	Example: Returns to Schooling	69
4.4	Ordinary Least Squares	70
4.5	Weighted Least Squares	81
4.6	Median and Quantile Regression	85
4.7	Model Misspecification	90
4.8	Instrumental Variables	95
4.9	Instrumental Variables in Practice	103
4.10	Practical Considerations	112
4.11	Bibliographic Notes	112
5	Maximum Likelihood and Nonlinear Least-Squares Estimation	116
5.1	Introduction	116
5.2	Overview of Nonlinear Estimators	117
5.3	Extremum Estimators	124
5.4	Estimating Equations	133
5.5	Statistical Inference	135
5.6	Maximum Likelihood	139
5.7	Quasi-Maximum Likelihood	146
5.8	Nonlinear Least Squares	150
5.9	Example: ML and NLS Estimation	159
5.10	Practical Considerations	163
5.11	Bibliographic Notes	163
6	Generalized Method of Moments and Systems Estimation	166
6.1	Introduction	166
6.2	Examples	167
6.3	Generalized Method of Moments	172
6.4	Linear Instrumental Variables	183
6.5	Nonlinear Instrumental Variables	192
6.6	Sequential Two-Step m-Estimation	200
6.7	Minimum Distance Estimation	202
6.8	Empirical Likelihood	203
6.9	Linear Systems of Equations	206
6.10	Nonlinear Sets of Equations	214
6.11	Practical Considerations	219
6.12	Bibliographic Notes	220

CONTENTS

7 Hypothesis Tests	223
7.1 Introduction	223
7.2 Wald Test	224
7.3 Likelihood-Based Tests	233
7.4 Example: Likelihood-Based Hypothesis Tests	241
7.5 Tests in Non-ML Settings	243
7.6 Power and Size of Tests	246
7.7 Monte Carlo Studies	250
7.8 Bootstrap Example	254
7.9 Practical Considerations	256
7.10 Bibliographic Notes	257
8 Specification Tests and Model Selection	259
8.1 Introduction	259
8.2 m-Tests	260
8.3 Hausman Test	271
8.4 Tests for Some Common Misspecifications	274
8.5 Discriminating between Nonnested Models	278
8.6 Consequences of Testing	285
8.7 Model Diagnostics	287
8.8 Practical Considerations	291
8.9 Bibliographic Notes	292
9 Semiparametric Methods	294
9.1 Introduction	294
9.2 Nonparametric Example: Hourly Wage	295
9.3 Kernel Density Estimation	298
9.4 Nonparametric Local Regression	307
9.5 Kernel Regression	311
9.6 Alternative Nonparametric Regression Estimators	319
9.7 Semiparametric Regression	322
9.8 Derivations of Mean and Variance of Kernel Estimators	330
9.9 Practical Considerations	333
9.10 Bibliographic Notes	333
10 Numerical Optimization	336
10.1 Introduction	336
10.2 General Considerations	336
10.3 Specific Methods	341
10.4 Practical Considerations	348
10.5 Bibliographic Notes	352

III Simulation-Based Methods

11 Bootstrap Methods	357
11.1 Introduction	357
11.2 Bootstrap Summary	358
11.3 Bootstrap Example	366
11.4 Bootstrap Theory	368
11.5 Bootstrap Extensions	373
11.6 Bootstrap Applications	376
11.7 Practical Considerations	382
11.8 Bibliographic Notes	382
12 Simulation-Based Methods	384
12.1 Introduction	384
12.2 Examples	385
12.3 Basics of Computing Integrals	387
12.4 Maximum Simulated Likelihood Estimation	393
12.5 Moment-Based Simulation Estimation	398
12.6 Indirect Inference	404
12.7 Simulators	406
12.8 Methods of Drawing Random Variates	410
12.9 Bibliographic Notes	416
13 Bayesian Methods	419
13.1 Introduction	419
13.2 Bayesian Approach	420
13.3 Bayesian Analysis of Linear Regression	435
13.4 Monte Carlo Integration	443
13.5 Markov Chain Monte Carlo Simulation	445
13.6 MCMC Example: Gibbs Sampler for SUR	452
13.7 Data Augmentation	454
13.8 Bayesian Model Selection	456
13.9 Practical Considerations	458
13.10 Bibliographic Notes	458

IV Models for Cross-Section Data

14 Binary Outcome Models	463
14.1 Introduction	463
14.2 Binary Outcome Example: Fishing Mode Choice	464
14.3 Logit and Probit Models	465
14.4 Latent Variable Models	475
14.5 Choice-Based Samples	478
14.6 Grouped and Aggregate Data	480
14.7 Semiparametric Estimation	482

CONTENTS

14.8	Derivation of Logit from Type I Extreme Value	486
14.9	Practical Considerations	487
14.10	Bibliographic Notes	487
15	Multinomial Models	490
15.1	Introduction	490
15.2	Example: Choice of Fishing Mode	491
15.3	General Results	495
15.4	Multinomial Logit	500
15.5	Additive Random Utility Models	504
15.6	Nested Logit	507
15.7	Random Parameters Logit	512
15.8	Multinomial Probit	516
15.9	Ordered, Sequential, and Ranked Outcomes	519
15.10	Multivariate Discrete Outcomes	521
15.11	Semiparametric Estimation	523
15.12	Derivations for MNL, CL, and NL Models	524
15.13	Practical Considerations	527
15.14	Bibliographic Notes	528
16	Tobit and Selection Models	529
16.1	Introduction	529
16.2	Censored and Truncated Models	530
16.3	Tobit Model	536
16.4	Two-Part Model	544
16.5	Sample Selection Models	546
16.6	Selection Example: Health Expenditures	553
16.7	Roy Model	555
16.8	Structural Models	558
16.9	Semiparametric Estimation	562
16.10	Derivations for the Tobit Model	566
16.11	Practical Considerations	568
16.12	Bibliographic Notes	569
17	Transition Data: Survival Analysis	573
17.1	Introduction	573
17.2	Example: Duration of Strikes	574
17.3	Basic Concepts	576
17.4	Censoring	579
17.5	Nonparametric Models	580
17.6	Parametric Regression Models	584
17.7	Some Important Duration Models	591
17.8	Cox PH Model	592
17.9	Time-Varying Regressors	597
17.10	Discrete-Time Proportional Hazards	600
17.11	Duration Example: Unemployment Duration	603

CONTENTS

17.12	Practical Considerations	608
17.13	Bibliographic Notes	608
18	Mixture Models and Unobserved Heterogeneity	611
18.1	Introduction	611
18.2	Unobserved Heterogeneity and Dispersion	612
18.3	Identification in Mixture Models	618
18.4	Specification of the Heterogeneity Distribution	620
18.5	Discrete Heterogeneity and Latent Class Analysis	621
18.6	Stock and Flow Sampling	625
18.7	Specification Testing	628
18.8	Unobserved Heterogeneity Example: Unemployment Duration	632
18.9	Practical Considerations	637
18.10	Bibliographic Notes	637
19	Models of Multiple Hazards	640
19.1	Introduction	640
19.2	Competing Risks	642
19.3	Joint Duration Distributions	648
19.4	Multiple Spells	655
19.5	Competing Risks Example: Unemployment Duration	658
19.6	Practical Considerations	662
19.7	Bibliographic Notes	663
20	Models of Count Data	665
20.1	Introduction	665
20.2	Basic Count Data Regression	666
20.3	Count Example: Contacts with Medical Doctor	671
20.4	Parametric Count Regression Models	674
20.5	Partially Parametric Models	682
20.6	Multivariate Counts and Endogenous Regressors	685
20.7	Count Example: Further Analysis	690
20.8	Practical Considerations	690
20.9	Bibliographic Notes	691
V	Models for Panel Data	
21	Linear Panel Models: Basics	697
21.1	Introduction	697
21.2	Overview of Models and Estimators	698
21.3	Linear Panel Example: Hours and Wages	708
21.4	Fixed Effects versus Random Effects Models	715
21.5	Pooled Models	720
21.6	Fixed Effects Model	726
21.7	Random Effects Model	734

CONTENTS

21.8	Modeling Issues	737
21.9	Practical Considerations	740
21.10	Bibliographic Notes	740
22	Linear Panel Models: Extensions	743
22.1	Introduction	743
22.2	GMM Estimation of Linear Panel Models	744
22.3	Panel GMM Example: Hours and Wages	754
22.4	Random and Fixed Effects Panel GMM	756
22.5	Dynamic Models	763
22.6	Difference-in-Differences Estimator	768
22.7	Repeated Cross Sections and Pseudo Panels	770
22.8	Mixed Linear Models	774
22.9	Practical Considerations	776
22.10	Bibliographic Notes	777
23	Nonlinear Panel Models	779
23.1	Introduction	779
23.2	General Results	779
23.3	Nonlinear Panel Example: Patents and R&D	762
23.4	Binary Outcome Data	795
23.5	Tobit and Selection Models	800
23.6	Transition Data	801
23.7	Count Data	802
23.8	Semiparametric Estimation	808
23.9	Practical Considerations	808
23.10	Bibliographic Notes	809
VI Further Topics		
24	Stratified and Clustered Samples	813
24.1	Introduction	813
24.2	Survey Sampling	814
24.3	Weighting	817
24.4	Endogenous Stratification	822
24.5	Clustering	829
24.6	Hierarchical Linear Models	845
24.7	Clustering Example: Vietnam Health Care Use	848
24.8	Complex Surveys	853
24.9	Practical Considerations	857
24.10	Bibliographic Notes	857
25	Treatment Evaluation	860
25.1	Introduction	860
25.2	Setup and Assumptions	862

CONTENTS

25.3	Treatment Effects and Selection Bias	865
25.4	Matching and Propensity Score Estimators	871
25.5	Differences-in-Differences Estimators	878
25.6	Regression Discontinuity Design	879
25.7	Instrumental Variable Methods	883
25.8	Example: The Effect of Training on Earnings	889
25.9	Bibliographic Notes	896
26	Measurement Error Models	899
26.1	Introduction	899
26.2	Measurement Error in Linear Regression	900
26.3	Identification Strategies	905
26.4	Measurement Errors in Nonlinear Models	911
26.5	Attenuation Bias Simulation Examples	919
26.6	Bibliographic Notes	920
27	Missing Data and Imputation	923
27.1	Introduction	923
27.2	Missing Data Assumptions	925
27.3	Handling Missing Data without Models	928
27.4	Observed-Data Likelihood	929
27.5	Regression-Based Imputation	930
27.6	Data Augmentation and MCMC	932
27.7	Multiple Imputation	934
27.8	Missing Data MCMC Imputation Example	935
27.9	Practical Considerations	939
27.10	Bibliographic Notes	940
A	Asymptotic Theory	943
A.1	Introduction	943
A.2	Convergence in Probability	944
A.3	Laws of Large Numbers	947
A.4	Convergence in Distribution	948
A.5	Central Limit Theorems	949
A.6	Multivariate Normal Limit Distributions	951
A.7	Stochastic Order of Magnitude	954
A.8	Other Results	955
A.9	Bibliographic Notes	956
B	Making Pseudo-Random Draws	957
	References	961
	Index	999

List of Figures

3.1	Social experiment with random assignment	<i>page</i> 50
4.1	Quantile regression estimates of slope coefficient	89
4.2	Quantile regression estimated lines	90
7.1	Power of Wald chi-square test	249
7.2	Density of Wald test on slope coefficient	253
9.1	Histogram for log wage	296
9.2	Kernel density estimates for log wage	296
9.3	Nonparametric regression of log wage on education	297
9.4	Kernel density estimates using different kernels	300
9.5	k -nearest neighbors regression	309
9.6	Nonparametric regression using Lowess	310
9.7	Nonparametric estimate of derivative of y with respect to x	317
11.1	Bootstrap estimate of the density of t -test statistic	368
12.1	Halton sequence draws compared to pseudo-random draws	411
12.2	Inverse transformation method for unit exponential draws	413
12.3	Accept–reject method for random draws	414
13.1	Bayesian analysis for mean parameter of normal density	424
14.1	Charter boat fishing: probit and logit predictions	466
15.1	Generalized random utility model	516
16.1	Tobit regression example	531
16.2	Inverse Mills ratio as censoring point c increases	540
17.1	Strike duration: Kaplan–Meier survival function	575
17.2	Weibull distribution: density, survivor, hazard, and cumulative hazard functions	585
17.3	Unemployment duration: Kaplan–Meier survival function	604
17.4	Unemployment duration: survival functions by unemployment insurance	605
17.5	Unemployment duration: Nelson–Aalen cumulated hazard function	606
17.6	Unemployment duration: cumulative hazard function by unemployment insurance	606

LIST OF FIGURES

18.1	Length-biased sampling under stock sampling: example	627
18.2	Unemployment duration: exponential model generalized residuals	633
18.3	Unemployment duration: exponential-gamma model generalized residuals	633
18.4	Unemployment duration: Weibull model generalized residuals	635
18.5	Unemployment duration: Weibull-IG model generalized residuals	636
19.1	Unemployment duration: Cox CR baseline survival functions	661
19.2	Unemployment duration: Cox CR baseline cumulative hazards	662
21.1	Hours and wages: pooled (overall) regression	712
21.2	Hours and wages: between regression	713
21.3	Hours and wages: within (fixed effects) regression	713
21.4	Hours and wages: first differences regression	714
23.1	Patents and R&D: pooled (overall) regression	793
25.1	Regression-discontinuity design: example	880
25.2	RD design: treatment assignment in sharp and fuzzy designs	883
25.3	Training impact: earnings against propensity score by treatment	892
27.1	Missing data: examples of missing regressors	924

List of Tables

1.1	Book Outline	<i>page</i> 11
1.2	Outline of a 20-Lecture 10-Week Course	15
1.3	Commonly Used Acronyms and Abbreviations	17
3.1	Features of Some Selected Social Experiments	51
3.2	Features of Some Selected Natural Experiments	54
4.1	Loss Functions and Corresponding Optimal Predictors	67
4.2	Least Squares Estimators and Their Asymptotic Variance	83
4.3	Least Squares: Example with Conditionally Heteroskedastic Errors	84
4.4	Instrumental Variables Example	103
4.5	Returns to Schooling: Instrumental Variables Estimates	111
5.1	Asymptotic Properties of M-Estimators	121
5.2	Marginal Effect: Three Different Estimates	122
5.3	Maximum Likelihood: Commonly Used Densities	140
5.4	Linear Exponential Family Densities: Leading Examples	148
5.5	Nonlinear Least Squares: Common Examples	151
5.6	Nonlinear Least-Squares Estimators and Their Asymptotic Variance	156
5.7	Exponential Example: Least-Squares and ML Estimates	161
6.1	Generalized Method of Moments: Examples	172
6.2	GMM Estimators in Linear IV Model and Their Asymptotic Variance	186
6.3	GMM Estimators in Nonlinear IV Model and Their Asymptotic Variance	195
6.4	Nonlinear Two-Stage Least-Squares Example	199
7.1	Test Statistics for Poisson Regression Example	242
7.2	Wald Test Size and Power for Probit Regression Example	253
8.1	Specification m-Tests for Poisson Regression Example	270
8.2	Nonnested Model Comparisons for Poisson Regression Example	284
8.3	Pseudo R^2 's: Poisson Regression Example	291
9.1	Kernel Functions: Commonly Used Examples	300
9.2	Semiparametric Models: Leading Examples	323
10.1	Gradient Method Results	339
10.2	Computational Difficulties: A Partial Checklist	350

LIST OF TABLES

11.1	Bootstrap Statistical Inference on a Slope Coefficient: Example	367
11.2	Bootstrap Theory Notation	369
12.1	Monte Carlo Integration: Example for x Standard Normal	392
12.2	Maximum Simulated Likelihood Estimation: Example	398
12.3	Method of Simulated Moments Estimation: Example	404
13.1	Bayesian Analysis: Essential Components	425
13.2	Conjugate Families: Leading Examples	428
13.3	Gibbs Sampling: Seemingly Unrelated Regressions Example	454
13.4	Interpretation of Bayes Factors	457
14.1	Fishing Mode Choice: Data Summary	464
14.2	Fishing Mode Choice: Logit and Probit Estimates	465
14.3	Binary Outcome Data: Commonly Used Models	467
15.1	Fishing Mode Multinomial Choice: Data Summary	492
15.2	Fishing Mode Multinomial Choice: Logit Estimates	493
15.3	Fishing Mode Choice: Marginal Effects for Conditional Logit Model	493
16.1	Health Expenditure Data: Two-Part and Selection Models	554
17.1	Survival Analysis: Definitions of Key Concepts	577
17.2	Hazard Rate and Survivor Function Computation: Example	582
17.3	Strike Duration: Kaplan–Meier Survivor Function Estimates	583
17.4	Exponential and Weibull Distributions: pdf, cdf, Survivor Function, Hazard, Cumulative Hazard, Mean, and Variance	584
17.5	Standard Parametric Models and Their Hazard and Survivor Functions	585
17.6	Unemployment Duration: Description of Variables	603
17.7	Unemployment Duration: Kaplan–Meier Survival and Nelson–Aalen Cumulated Hazard Functions	605
17.8	Unemployment Duration: Estimated Parameters from Four Parametric Models	607
17.9	Unemployment Duration: Estimated Hazard Ratios from Four Parametric Models	608
18.1	Unemployment Duration: Exponential Model with Gamma and IG Heterogeneity	634
18.2	Unemployment Duration: Weibull Model with and without Heterogeneity	635
19.1	Some Standard Copula Functions	654
19.2	Unemployment Duration: Competing and Independent Risk Estimates of Exponential Model with and without IG Frailty	659
19.3	Unemployment Duration: Competing and Independent Risk Estimates of Weibull Model with and without IG Frailty	660
20.1	Proportion of Zero Counts in Selected Empirical Studies	666
20.2	Summary of Data Sets Used in Recent Patent–R&D Studies	667
20.3	Contacts with Medical Doctor: Frequency Distribution	672
20.4	Contacts with Medical Doctor: Variable Descriptions	672
20.5	Contacts with Medical Doctor: Count Model Estimates	673
20.6	Contacts with Medical Doctor: Observed and Fitted Frequencies	674

LIST OF TABLES

21.1	Linear Panel Model: Common Estimators and Models	699
21.2	Hours and Wages: Standard Linear Panel Model Estimators	710
21.3	Hours and Wages: Autocorrelations of Pooled OLS Residuals	714
21.4	Hours and Wages: Autocorrelations of Within Regression Residuals	715
21.5	Pooled Least-Squares Estimators and Their Asymptotic Variances	721
21.6	Variances of Pooled OLS Estimator with Equicorrelated Errors	724
21.7	Hours and Wages: Pooled OLS and GLS Estimates	725
22.1	Panel Exogeneity Assumptions and Resulting Instruments	752
22.2	Hours and Wages: GMM-IV Linear Panel Model Estimators	755
23.1	Patents and R&D Spending: Nonlinear Panel Model Estimators	794
24.1	Stratification Schemes with Random Sampling within Strata	823
24.2	Properties of Estimators for Different Clustering Models	832
24.3	Vietnam Health Care Use: Data Description	850
24.4	Vietnam Health Care Use: FE and RE Models for Positive Expenditure	851
24.5	Vietnam Health Care Use: Frequencies for Pharmacy Visits	852
24.6	Vietnam Health Care Use: RE and FE Models for Pharmacy Visits	852
25.1	Treatment Effects Framework	865
25.2	Treatment Effects Measures: ATE and ATET	868
25.3	Training Impact: Sample Means in Treated and Control Samples	890
25.4	Training Impact: Various Estimates of Treatment Effect	891
25.5	Training Impact: Distribution of Propensity Score for Treated and Control Units Using DW (1999) Specification	894
25.6	Training Impact: Estimates of ATET	895
25.7	Training Evaluation: DW (2002) Estimates of ATET	896
26.1	Attenuation Bias in a Logit Regression with Measurement Error	919
26.2	Attenuation Bias in a Nonlinear Regression with Additive Measurement Error	920
27.1	Relative Efficiency of Multiple Imputation	935
27.2	Missing Data Imputation: Linear Regression Estimates with 10% Missing Data and High Correlation Using MCMC Algorithm	936
27.3	Missing Data Imputation: Linear Regression Estimates with 25% Missing Data and High Correlation Using MCMC Algorithm	937
27.4	Missing Data Imputation: Linear Regression Estimates with 10% Missing Data and Low Correlation Using MCMC Algorithm	937
27.5	Missing Data Imputation: Logistic Regression Estimates with 10% Missing Data and High Correlation Using MCMC Algorithm	938
27.6	Missing Data Imputation: Logistic Regression Estimates with 25% Missing Data and Low Correlation Using MCMC Algorithm	939
A.1	Asymptotic Theory: Definitions and Theorems	944
B.1	Continuous Random Variable Densities and Moments	957
B.2	Continuous Random Variable Generators	958
B.3	Discrete Random Variable Probability Mass Functions and Moments	959
B.4	Discrete Random Variable Generators	959

Preface

This book provides a detailed treatment of microeconometric analysis, the analysis of individual-level data on the economic behavior of individuals or firms. This type of analysis usually entails applying regression methods to cross-section and panel data.

The book aims at providing the practitioner with a comprehensive coverage of statistical methods and their application in modern applied microeconomics research. These methods include nonlinear modeling, inference under minimal distributional assumptions, identifying and measuring causation rather than mere association, and correcting departures from simple random sampling. Many of these features are of relevance to individual-level data analysis throughout the social sciences.

The ambitious agenda has determined the characteristics of this book. First, although oriented to the practitioner, the book is relatively advanced in places. A cookbook approach is inadequate because when two or more complications occur simultaneously – a common situation – the practitioner must know enough to be able to adapt available methods. Second, the book provides considerable coverage of practical data problems (see especially the last three chapters). Third, the book includes substantial empirical examples in many chapters to illustrate some of the methods covered. Finally, the book is unusually long. Despite this length we have been space-constrained. We had intended to include even more empirical examples, and abbreviated presentations will at times fail to recognize the accomplishments of researchers who have made substantive contributions.

The book assumes a good understanding of the linear regression model with matrix algebra. It is written at the mathematical level of the first-year economics Ph.D. sequence, comparable to Greene (2003). We have two types of readers in mind. First, the book can be used as a course text for a microeometrics course, typically taught in the second year of the Ph.D., or for data-oriented microeconomics field courses such as labor economics, public economics, and industrial organization. Second, the book can be used as a reference work for graduate students and applied researchers who despite training in microeconometrics will inevitably have gaps that they wish to fill.

For instructors using this book as an econometrics course text it is best to introduce the basic nonlinear cross-section and linear panel data models as early as possible,

initially skipping many of the methods chapters. The key methods chapter (Chapter 5) covers maximum-likelihood and nonlinear least-squares estimation. Knowledge of maximum likelihood and nonlinear least-squares estimators provides adequate background for the most commonly used nonlinear cross-section models (Chapters 14–17 and 20), basic linear panel data models (Chapter 21), and treatment evaluation methods (Chapter 25). Generalized method of moments estimation (Chapter 6) is needed especially for advanced linear panel data methods (Chapter 22).

For readers using this book as a reference work, the chapters have been written to be as self-contained as possible. The notable exception is that some command of general estimation results in Chapter 5, and occasionally Chapter 6, will be necessary. Most chapters on models are structured to begin with a discussion and example that is accessible to a wide audience.

The Web site www.econ.ucdavis.edu/faculty/cameron provides all the data and computer programs used in this book and related materials useful for instructional purposes.

This project has been long and arduous, and at times seemingly without an end. Its completion has been greatly aided by our colleagues, friends, and graduate students. We thank especially the following for reading and commenting on specific chapters: Bijan Borah, Kurt Brännäs, Pian Chen, Tim Cogley, Partha Deb, Massimiliano De Santis, David Drukker, Jeff Gill, Tue Gorgens, Shiferaw Gurmu, Lu Ji, Oscar Jorda, Roger Koenker, Chenghui Li, Tong Li, Doug Miller, Murat Munkin, Jim Prieger, Ahmed Rahmen, Sunil Sapra, Haruki Seitani, Yacheng Sun, Xiaoyong Zheng, and David Zimmer. Pian Chen gave detailed comments on most of the book. We thank Rajeev Dehejia, Bronwyn Hall, Cathy Kling, Jeffrey Kling, Will Manning, Brian McCall, and Jim Ziliak for making their data available for empirical illustrations. We thank our respective departments for facilitating our collaboration and for the production and distribution of the draft manuscript at various stages. We benefited from the comments of two anonymous reviewers. Guidance, advice, and encouragement from our Cambridge editor, Scott Parris, have been invaluable.

Our interest in econometrics owes much to the training and environments we encountered as students and in the initial stages of our academic careers. The first author thanks The Australian National University; Stanford University, especially Takeshi Amemiya and Tom MacCurdy; and The Ohio State University. The second author thanks the London School of Economics and The Australian National University.

Our interest in writing a book oriented to the practitioner owes much to our exposure to the research of graduate students and colleagues at our respective institutions, UC-Davis and IU-Bloomington.

Finally, we thank our families for their patience and understanding without which completion of this project would not have been possible.

A. Colin Cameron
Davis, California

Pravin K. Trivedi
Bloomington, Indiana

PART ONE

Preliminaries

Part 1 covers the essential components of microeconometric analysis – an economic specification, a statistical model and a data set.

Chapter 1 discusses the distinctive aspects of microeconomics, and provides an outline of the book. It emphasizes that discreteness of data, and nonlinearity and heterogeneity of behavioral relationships are key aspects of individual-level microeconomic models. It concludes by presenting the notation and conventions used throughout the book.

Chapters 2 and 3 set the scene for the remainder of the book by introducing the reader to key model and data concepts that shape the analyses of later chapters.

A key distinction in econometrics is between essentially descriptive models and data summaries at various levels of statistical sophistication and models that go beyond associations and attempt to estimate causal parameters. The classic definitions of causality in econometrics derive from the Cowles Commission simultaneous equations models that draw sharp distinctions between exogenous and endogenous variables, and between structural and reduced form parameters. Although reduced form models are very useful for some purposes, knowledge of structural or causal parameters is essential for policy analyses. Identification of structural parameters within the simultaneous equations framework poses numerous conceptual and practical difficulties. An increasingly-used alternative approach based on the potential outcome model, also attempts to identify causal parameters but it does so by posing limited questions within a more manageable framework. Chapter 2 attempts to provide an overview of the fundamental issues that arise in these and other alternative frameworks. Readers who initially find this material challenging should return to this chapter after gaining greater familiarity with specific models covered later in the book.

The empirical researcher's ability to identify causal parameters depends not only on the statistical tools and models but also on the type of data available. An experimental framework provides a standard for establishing causal connections. However, observational, not experimental, data form the basis of much of econometric inference. Chapter 3 surveys the pros and cons of three main types of data: observational data, data from social experiments, and data from natural experiments. The strengths and weaknesses of conducting causal inference based on each type of data are reviewed.

CHAPTER 1

Overview

1.1. Introduction

This book provides a detailed treatment of **microeconometric analysis**, the analysis of individual-level data on the economic behavior of individuals or firms. A broader definition would also include grouped data. Usually regression methods are applied to cross-section or panel data.

Analysis of individual data has a long history. Ernst Engel (1857) was among the earliest quantitative investigators of household budgets. Allen and Bowley (1935), Houthakker (1957), and Prais and Houthakker (1955) made important contributions following the same research and modeling tradition. Other landmark studies that were also influential in stimulating the development of microeconomics, even though they did not always use individual-level information, include those by Marschak and Andrews (1944) in production theory and by Wold and Jureen (1953), Stone (1953), and Tobin (1958) in consumer demand.

As important as the above earlier cited work is on household budgets and demand analysis, the material covered in this book has stronger connections with the work on discrete choice analysis and censored and truncated variable models that saw their first serious econometric applications in the work of McFadden (1973, 1984) and Heckman (1974, 1979), respectively. These works involved a major departure from the overwhelming reliance on linear models that characterized earlier work. Subsequently, they have led to significant methodological innovations in econometrics. Among the earlier textbook-level treatments of this material (and more) are the works of Maddala (1983) and Amemiya (1985). As emphasized by Heckman (2001), McFadden (2001), and others, many of the fundamental issues that dominated earlier work based on market data remain important, especially concerning the conditions necessary for identifiability of causal economic relations. Nonetheless, the style of microeconomics is sufficiently distinct to justify writing a text that is exclusively devoted to it.

Modern microeconomics based on individual-, household-, and establishment-level data owes a great deal to the greater availability of data from cross-section and longitudinal sample surveys and census data. In the past two decades, with the

expansion of electronic recording and collection of data at the individual level, data volume has grown explosively. So too has the available computing power for analyzing large and complex data sets. In many cases event-level data are available; for example, marketing science often deals with purchase data collected by electronic scanners in supermarkets, and industrial organization literature contains econometric analyses of airline travel data collected by online booking systems. There are now new branches of economics, such as social experimentation and experimental economics, that generate “experimental” data. These developments have created many new modeling opportunities that are absent when only aggregated market-level data are available. Meanwhile the explosive growth in the volume and types of data has also given rise to numerous methodological issues. Processing and econometric analysis of such large microdatabases, with the objective of uncovering patterns of economic behavior, constitutes the core of microeometrics. Econometric analysis of such data is the subject matter of this book.

Key precursors of this book are the books by Maddala (1983) and Amemiya (1985). Like them it covers topics that are presented only briefly, or not at all, in undergraduate and first-year graduate econometrics courses. Especially compared to Amemiya (1985) this book is more oriented to the practitioner. The level of presentation is nonetheless advanced in places, especially for applied researchers in disciplines that are less mathematically oriented than economics.

A relatively advanced presentation is needed for several reasons. First, the data are often discrete or censored, in which case **nonlinear methods** such as logit, probit, and Tobit models are used. This leads to statistical inference based on more difficult asymptotic theory.

Second, **distributional assumptions** for such data become critically important. One response is to develop highly parametric models that are sufficiently detailed to capture the complexities of data, but these models can be challenging to estimate. A more common response is to minimize parametric assumptions and perform statistical inference based on standard errors that are “robust” to complications such as heteroskedasticity and clustering. In such cases considerable knowledge can be needed to ensure valid statistical inference even if a standard regression package is used.

Third, economic studies often aim to determine **causation** rather than merely measure correlation, despite access to observational rather than experimental data. This leads to methods to isolate causation such as instrumental variables, simultaneous equations, measurement error correction, selection bias correction, panel data fixed effects, and differences-in-differences.

Fourth, microeconomic data are typically collected using cross-section and panel surveys, censuses, or social experiments. **Survey data** collected using these methods are subject to problems of complex survey methodology, departures from simple random sampling assumptions, and problems of sample selection, measurement errors, and incomplete, and/or missing data. Dealing with such issues in a way that can support valid population inferences from the estimated econometric models population requires use of advanced methods.

Finally, it is not unusual that two or more **complications occur simultaneously**, such as endogeneity in a logit model with panel data. Then a cookbook approach

becomes very difficult to implement. Instead, considerable understanding of the theory underlying the methods is needed, as the researcher may need to read econometrics journal articles and adapt standard econometrics software.

1.2. Distinctive Aspects of Microeconomics

We now consider several advantages of microeconomics that derive from its distinctive features.

1.2.1. Discreteness and Nonlinearity

The first and most obvious point is that microeconomic data are usually at a low level of aggregation. This has a major consequence for the functional forms used to analyze the variables of interest. In many, if not most, cases linear functional forms turn out to be simply inappropriate. More fundamentally, disaggregation brings to the forefront **heterogeneity** of individuals, firms, and organizations that should be properly controlled (modeled) if one wants to make valid inferences about the underlying relationships. We discuss these issues in greater detail in the following sections.

Although aggregation is not entirely absent in microdata, as for example when household- or establishment-level data are collected, the level of aggregation is usually orders of magnitude lower than is common in macro analyses. In the latter case the process of aggregation leads to smoothing, with many of the movements in opposite directions canceling in the course of summation. The aggregated variables often show smoother behavior than their components, and the relationships between the aggregates frequently show greater smoothness than the components. For example, a relation between two variables at a micro level may be piecewise linear with many nodes. After aggregation the relationship is likely to be well approximated by a smooth function. Hence an immediate consequence of disaggregation is the absence of features of continuity and smoothness both of the variables themselves and of the relationships between them.

Usually individual- and firm-level data cover a huge range of variation, both in the cross-section and time-series dimensions. For example, average weekly consumption of (say) beef is highly likely to be positive and smoothly varying, whereas that of an individual household in a given week may be frequently zero and may also switch to positive values from time to time. The average number of hours worked by female workers is unlikely to be zero, but many individual females have zero market hours of work (corner solutions), switching to positive values at other times in the course of their labor market history. Average household expenditure on vacations is usually positive, but many individual households may have zero expenditure on vacations in any given year. Average per capita consumption of tobacco products will usually be positive, but many individuals in the population have never consumed these products and never will, irrespective of price and income considerations. As Pudney (1989) has observed, microdata exhibit “holes, kinks and corners.” The holes correspond to nonparticipation in the activity of interest, kinks correspond to the switching behavior, and corners correspond

to the incidence of nonconsumption or nonparticipation at specific points of time. That is, discreteness and nonlinearity of response are intrinsic to microeconomics.

An important class of nonlinear models in microeconomics deals with **limited dependent variables** (Maddala, 1983). This class includes many models that provide suitable frameworks for analyzing discrete responses and responses with limited range of variation. Such tools of analyses are of course also available for analyzing macro-data, if required. The point is that they are indispensable in microeconomics and give it its distinctive feature.

1.2.2. Greater Realism

Macroeconomics is sometimes based on strong assumptions; the representative agent assumption is a leading example. A frequent appeal is made to microeconomic reasoning to justify certain specifications and interpretations of empirical results. However, it is rarely possible to say explicitly how these are affected by aggregation over time and micro units. Alternatively, very extreme aggregation assumptions are made. For example, aggregates are said to reflect the behavior of a hypothetical representative agent. Such assumptions also are not credible.

From the viewpoint of microeconomic theory, quantitative analysis founded on microdata may be regarded as more realistic than that based on aggregated data. There are three justifications for this claim. First, the measurement of the variables involved in such hypotheses is often more direct (though not necessarily free from measurement error) and has greater correspondence to the theory being tested. Second, hypotheses about economic behavior are usually developed from theories of individual behavior. If these hypotheses are tested using aggregated data, then many approximations and simplifying assumptions have to be made. The simplifying assumption of a representative agent causes a great loss of information and severely limits the scope of an empirical investigation. Because such assumptions can be avoided in microeconomics, and usually are, in principle the microdata provide a more realistic framework for testing microeconomic hypotheses. This is not a claim that the promise of microdata is necessarily achieved in empirical work. Such a claim must be assessed on a case-by-case basis. Finally, a realistic portrayal of economic activity should accommodate a broad range of outcomes and responses that are a consequence of individual heterogeneity and that are predicted by underlying theory. In this sense microeconomic data sets can support more realistic models.

Microeconomic data are often derived from household or firm surveys, typically encompassing a wide range of behavior, with many of the behavioral outcomes taking the form of discrete or categorical responses. Such data sets have many awkward features that call for special tools in the formulation and analysis that, although not entirely absent from macroeconomic work, nevertheless are less widely used.

1.2.3. Greater Information Content

The potential advantages of microdata sets can be realized if such data are informative. Because sample surveys often provide independent observations on thousands of

cross-sectional units, such data are thought to be more informative than the standard, usually highly serially correlated, macro time series typically consisting of at most a few hundred observations.

As will be explained in the next chapter, in practice the situation is not so clear-cut because the microdata may be quite noisy. At the individual level many (idiosyncratic) factors may play a large role in determining responses. Often these cannot be observed, leading one to treat them under the heading of a random component, which can be a very large part of observed variation. In this sense randomness plays a larger role in microdata than in macrodata. Of course, this affects measures of goodness of fit of the regressions. Students whose initial exposure to econometrics comes through aggregate time-series analysis are often conditioned to see large R^2 values. When encountering cross-section regressions for the first time, they express disappointment or even alarm at the “low explanatory power” of the regression equation. Nevertheless, there remains a strong presumption that, at least in certain dimensions, large microdata sets are highly informative.

Another qualification is that when one is dealing with purely cross-section data, very little can be said about the intertemporal aspects of relationships under study. This particular aspect of behavior can be studied using panel and transition data.

In many cases one is interested in the behavioral responses of a specific group of economic agents under some specified economic environment. One example is the impact of unemployment insurance on the job search behavior of young unemployed persons. Another example is the labor supply responses of low-income individuals receiving income support. Unless microdata are used such issues cannot be addressed directly in empirical work.

1.2.4. Microeconomic Foundations

Econometric models vary in the explicit role given to economic theory. At one end of the spectrum there are models in which the *a priori* theorizing may play a dominant role in the specification of the model and in the choice of an estimation procedure. At the other end of the spectrum are empirical investigations that make much less use of economic theory.

The goal of the analysis in the first case is to identify and estimate fundamental parameters, sometimes called deep parameters, that characterize individual taste and preferences and/or technological relationships. As a shorthand designation, we call this the **structural approach**. Its hallmark is a heavy dependence on economic theory and emphasis on causal inference. Such models may require many assumptions, such as the precise specification of a cost or production function or specification of the distribution of error terms. The empirical conclusions of such an exercise may not be robust with respect to the departures from the assumptions. In Section 2.4.4 we shall say more about this approach. At the present stage we simply emphasize that if the structural approach is implemented with aggregated data, it will yield estimates of the fundamental parameters only under very stringent (and possibly unrealistic) conditions. Microdata sets provide a more promising environment for the structural approach, essentially because they permit greater flexibility in model specification.

The goal of the analysis in the second case is to model relationship(s) between response variables of interest conditionally on variables the researcher takes as given, or exogenous. More formal definitions of **endogeneity** and **exogeneity** are given in Chapter 2. As a shorthand designation, we call this a **reduced form approach**. The essential point is that reduced form analysis does not always take into account all causal interdependencies. A regression model in which the focus is on the prediction of y given regressors \mathbf{x} , and not on the causal interpretation of the regression parameters, is often referred to as a reduced form regression. As will be explained in Chapter 2, in general the parameters of the reduced form model are functions of structural parameters. They may not be interpretable without some information about the structural parameters.

1.2.5. Disaggregation and Heterogeneity

It is sometimes said that many problems and issues of macroeconomics arise from serial correlation of macro time series, and those of microeconomics arise from heteroskedasticity of individual-level data. Although this is a useful characterization of the modeling effort in many microeconomic analyses, it needs amplification and is subject to important qualifications. In a range of microeconomic models, modeling of dynamic dependence may be an important issue.

The benefits of disaggregation, which were emphasized earlier in this section, come at a cost: As the data become more disaggregated the importance of controlling for interindividual heterogeneity increases. Heterogeneity, or more precisely unobserved heterogeneity, plays a very important role in microeconomics. Obviously, many variables that reflect interindividual heterogeneity, such as gender, race, educational background, and social and demographic factors, are directly observed and hence can be controlled for. In contrast, differences in individual motivation, ability, intelligence, and so forth are either not observed or, at best, imperfectly observed.

The simplest response is to ignore such heterogeneity, that is, to absorb it into the regression disturbance. After all this is how one treats the myriad small unobserved factors. This step of course increases the unexplained part of the variation. More seriously, ignoring persistent interindividual differences leads to **confounding** with other factors that are also sources of persistent interindividual differences. Confounding is said to occur when the individual contributions of different regressors (predictor variables) to the variation in the variable of interest cannot be statistically separated. Suppose, for example, that the factor x_1 (schooling) is said to be the source of variation in y (earnings), when another variable x_2 (ability), which is another source of variation, does not appear in the model. Then that part of total variation that is attributable to the second variable may be incorrectly attributed to the first variable. Intuitively, their relative importances are confounded. A leading source of confounding bias is the incorrect omission of regressors from the model and the inclusion of other variables that are proxies for the omitted variable.

Consider, for example, the case in which a program participation (0/1 dummy) variable D is included in the regression mean function with a vector of regressors \mathbf{x} ,

$$y = \mathbf{x}'\beta + \alpha D + u, \quad (1.1)$$

where u is an error term. The term “treatment” is used in biological and experimental sciences to refer to an administered regimen involving participants in some trial. In econometrics it commonly refers to participation in some activity that may impact an outcome of interest. This activity may be randomly assigned to the participants or may be self-selected by the participant. Thus, although it is acknowledged that individuals choose their years of schooling, one still thinks of years of schooling as a “treatment” variable. Suppose that program participation is taken to be a discrete variable. The coefficient α of the “treatment variable” measures the average impact of the program participation ($D = 1$), conditional on covariates. If one does not control for unobserved heterogeneity, then a potential ambiguity affects the interpretation of the results. If d is found to have a significant impact, then the following question arises: Is α significantly different from zero because D is correlated with some unobserved variable that affects y or because there is a causal relationship between D and y ? For example, if the program considered is university education, and the covariates do not include a measure of ability, giving a fully causal interpretation becomes questionable. Because the issue is important, more attention should be given to how to control for unobserved heterogeneity.

In some cases where dynamic considerations are involved the type of data available may put restrictions on how one can control for heterogeneity. Consider the example of two households, identical in all relevant respects except that one exhibits a systematically higher preference for consuming good A. One could control for this by allowing individual utility functions to include a heterogeneity parameter that reflects their different preferences. Suppose now that there is a theory of consumer behavior that claims that consumers become addicted to good A, in the sense that the more they consume of it in one period, the greater is the probability that they will consume more of it in the future. This theory provides another explanation of persistent interindividual differences in the consumption of good A. By controlling for heterogeneous preferences it becomes possible to test which source of persistence in consumption – preference heterogeneity or addiction – accounts for different consumption patterns. This type of problem arises whenever some dynamic element generates persistence in the observed outcomes. Several examples of this type of problem arise in various places in the book.

A variety of approaches for modeling heterogeneity coexist in microeconomics. A brief mention of some of these follows, with details postponed until later.

An extreme solution is to ignore all unobserved interindividual differences. If unobserved heterogeneity is uncorrelated with observed heterogeneity, and if the outcome being studied has no intertemporal dependence, then the aforementioned problems will not arise. Of course, these are strong assumptions and even with these assumptions not all econometric difficulties disappear.

One approach for handling heterogeneity is to treat it as a **fixed effect** and to estimate it as a coefficient of an individual specific 0/1 dummy variable. For example, in a cross-section regression, each micro unit is allowed its own dummy variable (intercept). This leads to an extreme proliferation of parameters because when a new individual is added to the sample, a new intercept parameter is also added. Thus this approach will not work if our data are cross sectional. The availability of multiple observations

per individual unit, most commonly in the form of panel data with T time-series observations for each of the N cross-section units, makes it possible to either estimate or eliminate the fixed effect, for example by first differencing if the model is linear and the fixed effect is additive. If the model is nonlinear, as is often the case, the fixed effect will usually not be additive and other approaches will need to be considered.

A second approach to modeling unobserved heterogeneity is through a **random effects** model. There are a number of ways in which the random effects model can be formulated. One popular formulation assumes that one or more regression parameters, often just the regression intercept, varies randomly across the cross section. In another formulation the regression error is given a component structure, with an individual specific random component. The random effects model then attempts to estimate the parameters of the distribution from which the random component is drawn. In some cases, such as demand analysis, the random term can be interpreted as random preference variation. Random effects models can be estimated using either cross-section or panel data.

1.2.6. Dynamics

A very common assumption in cross-section analysis is the absence of intertemporal dependence, that is, an absence of dynamics. Thus, implicitly it is assumed that the observations correspond to a stochastic equilibrium, with the deviation from the equilibrium being represented by serially independent random disturbances. Even in microeconomics for some data situations such an assumption may be too strong. For example, it is inconsistent with the presence of serially correlated unobserved heterogeneity. Dependence on lagged dependent variables also violates this assumption.

The foregoing discussion illustrates some of the potential limitations of a single cross-section analysis. Some limitations may be overcome if repeated cross sections are available. However, if there is dynamic dependence, the least problematic approach might well be to use panel data.

1.3. Book Outline

The book is split into six parts. Part 1 presents the issues involved in microeconometric modeling. Parts 2 and 3 present general theory for estimation and statistical inference for nonlinear regression models. Parts 4 and 5 specialize to the core models used in applied microeconomics for, respectively, cross-section and panel data. Part 6 covers broader topics that make considerable use of material presented in the earlier chapters.

The book outline is summarized in Table 1.1. The remainder of this section details each part in turn.

1.3.1. Part 1: Preliminaries

Chapters 2 and 3 expand on the special features of the **microeconometric** approach to modeling and **microeconomic data structures** within the more general statistical

Table 1.1. Book Outline

Part and Chapter	Background ^a	Example
1. Preliminaries		
1. Overview	—	
2. Causal and Noncausal Models	—	Simultaneous equations models
3. Microeconomic Data Structures	—	Observational data
2. Core Methods		
4. Linear Models	—	Ordinary least squares
5. Maximum Likelihood and Nonlinear Least-Squares Estimation	—	m-estimation or extremum estimation
6. Generalized Method of Moments and Systems Estimation	5	Instrumental variables
7. Hypothesis Tests	5	Wald, score, and likelihood ratio tests
8. Specification Tests and Model Selection	5,7	Conditional moment test
9. Semiparametric Methods	—	Kernel regression
10. Numerical Optimization	5	Newton–Raphson iterative method
3. Simulation-Based Methods		
11. Bootstrap Methods	7	Percentile t -method
12. Simulation-Based Methods	5	Maximum simulated likelihood
13. Bayesian Methods	—	Markov chain Monte Carlo
4. Models for Cross-Section Data		
14. Binary Outcome Models	5	Logit, probit for $y = (0, 1)$
15. Multinomial Models	5,14	Multinomial logit for $y = (1, \dots, m)$
16. Tobit and Selection Models	5,14	Tobit for $y = \max(y^*, 0)$
17. Transition Data: Survival Analysis	5	Cox proportional hazards for $y = \min(y^*, c)$
18. Mixture Models and Unobserved Heterogeneity	5,17	Unobserved heterogeneity
19. Models for Multiple Hazards	5,17	Multiple hazards
20. Models of Count Data	5	Poisson for $y = 0, 1, 2, \dots$
5. Models for Panel Data		
21. Linear Panel Models: Basics	—	Fixed and random effects
22. Linear Panel Models: Extensions	6,21	Dynamic and endogenous regressors
23. Nonlinear Panel Models	5,6,21,22	Panel logit, Tobit, and Poisson
6. Further Topics		
24. Stratified and Clustered Samples	5	Data $(y_{ij}, \mathbf{x}_{ij})$ correlated over j
25. Treatment Evaluation	5,21	Regressor $d = 1$ if in program
26. Measurement Error Models	5	Logit model with measurement errors
27. Missing Data and Imputation	5	Regression with missing observations

^a The background gives the essential chapter needed in addition to the treatment of ordinary and weighted LS in Chapter 4. Note that the first panel data chapter (Chapter 21) requires only Chapter 4.

arena of regression analysis. Many of the issues raised in these chapters are pursued throughout the book as the reader develops the necessary tools.

1.3.2. Part 2: Core Methods

Chapters 4–10 detail the main general methods used in classical estimation and statistical inference. The results given in Chapter 5 in particular are extensively used throughout the book.

Chapter 4 presents some results for the **linear regression model**, emphasizing those issues and methods that are most relevant for the rest of the book. Analysis is relatively straightforward as there is an explicit expression for linear model estimators such as ordinary least squares.

Chapters 5 and 6 present **estimation theory** that can be applied to nonlinear models for which there is usually no explicit solution for the estimator. Asymptotic theory is used to obtain the distribution of estimators, with emphasis on obtaining robust standard error estimates that rely on relatively weak distributional assumptions. A quite general treatment of estimation, along with specialization to nonlinear least-squares and maximum likelihood estimation, is presented in Chapter 5. The more challenging generalized method of moments estimator and specialization to instrumental variables estimation are given separate treatment in Chapter 6.

Chapter 7 presents **classical hypothesis testing** when estimators are nonlinear and the hypothesis being tested is possibly nonlinear in parameters. **Specification tests** in addition to hypothesis tests are the subject of Chapter 8.

Chapter 9 presents **semiparametric estimation** methods such as kernel regression. The leading example is flexible modeling of the conditional mean. For the patents example, the nonparametric regression model is $E[y|x] = g(x)$, where the function $g(\cdot)$ is unspecified and is instead estimated. Then estimation has an infinite-dimensional component $g(\cdot)$ leading to a nonstandard asymptotic theory. With additional regressors some further structure is needed and the methods are called semiparametric or seminonparametric.

Chapter 10 presents the **computational methods** used to compute a parameter estimate when the estimator is defined implicitly, usually as the solution to some first-order conditions.

1.3.3. Part 3: Simulation-Based Methods

Chapters 11–13 consider methods of estimation and inference that rely on simulation. These methods are generally more computationally intensive and, currently, less utilized than the methods presented in Part 2.

Chapter 11 presents the **bootstrap method** for statistical inference. This yields the empirical distribution of an estimator by obtaining new samples by simulation, such as by repeated resampling with replacement from the original sample. The bootstrap can provide a simple way to obtain standard errors when the formulas from asymptotic theory are complex, as is the case for some two-step estimators. Furthermore, if

implemented appropriately, the bootstrap can lead to better statistical inference in small samples.

Chapter 12 presents **simulation-based estimation methods**, developed for models that involve an integral over a probability distribution for which there is no closed-form solution. Estimation is still possible by making multiple draws from the relevant distribution and averaging.

Chapter 13 presents **Bayesian methods**, which combine a distribution for the observed data with a specified prior distribution for parameters to obtain a posterior distribution of the parameters that is the basis for estimation. Recent advances make computation possible even if there is no closed-form solution for the posterior distribution. Bayesian analysis can provide an approach to estimation and inference that is quite different from the classical approach. However, in many cases only the Bayesian tool kit is adopted to permit classical estimation and inference for problems that are otherwise intractable.

1.3.4. Part 4: Models for Cross-Section Data

Chapters 14–20 present the main nonlinear models for **cross-section data**. This part is the heart of the book and presents advanced topics such as models for limited dependent variables and sample selection. The classes of models are defined by the range of values taken by the dependent variable.

Binary data models for dependent variable that can take only two possible values, say $y = 0$ or $y = 1$, are presented in Chapter 14. In Chapter 15 an extension is made to **multinomial** models, for dependent variable that takes several discrete values. Examples include employment status (employed, unemployed, and out of the labor force) and mode of transportation to work (car, bus, or train). Linear models can be informative but are not appropriate, as they can lead to predicted probabilities outside the unit interval. Instead logit, probit, and related models are used.

Chapter 16 presents models with **censoring, truncation, sample selection**. Examples include annual hours of work, conditional on choosing to work, and hospital expenditures, conditional on being hospitalized. In these cases the data are incompletely observed with a bunching of observations at $y = 0$ and with the remaining $y > 0$. The model for the observed data can be shown to be nonlinear even if the underlying process is linear, and linear regression on the observed data can be very misleading. Simple corrections for censoring, truncation, or sample selection such as the Tobit model exist, but these are very dependent on distributional assumptions.

Models for **duration data** are presented in Chapters 17–19. An example is length of unemployment spell. Standard regression models include the exponential, Weibull, and Cox proportional hazards model. Additionally, as in Chapter 16, the dependent variable is often incompletely observed. For example, the data may be on the length of a current spell that is incomplete, rather than the length of a completed spell.

Chapter 20 presents **count data** models. Examples include various measures of health utilization such as number of doctor visits and number of days hospitalized. Again the model is nonlinear, as counts and hence the conditional mean are nonnegative. Leading parametric models include the Poisson and negative binomial.

1.3.5. Part 5: Models for Panel Data

Chapters 21–23 present methods for **panel data**. Here the data are observed in several time periods for each of the many individuals in the sample, so the dependent variable and regressors are indexed by both individual and time. Any analysis needs to control for the likely positive correlation of error terms in different time periods for a given individual. Additionally, panel data can provide sufficient data to control for unobserved time-invariant individual-specific effects, permitting identification of causation under weaker assumptions than those needed if only cross-section data are available.

The basic linear panel data model is presented in Chapter 21, with emphasis on **fixed effects** and **random effects** models. Extensions of linear models to permit lagged dependent variables and endogenous regressors are presented in Chapter 22. Panel methods for the nonlinear models of Part 4 are presented in Chapter 23.

The panel data methods are placed late in the book to permit a unified self-contained treatment. Chapter 21 could have been placed immediately after Chapter 4 and is written in an accessible manner that relies on little more than knowledge of least-squares estimation.

1.3.6. Part 6: Further Topics

This part considers important topics that can generally relate to any and all models covered in Parts 4 and 5. Chapter 24 deals with modeling of clustered data in several different models. Chapter 25 discusses treatment evaluation. Treatment evaluation is a general term that can cover a wide variety of models in which the focus is on measuring the impact of some “treatment” that is either exogenously or randomly assigned to an individual on some measure of interest, denoted an “outcome variable.” Chapter 26 deals with the consequences of measurement errors in outcome and/or regressor variables, with emphasis on some leading nonlinear models. Chapter 27 considers some methods of handling missing data in linear and nonlinear regression models.

1.4. How to Use This Book

The book assumes a basic understanding of the linear regression model with matrix algebra. It is written at the mathematical level of the first-year economics Ph.D. sequence, comparable to Greene (2003).

Although some of the material in this book is covered in a first-year sequence, most of it appears in second-year econometrics Ph.D. courses or in data-oriented microeconomics field courses such as labor economics, public economics, or industrial organization. This book is intended to be used as both an econometrics text and as an adjunct for such field courses. More generally, the book is intended to be useful as a reference work for applied researchers in economics, in related social sciences such as sociology and political science, and in epidemiology.

For readers using this book as a reference work, the models chapters have been written to be as self-contained as possible. For the specific models presented in Parts 4

Table 1.2. Outline of a 20-Lecture 10-Week Course

Lectures	Chapter	Topic
1–3	4, Appx. A	Review of linear models and asymptotic theory
4–7	5	Estimation: m-estimation, ML, and NLS
8	10	Estimation: numerical optimization
9–11	14, 15	Models: binary and multinomial
12–14	16	Models: censored and truncated
15	6	Estimation: GMM
16	7	Testing: hypothesis tests
17–19	21	Models: basic linear panel
20	9	Estimation: semiparametric

and 5 it will generally be sufficient to read the relevant chapter in isolation, except that some command of the general estimation results in Chapter 5 and in some cases Chapter 6 will be necessary. Most chapters are structured to begin with a discussion and example that is accessible to a wide audience.

For instructors using this book as a course text it is best to introduce the basic nonlinear cross-section and linear panel data models as early as possible, skipping many of the methods chapters. The most commonly used nonlinear cross-section models are presented in Chapters 14–16; these require knowledge of maximum likelihood and least-squares estimation, presented in Chapter 5. Chapter 21 on linear panel data models requires even less preparation, essentially just Chapter 4.

Table 1.2 provides an outline for a one-quarter second-year graduate course taught at the University of California, Davis, immediately following the required first-year statistics and econometrics sequence. A quarter provides sufficient time to cover the basic results given in the first half of the chapters in this outline. With additional time one can go into further detail or cover a subset of Chapters 11–13 on computationally intensive estimation methods (simulation-based estimation, the bootstrap, which is also briefly presented in Chapter 7, and Bayesian methods); additional cross-section models (durations and counts) presented in Chapters 17–20; and additional panel data models (linear model extensions and nonlinear models) given in Chapters 22 and 23.

At Indiana University, Bloomington, a 15-week semester-long field course in microeometrics is based on material in most of Parts 4 and 5. The prerequisite courses for this course cover material similar to that in Part 2.

Some exercises are provided at the end of each chapter after the first three introductory chapters. These exercises are usually learning-by-doing exercises; some are purely methodological whereas others entail analysis of generated or actual data. The level of difficulty of the questions is mostly related to the level of difficulty of the topic.

1.5. Software

There are many software packages available for data analysis. Popular packages with strong microeconometric capabilities include LIMDEP, SAS, and STATA, all of which

offer an impressive range of canned routines and additionally support user-defined procedures using a matrix programming language. Other packages that are also widely used include EVIEW, PCGIVE, and TSP. Despite their time-series orientation, these can support some cross-section data analysis. Users who wish to do their own programming also have available a variety of options including GAUSS, MATLAB, OX, and SAS/IML. The latest detailed information about these packages and many others can be efficiently located via an Internet browser and a search engine.

1.6. Notation and Conventions

Vector and matrix algebra are used extensively.

Vectors are defined as column vectors and represented using lowercase bold. For example, for linear regression the regressor vector \mathbf{x} is a $K \times 1$ column vector with j th entry x_j and the parameter vector $\boldsymbol{\beta}$ is a $K \times 1$ column vector with j th entry β_j , so

$$\begin{matrix} \mathbf{x} \\ (K \times 1) \end{matrix} = \begin{bmatrix} x_1 \\ \vdots \\ x_K \end{bmatrix} \quad \text{and} \quad \begin{matrix} \boldsymbol{\beta} \\ (K \times 1) \end{matrix} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \end{bmatrix}.$$

Then the linear regression model $y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + u$ is expressed as $y = \mathbf{x}'\boldsymbol{\beta} + u$. At times a subscript i is added to denote the typical i th observation. The linear regression equation for the i th observation is then

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + u_i.$$

The sample is one of N observations, $\{(y_i, \mathbf{x}_i), i = 1, \dots, N\}$. In this book observations are usually assumed to be independent over i .

Matrices are represented using uppercase bold. In matrix notation the sample is (\mathbf{y}, \mathbf{X}) , where \mathbf{y} is an $N \times 1$ vector with i th entry y_i and \mathbf{X} is a matrix with i th row \mathbf{x}'_i , so

$$\begin{matrix} \mathbf{y} \\ (N \times 1) \end{matrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \quad \text{and} \quad \begin{matrix} \mathbf{X} \\ (N \times \dim(\mathbf{x})) \end{matrix} = \begin{bmatrix} \mathbf{x}'_1 \\ \vdots \\ \mathbf{x}'_N \end{bmatrix}.$$

The linear regression model upon stacking all N observations is then

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

where \mathbf{u} is an $N \times 1$ column vector with i th entry u_i .

Matrix notation is compact but at times it is clearer to write products of matrices as summations of products of vectors. For example, the OLS estimator can be equivalently written in either of the following ways:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \left(\sum_{i=1}^N \mathbf{x}_i \mathbf{x}'_i \right)^{-1} \sum_{i=1}^N \mathbf{x}_i y_i.$$

Table 1.3. Commonly Used Acronyms and Abbreviations

Linear	OLS Ordinary least squares GLS Generalized least squares FGLS Feasible generalized least squares IV Instrumental variables 2SLS Two-stage least squares 3SLS Three-stage least squares
Nonlinear	NLS Nonlinear least squares FGNLS Feasible generalized nonlinear least squares NIV Nonlinear instrumental variables NL2SLS Nonlinear two-stage least squares NL3SLS Nonlinear three-stage least squares
General	LS Least squares ML Maximum likelihood QML Quasi-maximum likelihood GMM Generalized method of moments GEE Generalized estimating equations

Generic notation for a parameter is the $q \times 1$ vector θ . The regression parameters are represented by the $K \times 1$ vector β , which may equal θ or may be a subset of θ depending on the context.

The book uses many abbreviations and acronyms. Table 1.3 summarizes abbreviations used for some common estimation methods, ordered by whether the estimator is developed for linear or nonlinear regression models. We also use the following: dgp (data-generating process), iid (independently and identically distributed), pdf (probability density function), cdf (cumulative distribution function), L (likelihood), $\ln L$ (log-likelihood), FE (fixed effects), and RE (random effects).

CHAPTER 2

Causal and Noncausal Models

2.1. Introduction

Microeometrics deals with the theory and applications of methods of data analysis developed for microdata pertaining to individuals, households, and firms. A broader definition might also include regional- and state-level data. Microdata are usually either cross sectional, in which case they refer to conditions at the same point in time, or longitudinal (panel) in which case they refer to the same observational units over several periods. Such observations are generated from both nonexperimental setups, such as censuses and surveys, and quasi-experimental or experimental setups, such as social experiments implemented by governments with the participation of volunteers.

A microeconomic model may be a full specification of the probability distribution of a set of microeconomic observations; it may also be a partial specification of some distributional properties, such as moments, of a subset of variables. The mean of a single dependent variable conditional on regressors is of particular interest.

There are several objectives of microeconomics. They include both data description and causal inference. The first can be defined broadly to include moment properties of response variables, or regression equations that highlight associations rather than causal relations. The second category includes causal relationships that aim at measurement and/or empirical confirmation or refutation of conjectures and propositions regarding microeconomic behavior. The type and style of empirical investigations therefore span a wide spectrum. At one end of the spectrum can be found very highly structured models, derived from detailed specification of the underlying economic behavior, that analyze **causal** (behavioral) or **structural relationships** for interdependent microeconomic variables. At the other end are **reduced form** studies that aim to uncover correlations and associations among variables, without necessarily relying on a detailed specification of all relevant interdependencies. Both approaches share the common goal of uncovering important and striking relationships that could be helpful in understanding microeconomic behavior, but they differ in the extent to which they rely on economic theory to guide their empirical investigations.

2.1. INTRODUCTION

As a subdiscipline microeconomics is newer than macroeconomics, which is concerned with modeling of market and aggregate data. A great deal of the early work in applied econometrics was based on aggregate time-series data collected by government agencies. Much of the early work on statistical demand analysis up until about 1940 used market rather than individual or household data (Hendry and Morgan, 1996). Morgan's (1990) book on the history of econometric ideas makes no reference to microeconometric work before the 1940s, with one important exception. That exception is the work on household budget data that was instigated by concern with the living standards of the less well-off in many countries. This led to the collection of household budget data that provided the raw material for some of the earlier microeconometric studies such as those pioneered by Allen and Bowley (1935). Nevertheless, it is only since the 1950s that microeconomics has emerged as a distinctive and recognized subdiscipline. Even into the 1960s the core of microeconomics consisted of demand analyses based on household surveys.

With the award of the year 2000 Nobel Prize in Economics to James Heckman and Daniel McFadden for their contributions to microeconomics, the subject area has achieved clear recognition as a distinct subdiscipline. The award cited Heckman "for his development of theory and methods for analyzing selective samples" and McFadden "for his development of theory and methods for analyzing discrete choice." Examples of the type of topics that microeconomics deals with were also mentioned in the citation: "... what factors determine whether an individual decides to work and, if so, how many hours? How do economic incentives affect individual choices regarding education, occupation or place of residence? What are the effects of different labor-market and educational programs on an individual's income and employment?"

Applications of microeconometric methods can be found not only in every area of microeconomics but also in other cognate social sciences such as political science, sociology, and geography.

Beginning with the 1970s and especially within the past two decades revolutionary advances in our capacity for handling large data sets and associated computations have taken place. These, together with the accompanying explosion in the availability of large microeconomic data sets, have greatly expanded the scope of microeconomics. As a result, although empirical demand analysis continues to be one of the most important areas of application for microeconometric methods, its style and content have been heavily influenced by newer methods and models. Further, applications in economic development, finance, health, industrial organization, labor and public economics, and applied microeconomics generally are now commonplace, and these applications will be encountered at various places in this book.

The primary focus of this book is on the newer material that has emerged in the past three decades. Our goal is to survey concepts, models, and methods that we regard as standard components of a modern microeconometrician's tool kit. Of course, the notion of standard methods and models is inevitably both subjective and elastic, being a function of the presumed clientele of this book as well as the authors' own backgrounds. There may also be topics we regard as too advanced for an introductory book such as this that others would place in a different category.

Microeconomics focuses on the complications of nonlinear models and on obtaining estimates that can be given a structural interpretation. Much of this book, especially Parts 2–4, presents methods for nonlinear models. These nonlinear methods overlap with many areas of applied statistics including biostatistics. By contrast, the distinguishing feature of econometrics is the **emphasis placed on causal modeling**. This chapter introduces the key concepts related to causal (and noncausal) modeling, concepts that are germane to both linear and nonlinear models.

Sections 2.2 and 2.3 introduce the key concepts of structure and exogeneity. Section 2.4 uses the linear simultaneous equations model as a specific illustration of a structural model and connects it with the other important concepts of reduced form models. Identification definitions are given in Section 2.5. Section 2.6 considers single-equation structural models. Section 2.7 introduces the potential outcome model and compares the causal parameters and interpretations in the potential outcome model with those in the simultaneous equations model. Section 2.8 provides a brief discussion of modeling and estimation strategies designed to handle computational and data challenges.

2.2. Structural Models

Structure consists of

1. a set of variables \mathbf{W} (“data”) partitioned for convenience as $[\mathbf{Y} \ \mathbf{Z}]$;
2. a joint probability distribution of \mathbf{W} , $F(\mathbf{W})$;
3. an a priori ordering of \mathbf{W} according to hypothetical cause-and-effect relationships and specification of a priori restrictions on the hypothesized model; and
4. a parametric, semiparametric, or nonparametric specification of functional forms and the restrictions on the parameters of the model.

This general description of a structural model is consistent with a well-established Cowles Commission definition of a structure. For example, Sargan (1988, p. 27) states:

A model is the specification of the probability distribution for a set of observations.

A structure is the specification of the parameters of that distribution. Therefore, a structure is a model in which all the parameters are assigned numerical values.

We consider the case in which the modeling objective is to explain the values of observable vector-valued variable \mathbf{y} , $\mathbf{y}' = (y_1, \dots, y_G)$. Each element of \mathbf{y} is a function of some other elements of \mathbf{y} and of explanatory variables \mathbf{z} and a purely random disturbance u . Note that the variables y are assumed to be interdependent. By contrast, interdependence between z_i is not modeled. The i th observation satisfies the set of implicit equations

$$\mathbf{g}(\mathbf{y}_i, \mathbf{z}_i, \mathbf{u}_i | \boldsymbol{\theta}) = \mathbf{0}, \quad (2.1)$$

where \mathbf{g} is a known function. We refer to this as the **structural model**, and we refer to $\boldsymbol{\theta}$ as structural parameters. This corresponds to property 4 given earlier in this section.

Assume that there is a unique solution for \mathbf{y}_i for every $(\mathbf{z}_i, \mathbf{u}_i)$. Then we can write the equations in an explicit form with \mathbf{y} as function of (\mathbf{z}, \mathbf{u}) :

$$\mathbf{y}_i = \mathbf{f}(\mathbf{z}_i, \mathbf{u}_i | \boldsymbol{\pi}). \quad (2.2)$$

This is referred to as the **reduced form** of the structural model, where $\boldsymbol{\pi}$ is a vector of reduced form parameters that are functions of $\boldsymbol{\theta}$. The reduced form is obtained by solving the structural model for the endogenous variables \mathbf{y}_i , given $(\mathbf{z}_i, \mathbf{u}_i)$. The reduced form parameters $\boldsymbol{\pi}$ are functions of $\boldsymbol{\theta}$.

If the objective of modeling is inference about elements of $\boldsymbol{\theta}$, then (2.1) provides a direct route. This involves estimation of the structural model. However, because elements of $\boldsymbol{\pi}$ are functions of $\boldsymbol{\theta}$, (2.2) also provides an indirect route to inference on $\boldsymbol{\theta}$. If $\mathbf{f}(\mathbf{z}_i, \mathbf{u}_i | \boldsymbol{\pi})$ has a known functional form, and if it is additively separable in \mathbf{z}_i and \mathbf{u}_i , such that we can write

$$\mathbf{y}_i = \mathbf{g}(\mathbf{z}_i | \boldsymbol{\pi}) + \mathbf{u}_i = E[\mathbf{y}_i | \mathbf{z}_i] + \mathbf{u}_i, \quad (2.3)$$

then the regression of \mathbf{y} on \mathbf{z} is a natural prediction function for \mathbf{y} given \mathbf{z} . In this sense the reduced form equation has a useful role for making conditional predictions of \mathbf{y}_i given $(\mathbf{z}_i, \mathbf{u}_i)$. To generate predictions of the left-hand-side variable for assigned values of the right-hand-side variables of (2.2) requires estimates of $\boldsymbol{\pi}$, which may be computationally simpler.

An important extension of (2.3) is the **transformation model**, which for scalar y takes the form

$$\Lambda(y) = \mathbf{z}'\boldsymbol{\pi} + \mathbf{u}, \quad (2.4)$$

where $\Lambda(y)$ is a transformation function (e.g., $\Lambda(y) = \ln(y)$ or $\Lambda(y) = y^{1/2}$). In some cases the transformation function may depend on unknown parameters. A transformation model is distinct from a regression, but it too can be used to make estimates of $E[y|\mathbf{z}]$. An important example is the accelerated failure time model analyzed in Chapter 17.

One of the most important, and potentially controversial, steps in the specification of the structural model is property 3, in which an a priori ordering of variables into causes and effects is assigned. In essence this involves drawing a distinction between those variables whose variation the model is designed to explain and those whose variation is externally determined and hence lie outside the scope of investigation. In microeconomics, examples of the former are years of schooling and hours worked; examples of the latter are gender, ethnicity, age, and similar demographic variables. The former, denoted \mathbf{y} , are referred to as **endogenous** and the latter, denoted \mathbf{z} , are called **exogenous** variables.

Exogeneity of a variable is an important simplification because in essence it justifies the decision to treat that variable as ancillary, and not to model that variable because the parameters of that relationship have no direct bearing on the variable under study. This important notion needs a more formal definition, which we now provide.

2.3. Exogeneity

We begin by considering the representation of a general finite dimensional parametric case in which the joint distribution of \mathbf{W} , with parameters θ partitioned as $(\theta_1 \theta_2)$, is factored into the conditional density of \mathbf{Y} given \mathbf{Z} , and the marginal distribution of \mathbf{Z} , giving

$$f_J(\mathbf{W}|\theta) = f_C(\mathbf{Y}|\mathbf{Z}, \theta) \times f_M(\mathbf{Z}|\theta). \quad (2.5)$$

A special case of this result occurs if

$$f_J(\mathbf{W}|\theta) = f_C(\mathbf{Y}|\mathbf{Z}, \theta_1) \times f_M(\mathbf{Z}|\theta_2),$$

where θ_1 and θ_2 are functionally independent. Then we say that \mathbf{Z} is exogenous with respect to θ_1 ; this means that knowledge of $f_M(\mathbf{Z}|\theta_2)$ is not required for inference on θ_1 , and hence we can validly condition the distribution of \mathbf{Y} on \mathbf{Z} .

Models can always be reparameterized. So next consider the case in which the model is reparameterized in terms of parameters φ , with one-to-one transformation of θ , say $\varphi = h(\theta)$, where φ is partitioned into (φ_1, φ_2) . This reparametrization may be of interest if, for example, φ_1 is structurally invariant to a class of policy interventions. Suppose φ_1 is the parameter of interest. In such a case one is interested in the exogeneity of \mathbf{Z} with respect to φ_1 . Then, the condition for exogeneity is that

$$f_J(\mathbf{W}|\varphi) = f_C(\mathbf{Y}|\mathbf{Z}, \varphi_1) \times f_M(\mathbf{Z}|\varphi_2), \quad (2.6)$$

where φ_1 is independent of φ_2 .

Finally consider the case in which the interest is in a parameter λ that is a function of φ , say $h(\varphi)$. Then for exogeneity of \mathbf{Z} with respect to λ , we need two conditions: (i) λ depends only on φ_1 , i.e., $\lambda = h(\varphi_1)$, and hence only the conditional distribution is of interest; and (ii) φ_1 and φ_2 are “variation free” which means that the parameters of the joint distribution are not subject to cross-restrictions, i.e. $(\varphi_1, \varphi_2) \in \Phi_1 \times \Phi_2 = \{\varphi_1 \in \Phi_1, \varphi_2 \in \Phi_2\}$.

The factorization in (2.5)-(2.6) plays an important role in the development of the exogeneity concept. Of special interest in this book are the following three concepts related to exogeneity: (1) weak exogeneity; (2) Granger noncausality; (3) strong exogeneity.

Definition 2.1 (Weak Exogeneity): \mathbf{Z} is **weakly exogenous for λ** if (i) and (ii) hold.

If the marginal model parameters are uninformative for inference on λ , then inference on λ can proceed on the basis of the conditional distribution $f(\mathbf{Y}|\mathbf{Z}, \varphi_1)$ alone. The operational implication is that weakly exogenous variables can be taken as given if one’s main interest is in inference on λ or φ_1 . This does not mean that there is no statistical model for \mathbf{Z} ; it means that the parameters of that model play no role in the inference on φ_1 , and hence are irrelevant.

2.3.1. Conditional Independence

Originally, the Granger causality concept was defined in the context of prediction in a time-series environment. More generally, it can be interpreted as a form of **conditional independence** (Holland, 1986, p. 957).

Partition \mathbf{z} into two subsets \mathbf{z}_1 and \mathbf{z}_2 ; let $\mathbf{W} = [\mathbf{y}, \mathbf{z}_1, \mathbf{z}_2]$ be the matrices of variables of interest. Then \mathbf{z}_1 and \mathbf{y} are conditionally independent given \mathbf{z}_2 if

$$f(\mathbf{y}|\mathbf{z}_1, \mathbf{z}_2) = f(\mathbf{y}|\mathbf{z}_2). \quad (2.8)$$

This is stronger than the **mean independence** assumption, which would imply

$$\mathbb{E}[\mathbf{y}|\mathbf{z}_1, \mathbf{z}_2] = \mathbb{E}[\mathbf{y}|\mathbf{z}_2]. \quad (2.9)$$

Then \mathbf{z}_1 has no predictive value for \mathbf{y} , after conditioning on \mathbf{z}_2 . In a predictive sense this means that \mathbf{z}_1 does not Granger-cause \mathbf{y} .

In a time-series context, \mathbf{z}_1 and \mathbf{z}_2 would be mutually exclusive lagged values of subsets of \mathbf{y} .

Definition 2.2 (Strong Exogeneity): \mathbf{z}_1 is **strongly exogenous** for φ if it is weakly exogenous for φ and does not Granger-cause \mathbf{y} so (2.8) holds.

2.3.2. Exogenizing Variables

Exogeneity is a strong assumption. It is a property of random variables relative to parameters of interest. Hence a variable may be validly treated as exogenous in one structural model but not in another; the key issue is the parameters that are the subject of inference. Arbitrary imposition of this property will have some undesirable consequences that will be discussed in Section 2.4.

The exogeneity assumption may be justified by a priori theorizing, in which case it is a part of the maintained hypothesis of the model. It may in some cases be justified as a valid approximation, in which case it may be subject to testing, as discussed in Section 8.4.3. In cross-section analysis it may be justified as being a consequence of a natural experiment or a quasi-experiment in which the value of the variable is determined by an external intervention; for example, government or regulatory authority may determine the setting of a tax rate or a policy parameter. Of special interest is the case in which an external intervention results in a change in the value of an important policy variable. Such a natural experiment is tantamount to exogenization of some variable. As we shall see in Chapter 3, this creates a quasi-experimental opportunity to study the impact of a variable in the absence of other complicating factors.

2.4. Linear Simultaneous Equations Model

An important special case of the general structural model specified in (2.1) is the linear simultaneous equation model developed by the Cowles Commission econometricians. Comprehensive treatment of this model is available in many textbooks (e.g., Sargan,

1988). The treatment here is brief and selective; also see Section 6.9.6. The objective is to bring into the discussion several key ideas and concepts that have more general relevance. Although the analysis is restricted to linear models, many insights are routinely applied to nonlinear models.

2.4.1. The SEM Setup

The **linear simultaneous equations model** (SEM) setup is as follows:

$$\begin{aligned} y_{1i}\beta_{11} + \cdots + y_{Gi}\beta_{1G} + z_{1i}\gamma_{11} + \cdots + z_{Ki}\gamma_{1K} &= u_{1i} \\ \vdots &\quad \vdots = \vdots \\ y_{1i}\beta_{G1} + \cdots + y_{Gi}\beta_{GG} + z_{1i}\gamma_{G1} + \cdots + z_{Ki}\gamma_{GK} &= u_{Gi}, \end{aligned}$$

where i is the observation subscript.

A clear a priori distinction or preordering is made between endogenous variables, $\mathbf{y}'_i = (y_{1i}, \dots, y_{Gi})$, and exogenous variables, $\mathbf{z}'_i = (z_{1i}, \dots, z_{Ki})$. By definition the exogenous variables are uncorrelated with the purely random disturbances (u_{1i}, \dots, u_{Gi}). In its unrestricted form every variable enters every equation.

In matrix notation, the G -equation SEM for the i th equation is written as

$$\mathbf{y}'_i \mathbf{B} + \mathbf{z}'_i \boldsymbol{\Gamma} = \mathbf{u}'_i, \quad (2.10)$$

where \mathbf{y}_i , \mathbf{B} , \mathbf{z}_i , $\boldsymbol{\Gamma}$, and \mathbf{u}_i have dimensions $G \times 1$, $G \times G$, $K \times 1$, $K \times G$, and $G \times 1$, respectively. For specified values of $(\mathbf{B}, \boldsymbol{\Gamma})$ and $(\mathbf{z}_i, \mathbf{u}_i)$ G linear simultaneous equations can in principle be solved for \mathbf{y}_i .

The standard assumptions of SEM are as follows:

1. \mathbf{B} is nonsingular and has rank G .
2. $\text{rank}[\mathbf{Z}] = K$. The $N \times K$ matrix \mathbf{Z} is formed by stacking \mathbf{z}'_i , $i = 1, \dots, N$.
3. $\text{plim } N^{-1} \mathbf{Z}' \mathbf{Z} = \boldsymbol{\Sigma}_{zz}$ is a symmetric $K \times K$ positive definite matrix.
4. $\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$; that is, $E[\mathbf{u}_i] = \mathbf{0}$ and $E[\mathbf{u}_i \mathbf{u}_i'] = \boldsymbol{\Sigma} = [\sigma_{ij}]$, where $\boldsymbol{\Sigma}$ is a symmetric $G \times G$ positive definite matrix.
5. The errors in each equation are serially independent.

In this model the structure (or structural parameters) consists of $(\mathbf{B}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma})$. Writing

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}'_1 \\ \vdots \\ \mathbf{y}'_N \end{bmatrix}, \quad \mathbf{Z} = \begin{bmatrix} \mathbf{z}'_1 \\ \vdots \\ \mathbf{z}'_N \end{bmatrix}, \quad \mathbf{U} = \begin{bmatrix} \mathbf{u}'_1 \\ \vdots \\ \mathbf{u}'_N \end{bmatrix}$$

allows us to express the **structural model** more compactly as

$$\mathbf{YB} + \mathbf{Z}\boldsymbol{\Gamma} = \mathbf{U}, \quad (2.11)$$

where the arrays \mathbf{Y} , \mathbf{B} , \mathbf{Z} , $\boldsymbol{\Gamma}$, and \mathbf{U} have dimensions $N \times G$, $G \times G$, $N \times K$, $K \times G$, and $N \times G$, respectively. Solving for all the endogenous variables in terms of all

2.4. LINEAR SIMULTANEOUS EQUATIONS MODEL

the exogenous variables, we obtain the **reduced form of the SEM**:

$$\begin{aligned} \mathbf{Y} + \mathbf{Z}\boldsymbol{\Gamma}\mathbf{B}^{-1} &= \mathbf{U}\mathbf{B}^{-1}, \\ \mathbf{Y} &= \mathbf{Z}\boldsymbol{\Pi} + \mathbf{V}, \end{aligned} \quad (2.12)$$

where $\boldsymbol{\Pi} = -\boldsymbol{\Gamma}\mathbf{B}^{-1}$ and $\mathbf{V} = \mathbf{U}\mathbf{B}^{-1}$. Given Assumption 4, $\mathbf{v}_i \sim \mathcal{N}[\mathbf{0}, \mathbf{B}^{-1'}\boldsymbol{\Sigma}\mathbf{B}^{-1}]$.

In the SEM framework the structural model has primacy for several reasons. First, the equations themselves have interpretations as economic relationships such as demand or supply relations, production functions, and so forth, and they are subject to restrictions of economic theory. Consequently, \mathbf{B} and $\boldsymbol{\Gamma}$ are parameters that describe economic behavior. Hence a priori theory can be invoked to form expectations about the sign and size of individual coefficients. By contrast, the unrestricted reduced form parameters are potentially complicated functions of the structural parameters, and as such it may be difficult to evaluate them postestimation. This consideration may have little weight if the goal of econometric modeling is prediction rather than inference on parameters with behavioral interpretation.

Consider, without loss of generality, the first equation in the model (2.11), with y_1 as the dependent variable. In addition, some of the remaining $G - 1$ endogenous variables and $K - 1$ exogenous variables may be absent from this equation. From (2.12) we see that in general the endogenous variables \mathbf{Y} depend stochastically on \mathbf{V} , which in turn is a function of the structural errors \mathbf{U} . Therefore, in general $\text{plim } N^{-1}\mathbf{Y}'\mathbf{U} \neq \mathbf{0}$. Generally, the application of the least-squares estimator in the simultaneous equation setting yields inconsistent estimates. This is a well-known and basic result from the simultaneous equations literature, often referred to as the “simultaneous equations bias” problem. The vast literature on simultaneous equations models deals with identification and consistent estimation when the least-squares approach fails; see Sargan (1988) and Schmidt (1976), and Section 6.9.6.

The reduced form of SEM expresses every endogenous variable as a linear function of all exogenous variables and all structural disturbances. The reduced form disturbances are linear combinations of the structural disturbances. From the reduced form for the i th observation

$$E[\mathbf{y}_i | \mathbf{z}_i] = \mathbf{z}'_i \boldsymbol{\Pi}, \quad (2.13)$$

$$V[\mathbf{y}_i | \mathbf{z}_i] = \boldsymbol{\Omega} \equiv \mathbf{B}^{-1'}\boldsymbol{\Sigma}\mathbf{B}^{-1}. \quad (2.14)$$

The reduced form parameters $\boldsymbol{\Pi}$ are derived parameters defined as functions of the structural parameters. If $\boldsymbol{\Pi}$ can be consistently estimated then the reduced form can be used to make predictive statements about variations in \mathbf{Y} due to exogenous changes in \mathbf{Z} . This is possible even if \mathbf{B} and $\boldsymbol{\Gamma}$ are not known. Given the exogeneity of \mathbf{Z} , the full set of reduced form regressions is a multivariate regression model that can be estimated consistently by least squares. The reduced form provides a basis for making conditional predictions of \mathbf{Y} given \mathbf{Z} .

The restricted reduced form is the unrestricted reduced form model subject to restrictions. If these are the same restrictions as those that apply to the structure, then structural information can be recovered from the reduced form.

In the SEM framework, the unknown structural parameters, the nonzero elements of \mathbf{B} , $\boldsymbol{\Gamma}$, and $\boldsymbol{\Sigma}$, play a key role because they reflect the causal structure of the model. The interdependence between endogenous variables is described by \mathbf{B} , and the responses of endogenous variables to exogenous shocks in \mathbf{Z} is reflected in the parameter matrix $\boldsymbol{\Gamma}$. In this setup the causal parameters of interest are those that measure the direct marginal impact of a change in an explanatory variable, y_j or z_k on the outcome of interest y_l , $l \neq j$, and functions of such parameters and data. The elements of $\boldsymbol{\Sigma}$ describe the dispersion and dependence properties of the random disturbances, and hence they measure some properties of the way the data are generated.

2.4.2. Causal Interpretation in SEM

A simple example will illustrate the causal interpretation of parameters in SEM. The structural model has two continuous endogenous variables y_1 and y_2 , a single continuous exogenous variable z_1 , one stochastic relationship linking y_1 and y_2 , and one definitional identity linking all three variables in the model:

$$\begin{aligned} y_1 &= \gamma_1 + \beta_1 y_2 + u_1, \quad 0 < \beta_1 < 1, \\ y_2 &= y_1 + z_1. \end{aligned}$$

In this model u_1 is a stochastic disturbance, independent of z_1 , with a well-defined distribution. The parameter β_1 is subject to an inequality constraint that is also a part of the model specification. The variable z_1 is exogenous and therefore its variation is induced by external sources that we may regard as interventions. These interventions have a direct impact on y_2 through the identity and also an indirect one through the first equation. The impact is measured by the reduced form of the model, which is

$$\begin{aligned} y_1 &= \frac{\gamma_1}{1 - \beta_1} + \frac{\beta_1}{1 - \beta_1} z_1 + \frac{1}{1 - \beta_1} u_1 \\ &= E[y_1|z_1] + v_1, \\ y_2 &= \frac{\gamma_1}{1 - \beta_1} + \frac{1}{1 - \beta_1} z_1 + \frac{1}{1 - \beta_1} u_1 \\ &= E[y_2|z_1] + v_1, \end{aligned}$$

where $v_1 = u_1/(1 - \beta_1)$. The reduced form coefficients $\beta_1/(1 - \beta_1)$ and $1/(1 - \beta_1)$ have a causal interpretation. Any externally induced unit change in z_1 will cause the value of y_1 and y_2 to change by these amounts. Note that in this model y_1 and y_2 also respond to u_1 . In order not to confound the impact of the two sources of variation we require that z_1 and u_1 are independent.

Also note that

$$\begin{aligned} \frac{\partial y_1}{\partial y_2} &= \beta_1 = \frac{\beta_1}{1 - \beta_1} \div \frac{1}{1 - \beta_1} \\ &= \frac{\partial y_1}{\partial z_1} \div \frac{\partial y_2}{\partial z_1}. \end{aligned}$$

In what sense does β_1 measure the causal effect of y_2 on y_1 ? To see a possible difficulty, observe that y_1 and y_2 are interdependent or jointly determined, so it is unclear in what sense y_2 “causes” y_1 . Although z_1 (and u_1) is the ultimate cause of changes in the reduced form sense, y_2 is a proximate or an intermediate cause of y_1 . That is, the first structural equation provides a snapshot of the impact of y_2 on y_1 , whereas the reduced form gives the (equilibrium) impact after allowing for all interactions between the endogenous variables to work themselves out. In a SEM framework even endogenous variables are viewed as causal variables, and their coefficients as causal parameters. This approach can cause puzzlement for those who view causality in an experimental setting where independent sources of variation are the causal variables. The SEM approach makes sense if y_2 has an independent and exogenous source of variation, which in this model is z_1 . Hence the marginal response coefficient β_1 is a function of how y_1 and y_2 respond to a change in z_1 , as the immediately preceding equation makes clear.

Of course this model is but a special case. More generally, we may ask under what conditions will the SEM parameters have a meaningful causal interpretation. We return to this issue when discussing identification concepts in Section 2.5.

2.4.3. Extensions to Nonlinear and Latent Variable Models

If the simultaneous model is **nonlinear in parameters** only, the structural model can be written as

$$\mathbf{YB}(\boldsymbol{\theta}) + \mathbf{Z}\Gamma(\boldsymbol{\theta}) = \mathbf{U}, \quad (2.15)$$

where $\mathbf{B}(\boldsymbol{\theta})$ and $\Gamma(\boldsymbol{\theta})$ are matrices whose elements are functions of the structural parameters $\boldsymbol{\theta}$. An explicit reduced form can be derived as before.

If **nonlinearity** is instead **in variables** then an explicit (analytical) reduced form may not be possible, although linearized approximations or numerical solutions of the dependent variables, given (\mathbf{z}, \mathbf{u}) , can usually be obtained.

Many microeconometric models involve **latent** or **unobserved variables** as well as observed endogenous variables. For example, search and auction theory models use the concept of reservation wage or reservation price, choice models invoke indirect utility, and so forth. In the case of such models the structural model (2.1) may be replaced by

$$\mathbf{g}(\mathbf{y}_i^*, \mathbf{z}_i, \mathbf{u}_i | \boldsymbol{\theta}) = \mathbf{0}, \quad (2.16)$$

where the latent variables \mathbf{y}_i^* replace the observed variables \mathbf{y}_i . The corresponding reduced form solves for \mathbf{y}_i^* in terms of $(\mathbf{z}_i, \mathbf{u}_i)$, yielding

$$\mathbf{y}_i^* = \mathbf{f}(\mathbf{z}_i, \mathbf{u}_i | \boldsymbol{\pi}). \quad (2.17)$$

This reduced form has limited usefulness as \mathbf{y}_i^* is not fully observed. However, if we have functions $\mathbf{y}_i = \mathbf{h}(\mathbf{y}_i^*)$ that relate observable with latent counterparts of \mathbf{y}_i , then the reduced form in terms of observables is

$$\mathbf{y}_i = \mathbf{h}(\mathbf{f}(\mathbf{z}_i, \mathbf{u}_i | \boldsymbol{\pi})). \quad (2.18)$$

See Section 16.8.2 for further details.

When the structural model involves nonlinearities in variables, or when latent variables are involved, an explicit derivation of the functional form of this reduced form may be difficult to obtain. In such cases practitioners use approximations. By citing mathematical or computational convenience, a specific functional form may be used to relate an endogenous variable to all exogenous variables, and the result would be referred to as a “reduced form type relationship.”

2.4.4. Interpretations of Structural Relationships

Marschak (1953, p. 26) in an influential essay gave the following definition of a structure:

Structure was defined as a set of conditions which did not change while observations were being made but which might change in future. If a specified change of structure is expected or intended, prediction of variables of interest to the policy maker requires some knowledge of past structure. . . . In economics, the conditions that constitute a structure are (1) a set of relations describing human behavior and institutions as well as technological laws and involving, in general, nonobservable random disturbances and nonobservable random errors of measurement; (2) the joint probability distribution of these random quantities.

Marschak argued that the structure was fundamental for a quantitative evaluation or tests of economic theory and that the choice of the best policy requires knowledge of the structure.

In the SEM literature a structural model refers to “autonomous” (not “derived”) relationships. There are other closely related concepts of a structure. One such concept refers to “deep parameters,” by which is meant technology and preference parameters that are invariant to interventions.

In recent years an alternative usage of the term structure has emerged, one that refers to econometric models based on the hypothesis of dynamic stochastic optimization by rational agents. In this approach the starting point for any structural estimation problem is the first-order necessary conditions that define the agent’s optimizing behavior. For example, in a standard problem of maximizing utility subject to constraints, the behavioral relations are the deterministic first-order marginal utility conditions. If the relevant functional forms are explicitly stated, and stochastic errors of optimization are introduced, then the first-order conditions define a behavioral model whose parameters characterize the utility function – the so-called deep or policy-invariant parameters. Examples are given in Sections 6.2.7 and 16.8.1.

Two features of this **highly structural approach** should be mentioned. First, they rely on a priori economic theory in a serious manner. Economic theory is not used simply to generate a list of relevant variables that one can use in a more or less arbitrarily specified functional form. Rather, the underlying economic theory has a major (but not exclusive) role in specification, estimation, and inference. The second feature is that identification, specification, and estimation of the resulting model can be very complicated, because the agent’s optimization problem is potentially very complex,

especially if dynamic optimization under uncertainty is postulated and discreteness and discontinuities are present; see Rust (1994).

2.5. Identification Concepts

The goal of the SEM approach is to consistently estimate $(\mathbf{B}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma})$ and conduct statistical inference. An important precondition for consistent estimation is that the model should be identified. We briefly discuss the important twin concepts of **observational equivalence** and **identifiability** in the context of parametric models.

Identification is concerned with determination of a parameter given sufficient observations. In this sense, it is an asymptotic concept. Statistical uncertainty necessarily affects any inference based on a finite number of observations. By hypothetically considering the possibility that sufficient number of observations are available, it is possible to consider whether it is logically possible to determine a parameter of interest either in the sense of its point value or in the sense of determining the set of which the parameter is an element. Therefore, identification is a fundamental consideration and logically occurs prior to and is separate from statistical estimation. A great deal of econometric literature on identification focuses on point identification. This is also the emphasis of this section. However, **set identification**, or **bounds identification**, is an important approach that will be used in selected places in this book (e.g., Chapters 25 and 27; see Manski, 1995).

Definition 2.3 (Observational Equivalence): Two structures of a model defined as joint probability distribution function $\Pr[\mathbf{x}|\boldsymbol{\theta}]$, $\mathbf{x} \in \mathbf{W}$, $\boldsymbol{\theta} \in \Theta$, are **observationally equivalent** if $\Pr[\mathbf{x}|\boldsymbol{\theta}^1] = \Pr[\mathbf{x}|\boldsymbol{\theta}^2] \forall \mathbf{x} \in \mathbf{W}$.

Less formally, if, given the data, two structural models imply identical joint probability distributions of the variables, then the two structures are observationally equivalent. The existence of multiple observationally equivalent structures implies the failure of identification.

Definition 2.4 (Identification): A structure $\boldsymbol{\theta}^0$ is **identified** if there is no other observationally equivalent structure in Θ .

A simple example of nonidentification occurs when there is perfect collinearity between regressors in the linear regression $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$. Then we can identify the linear combination $\mathbf{C}\boldsymbol{\beta}$, where $\text{rank}[\mathbf{C}] < \text{rank}[\boldsymbol{\beta}]$, but we cannot identify $\boldsymbol{\beta}$ itself.

This definition concerns uniqueness of the structure. In the context of the SEM we have given, this definition means that identification requires that there is a unique triple $(\mathbf{B}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma})$ consistent with the observed data. In SEM, as in other cases, identification involves being able to obtain unique estimates of structural parameters given the sample moments of the data. For example, in the case of the reduced form (2.12), under the stated assumptions the least-squares estimator provides unique estimates of $\boldsymbol{\Pi}$, that is, $\widehat{\boldsymbol{\Pi}} = [\mathbf{Z}'\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{Y}$, and identification of $\mathbf{B}, \boldsymbol{\Gamma}$ requires that there is a solution

for the unknown elements of $\boldsymbol{\Gamma}$ and \mathbf{B} from the equations $\boldsymbol{\Pi} + \boldsymbol{\Gamma}\mathbf{B}^{-1} = \mathbf{0}$, given a priori restrictions on the model. A unique solution implies just identification of the model.

A complete model is said to be identified if all the model parameters are identified. It is possible that for some models only a subset of parameters is identified. In some situations it may be important to be able to identify some function of parameters, and not necessarily all the individual parameters. Identification of a function of parameters means that function can be recovered uniquely from $F(\mathbf{W}|\Theta)$.

How does one ensure that the structures of alternative model specifications can be “ruled out”? In SEM the solution to this problem depends on augmenting the sample information by a priori restrictions on $(\mathbf{B}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma})$. The a priori restrictions must introduce sufficient additional information into the model to rule out the existence of other observationally equivalent structures.

The need for a priori restrictions is demonstrated by the following argument. First note that given the assumptions of Section 2.4.1 the reduced form, defined by $(\boldsymbol{\Pi}, \boldsymbol{\Omega})$, is always unique. Initially suppose there are no restrictions on $(\mathbf{B}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma})$. Next suppose that there are two observationally equivalent structures $(\mathbf{B}_1, \boldsymbol{\Gamma}_1, \boldsymbol{\Sigma}_1)$ and $(\mathbf{B}_2, \boldsymbol{\Gamma}_2, \boldsymbol{\Sigma}_2)$. Then

$$\boldsymbol{\Pi} = -\boldsymbol{\Gamma}_1\mathbf{B}_1^{-1} = -\boldsymbol{\Gamma}_2\mathbf{B}_2^{-1}. \quad (2.19)$$

Let \mathbf{H} be a $G \times G$ nonsingular matrix. Then $\boldsymbol{\Gamma}_1\mathbf{B}_1^{-1} = \boldsymbol{\Gamma}_1\mathbf{H}\mathbf{H}^{-1}\mathbf{B}_1^{-1} = \boldsymbol{\Gamma}_2\mathbf{B}_2^{-1}$, which means that $\boldsymbol{\Gamma}_2 = \boldsymbol{\Gamma}_1\mathbf{H}$, $\mathbf{B}_2 = \mathbf{B}_1\mathbf{H}$. Thus the second structure is a linear transformation of the first.

The SEM solution to this problem is to introduce restrictions on $(\mathbf{B}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma})$ such that we can rule out the existence of linear transformations that lead to observationally equivalent structures. In other words, the restrictions on $(\mathbf{B}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma})$ must be such that there is no matrix \mathbf{H} that would yield another structure with the same reduced form; given $(\boldsymbol{\Pi}, \boldsymbol{\Omega})$ there will be a unique solution to the equations $\boldsymbol{\Pi} = -\boldsymbol{\Gamma}\mathbf{B}^{-1}$ and $\boldsymbol{\Omega} \equiv (\mathbf{B}^{-1})'\boldsymbol{\Sigma}\mathbf{B}^{-1}$.

In practice a variety of restrictions can be imposed including (1) normalizations, such as setting diagonal elements of \mathbf{B} equal to 1, (2) zero (exclusion) and linear homogeneous and inhomogeneous restrictions, and (3) covariance and inequality restrictions. Details of the necessary and sufficient conditions for identification in linear and nonlinear models can be found in many texts including Sargan (1988).

Meaningful imposition of identifying restrictions requires that the a priori restrictions imposed should be valid a posteriori. This idea is pursued further in several chapters where identification issues are considered (e.g., Section 6.9).

Exclusion restrictions essentially state that the model contains some variables that have zero impact on some endogenous variables. That is, certain directions of causation are ruled out a priori. This makes it possible to identify other directions of causation. For example, in the simple two-variable example given earlier, z_1 did not enter the y_1 -equation, making it possible to identify the direct impact of y_2 on y_1 . Although exclusion restrictions are the simplest to apply, in parametric models identification can also be secured by inequality restrictions and covariance restrictions.

If there are no restrictions on Σ , and the diagonal elements of \mathbf{B} are normalized to 1, then a **necessary condition** for identification is the **order condition**, which states that the number of excluded exogenous variables must at least equal the number of included endogenous variables. A **sufficient condition** is the **rank condition** given in many texts that ensures for the j th equation parameters $\Pi\Gamma_j = -\mathbf{B}_j$ yields a unique solution for (Γ_j, \mathbf{B}_j) given Π .

Given identification, the term **just (exact) identification** refers to the case when the order condition is exactly satisfied; **overidentification** refers to the case when the number of restrictions on the system exceeds that required for exact identification.

Identification in nonlinear SEM has been discussed in Sargan (1988), who also gives references to earlier related work.

2.6. Single-Equation Models

Without loss of generality consider the first equation of a linear SEM subject to normalization $\beta_{11} = 1$. Let $y = y_1$, let \mathbf{y}_1 denote the endogenous components of \mathbf{y} other than y_1 , and let \mathbf{z}_1 denote the exogenous components of \mathbf{z} with

$$y = \mathbf{y}'_1 \boldsymbol{\alpha} + \mathbf{z}'_1 \boldsymbol{\gamma} + u. \quad (2.20)$$

Many studies skip the formal steps involved in going from a system to a single equation and begin by writing the regression equation

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

where some components of \mathbf{x} are endogenous (implicitly \mathbf{y}_1) and others are exogenous (implicitly \mathbf{z}_1). The focus lies then on estimating the impact of changes in key regressor(s) that may be endogenous or exogenous, depending on the assumptions. Instrumental variable or two-stage least-squares estimation is the most obvious estimation strategy (see Sections 4.8, 6.4, and 6.5).

In the SEM approach it is natural to specify at least some of the remaining equations in the model, even if they are not the focus of inquiry. Suppose \mathbf{y}_1 has dimension 1. Then the first possibility is to specify the structural equation for y_1 and for the other endogenous variables that may appear in this structural equation for y_1 . A second possibility is to specify the reduced form equation for y_1 . This will show exogenous variables that affect y_1 but do not directly affect y . An advantage is that in such a setting instrumental variables emerge naturally. However, in recent empirical work using instrumental variables in a single-equation setting, even the formal step of writing down a reduced form for the endogenous right-hand-side variable is avoided.

2.7. Potential Outcome Model

Motivation for causal inference in econometric models is especially strong when the focus is on the impact of public policy and/or private decision variables on some

specific outcomes. Specific examples include the impact of transfer payments on labor supply, the impact of class size on student learning, and the impact of health insurance on utilization of health care. In many cases the causal variables themselves reflect individual decisions and hence are potentially endogenous. When, as is usually the case, econometric estimation and inference are based on **observational data**, identification of and inference on causal parameters pose many challenges. These challenges can become potentially less serious if the causal issues are addressed using data from a controlled **social experiment** with a proper statistical design. Although such experiments have been implemented (see Section 3.3 for examples and details) they are generally expensive to organize and run. Therefore, it is more attractive to implement causal modeling using data generated by a **natural experiment** or in a quasi-experimental setting. Section 3.4 discusses the pros and cons of these data structures; but for present purposes one should think of a natural or **quasi experiment** as a setting in which some causal variable changes exogenously and independently of other explanatory variables, making it relatively easier to identify causal parameters.

A major obstacle for causality modeling stems from the *fundamental problem of causal inference* (Holland, 1986). Let X be the hypothesized cause and Y the outcome. By manipulating the value of X we can change the value of Y . Suppose the value of X is changed from x_1 to x_2 . Then a measure of the causal impact of the change on Y is formed by comparing the two values of Y : y_2 , which results from the change, and y_1 , which would have resulted had no change in x occurred. However, if X did change, then the value of Y , in the absence of the change, would not be observed. Hence nothing more can be said about causal impact without some hypothesis about what value Y would have assumed in the absence of the change in X . The latter is referred to as a **counterfactual**, which means hypothetical unobserved value. Briefly stated, all causal inference involves comparison of a factual with a counterfactual outcome. In the conventional econometric model (e.g., SEM) a counterfactual does not need to be explicitly stated.

A relatively newer strand in the microeconomic literature – **program evaluation** or **treatment evaluation** – provides a statistical framework for the estimation of causal parameters. In the statistical literature this framework is also known as the **Rubin causal model (RCM)** in recognition of a key early contribution by Rubin (1974, 1978), who in turn cites R.A. Fisher as originator of the approach. Although, following recent convention, we refer to this as the Rubin causal model, Neyman (Splawa-Neyman) also proposed a similar statistical model in an article published in Polish in 1923; see Neyman (1990). Models involving counterfactuals have been independently developed in econometrics following the seminal work of Roy (1951). In the remainder of this section the salient features of RCM will be analyzed.

Causal parameters based on counterfactuals provide statistically meaningful and operational definitions of causality that in some respects differ from the traditional Cowles foundation definition. First, in ideal settings this framework leads to considerable simplicity of econometric methods. Second, this framework typically focuses on

the *fewer* causal parameters that are thought to be most relevant to policy issues that are examined. This contrasts with the traditional econometric approach that focuses simultaneously on all structural parameters. Third, the approach provides additional insights into the properties of causal parameters estimated by the standard structural methods.

2.7.1. The Rubin Causal Model

The term “treatment” is used interchangeably with “cause.” In medical studies of new drug evaluation, involving groups of those who receive the treatment and those who do not, the drug response of the treated is compared with that of the untreated. A measure of causal impact is the average difference in the outcomes of the treated and the nontreated groups. In economics, the term treatment is used very broadly. Essentially it covers variables whose impact on some outcome is the object of study. Examples of treatment–outcome pairs include schooling and wages, class size and scholastic performance, and job training and earnings. Note that a treatment need not be exogenous, and in many situations it is an endogenous (choice) variable.

Within the framework of a **potential outcome model (POM)**, which assumes that every element of the target population is potentially exposed to the treatment, the triple (y_{1i}, y_{0i}, D_i) , $i = 1, \dots, N$, forms the basis of treatment evaluation. The categorical variable D takes the values 1 and 0, respectively, when treatment is or is not received; y_{1i} measures the response for individual i receiving treatment, and y_{0i} measures that when not receiving treatment. That is,

$$y_i = \begin{cases} y_{1i} & \text{if } D_i = 1, \\ y_{0i} & \text{if } D_i = 0. \end{cases} \quad (2.21)$$

Since the receipt and nonreceipt of treatment are mutually exclusive states for individual i , only one of the two measures is available for any given i , the unavailable measure being the counterfactual. The effect of the cause D on outcome of individual i is measured by $(y_{1i} - y_{0i})$. The average causal effect of $D_i = 1$, relative to $D_i = 0$, is measured by the **average treatment effect (ATE)**:

$$\text{ATE} = E[y|D = 1] - E[y|D = 0], \quad (2.22)$$

where expectations are with respect to the probability distribution over the target population. Unlike the conventional structural model that emphasizes marginal effects, the POM framework emphasizes ATE and parameters related to it.

The experimental approach to the estimation of ATE-type parameters involves a **random assignment** of treatment followed by a comparison of the outcomes with a set of nontreated cases that serve as controls. Such an experimental design is explained in greater detail in Chapter 3. Random assignment implies that individuals exposed to treatment are chosen randomly, and hence the treatment assignment does not depend on the outcome and is uncorrelated with the attributes of treated subjects. Two major simplifications follow. The treatment variable can be treated as exogenous and its coefficient in a linear regression will not suffer from omitted variable bias if some

relevant variables are unavoidably omitted from the regression. Under certain conditions, discussed at greater length in Chapters 3 and 25, the mean difference between the outcomes of the treated and the control groups will provide an estimate of ATE. The payoff to the well-designed experiment is the relative simplicity with which causal statements can be made. Of course, to ensure high statistical precision for the treatment effect estimate, one should still control for those attributes that also independently influence the outcomes.

Because random assignment of treatment is generally not feasible in economics, estimation of ATE-type parameters must be based on observational data generated under nonrandom treatment assignment. Then the consistent estimation of ATE will be threatened by several complications that include, for example, possible correlation between the outcomes and treatment, omitted variables, and endogeneity of the treatment variable. Some econometricians have suggested that the absence of randomization comprises the major impediment to convincing statistical inference about causal relationships.

The potential outcome model can lead to causal statements if the counterfactual can be clearly stated and made operational. An explicit statement of the counterfactual, with a clear implication of what should be compared, is an important feature of this model. If, as may be the case with observational data, there is lack of a clear distinction between observed and counterfactual quantities, then the answer to the question of who is affected by the treatment remains unclear. ATE is a measure that weights and combines marginal responses of specific subpopulations. Specific assumptions are required to operationalize the counterfactual. Information on both treated and untreated units that can be observed is needed to estimate ATE. For example, it is necessary to identify the untreated group that proxies the treated group if the treatment were not applied. It is not necessarily true that this step can always be implemented. The exact way in which the treated are selected involves issues of sampling design that are also discussed in Chapters 3 and 25.

A second useful feature of the POM is that it identifies opportunities for causal modeling created by natural or quasi-experiments. When data are generated in such settings, and provided certain other conditions are satisfied, causal modeling can occur without the full complexities of the SEM framework. This issue is analyzed further in Chapters 3 and 25.

Third, unlike the structural form of the SEM where all variables other than that being explained can be labeled as “causes,” in the POM not all explanatory variables can be regarded as causal. Many are simply attributes of the units that must be controlled for in regression analysis, and attributes are not causes (Holland, 1986). Causal parameters must relate to variables that are actually or potentially, and directly or indirectly, subject to intervention.

Finally, identifiability of the ATE parameter may be an easier research goal and hence feasible in situations where the identifiability of a full SEM may not be (Angrist, 2001). Whether this is so has to be determined on a case-by-case basis. However, many available applications of the POM typically employ a limited, rather than full, information framework. However, even within the SEM framework the use of a limited information framework is also feasible, as was previously discussed.

2.8. Causal Modeling and Estimation Strategies

In this section we briefly sketch some of the ways in which econometricians approach the modeling of causal relationships. These approaches can be used within both SEM and POM frameworks, but they are typically identified with the former.

2.8.1. Identification Frameworks

Full-Information Structural Models

One variant of this approach is based on the parametric specification of the joint distribution of endogenous variables conditional on exogenous variables. The relationships are not necessarily derived from an optimizing model of behavior. Parametric restrictions are placed to ensure identification of the model parameters that are the target of statistical inference. The entire model is estimated simultaneously using maximum likelihood or moments-based estimation. We call this approach the **full-information structural approach**. For well-specified models this is an attractive approach but in general its potential limitation is that it may contain some equations that are poorly specified. Under joint estimation the effects of localized misspecification may also affect other estimates.

Statistically we may interpret the full-information approach as one in which the joint probability distribution of endogenous variables, given the exogenous variables, forms the basis of inference about causality. The jointness may derive from contemporaneous or dynamic interdependence between endogenous variables and/or the disturbances on the equations.

Limited-Information Structural Models

By contrast, when the central object of statistical inference is estimation of one or two key parameters, a **limited-information** approach may be used. A feature of this approach is that, although one equation is the focus of inference, the joint dependence between it and other endogenous variables is exploited. This requires that explicit assumptions are made about some features of the model that are not the main object of inference. Instrumental variable methods, sequential multistep methods, and limited information maximum likelihood methods are specific examples of this approach. To implement the approach one typically works with one (or more) structural equations and some implicitly or explicitly stated reduced form equations. This contrasts with the full-information approach where all equations are structural. The limited-information approach is often computationally more tractable than the full-information one.

Statistically we may interpret the limited-information approach as one in which the joint distribution is factored into the product of a conditional model for the endogenous variable(s) of interest, say \mathbf{y}_1 , and a marginal model for other endogenous variables, say \mathbf{y}_2 , which are in the set of the conditioning variables, as in

$$f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = g(\mathbf{y}_1|\mathbf{x}, \mathbf{y}_2, \boldsymbol{\theta}_1)h(\mathbf{y}_2|\mathbf{x}, \boldsymbol{\theta}_2), \quad \boldsymbol{\theta} \in \Theta. \quad (2.23)$$

Modeling may be based on the component $g(\mathbf{y}_1|\mathbf{x}, \mathbf{y}_2, \boldsymbol{\theta}_1)$ with minimal attention to $h(\mathbf{y}_2|\mathbf{x}, \boldsymbol{\theta}_2)$ if $\boldsymbol{\theta}_2$ are regarded as **nuisance parameters**. Of course, such a factorization is not unique, and hence the limited-information approach can have several variants.

Identified Reduced Forms

A third variant of the SEM approach works with an **identified reduced form**. Here too one is interested in structural parameters. However, it may be convenient to estimate these from the reduced form subject to restrictions. In time series the identified vector autoregressions provide an example.

2.8.2. Identification Strategies

There are numerous potential ways in which the identification of key model parameters can be jeopardized. Omitted variables, functional form misspecifications, measurement errors in explanatory variables, using data unrepresentative of the population, and ignoring endogeneity of explanatory variables are leading examples. Microeometrics contains many specific examples of how these challenges can be tackled. Angrist and Krueger (2000) provide a comprehensive survey of popular identification strategies in labor economics, with emphasis on the POM framework. Most of the issues are developed elsewhere in the book, but a brief mention is made here.

Exogenization

Data are sometimes generated by natural experiments and quasi-experiments. The idea here is simply that a policy variable may exogenously change for some subpopulation while it remains the same for other subpopulations. For example, minimum wage laws in one state may change while they remain unchanged in a neighboring state. Such events naturally create treatment and control groups. If the natural experiment approximates a randomized treatment assignment, then exploiting such data to estimate structural parameters can be simpler than estimation of a larger simultaneous equations model with endogenous treatment variables. It is also possible that the treatment variable in a natural experiment can be regarded as exogenous, but the treatment itself is not randomly assigned.

Elimination of Nuisance Parameters

Identification may be threatened by the presence of a large number of nuisance parameters. For example, in a cross-section regression model the conditional mean function $E[y_i|\mathbf{x}_i]$ may involve an individual specific fixed effect α_i , assumed to be correlated with the regression error. This effect cannot be identified without many observations on each individual (i.e., panel data). However, with just a short panel it could be eliminated by a transformation of the model. Another example is the presence of time-invariant unobserved exogenous variables that may be common to groups of individuals.

An example of a transformation that eliminates fixed effects is taking differences and working with the differences-in-differences form of the model.

Controlling for Confounders

When variables are omitted from a regression, and when omitted factors are correlated with the included variables, a confounding bias results. For example, in a regression with earnings as a dependent variable and schooling as an explanatory variable, individual ability may be regarded as an omitted variable because only imperfect proxies for it are typically available. This means that potentially the coefficient of the schooling variable may not be identified. One possible strategy is to introduce **control variables** in the model; the general approach is called the **control function approach**. These variables are an attempt to approximate the influence of the omitted variables. For example, various types of scholastic achievement scores may serve as controls for ability.

Creating Synthetic Samples

Within the POM framework a causal parameter may be unidentified because no suitable comparison or control group can provide the benchmark for estimation. A potential solution is to create a synthetic sample that includes a comparison group that are proxies for controls. Such a sample is created by **matching** (discussed in Chapter 25). If treated samples can be augmented by well-matched controls, then identification of causal parameters can be achieved in the sense that a parameter related to ATE can be estimated.

Instrumental Variables

If identification is jeopardized because the treatment variable is endogenous, then a standard solution is to use valid instrumental variables. This is easier said than done. The choice of the instrumental variable as well as the interpretation of the results obtained must be done carefully because the results may be sensitive to the choice of instruments. The approach is analyzed in Sections 4.8, 4.9, 6.4, 6.5, and 25.7, as well as in several other places in the book as the need arises. Again a natural experiment may provide a valid instrument.

Reweighting Samples

Sample-based inferences about the population are only valid if the sample data are representative of the population. The problem of sample selection or biased sampling arises when the sample data are not representative, in which case the population parameters are not identified. This problem can be approached as one that requires correction for sample selection (Chapter 16) or one that requires reweighting of the sample information (Chapter 24).

2.9. Bibliographic Notes

- 2.1** The 2001 Nobel lectures by Heckman and McFadden are excellent sources for both historical and current information about the developments in microeconomics. Heckman's lecture is remarkable for its comprehensive scope and offers numerous insights into many aspects of microeconomics. His discussion of heterogeneity has many points of contact with several topics covered in this book.
- 2.2** Marschak (1953) gives a classic statement of the primacy of structural modeling for policy evaluation. He makes an early mention of the idea of parameter invariance.
- 2.3** Engle, Hendry, and Richard (1983) provide definitions of weak and strong exogeneity in terms of the distribution of observable variables. They make links with previous literature on exogeneity concepts.
- 2.4** and **2.5** The term "identification" was used by Koopmans (1949). Point identification in linear parametric models is covered in most textbooks including those by Sargan (1988) who gives a comprehensive and succinct treatment, Davidson and MacKinnon (2004), and Greene (2003). Gouriéroux and Monfort (1989, chapter 3.4) provide a different perspective using Fisher and Kullback information measures. Bounds identification in several leading cases is developed in Manski (1995).
- 2.6** Heckman (2000) provides a historical overview and modern interpretations of causality in the traditional econometric model. Causality concepts within the POM framework are carefully and incisively analyzed by Holland (1986), who also relates them to other definitions. A sample of the statisticians' viewpoints of causality from a historical perspective can be found in Freedman (1999). Pearl (2000) gives insightful schematic exposition of the idea of "treating causation as a summary of behavior under interventions," as well as numerous problems of inferring causality in a nonexperimental situation.
- 2.7** Angrist and Krueger (1999) survey solutions to identification pitfalls with examples from labor economics.

CHAPTER 3

Microeconomic Data Structures

3.1. Introduction

This chapter surveys issues concerning the potential usefulness and limitations of different types of microeconomic data. By far the most common data structure used in microeconomics is survey or census data. These data are usually called **observational data** to distinguish them from **experimental data**.

This chapter discusses the potential limitation of the aforementioned data structures. The inherent limitations of observational data may be further compounded by the manner in which the data are collected, that is, by the sample frame (the way the sample is generated), sample design (simple random sample versus stratified random sample), and sample scope (cross-section versus longitudinal data). Hence we also discuss sampling issues in connection with the use of observational data. Some of this terminology is new at this stage but will be explained later in this chapter.

Microeconomics goes beyond the analysis of survey data under the assumptions of simple random sampling. This chapter considers extensions. Section 3.2 outlines the structure of multistage sample surveys and some common forms of departure from random sampling; a more detailed analysis of their statistical implications is provided in later chapters. It also considers some commonly occurring complications that result in the data not being necessarily representative of the population. Given the deficiencies of observational data in estimating causal parameters, there has been an increased attempt at exploiting experimental and quasi-experimental data and frameworks. Section 3.3 examines the potential of data from social experiments. Section 3.4 considers the modeling opportunities arising from a special type of observational data, generated under quasi-experimental conditions, that naturally provide treated and untreated subjects and hence are called natural experiments. Section 3.5 covers practical issues of microdata management.

3.2. Observational Data

The major source of microeconomic observational data is surveys of households, firms, and government administrative data. Census data can also be used to generate samples. Many other samples are often generated at points of contact between transacting parties. For example, marketing data may be generated at the point of sale and/or surveys among (actual or potential) purchasers. The Internet (e.g., online auctions) is also a source of data.

There is a huge literature on sample surveys from the viewpoint of both survey statisticians and users of survey data. The first discusses how to sample from the population and the results from different sampling designs, and the second deals with the issues of estimation and inference that arise when survey data are collected using different sampling designs. A key issue is how well the sample represents the population. This chapter deals with both strands of the literature in an introductory fashion. Many additional details are given in Chapter 24.

3.2.1. Nature of Survey Data

The term observational data usually refers to survey data collected by sampling the relevant population of subjects without any attempt to control the characteristics of the sampled data. Let t denote the time subscript, let \mathbf{w} denote a set of variables of interest. In the present context t can be a point in time or time interval. Let \mathcal{S}_t denote a sample from population probability distribution $F(\mathbf{w}_t|\boldsymbol{\theta}_t)$; \mathcal{S}_t is a draw from $F(\mathbf{w}_t|\boldsymbol{\theta}_t)$, where $\boldsymbol{\theta}$ is a parameter vector. The population should be thought of as a set of points with characteristics of interest, and for simplicity we assume that the form of the probability distribution F is known. A simple random sampling scheme allows every element of the population to have an equal probability of being included in the sample. More complex sampling schemes will be considered later.

The abstract concept of a **stationary population** provides a useful benchmark. If the moments of the characteristics of the population are constant, then we can write $\boldsymbol{\theta}_t = \boldsymbol{\theta}$, for all t . This is a strong assumption because it implies that the moments of the characteristics of the population are time-invariant. For example, the age–sex distribution should be constant. More realistically, some population characteristics would not be constant. To handle such a possibility, (the parameters of) each population may be regarded as a draw from a **superpopulation** with constant characteristics. Specifically, we think of each $\boldsymbol{\theta}_t$ as a draw from a probability distribution with constant (hyper)parameter $\boldsymbol{\theta}$. The terms superpopulation and hyperparameters occur frequently in the literature on hierarchical models discussed in Chapter 24. Additional complications arise if $\boldsymbol{\theta}_t$ has an evolutionary component, for example through dependence on t , or if successive values are interdependent. Using hierarchical models, discussed in Chapters 13 and 26, provides one approach for modeling the relation between hyperparameters and subpopulation characteristics.

3.2.2. Simple Random Samples

As a benchmark for subsequent discussion, consider simple random sampling in which the probability of sampling unit i from a population of size N , with N large, is $1/N$ for all i . Partition \mathbf{w} as $[y : \mathbf{x}]$. Suppose our interest is in modeling y , a possibly vector-valued outcome variable, conditional on the exogenous covariate vector \mathbf{x} , whose joint distribution is denoted $f_J(y, \mathbf{x})$. This can be always be factored as the product of the conditional distribution $f_C(y|\mathbf{x}, \boldsymbol{\theta})$ and the marginal distribution $f_M(\mathbf{x})$:

$$f_J(y, \mathbf{x}) = f_C(y|\mathbf{x}, \boldsymbol{\theta})f_M(\mathbf{x}). \quad (3.1)$$

Simple random sampling involves drawing the (y, \mathbf{x}) combinations uniformly from the entire population.

3.2.3. Multistage Surveys

One alternative is a **stratified multistage cluster sampling**, also referred to as a **complex survey** method. Large-scale surveys like the Current Population Survey (CPS) and the Panel Survey of Income Dynamics (PSID) take this approach. Section 24.2 provides additional detail on the structure of the CPS.

The complex survey design has advantages. It is more cost effective because it reduces geographical dispersion, and it becomes possible to sample certain subpopulations more intensively. For example, “oversampling” of small subpopulations exhibiting some relevant characteristic becomes feasible whereas a random sample of the population would produce too few observations to support reliable results. A disadvantage is that stratified sampling will reduce interindividual variation, which is essential for greater precision.

The sample survey literature focuses on **multistage surveys** that sequentially partition the population into the following categories:

1. **Strata**: Nonoverlapping subpopulations that exhaust the population.
2. **Primary sampling units** (PSUs): Nonoverlapping subsets of the strata.
3. **Secondary sampling units** (SSUs): Sub-units of the PSU, which may in turn be partitioned, and so on.
4. **Ultimate sampling unit** (USU): The final unit chosen for interview, which could be a household or a collection of households (a segment).

As an example, the strata may be the various states or provinces in a country, the PSU may be regions within the state or province, and the USU may be a small cluster of households in the same neighborhood.

Usually all strata are surveyed so that, for example, all states will be included in the sample with certainty. But not all of the PSUs and their subdivisions are surveyed, and they may be sampled at different rates. In **two-stage sampling** the surveyed PSUs are drawn at random and the USU is then drawn at random from the selected PSUs. In **multistage sampling** intermediate sampling units such as SSUs also appear.

A consequence of these sampling methods is that different households will have different probabilities of being sampled. The sample is then unrepresentative of the population. Many surveys provide **sampling weights** that are intended to be inversely proportional to the probability of being sampled, in which case these weights can be used to obtain unbiased estimators of population characteristics.

Survey data may be clustered due to, for example, sampling of many households in the same small neighborhood. Observations in the same cluster are likely to be dependent or correlated because they may depend on some observable or unobservable factor that could affect all observations in a stratum. For example, a suburb may be dominated by high-income households or by households that are relatively homogeneous in some dimension of their preferences. Data from these households will tend to be correlated, at least unconditionally, though it is possible that such correlation is negligible after conditioning on observable characteristics of the households. Statistical inference ignoring correlation between sampled observations yields erroneous estimates of variances that are smaller than those from the correct formula. These issues are covered in greater depth in Section 24.5. Two-stage and multistage samples potentially further complicate the computation of standard errors.

In summary, (1) stratification with different sampling rates within strata means that the sample is unrepresentative of the population; (2) sampling weights inversely proportional to the probability of being sampled can be used to obtain unbiased estimation of population characteristics; and (3) clustering may lead to correlation of observations and understatement of the true standard errors of estimators unless appropriate adjustments are made.

3.2.4. Biased Samples

If a random sample is drawn then the probability distribution for the data is the same as the population distribution. Certain departures from random sampling cause a divergence between the two; this is referred to as **biased sampling**. The data distribution differs from the population distribution in a manner that depends on the nature of the deviation from random sampling. Deviation from random sampling occurs because it is sometimes more convenient or cost effective to obtain the data from a subpopulation even though it is not representative of the entire population. We now consider several examples of such departures, beginning with a case in which there is no departure from randomness.

Exogenous Sampling

Exogenous sampling from survey data occurs if the analyst segments the available sample into subsamples based only on a set of exogenous variables \mathbf{x} , but not on the response variable. For example, in a study of hospitalizations in Germany, Geil et al. (1997) segmented the data into two categories, those with and without chronic conditions. Classification by income categories is also common. Perhaps it is more accurate to depict this type of sampling as exogenous subsampling because it is done by reference to an existing sample that has already been collected. Segmenting an existing

sample by gender, health, or socioeconomic status is very common. Under the assumptions of exogenous sampling the probability distribution of the exogenous variables is independent of y and contains no information about the population parameters of interest, θ . Therefore, one may ignore the marginal distribution of the exogenous variables and simply base estimation on the conditional distribution $f(y|\mathbf{x}, \theta)$. Of course, the assumption may be wrong and the observed distribution of the outcome variable may depend on the selected segmenting variable, which may be correlated with the outcome, thus causing departure from exogenous sampling.

Response-Based Sampling

Response-based sampling occurs if the probability of an individual being included in the sample depends on the responses or choices made by that individual. In this case sample selection proceeds in terms of rules defined in terms of the endogenous variable under study.

Three examples are as follows: (1) In a study of the effect of negative income tax or Aid to Families with Dependent Children (AFDC) on labor supply only those below the poverty line are surveyed. (2) In a study of determinants of public transport modal choice, only users of public transport (a subpopulation) are surveyed. (3) In a study of the determinants of number of visits to a recreational site, only those with at least one visit are included.

Lower survey costs provide an important motivation for using choice-based samples in preference to simple random samples. It would require a very large random sample to generate enough observations (information) about a relatively infrequent outcome or choice, and hence it is cheaper to collect a sample from those who have actually made the choice.

The practical significance of this is that consistent estimation of population parameters θ can no longer be carried out using the conditional population density $f(y|\mathbf{x})$ alone. The effect of the sampling scheme must also be taken into account. This topic is discussed further in Section 24.4.

Length-Biased Sampling

Length-biased sampling illustrates how biases may result from sampling one population to make inferences about a different population. Strictly speaking, it is not so much an example of departure from randomness in sampling as one of sampling the “wrong” population.

Econometric studies of transitions model the time spent in origin state j by individual i before transiting to another destination state s . An example is when j corresponds to unemployment and s to employment. The data used in such studies can come from one of several possible sources. One source is sampling individuals who are unemployed on a particular date, another is to sample those who are in the labor force regardless of their current state, and a third is to sample individuals who are either entering or leaving unemployment during a specified period of time. Each type of sampling scheme is based on a different concept of the relevant population. In the

first case the relevant population is the stock of unemployed individuals, in the second the labor force, and in the third individuals with transitioning employment status. This topic is discussed further in Section 18.6.

Suppose that the purpose of the survey is to calculate a measure of the average duration of unemployment. This is the average length of time a randomly chosen individual will spend in unemployment if he or she becomes unemployed. The answer to this apparently straightforward question may vary depending on how the sample data are obtained. The flow distribution of completed durations is in general quite different from the stock distribution. When we sample the stock, the probability of being in the sample is higher for individuals with longer durations. When we sample the flow out of the state, the probability does not depend on the time spent in the state. This is the well-known example of length-biased sampling in which the estimate obtained by sampling the stock is a biased estimate of the average length of an unemployment spell of a random entrant to unemployment.

The following simple schematic diagram may clarify the point:



Here we use the symbol \bullet to denote slow movers and the symbol \circ to denote fast movers. Suppose the two types are equally represented in the flow, but the slow movers stay in the stock longer than the fast movers. Then the stock population has a higher proportion of slow movers. Finally, the exit population has a higher proportion of fast movers. The argument will generalize to other types of heterogeneity.

The point of this example is not that flow sampling is a better thing to do than stock sampling. Rather, it is that, depending on what the question is, stock sampling may not yield a random sample of the relevant population.

3.2.5. Bias due to Sample Selection

Consider the following problem. A researcher is interested in measuring the effect of training, denoted z (treatment), on posttraining wages, denoted y (outcome), given the worker's characteristics, denoted x . The variable z takes the value 1 if the worker has received training and is 0 otherwise. Observations are available on (x, D) for all workers but on y only for those who received training ($D = 1$). One would like to make inferences about the average impact of training on the posttraining wage of a randomly chosen worker with known characteristics who is currently untrained ($D = 0$). The problem of **sample selection** concerns the difficulty of making such an inference.

Manski (1995), who views this as a problem of identification, defines the selection problem formally as follows:

This is the problem of identifying conditional probability distributions from random sample data in which the realizations of the conditioning variables are always observed but realizations of the outcomes are censored.

3.2. OBSERVATIONAL DATA

Suppose y is the outcome to be predicted, and the conditioning variables are denoted by x . The variable z is a censoring indicator that takes the value 1 if the outcome y is observed and 0 otherwise. Because the variables (D, x) are always observed, but y is observed only when $D = 1$, Manski views this as a *censored sampling process*. The censored sampling process does not identify $\Pr[y|x]$, as can be seen from

$$\Pr[y|x] = \Pr[y|x, D = 1]\Pr[D = 1|x] + \Pr[y|x, D = 0]\Pr[D = 0|x]. \quad (3.2)$$

The sampling process can identify three of the four terms on the right-hand side, but provides no information about the term $\Pr[y|x, D = 0]$. Because

$$E[y|x] = E[y|x, D = 1] \cdot \Pr[D = 1|x] + E[y|x, D = 0] \cdot \Pr[D = 0|x],$$

whenever the censoring probability $\Pr[D = 0|x]$ is positive, the available empirical evidence places no restrictions on $E[y|x]$. Consequently, the censored-sampling process can identify $\Pr[y|x]$ only for some unknown value of $\Pr[y|x, D = 0]$. To learn anything about the $E[y|x]$, restrictions will need to be placed on $\Pr[y|x]$.

The alternative approaches for solving this problem are discussed in Section 16.5.

3.2.6. Quality of Survey Data

The quality of sample data depends not only on the sample design and the survey instrument but also on the survey responses. This observation applies especially to observational data. We consider several ways in which the quality of the sample data may be compromised. Some of the problems (e.g., attrition) can also occur with other types of data. This topic overlaps with that of biased sampling.

Problem of Survey Nonresponse

Surveys are normally voluntary, and incentive to participate may vary systematically according to household characteristics and type of question asked. Individuals may refuse to answer some questions. If there is a systematic relationship between refusal to answer a question and the characteristics of the individual, then the issue of the representativeness of a survey after allowing for **nonresponse** arises. If nonresponse is ignored, and if the analysis is carried out using the data from respondents only, how will the estimation of parameters of interest be affected?

Survey nonresponse is a special case of the selection problem mentioned in the preceding section. Both involve biased samples. To illustrate how it leads to distorted inference consider the following model:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \mid \mathbf{x}, \mathbf{z} \sim \mathcal{N} \left(\begin{bmatrix} \mathbf{x}'\boldsymbol{\beta} \\ \mathbf{z}'\boldsymbol{\gamma} \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right), \quad (3.3)$$

where y_1 is a continuous random variable of interest (e.g., expenditure) that depends on \mathbf{x} , and y_2 is a latent variable that measures the “propensity to participate” in a survey

and depends on \mathbf{z} . The individual participates if $y_2 > 0$; otherwise the individual does not. The variables \mathbf{x} and \mathbf{z} are assumed to be exogenous. The formulation allows y_1 and y_2 to be correlated.

Suppose we estimate β from the data supplied by participants by least squares. Is this estimator unbiased in the presence of nonparticipation? The answer is that if nonparticipation is random and independent of y_1 , the variable of interest, then there is no bias, but otherwise there will be.

The argument is as follows:

$$\hat{\beta} = [\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'y_1,$$

$$E[\hat{\beta} - \beta] = E\left[[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'E[y_1 - \mathbf{X}\beta|\mathbf{X}, \mathbf{Z}, y_2 > 0]\right],$$

where the first line gives the least-squares formula for the estimates of β and the second line gives its bias. If y_1 and y_2 are independent, conditional on \mathbf{X} and \mathbf{Z} , $\sigma_{12} = 0$, then

$$E[y_1 - \mathbf{X}\beta|\mathbf{X}, \mathbf{Z}, y_2 > 0] = E[y_1 - \mathbf{X}\beta|\mathbf{X}, \mathbf{Z}] = \mathbf{0},$$

and there is no bias.

Missing and Mismeasured Data

Survey respondents dealing with an extensive questionnaire will not necessarily answer every question and even if they do, the answers may be deliberately or fortuitously false. Suppose that the sample survey attempts to obtain a vector of responses denoted as $\mathbf{x}_i = (x_{i1}, \dots, x_{iK})$ from N individuals, $i = 1, \dots, N$. Suppose now that if an individual fails to provide information on any one or more elements of \mathbf{x}_i , then the entire vector is discarded. The first problem resulting from **missing data** is that the sample size is reduced. The second potentially more serious problem is that missing data can potentially lead to biases similar to the selection bias. If the data are missing in a systematic manner, then the sample that is left to analyze may not be representative of the population. A form of selection bias may be induced by any systematic pattern of nonresponse. For example, high-income respondents may systematically not respond to questions about income. Conversely, if the data are missing completely at random then discarding incomplete observations will reduce precision but not generate biases. Chapter 27 discusses the missing-data problem and solutions in greater depth.

Measurement errors in survey responses are a pervasive problem. They can arise from a variety of causes, including incorrect responses arising from carelessness, deliberate misreporting, faulty recall of past events, incorrect interpretation of questions, and data-processing errors. A deeper source of measurement error is due to the measured variable being at best an imperfect **proxy** for the relevant theoretical concept. The consequences of such measurement errors is a major topic and is discussed in Chapter 26.

Sample Attrition

In panel data situations the survey involves repeated observations on a set of individuals. In this case we can have

- full response in all periods (full participation),
- nonresponse in the first period and in all subsequent periods (nonparticipation), or
- partial response in the sense of response in the initial periods but nonresponse in later periods (incomplete participation) – a situation referred to as **sample attrition**.

Sample attrition leads to missing data, and the presence of any nonrandom pattern of “missingness” will lead to the sample selection type problems already mentioned. This can be interpreted as a special case of the sample selection problem. Sample attrition is discussed briefly in Sections 21.8.5 and 23.5.2.

3.2.7. Types of Observational Data

Cross-section data are obtained by observing \mathbf{w} , for the sample \mathcal{S}_t for some t . Although it is usually impractical to sample all households at the same point of time, cross-section data are still a snapshot of characteristics of each element of a subset of the population that will be used to make inferences about the population. If the population is stationary, then inferences made about $\boldsymbol{\theta}_t$ using \mathcal{S}_t may be valid also for $t' \neq t$. If there is significant dependence between past and current behavior, then longitudinal data are required to identify the relationship of interest. For example, past decisions may affect current outcomes; inertia or habit persistence may account for current purchases, but such dependence cannot be modeled if the history of purchases is not available. This is one of the limitations imposed by cross-section data.

Repeated cross-section data are obtained by a sequence of independent samples \mathcal{S}_t taken from $F(\mathbf{w}_t | \boldsymbol{\theta}_t)$, $t = 1, \dots, T$. Because the sample design does not attempt to retain the same units in the sample, information about dynamic dependence in behavior is lost. If the population is stationary then repeated cross-section data are obtained by a sampling process somewhat akin to sampling with replacement from the constant population. If the population is nonstationary, repeated cross sections are related in a manner that depends on how the population is changing over time. In such a case the objective is to make inferences about the underlying constant (hyper)parameters. The analysis of repeated cross sections is discussed in Section 22.7.

Panel or longitudinal data are obtained by initially selecting a sample \mathcal{S} and then collecting observations for a sequence of time periods, $t = 1, \dots, T$. This can be achieved by interviewing subjects and collecting both present and past data at the same time, or by tracking the subjects once they have been inducted into the survey. This produces a sequence of data vectors $\{\mathbf{w}_1, \dots, \mathbf{w}_T\}$ that are used to make inferences about either the behavior of the population or that of the particular sample of individuals. The appropriate methodology in each case may not be the same. If the data are drawn from a nonstationary population, the appropriate objective should be inference on (hyper)parameters of the superpopulation.

Some limitations of these types of data are immediately obvious. Cross-section samples and repeated cross-sections do not in general provide suitable data for modeling intertemporal dependence in outcomes. Such data are only suitable for modeling static relationships. In contrast, longitudinal data, especially if they span a sufficiently long time period, are suitable for modeling both static and dynamic relationships.

Longitudinal data are not free from problems. The first issue is representativeness of the panel. Problems of inference regarding population behavior using longitudinal data become more difficult if the population is not stationary. For analyzing dynamics of behavior, retaining original households in the panel for as long as possible is an attractive option. In practice, longitudinal data sets suffer from the problem of “sample attrition,” perhaps due to “sample fatigue.” This simply means that survey respondents do not continue to provide responses to questionnaires. This creates two problems: (1) The panel becomes unbalanced and (2) there is the danger that the retained household may not be “typical” and that the sample becomes unrepresentative of the population. When the available sample data are not a random draw from the population, results based on different types of data will be susceptible to biases to different degrees. The problem of “sample fatigue” arises because over time it becomes more difficult to retain individuals within the panel or they may be “lost” (censored) for some other reason, such as a change of location. These issues are dealt with later in the book. Analysis of longitudinal data may nevertheless provide information about some aspects of the behavior of the sampled units, although extrapolation to population behavior may not be straightforward.

3.3. Data from Social Experiments

Observational and experimental data are distinct because an experimental environment can in principle be closely monitored and controlled. This makes it possible to vary a causal variable of interest, holding other covariates at controlled settings. In contrast, observational data are generated in an uncontrolled environment, leaving open the possibility that the presence of confounding factors will make it more difficult to identify the causal relationship of interest. For example, when one attempts to study the earnings–schooling relationship using observational data, one must accept that the years of schooling of an individual is itself an outcome of an individual’s decision-making process, and hence one cannot regard the level of schooling as if it had been set by a hypothetical experimenter.

In social sciences, data analogous to experimental data come from either **social experiments**, defined and described in greater detail in the following, or from “laboratory” experiments on small groups of voluntary participants that mimic the behavior of economic agents in the real-life counterpart of the experiment. Social experiments are relatively uncommon, and yet experimental concepts, methods, and data serve as a benchmark for evaluating econometric studies based on observational data.

This section provides a brief account of the methodology of social experiments, the nature of the data emanating from them, and some problems and issues of econometric methodology that they generate.

The central feature of the experimental methodology involves a comparison between the outcomes of the randomly selected experimental group that is subjected to a “**treatment**” with those of a **control** (comparison) group. In a good experiment considerable care is exercised in matching the control and experimental (“treated”) groups, and in avoiding potential biases in outcomes. Such conditions may not be realized in observational environments, thereby leading to a possible lack of identification of causal parameters of interest. Sometimes, however, experimental conditions may be approximately replicated in observational data. Consider, for example, two contiguous regions or states, one of which pursues a different minimum-wage policy from the other, creating the conditions of a **natural experiment** in which observations from the “treated” state can be compared with those from the “control” state. The data structure of a natural experiment has also attracted attention in econometrics.

A social experiment involves exogenous variations in the economic environment facing the set of experimental subjects, which is partitioned into one subset that receives the experimental treatment and another that serves as a control group. In contrast to observational studies in which changes in exogenous and endogenous factors are often confounded, a well-designed social experiment aims to isolate the role of treatment variables. In some experimental designs there may be no explicit **control group**, but varying levels of the treatment are applied, in which case it becomes possible in principle to estimate the entire **response surface** of experimental outcomes.

The primary object of a social experiment is to estimate the impact of an actual or potential social program. The potential outcome model of Section 2.7 provides a relevant background for modeling the impact of social experiments. Several alternative measures of impact have been proposed and these will be discussed in the chapter on program evaluation (Chapter 25).

Burtless (1995) summarizes the case for social experiments, while noting some potential limitations. In a companion article Heckman and Smith (1995) focus on limitations of actual social experiments that have been implemented. The remaining discussion in this section borrows significantly from these papers.

3.3.1. Leading Features of Social Experiments

Social experiments are motivated by policy issues about how subjects would react to a type of policy that has never been tried and hence one for which no observed response data exist. The idea of a social experiment is to enlist a group of willing participants, some of whom are randomly assigned to a treatment group and the rest to a control group. The difference between the responses of those in the treatment group, subjected to the policy change, and those in the control group, who are not, is the estimated effect of the policy. Schematically the standard experimental design is as depicted in Figure 3.1.

The term “**experimental**s” refers to the group receiving treatments, “**controls**” to the group not receiving treatment, and “**random assignment**” to the process of assigning individuals to the two groups.

Randomized trials were introduced in statistics by R. A. Fisher (1928) and his co-workers. A typical agricultural experiment would consist of a trial in which a new

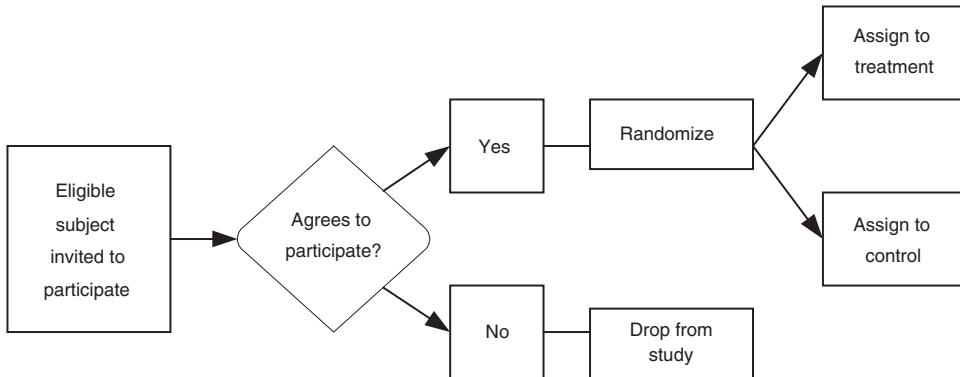


Figure 3.1: Social experiment with random assignment.

treatment such as fertilizer application would be applied to plants growing on randomly chosen blocks of land and then the responses would be compared with those of a control group of plants, similar to the experimentals in all relevant respects but not given experimental treatment. If the effect of all other differences between the experimental and control groups can be eliminated, the estimated difference between the two sets of responses can be attributed to the treatment. In the simplest situation one can concentrate on a comparison of the mean outcome of the treated group and of the untreated group.

Although in agricultural and biomedical sciences, the randomized experiments methodology has been long established, in economics and social sciences it is new. It is attractive for studying responses to policy changes for which no observational data exist, perhaps because the policy changes of interest have never occurred. Randomized experiments also permit a greater variation in policy variables and parameters than are present in observational data, thereby making it easier to identify and study responses to policy changes. In many cases the social experiment may try out a policy that has never been tried, so the observational data remain completely silent on its potential impact.

Social experiments are still rather rare outside the United States, partly because they are expensive to run. In the United States a number of such experiments have taken place since the early 1970s. Table 3.1 summarizes features of some relatively well-known examples; for a more extensive coverage see Burtless (1995).

An experiment may produce either cross-section or longitudinal data, although cost considerations will usually limit the time dimension well below what is typical in observational data. When an experiment lasts several years and has multiple stages and/or geographical locations, as in the case of RHIE, interim analyses based on “incomplete” data are not uncommon (Newhouse et al., 1993).

3.3.2. Advantages of Social Experiments

Burtless (1995) surveys the advantages of social experiments with great clarity. The key advantage stems from randomized trials that remove any correlation between the observed and unobserved characteristics of program participants. Hence the

3.3. DATA FROM SOCIAL EXPERIMENTS

Table 3.1. Features of Some Selected Social Experiments

Experiment	Tested Treatments	Target Population
Rand Health Insurance Experiment (RHIE), 1974–1982	Health insurance plans with varying copayment rate and differing levels of maximum out-of-pocket expenses	Low- and moderate-level income persons and families
Negative Income Tax (NIT), 1968–1978	NIT plans with alternative income guarantees and tax rates	Low- and moderate-level income persons and families with nonaged head of household
Job Training Partnership Act (JTPA), (1986–1994)	Job search assistance, on-the-job training, classroom training financed under JTPA	Out-of-school youths and disadvantaged adults

contribution of the treatment to the outcome difference between the treated and control groups can be estimated without confounding bias even if one cannot control for the confounding variables. The presence of correlation between treatment and confounding variables often plagues observational studies and complicates causal inference. By contrast, an experimental study conducted under ideal circumstances can produce a consistent estimate of the average difference in outcomes of the treated and nontreated groups without much computational complexity.

If, however, an outcome depends on treatment as well as other observable factors, then controlling for the latter will in general improve the precision of the impact estimate.

Even if observational data are available, the generation and use of experimental data has great appeal because it offers the possibility of **exogenizing** a policy variable, and randomization of treatments can potentially lead to great simplification of statistical analysis. Conclusions based on observational data often lack generality because they are based on a nonrandom sample from the population – the problem of selection bias. An example is the aforementioned RHIE study whose major focus is on the price responsiveness of the demand for health services. Availability of health insurance affects the user price of health services and thereby its use. An important policy issue is the extent to which “overutilization” of health services would result from subsidized health insurance. One can, of course, use observational data to model the relation between the demand for health services and the level of insurance. However, such analyses are subject to the criticism that the level of health insurance should not be treated as exogenous. Theoretical analyses show that the demand for health insurance and health care are jointly determined, so causation is not unidirectional. This fact can potentially make it difficult to identify the role of health insurance. Treating health insurance as exogenous biases the estimate of price responsiveness. However, in an experimental setup the participating households could be assigned an insurance policy, making it an exogenous variable. The role of insurance is then identifiable. Once the key variable of interest is exogenized, the direction of causation becomes clear and the impact of

the treatment can be studied unambiguously. Furthermore, if the experiment is free from some of the problems that we mention in the following, this greatly simplifies statistical analysis relative to what is often necessary in survey data.

3.3.3. Limitations of Social Experiments

The application of a nonhuman methodology, initially that is, one developed for and applied to nonhuman subjects, to human subjects has generated a lively debate in the literature. See especially Heckman and Smith (1995), who argue that many social experiments may suffer from limitations that apply to observational studies. These issues concern general points such as the merits of experimental versus observational methodology, as well as specific issues concerning the biases and problems inherent in the use of human subjects. Several of the issues are covered in more detail in later chapters but a brief overview follows.

Social experiments are very costly to run. Sometimes, perhaps often, they do not correspond to “clean” randomized trials. Hence the results from such experiments are not always unambiguous and easily interpretable, or free from biases. If the treatment variable has many alternative settings of interest, or if extrapolation is an important objective, then a very large sample must be collected to ensure sufficient data variation and to precisely gauge the effect of treatment variation. In that case the cost of the experiment will also increase. If the cost factor prevents a large enough experiment, its utility relative to observational studies may be questionable; see the papers by Rosen and Stafford in Hausman and Wise (1985).

Unfortunately the design of some social experiments is flawed. Hausman and Wise (1985) argue that the data from the New Jersey negative income tax experiment was subject to endogenous stratification, which they describe as follows:

... [T]he reason for an experiment is, by randomization, to eliminate correlation between the treatment variable and other determinants of the response variable that is under study. In each of the income-maintenance experiments, however, the experimental sample was selected in part on the basis of the dependent variable, and the assignment to treatment versus control group was based in part on the dependent variable as well. In general, the group eligible for selection – based on family status, race, age of family head, etc. – was stratified on the basis of income (and other variables) and persons were selected from within the strata. (Hausman and Wise, 1985, pp. 190–191)

The authors conclude that, in the presence of endogenous stratification, unbiased estimation of treatment effects is not straightforward. Unfortunately, a fully randomized trial in which treatment assignment within a randomly selected experimental group from the population is independent of income would be much more costly and may not be feasible.

There are several other issues that detract from the ideal simplicity of a randomized experiment. First, if experimental sites are selected randomly, cooperation of administrators and potential participants at that site would be required. If this is not forthcoming, then alternative treatment sites where such cooperation is obtainable

3.3. DATA FROM SOCIAL EXPERIMENTS

will be substituted, thereby compromising the random assignment principle; see Hotz (1992).

A second problem is that of sample selection, which is relevant because participation is voluntary. For ethical reasons there are many experiments that simply cannot be done (e.g., random assignment of students to years of education). Unlike medical experiments that can achieve the gold standard of a double-blind protocol, in social experiments experimenters and subjects know whether they are in treatment or control groups. Furthermore, those in control groups may obtain treatment, (e.g., training) from alternative sources. If the decision to participate is uncorrelated with either x or ε , the analysis of the experimental data is simplified.

A third problem is sample attrition caused by subjects dropping out of the experiment after it has started. Even if the initial sample was random the effect of nonrandom attrition may well lead to a problem similar to the attrition bias in panels. Finally, there is the problem of **Hawthorne effect**. The term originates in social psychology research conducted jointly by the Harvard Graduate School of Business Administration and the management of the Western Electric Company at the latter's Hawthorne works in Chicago from 1926 to 1932. Human subjects, unlike inanimate objects, may change or adapt their behavior while participating in the experiment. In this case the variation in the response observed under experimental conditions cannot be attributed solely to treatment.

Heckman and Smith (1995) mention several other difficulties in implementing a randomized treatment. Because the administration of a social experiment involves a bureaucracy, there is a potential for biases. **Randomization bias** occurs if the assignment introduces a systematic difference between the experimental participant and the participant during its normal operation. Heckman and Smith document the possibilities of such bias in actual experiments. Another type of bias, called **substitution bias**, is introduced when the controls may be receiving some form of treatment that substitutes for the experimental treatment. Finally, analysis of social experiments is inevitably of a partial equilibrium nature. One cannot reliably extrapolate the treatment effects to the entire population because the *ceteris paribus* assumption will not hold when the entire population is involved.

Specifically, the key issue is whether one can extrapolate the results from the experiment to the population at large. If the experiment is conducted as a pilot program on a small scale, but the intention is to predict the impact of policies that are more broadly applied, then the obvious limitation is that the pilot program cannot incorporate the broader impact of the treatment. A broadly applied treatment may change the economic environment sufficiently to invalidate the predictions from a partial equilibrium setup. So the treatment will not be like the actual policy that it mimics.

In summary, social experiments, in principle, could yield data that are easier to analyze and to understand in terms of cause and effect than observational data. Whether this promise is realized depends on the experimental design. A poor experimental design generates its own statistical complications, which affect the precision of the conclusions. Social experiments differ fundamentally from those in biology and agriculture because human subjects and treatment administrators tend to be both active and forward-looking individuals with personal preferences, rather than

Table 3.2. Features of Some Selected Natural Experiments

Experiment	Treatments Studied	Reference
Outcomes for identical twins with different schooling levels	Differences in returns to schooling through correlation between schooling and wages	Ashenfelter and Krueger (1994)
Transition to National Health Insurance in Canada as Saskatchewan moves to NHI and other states follow several years later	Labor market effects of NHI based on comparison of provinces with and without NHI	Gruber and Hanratty (1995)
New Jersey increases minimum wage while neighboring Pennsylvania does not	Minimum wage effects on employment	Card and Krueger (1994)

passive administrators of a standard protocol or willing recipients of randomly assigned treatment.

3.4. Data from Natural Experiments

Sometimes, however, a researcher may have available data from a “**natural experiment**.” A natural experiment occurs when a subset of the population is subjected to an exogenous variation in a variable, perhaps as a result of a policy shift, that would ordinarily be subject to endogenous variation. Ideally, the source of the variation is well understood.

In microeconomics there are broadly two ways in which the idea of a natural experiment is exploited. For concreteness consider the simple regression model

$$y = \beta_1 + \beta_2 x + u, \quad (3.4)$$

where x is an endogenous treatment variable correlated with u .

Suppose that there is an exogenous intervention that changes x . Examples of such external intervention are administrative rules, unanticipated legislation, natural events such as twin births, weather-related shocks, and geographical variation; see Table 3.2 for examples. Exogenous intervention creates an opportunity for evaluating its impact by comparing the behavior of the impacted group both pre- and postintervention, or with that of a nonimpacted group postintervention. That is, “natural” comparison groups are generated by the event that facilitates estimation of the β_2 . Estimation is simplified because x can be treated as exogenous.

The second way in which a natural experiment can assist inference is by generating natural instrumental variables. Suppose z is a variable that is correlated with x , or perhaps causally related to x , and uncorrelated with u . Then an **instrumental variable** estimator of β_2 , expressed in terms of sample covariances, is

$$\hat{\beta}_2 = \frac{\text{Cov}[z, y]}{\text{Cov}[z, x]} \quad (3.5)$$

(see Section 4.8.5). In an observational data setup an instrumental variable with the right properties may be difficult to find, but it could arise naturally in a favorable natural experiment. Then estimation would be simplified. We consider the first case in the next section; the topic of naturally generated instruments will be covered in Chapter 25.

3.4.1. Natural Exogenous Interventions

Such data are less expensive to collect and they also allow the researcher to evaluate the role of some specific factor in isolation, as in a controlled experiment, because “nature” holds constant variations attributed to other factors that are not of direct interest. Such natural experiments are attractive because they generate treatment and control groups inexpensively and in a real-world setting. Whether a natural experiment can support convincing inference depends, in part, on whether the supposed natural intervention is genuinely exogenous, whether its impact is sufficiently large to be measurable, and whether there are good treatment and control groups. Just because a change is legislated, for example, does not mean that it is an exogenous intervention. However, in appropriate cases, opportunistic exploitation of such data sets can yield valuable empirical insights.

Investigations based on natural experiments have several potential limitations whose importance in any given study can only be assessed through a careful consideration of the relevant theory, facts, and institutional setting. Following Campbell (1969) and Meyer (1995), these are grouped into limitations that affect a study’s internal validity (i.e., the inferences about policy impact drawn from the study) and those that affect a study’s external validity (i.e., the generalization of the conclusions to other members of the population).

Consider an investigation of a policy change in which conclusions are drawn from a comparison of pre- and postintervention data, using the regression method briefly described in the following and in greater detail in Chapter 25. In any study there will be omitted variables that may have also changed in the time interval between policy change and its impact. The characteristics of sampled individuals such as age, health status, and their actual or anticipated economic environment may also change. These omitted factors will directly affect the measured impact of the policy change. Whether the results can be generalized to other members of the population will depend on the absence of bias due to nonrandom sampling, existence of significant interaction effects between the policy change and its setting, and an absence of the role of historical factors that would cause the impact to vary from one situation to another. Of course, these considerations are not unique to data from natural experiments; rather, the point is that the latter are not necessarily free from these problems.

3.4.2. Differences in Differences

One simple regression method is based on a comparison of outcomes in one group before and after a policy intervention. For example, consider

$$y_{it} = \alpha + \beta D_t + \varepsilon_{it}, \quad i = 1, \dots, N, \quad t = 0, 1,$$

where $D_t = 1$ in period 1 (postintervention), $D_t = 0$ in period 0 (preintervention), and y_{it} measures the outcome. The regression estimated from the pooled data will yield an estimate of policy impact parameter β . This is easily shown to be equal to the average difference in the pre- and postintervention outcome,

$$\begin{aligned}\widehat{\beta} &= N^{-1} \sum_i (y_{i1} - y_{i0}) \\ &= \bar{y}_1 - \bar{y}_0.\end{aligned}$$

The one-group before and after design makes the strong assumption that the group remains comparable over time. This is required for identifiability of β . If, for example, we allowed α to vary between the two periods, β would no longer be identified. Changes in α are confounded with the policy impact.

One way to improve on the previous design is to include an additional untreated comparison group, that is, one not impacted by policy, and for which the data are available in both periods. Using Meyer's (1995) notation, the relevant regression now is

$$y_{it}^j = \alpha + \alpha_1 D_t + \alpha^1 D^j + \beta D_t^j + \varepsilon_{it}^j, \quad i = 1, \dots, N, \quad t = 0, 1,$$

where j is the group superscript, $D^j = 1$ if j equals 1 and $D^j = 0$ otherwise, $D_t^j = 1$ if both j and t equal 1 and $D_t^j = 0$ otherwise, and ε is a zero-mean constant-variance error term. The equation does not include covariates but they can be added, and those that do not vary are already subsumed under α . This relation implies that, for the treated group, we have preintervention

$$y_{i0}^1 = \alpha + \alpha^1 D^1 + \varepsilon_{i0}^1$$

and postintervention

$$y_{i1}^1 = \alpha + \alpha_1 + \alpha^1 D^1 + \beta + \varepsilon_{i1}^1.$$

The impact is therefore

$$y_{i1}^1 - y_{i0}^1 = \alpha_1 + \beta + \varepsilon_{i1}^1 - \varepsilon_{i0}^1. \quad (3.6)$$

The corresponding equations for the untreated group are

$$y_{i0}^0 = \alpha + \varepsilon_{i0}^0,$$

$$y_{i1}^0 = \alpha + \alpha_1 + \varepsilon_{i1}^0,$$

and hence the difference is

$$y_{i1}^0 - y_{i0}^0 = \alpha_1 + \varepsilon_{i1}^0 - \varepsilon_{i0}^0. \quad (3.7)$$

Both the first-difference equations include the period-1 specific effect α_1 , which can be eliminated by taking the difference between Equations (3.6) and (3.7):

$$(y_{i1}^1 - y_{i0}^1) - (y_{i1}^0 - y_{i0}^0) = \beta + (\varepsilon_{i1}^1 - \varepsilon_{i0}^1) - (\varepsilon_{i1}^0 - \varepsilon_{i0}^0). \quad (3.8)$$

Assuming that $E[(\varepsilon_{i1}^1 - \varepsilon_{i0}^1) - (\varepsilon_{i1}^0 - \varepsilon_{i0}^0)]$ equals zero, we can obtain an unbiased estimate of β by the sample average of $(y_{i1}^1 - y_{i0}^1) - (y_{i1}^0 - y_{i0}^0)$. This method uses

differences in differences. If time-varying covariates are present, they can be included in the relevant equations and their differences will appear in the regression equation (3.8).

For simplicity our analysis ignored the possibility that there remain observable differences in the distribution of characteristics between the treatment and control groups. If so, then such differences must be controlled for. The standard solution is to include such controlling variables in the regression.

An example of a study based on a natural experiment is that of Ashenfelter and Krueger (1994). They estimate the returns to schooling by contrasting the wage rates of identical twins with different schooling levels. In this case running a regular experiment in which individuals are exogenously assigned different levels of schooling is simply not feasible. Nonetheless, some experimental-type controls are needed. As the authors explain:

Our goal is to ensure that the correlation we observe between schooling and wage rates is not due to a correlation between schooling and a worker's ability or other characteristics. We do this by taking advantage of the fact that monozygotic twins are genetically identical and have similar family backgrounds.

Data on twins have served as a basis for a number of other econometric studies (Rosenzweig and Wolpin, 1980; Bronars and Grogger, 1994). Since the twinning probability in the population is not high, an important issue is generating a sufficiently large representative sample, allowing for some nonresponse. One source of such data is the census. Another source is the "twins festivals" that are held in the United States. Ashenfelter and Krueger (1994, p. 1158) report that their data were obtained from interviews conducted at the 16th Annual Twins Day Festival, Twinsburg, Ohio, August 1991, which is the largest gathering of twins, triplets, and quadruplets in the world.

The attraction of using the twins data is that the presence of common effects from both observable and unobservable factors can be eliminated by modeling the *differences* between the outcomes of the twins. For example, Ashenfelter and Krueger estimate a regression model of the difference in the log of wage rates between the first and the second twin. The first differencing operation eliminates the effects of age, gender, ethnicity, and so forth. The remaining explanatory variables are differences between schooling levels, which is the variable of main interest, and variables such as differences in years of tenure and marital status.

3.4.3. Identification through Natural Experiments

The natural experiments school has had a useful impact on econometric practice. By encouraging the opportunistic exploitation of quasi-experimental data, and by using modeling frameworks such as the POM of Chapter 2, econometric practice bridges the gap between observational and experimental data. The notions of parameter identification rooted in the SEM framework are broadened to include identification of measures that are interesting from a policy viewpoint. The main advantage of using data from a natural experiment is that a policy variable of interest might be validly treated as exogenous. However, in using data from natural experiments, as in the case of social

experiments, the choice of control groups plays a critical role in determining the reliability of the conclusions. Several potential problems that affect a social experiment, such as selectivity and attrition bias, will also remain potential problems in the case of natural experiments. Only a subset of interesting policy problems may lend themselves to analysis within the natural experiment framework. The experiment may apply only to a small part of the population, and the conditions under which it occurs may not replicate themselves easily. An example given in Section 22.6 illustrates this point in the context of difference in differences.

3.5. Practical Considerations

Although there has been an explosion in the number and type of microdata sets that are available, certain well-established databases have supported numerous studies. We provide a very partial list of some of very well known U.S. micro databases. For further details, see the respective Web sites for these data sets or the data clearinghouses mentioned in the following. Many of these allow you to download the data directly.

3.5.1. Some Sources of Microdata

Panel Study in Income Dynamics (PSID): Based at the Survey Research Center at the University of Michigan, PSID is a national survey that has been running since 1968. Today it covers over 40,000 individuals and collects economic and demographic data. These data have been used to support a wide variety of microeconomic analyses. Brown, Duncan and Stafford (1996) summarize recent developments in PSID data.

Current Population Survey (CPS): This is a monthly national survey of about 50,000 households that provides information on labor force characteristics. The survey has been conducted for more than 50 years. Major revisions in the sample have followed each of the decennial censuses. For additional details about this survey see Section 24.2. It is the basis of many federal government statistics on earnings and unemployment. It is also an important source of microdata that have supported numerous studies especially of labor markets. The survey was redesigned in 1994 (Polivka, 1996).

National Longitudinal Survey (NLS): The NLS has four original cohorts: NLS Older Men, NLS Young Men, NLS Mature Women, and NLS Young Women. Each of the original cohorts is a national yearly survey of over 5,000 individuals who have been repeatedly interviewed since the mid-1960s. Surveys collect information on each respondent's work experiences, education, training, family income, household composition, marital status, and health. Supplementary data on age, sex, etc. are available.

National Longitudinal Surveys of Youth (NLSY): The NLSY is a national annual survey of 12,686 young men and young women who were 14 to 22 years of age when they were first surveyed in 1979. It contains three subsamples. The data

provide a unique opportunity to study the life-course experiences of a large sample of young adults who are representative of American men and women born in the late 1950s and early 1960s. A second NLSY began in 1997.

Survey of Income and Program Participation (SIPP): SIPP is a longitudinal survey of around 8,000 housing units per month. It covers income sources, participation in entitlement programs, correlation between these items, and individual attachments to the job market over time. It is a multipanel survey with a new panel being introduced at the beginning of each calendar year. The first panel of SIPP was initiated in October 1983. Compared with CPS, SIPP has fewer employed and more unemployed persons.

Health and Retirement Study (HRS): The HRS is a longitudinal national study. The baseline consists of interviews with members of 7,600 households in 1992 (respondents aged from 51 to 61) with follow-ups every two years for 12 years. The data contain a wealth of economic, demographic, and health information.

World Bank's Living Standards Measurement Study (LSMS): The World Bank's LSMS household surveys collect data "on many dimensions of household well-being that can be used to assess household welfare, understand household behavior, and evaluate the effects of various government policies on the living conditions of the population" in many developing countries. Many examples of the use of these data can be found in Deaton (1997) and in the economic development literature. Grosh and Glewwe (1998) outline the nature of the data and provide references to research studies that have used them.

Data clearinghouses: The Interuniversity Consortium for Political and Social Research (ICPSR) provides access to many data sets, including the PSID, CPS, NLS, SIPP, National Medical Expenditure Survey (NMES), and many others. The U.S. Bureau of Labor Statistics handles the CPS and NLS surveys. The U.S. Bureau of Census handles the SIPP. The U.S. National Center for Health Statistics provides access to many health data sets. A useful gateway to European data archives is the Council of European Social Science Data Archives (CESSDA), which provides links to several European national data archives.

Journal data archives: For some purposes, such as replication of published results for classroom work, you can get the data from journal archives. Two archives in particular have well-established procedures for data uploads and downloads using an Internet browser. The *Journal of Business and Economic Statistics* archives data used in most but not all articles published in that journal. The *Journal of Applied Econometrics* data archive is also organized along similar lines and contains data pertaining to most articles published since 1994.

3.5.2. Handling Microdata

Microeconomic data sets tend to be quite large. Samples of several hundreds or thousands are common and even those of tens of thousands are not unusual. The distributions of outcomes of interest are often nonnormal, in part because one is often dealing

with discrete data such as binary outcomes, or with data that have limited variation such as proportions or shares, or with truncated or censored continuous outcomes. Handling large nonnormal data sets poses some problems of summarizing and reporting the important features of data. Often it is useful to use one computing environment (program) for data extraction, reduction, and preparation and a different one for model estimation.

3.5.3. Data Preparation

The most basic feature of microeconometric analysis is that the process of arriving at the sample finally used in the econometric investigation is likely to be a long one. It is important to accurately document decisions and choices made by the investigator in the process of “cleaning up” the data. Let us consider some specific examples.

One of the most common features of sample survey data is **nonresponse** or partial response. The problems of nonresponse have already been discussed. Partial response usually means that some parts of survey questionnaires were not answered. If this means that some of the required information is not available, the observations in question are deleted. This is called **listwise deletion**. If this problem occurs in a significant number of cases, it should be properly analyzed and reported because it could lead to an unrepresentative sample and biases in estimation. The issue is analyzed in Chapter 27. For example, consider a question in a household survey to which high-income households do not respond, leading to a sample in which these households are underrepresented. Hence the end effect is no different from one in which there is a full response but the sample is not representative.

A second problem is *measurement error* in reported data. Microeconomic data are typically noisy. The extent, type, and seriousness of measurement error depends on the type of survey cross section or panel, the individual who responds to the survey, and the variable about which information is sought. For example, self-reported income data from panel surveys are strongly suspected to have serially correlated measurement error. In contrast, reported expenditure magnitudes are usually thought to have a smaller measurement error. Deaton (1997) surveys some of the sources of measurement error with special reference to the World Bank’s *Living Standards Measurement Survey*, although several of the issues raised have wider relevance. The biases from measurement error depend on what is done to the data in terms of transformations (e.g., first differencing) and the estimator used. Hence to make informative statements about the seriousness of biases from measurement error, one must analyze well-defined models. Later chapters will give examples of the impact of measurement error in specific contexts.

3.5.4. Checking Data

In large data sets it is easy to have erroneous data resulting from keyboard and coding errors. One should therefore apply some elementary checks that would reveal the existence of problems. One can check the data before analyzing it by examining some

descriptive statistics. The following techniques are useful. First, use summary statistics (min, max, mean, and median) to make sure that the data are in the proper interval and on the proper scale. For instance, categorical variables should be between zero and one, counts should be greater than or equal to zero. Sometimes missing data are coded as -999, or some other integer, so take care not to treat these entries as data. Second, one should know whether changes are fractional or on a percentage scale. Third, use box and whisker plots to identify problematic observations. For instance, using box and whisker plots one researcher found a country that had negative population growth (owing to a war) and another country that had recorded investment as more than GDP (because foreign aid had been excluded from the GDP calculation). Checking observations before proceeding with estimation may also suggest normalizing transformations and/or distributional assumptions with features appropriate for modeling a particular data set. Third, screening data may suggest appropriate data transforms. For example, box and whisker plots and histograms could suggest which variables might be better modeled via a log or power transform. Finally, it may be important to check the scales of measurement. For some purposes, such as the use of nonlinear estimators, it may be desirable to scale variables so that they have roughly similar scale. Summary statistics can be used to check that the means, variances, and covariances of the variables indicate proper scaling.

3.5.5. Presenting Descriptive Statistics

Because microdata sets are usually large, it is essential to provide the reader with an initial table of descriptive statistics, usually mean, standard deviation, minimum, and maximum for every variable. In some cases unexpectedly large or small values may reveal the presence of a gross recording error or erroneous inclusion of an incorrect data point. Two-way scatter diagrams are usually not helpful, but tabulation of categorical variables (contingency tables) can be. For discrete variables histograms can be useful and for continuous variables density plots can be informative.

3.6. Bibliographic Notes

- 3.2 Deaton (1997) provides an introduction to sample surveys especially for developing economies. Several specific references to complex surveys are provided in Chapter 24. Beckett et al. (1988) investigate the importance of the issue of representativeness of the PSID.
- 3.3 The collective volume edited by Hausman and Wise (1985) contains several papers on individual social experiments including the RHIE, NIT, and Time-of-Use pricing experiments. Several studies question the usefulness of the experimental data and there is extensive discussion of the flaws in experimental designs that preclude clear conclusions. Pros and cons of social experiments versus observational data are discussed in an excellent pair of papers by Burtless (1995) and Heckman and Smith (1995).
- 3.4 A special issue of the *Journal of Business and Economic Statistics* (1995) carries a number of articles that use the methodology of quasi- or natural experiments. The collection includes an article by Meyer who surveys the issues in and the methodology of econometric

studies that use data from natural experiments. He also provides a valuable set of guidelines on the credible use of natural variation in making inferences about the impact of economic policies, partly based on the work of Campbell (1969). Kim and Singal (1993) study the impact of changes in market concentration on price using the data generated by airline mergers. Rosenzweig and Wolpin (2000) review an extensive literature based on natural experiments such as identical twins. Isacsson (1999) uses the twins approach to study returns to schooling using Swedish data. Angrist and Lavy (1999) study the impact of class size on test scores using data from schools that are subject to “Maimonides’ Rule” (briefly reviewed in Section 25.6), which states that class size should not exceed 40. The rule generates an instrument.

PART TWO

Core Methods

Part 2 presents the core estimation methods – least squares, maximum likelihood and method of moments – and associated methods of inference for nonlinear regression models that are central in microeconomics. The material also includes modern topics such as quantile regression, sequential estimation, empirical likelihood, semiparametric and nonparametric regression, and statistical inference based on the bootstrap. In general the discussion is at a level intended to provide enough background and detail to enable the practitioner to read and comprehend articles in the leading econometrics journals and, where needed, subsequent chapters of this book. We presume prior familiarity with linear regression analysis.

The essential estimation theory is presented in three chapters. Chapter 4 begins with the linear regression model. It then covers at an introductory level quantile regression, which models distributional features other than the conditional mean. It provides a lengthy expository treatment of instrumental variables estimation, a major method of causal inference. Chapter 5 presents the most commonly-used estimation methods for nonlinear models, beginning with the topic of m-estimation, before specialization to maximum likelihood and nonlinear least squares regression. Chapter 6 provides a comprehensive treatment of generalized method of moments, which is a quite general estimation framework that is applicable for linear and nonlinear models in single-equation and multi-equation settings. The chapter emphasizes the special case of instrumental variables estimation.

We then turn to model testing. Chapter 7 covers both the classical and bootstrap approaches to hypothesis testing, while Chapter 8 presents relatively more modern methods of model selection and specification analysis. Because of their importance the computationally-intensive bootstrap methods are also the subject of a more detailed chapter, Chapter 11 in Part 3. A distinctive feature of this book is that, as much as possible, testing procedures are presented in a unified manner in just these three chapters. The procedures are then illustrated in specific applications throughout the book.

Chapter 9 is a stand-alone chapter that presents nonparametric and semiparametric estimation methods that place a flexible structure on the econometric model.

PART TWO: CORE METHODS

Chapter 10 presents the computational methods used to compute the nonlinear estimators presented in chapters 5 and 6. This material becomes especially relevant to the practitioner if an estimator is not automatically computed by an econometrics package, or if numerical difficulties are encountered in model estimation.

CHAPTER 4

Linear Models

4.1. Introduction

A great deal of empirical microeconomics research uses linear regression and its various extensions. Before moving to nonlinear models, the emphasis of this book, we provide a summary of some important results for the single-equation linear regression model with cross-section data. Several different estimators in the linear regression model are presented.

Ordinary least-squares (OLS) estimation is especially popular. For typical microeconomic cross-section data the model error terms are likely to be heteroskedastic. Then statistical inference should be robust to heteroskedastic errors and efficiency gains are possible by use of weighted rather than ordinary least squares.

The OLS estimator minimizes the sum of squared residuals. One alternative is to minimize the sum of the absolute value of residuals, leading to the least absolute deviations estimator. This estimator is also presented, along with extension to quantile regression.

Various model misspecifications can lead to inconsistency of least-squares estimators. In such cases inference about economically interesting parameters may require more advanced procedures and these are pursued at considerable length and depth elsewhere in the book. One commonly used procedure is instrumental variables regression. The current chapter provides an introductory treatment of this important method and additionally addresses the complication of weak instruments.

Section 4.2 provides a definition of regression and presents various loss functions that lead to different estimators for the regression function. An example is introduced in Section 4.3. Some leading estimation procedures, specifically ordinary least squares, weighted least squares, and quantile regression, are presented in, respectively, Sections 4.4, 4.5, and 4.6. Model misspecification is considered in Section 4.7. Instrumental variables regression is presented in Sections 4.8 and 4.9. Sections 4.3–4.5, 4.7, and 4.8 cover standard material in introductory courses, whereas Sections 4.2, 4.6, and 4.9 introduce more advanced material.

4.2. Regressions and Loss Functions

In modern microeconomics the term **regression** refers to a bewildering range of procedures for studying the relationship between an outcome variable y and a set of regressors \mathbf{x} . It is helpful, therefore, to state at the beginning the motivation and justification for some of the leading types of regressions.

For exposition it is convenient to think of the purpose of regression to be **conditional prediction** of y given \mathbf{x} . In practice, regression models are also used for other purposes, most notably causal inference. Even then a prediction function constitutes a useful data summary and is still of interest. In particular, see Section 4.2.3 for the distinction between linear prediction and causal inference based on a linear causal mean.

4.2.1. Loss Functions

Let \hat{y} denote the **predictor** defined as a function of \mathbf{x} . Let $e \equiv y - \hat{y}$ denote the **prediction error**, and let

$$L(e) = L(y - \hat{y}) \quad (4.1)$$

denote the **loss** associated with the error e . As in decision analysis we assume that the predictor forms the basis of some decision, and the prediction error leads to disutility on the part of the decision maker that is captured by $L(e)$, whose precise functional form is a choice of the decision maker. The loss function has the property that it is increasing in $|e|$.

Treating (y, \hat{y}) as random, the decision maker minimizes the expected value of the loss function, denoted $E[L(e)]$. If the predictor depends on \mathbf{x} , a K -dimensional vector, then **expected loss** is expressed as

$$E [L((y - \hat{y})|\mathbf{x})]. \quad (4.2)$$

The choice of the loss function should depend in a substantive way on the losses associated with prediction errors. In some situations, such as weather forecasting, there may be a sound basis for choosing one loss function over another.

In econometrics, there is often no clear guide and the convention is to specify quadratic loss. Then (4.1) specializes to $L(e) = e^2$ and by (4.2) the optimal predictor minimizes the expected loss $E[L(e|\mathbf{x})] = E[e^2|\mathbf{x}]$. It follows that in this case the minimum mean-squared prediction error criterion is used to compare predictors.

4.2.2. Optimal Prediction

The decision theory approach to choosing the **optimal predictor** is framed in terms of **minimizing expected loss**,

$$\min_{\hat{y}} E [L(y - \hat{y})|\mathbf{x}].$$

Thus the optimality property is relative to the loss function of the decision maker.

Table 4.1. Loss Functions and Corresponding Optimal Predictors

Type of Loss Function	Definition	Optimal Predictor
Squared error loss	$L(e) = e^2$	$E[y x]$
Absolute error loss	$L(e) = e $	$\text{med}[y x]$
Asymmetric absolute loss	$L(e) = \begin{cases} (1 - \alpha) e & \text{if } e < 0 \\ \alpha e & \text{if } e \geq 0 \end{cases}$	$q_\alpha[y x]$
Step loss	$L(e) = \begin{cases} 0 & \text{if } e < 0 \\ 1 & \text{if } e \geq 0 \end{cases}$	$\text{mod}[y x]$

Four leading examples of loss function, and the associated optimal predictor function, are given in Table 4.1. We provide a brief presentation for each in turn. A detailed analysis is given in Manski (1988a).

The most well known loss function is the **squared error loss** (or mean-square loss) function. Then the optimal predictor of y is the **conditional mean** function, $E[y|x]$. In the most general case no structure is placed on $E[y|x]$ and estimation is by nonparametric regression (see Chapter 9). More often a model for $E[y|x]$ is specified, with $E[y|x] = g(\mathbf{x}, \boldsymbol{\beta})$, where $g(\cdot)$ is a specified function and $\boldsymbol{\beta}$ is a finite-dimensional vector of parameters that needs to be estimated. The optimal prediction is $\hat{y} = g(\mathbf{x}, \hat{\boldsymbol{\beta}})$, where $\hat{\boldsymbol{\beta}}$ is chosen to minimize the in-sample loss

$$\sum_{i=1}^N L(e_i) = \sum_{i=1}^N e_i^2 = \sum_{i=1}^N (y_i - g(\mathbf{x}_i, \boldsymbol{\beta}))^2.$$

The loss function is the sum of squared residuals, so estimation is by nonlinear least squares (see Section 5.8). If the conditional mean function $g(\cdot)$ is restricted to be linear in \mathbf{x} and $\boldsymbol{\beta}$, so that $E[y|x] = \mathbf{x}'\boldsymbol{\beta}$, then the optimal predictor is $\hat{y} = \mathbf{x}'\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the ordinary least-squares estimator detailed in Section 4.4.

If the loss criterion is **absolute error loss**, then the optimal predictor is the **conditional median**, denoted $\text{med}[y|x]$. If the conditional median function is linear, so that $\text{med}[y|x] = \mathbf{x}'\boldsymbol{\beta}$, then the optimal predictor is $\hat{y} = \mathbf{x}'\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the least absolute deviations estimator that minimizes $\sum_i |y_i - \mathbf{x}'_i\boldsymbol{\beta}|$. This estimator is presented in Section 4.6.

Both the squared error and absolute error loss functions are symmetric, so the same penalty is imposed for prediction error of a given magnitude regardless of the direction of the prediction error. **Asymmetric absolute error loss** instead places a penalty of $(1 - \alpha)|e|$ on overprediction and a different penalty $\alpha|e|$ on underprediction. The asymmetry parameter α is specified. It lies in the interval $(0, 1)$ with symmetry when $\alpha = 0.5$ and increasing asymmetry as α approaches 0 or 1. The optimal predictor can be shown to be the **conditional quantile**, denoted $q_\alpha[y|x]$; a special case is the conditional median when $\alpha = 0.5$. Conditional quantiles are defined in Section 4.6, which presents quantile regression (Koenker and Bassett, 1978).

The last loss function given in Table 4.1 is **step loss**, which bases the loss simply on the sign of the prediction error regardless of the magnitude. The optimal predictor is the

conditional mode, denoted $\text{mod}[y|\mathbf{x}]$. This provides motivation for mode regression (Lee, 1989).

Maximum likelihood does not fall as easily into the prediction framework of this section. It can, however, be given an expected loss interpretation in terms of predicting the density and minimizing Kullback–Liebler information (see Section 5.7).

The results just stated imply that the econometrician interested in estimating a prediction function from the data (y, \mathbf{x}) should choose the prediction function according to the loss function. The use of the popular linear regression implies, at least implicitly, that the decision maker has a quadratic loss function and believes that the conditional mean function is linear. However, if one of the other three loss functions is specified, then the optimal predictor will be based on one of the three other types of regressions. In practice there can be no clear reason for preferring a particular loss function.

Regressions are often used as data summaries, rather than for prediction per se. Then it can be useful to consider a range of estimators, as alternative estimators may provide useful information about the sensitivity of estimates. Manski (1988a, 1991) has pointed out that the quadratic and absolute error loss functions are both convex. If the conditional distribution of $y|\mathbf{x}$ is symmetric then the conditional mean and median estimators are both consistent and can be expected to be quite close. Furthermore, if one avoids assumptions about the distribution of $y|\mathbf{x}$, then differences in alternative estimators provide a way of learning about the data distribution.

4.2.3. Linear Prediction

The optimal predictor under squared error loss is the conditional mean $E[y|\mathbf{x}]$. If this conditional mean is linear in \mathbf{x} , so that $E[y|\mathbf{x}] = \mathbf{x}'\beta$, the parameter β has a structural or causal interpretation and consistent estimation of β by OLS implies consistent estimation of $E[y|\mathbf{x}] = \mathbf{x}'\beta$. This permits meaningful policy analysis of effects of changes in regressors on the conditional mean.

If instead the conditional mean is nonlinear in \mathbf{x} , so that $E[y|\mathbf{x}] \neq \mathbf{x}'\beta$, the structural interpretation of OLS disappears. However, it is still possible to interpret β as the best linear predictor under squared error loss. Differentiation of the expected loss $E[(y - \mathbf{x}'\beta)^2]$ with respect to β yields first-order conditions $-2E[\mathbf{x}(y - \mathbf{x}'\beta)] = \mathbf{0}$, so the optimal linear predictor is $\hat{\beta} = (E[\mathbf{x}\mathbf{x}'])^{-1}E[\mathbf{x}y]$ with sample analogue the OLS estimator.

Usually we specialize to models with intercept. In a change of notation we define \mathbf{x} to denote regressors excluding the intercept, and we replace $\mathbf{x}'\beta$ by $\alpha + \mathbf{x}'\gamma$. The first-order conditions with respect to α and γ are that $-2E[u] = 0$ and $-2E[\mathbf{x}u] = \mathbf{0}$, where $u = y - (\alpha + \mathbf{x}'\gamma)$. These imply that $E[u] = 0$ and $\text{Cov}[\mathbf{x}, u] = \mathbf{0}$. Solving yields

$$\begin{aligned}\gamma &= (V[\mathbf{x}])^{-1} \text{Cov}[\mathbf{x}, y], \\ \alpha &= E[y] - E[\mathbf{x}']\gamma;\end{aligned}\tag{4.3}$$

see, for example, Goldberger (1991, p. 52).

From the derivation of (4.3) it should be clear that for data (y, \mathbf{x}) we can always write a linear regression model

$$y = \alpha + \mathbf{x}'\gamma + u,\tag{4.4}$$

where the parameters α and γ are defined in (4.3) and the error term u satisfies $E[u] = 0$ and $\text{Cov}[\mathbf{x}, u] = \mathbf{0}$.

A linear regression model can therefore always be given the nonstructural or reduced form interpretation as the **best linear prediction** (or linear projection) under squared error loss. However, for the conditional mean to be linear in \mathbf{x} , so that $E[y|\mathbf{x}] = \alpha + \mathbf{x}'\gamma$, requires the assumption that $E[u|\mathbf{x}] = 0$, in addition to $E[u] = 0$ and $\text{Cov}[\mathbf{x}, u] = \mathbf{0}$.

This distinction is of practical importance. For example, if $E[u|\mathbf{x}] = 0$, so that $E[y|\mathbf{x}] = \alpha + \mathbf{x}'\gamma$, then the probability limit of a least-squares (LS) estimator $\hat{\gamma}$ is γ regardless of whether the LS estimator is weighted or unweighted, or whether the sample is obtained by simple random sampling or by exogenous stratified sampling. If instead $E[y|\mathbf{x}] \neq \alpha + \mathbf{x}'\gamma$ then these different LS estimators may have different probability limits. This example is discussed further in Section 24.3.

A structural interpretation of OLS requires that the conditional mean of the error term, given regressors, equals zero.

4.3. Example: Returns to Schooling

A leading linear regression application from labor economics concerns measuring the impact of education on wages or earnings.

A typical returns to schooling model specifies

$$\ln w_i = \alpha s_i + \mathbf{x}'_{2i}\beta + u_i, \quad i = 1, \dots, N, \quad (4.5)$$

where w denotes hourly wage or annual earnings, s denotes years of completed schooling, and \mathbf{x}_2 denotes control variables such as work experience, gender, and family background. The subscript i denotes the i th person in the sample. Since the dependent variable is log wage, the model is a log-linear model and the coefficient α measures the proportionate change in earnings associated with a one-year increase in education.

Estimation of this model is most often by ordinary least squares. The transformation to $\ln w$ in practice ensures that errors are approximately homoskedastic, but it is still best to obtain heteroskedastic consistent standard errors as detailed in Section 4.4. Estimation can also be by quantile regression (see Section 4.6), if interest lies in distributional issues such as behavior in the lower quartile.

The regression (4.5) can be used immediately in a descriptive manner. For example, if $\hat{\alpha} = 0.10$ then a one-year increase in schooling is associated with 10% higher earnings, controlling for all the factors included in \mathbf{x}_2 . It is important to add the last qualifier as in this example the estimate $\hat{\alpha}$ usually becomes smaller as \mathbf{x}_2 is expanded to include additional controls likely to influence earnings.

Policy interest lies in determining the impact of an *exogenous change* in schooling on earnings. However, schooling is not randomly assigned; rather, it is an outcome that depends on choices made by the individual. Human capital theory treats schooling as investment by individuals in themselves, and α is interpreted as a measure of return to human capital. The regression (4.5) is then a regression of one endogenous variable, y , on another, s , and so does not measure the causal impact of an exogenous change

in s . The conditional mean function here is not causally meaningful because one is conditioning on a factor, schooling, that is *endogenous*. Indeed, unless we can argue that s is itself a function of variables at least one of which can vary independently of u , it is unclear just what it means to regard α as a causal parameter.

Such concern about endogenous regressors with observational data on individuals pervades microeconomic analysis. The standard assumptions of the linear regression model given in Section 4.4 are that regressors are exogenous. The consequences of endogenous regressors are considered in Section 4.7. One method to control for endogenous regressors, instrumental variables, is detailed in Section 4.8. A recent extensive review of ways to control for endogeneity in this wage–schooling example is given in Angrist and Krueger (1999). These methods are summarized in Section 2.8 and presented throughout this book.

4.4. Ordinary Least Squares

The simplest example of regression is the OLS estimator in the linear regression model.

After first defining the model and estimator, a quite detailed presentation of the asymptotic distribution of the OLS estimator is given. The exposition presumes previous exposure to a more introductory treatment. The model assumptions made here permit stochastic regressors and heteroskedastic errors and accommodate data that are obtained by exogenous stratified sampling.

The key result of how to obtain heteroskedastic-robust standard errors of the OLS estimator is given in Section 4.4.5.

4.4.1. Linear Regression Model

In a standard cross-section regression model with N observations on a scalar dependent variable and several regressors, the data are specified as (\mathbf{y}, \mathbf{X}) , where \mathbf{y} denotes observations on the dependent variable and \mathbf{X} denotes a matrix of explanatory variables.

The general regression model with additive errors is written in vector notation as

$$\mathbf{y} = E[\mathbf{y}|\mathbf{X}] + \mathbf{u}, \quad (4.6)$$

where $E[\mathbf{y}|\mathbf{X}]$ denotes the conditional expectation of the random variable \mathbf{y} given \mathbf{X} , and \mathbf{u} denotes a vector of unobserved random errors or disturbances. The right-hand side of this equation decomposes \mathbf{y} into two components, one that is deterministic given the regressors and one that is attributed to random variation or noise. We think of $E[\mathbf{y}|\mathbf{X}]$ as a conditional prediction function that yields the average value, or more formally the expected value, of \mathbf{y} given \mathbf{X} .

A **linear regression model** is obtained when $E[\mathbf{y}|\mathbf{X}]$ is specified to be a linear function of \mathbf{X} . Notation for this model has been presented in detail in Section 1.6. In vector notation the i th observation is

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + u_i, \quad (4.7)$$

where \mathbf{x}_i is a $K \times 1$ **regressor vector** and β is a $K \times 1$ **parameter vector**. At times it is simpler to drop the subscript i and write the model for typical observation as $y = \mathbf{x}'\beta + u$. In matrix notation the N observations are stacked by row to yield

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}, \quad (4.8)$$

where \mathbf{y} is an $N \times 1$ **vector of dependent variables**, \mathbf{X} is an $N \times K$ **regression matrix**, and \mathbf{u} is an $N \times 1$ **error vector**.

Equations (4.7) and (4.8) are equivalent expressions for the linear regression model and will be used interchangeably. The latter is more concise and is usually the most convenient representation.

In this setting y is referred to as the **dependent variable** or **endogenous variable** whose variation we wish to study in terms of variation in \mathbf{x} and u ; u is referred to as the **error term** or **disturbance term**; and \mathbf{x} is referred to as **regressors** or **predictors** or **covariates**. If Assumption 4 in Section 4.4.6 holds, then all components of \mathbf{x} are **exogenous variables** or **independent variables**.

4.4.2. OLS Estimator

The OLS estimator is defined to be the estimator that minimizes the sum of squared errors

$$\sum_{i=1}^N u_i^2 = \mathbf{u}'\mathbf{u} = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta). \quad (4.9)$$

Setting the derivative with respect to β equal to $\mathbf{0}$ and solving for β yields the OLS estimator,

$$\hat{\beta}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \quad (4.10)$$

see Exercise 4.5 for a more general result, where it is assumed that the matrix inverse of $\mathbf{X}'\mathbf{X}$ exists. If $\mathbf{X}'\mathbf{X}$ is of less than full rank, the inverse can be replaced by a generalized inverse. Then OLS estimation still yields the optimal linear predictor of y given \mathbf{x} if squared error loss is used, but many different linear combinations of \mathbf{x} will yield this optimal predictor.

4.4.3. Identification

The OLS estimator can always be computed, provided that $\mathbf{X}'\mathbf{X}$ is nonsingular. The more interesting issue is what $\hat{\beta}_{OLS}$ tells us about the data.

We focus on the ability of the OLS estimator to permit identification (see Section 2.5) of the conditional mean $E[y|\mathbf{X}]$. For the linear model the parameter β is identified if

1. $E[\mathbf{y}|\mathbf{X}] = \mathbf{X}\beta$ and
2. $\mathbf{X}\beta^{(1)} = \mathbf{X}\beta^{(2)}$ if and only if $\beta^{(1)} = \beta^{(2)}$.

The first condition that the conditional mean is correctly specified ensures that β is of intrinsic interest; the second assumption implies that $\mathbf{X}'\mathbf{X}$ is nonsingular, which is the same condition needed to compute the unique OLS estimate (4.10).

4.4.4. Distribution of the OLS Estimator

We focus on the asymptotic properties of the OLS estimator. Consistency is established and then the limit distribution is obtained by rescaling the OLS estimator. Statistical inference then requires consistent estimation of the variance matrix of the estimator. The analysis makes extensive use of asymptotic theory, which is summarized in Appendix A.

Consistency

The properties of an estimator depend on the process that actually generated the data, the **data generating process (dgp)**. We assume the dgp is $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$, so that the model (4.8) is correctly specified. In some places, notably Chapters 5 and 6 and Appendix A the subscript 0 is added to β , so the dgp is $\mathbf{y} = \mathbf{X}\beta_0 + \mathbf{u}$. See Section 5.2.3 for discussion.

Then

$$\begin{aligned}\hat{\beta}_{OLS} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{u}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u},\end{aligned}$$

and the OLS estimator can be expressed as

$$\hat{\beta}_{OLS} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}. \quad (4.11)$$

To prove consistency we rewrite (4.11) as

$$\hat{\beta}_{OLS} = \beta + (N^{-1}\mathbf{X}'\mathbf{X})^{-1}N^{-1}\mathbf{X}'\mathbf{u}. \quad (4.12)$$

The reason for renormalization in the right-hand side is that $N^{-1}\mathbf{X}'\mathbf{X} = N^{-1}\sum_i \mathbf{x}_i\mathbf{x}_i'$ is an average that converges in probability to a finite nonzero matrix if \mathbf{x}_i satisfies assumptions that permit a law of large numbers to be applied to $\mathbf{x}_i\mathbf{x}_i'$ (see Section 4.4.8 for detail). Then

$$\text{plim } \hat{\beta}_{OLS} = \beta + (\text{plim } N^{-1}\mathbf{X}'\mathbf{X})^{-1}(\text{plim } N^{-1}\mathbf{X}'\mathbf{u}),$$

using Slutsky's Theorem (Theorem A.3). The OLS estimator is **consistent** for β (i.e., $\text{plim } \hat{\beta}_{OLS} = \beta$) if

$$\text{plim } N^{-1}\mathbf{X}'\mathbf{u} = \mathbf{0}. \quad (4.13)$$

If a law of large numbers can be applied to the average $N^{-1}\mathbf{X}'\mathbf{u} = N^{-1}\sum_i \mathbf{x}_i u_i$ then a necessary condition for (4.13) to hold is that $E[\mathbf{x}_i u_i] = \mathbf{0}$.

Limit Distribution

Given consistency, the limit distribution of $\widehat{\beta}_{OLS}$ is degenerate with all the mass at β . To obtain the limit distribution we multiply $\widehat{\beta}_{OLS}$ by \sqrt{N} , as this rescaling leads to a random variable that under standard cross-section assumptions has nonzero yet finite variance asymptotically. Then (4.11) becomes

$$\sqrt{N}(\widehat{\beta}_{OLS} - \beta) = (N^{-1}\mathbf{X}'\mathbf{X})^{-1} N^{-1/2} \mathbf{X}' \mathbf{u}. \quad (4.14)$$

The proof of consistency assumed that $\text{plim } N^{-1}\mathbf{X}'\mathbf{X}$ exists and is finite and nonzero. We assume that a central limit theorem can be applied to $N^{-1/2}\mathbf{X}'\mathbf{u}$ to yield a multivariate normal limit distribution with finite, nonsingular covariance matrix. Applying the product rule for limit normal distributions (Theorem A.17) implies that the product in the right-hand side of (4.14) has a limit normal distribution. Details are provided in Section 4.4.8.

This leads to the following proposition, which permits regressors to be stochastic and does not restrict model errors to be homoskedastic and uncorrelated.

Proposition 4.1 (Distribution of OLS Estimator). *Make the following assumptions:*

- (i) *The dgp is model (4.8), that is, $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$.*
- (ii) *Data are independent over i with $E[\mathbf{u}|\mathbf{X}] = \mathbf{0}$ and $E[\mathbf{u}\mathbf{u}'|\mathbf{X}] = \Omega = \text{Diag}[\sigma_i^2]$.*
- (iii) *The matrix \mathbf{X} is of full rank so that $\mathbf{X}\beta^{(1)} = \mathbf{X}\beta^{(2)}$ iff $\beta^{(1)} = \beta^{(2)}$.*
- (iv) *The $K \times K$ matrix*

$$\mathbf{M}_{\mathbf{xx}} = \text{plim } N^{-1}\mathbf{X}'\mathbf{X} = \text{plim } \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i' = \lim \frac{1}{N} \sum_{i=1}^N E[\mathbf{x}_i \mathbf{x}_i'] \quad (4.15)$$

exists and is finite nonsingular.

- (v) *The $K \times 1$ vector $N^{-1/2}\mathbf{X}'\mathbf{u} = N^{-1/2} \sum_{i=1}^N \mathbf{x}_i \mathbf{u}_i \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{M}_{\mathbf{x}\Omega\mathbf{x}}]$, where*

$$\mathbf{M}_{\mathbf{x}\Omega\mathbf{x}} = \text{plim } N^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X} = \text{plim } \frac{1}{N} \sum_{i=1}^N u_i^2 \mathbf{x}_i \mathbf{x}_i' = \lim \frac{1}{N} \sum_{i=1}^N E[u_i^2 \mathbf{x}_i \mathbf{x}_i']. \quad (4.16)$$

Then the OLS estimator $\widehat{\beta}_{OLS}$ defined in (4.10) is consistent for β and

$$\sqrt{N}(\widehat{\beta}_{OLS} - \beta) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{M}_{\mathbf{xx}}^{-1} \mathbf{M}_{\mathbf{x}\Omega\mathbf{x}} \mathbf{M}_{\mathbf{xx}}^{-1}]. \quad (4.17)$$

Assumption (i) is used to obtain (4.11). Assumption (ii) ensures $E[\mathbf{y}|\mathbf{X}] = \mathbf{X}\beta$ and permits heteroskedastic errors with variance σ_i^2 , more general than the homoskedastic uncorrelated errors that restrict $\Omega = \sigma^2 \mathbf{I}$. Assumption (iii) rules out perfect collinearity among the regressors. Assumption (iv) leads to the rescaling of $\mathbf{X}'\mathbf{X}$ by N^{-1} in (4.12) and (4.14). Note that by a law of large numbers $\text{plim} = \lim E$ (see Appendix Section A.3).

The essential condition for consistency is (4.13). Rather than directly assume this we have used the stronger assumption (v) which is needed to obtain result (4.17).

Given that $N^{-1/2}\mathbf{X}'\mathbf{u}$ has a limit distribution with zero mean and finite variance, multiplication by $N^{-1/2}$ yields a random variable that converges in probability to zero and so (4.13) holds as desired. Assumption (v) is required, along with assumption (iv), to obtain the limit normal result (4.17), which by Theorem A.17 then follows immediately from (4.14). More primitive assumptions on u_i and \mathbf{x}_i that ensure (iv) and (v) are satisfied are given in Section 4.4.6, with formal proof in Section 4.4.8.

Asymptotic Distribution

Proposition 4.1 gives the **limit distribution** of $\sqrt{N}(\widehat{\beta}_{OLS} - \beta)$, a rescaling of $\widehat{\beta}_{OLS}$. Many practitioners prefer to see asymptotic results written directly in terms of the distribution of $\widehat{\beta}_{OLS}$, in which case the distribution is called an *asymptotic distribution*. This asymptotic distribution is interpreted as being applicable *in large samples*, meaning samples large enough for the limit distribution to be a good approximation but not so large that $\widehat{\beta}_{OLS} \xrightarrow{P} \beta$ as then its asymptotic distribution would be degenerate. The discussion mirrors that in Appendix A.6.4.

The asymptotic distribution is obtained from (4.17) by division by \sqrt{N} and addition of β . This yields the **asymptotic distribution**

$$\widehat{\beta}_{OLS} \xrightarrow{a} \mathcal{N}[\beta, N^{-1}\mathbf{M}_{xx}^{-1}\mathbf{M}_{x\Omega x}\mathbf{M}_{xx}^{-1}], \quad (4.18)$$

where the symbol \xrightarrow{a} means is “asymptotically distributed as.” The variance matrix in (4.18) is called the **asymptotic variance matrix** of $\widehat{\beta}_{OLS}$ and is denoted $V[\widehat{\beta}_{OLS}]$. Even simpler notation drops the limits and expectations in the definitions of \mathbf{M}_{xx} and $\mathbf{M}_{x\Omega x}$ and the asymptotic distribution is denoted

$$\widehat{\beta}_{OLS} \xrightarrow{a} \mathcal{N}[\beta, (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}], \quad (4.19)$$

and $V[\widehat{\beta}_{OLS}]$ is defined to be the variance matrix in (4.19).

We use both (4.18) and (4.19) to represent the asymptotic distribution in later chapters. Their use is for convenience of presentation. Formal asymptotic results for statistical inference are based on the limit distribution rather than the asymptotic distribution.

For implementation, the matrices \mathbf{M}_{xx} and $\mathbf{M}_{x\Omega x}$ in (4.17) or (4.18) are replaced by consistent estimates $\widehat{\mathbf{M}}_{xx}$ and $\widehat{\mathbf{M}}_{x\Omega x}$. Then the **estimated asymptotic variance matrix** of $\widehat{\beta}_{OLS}$ is

$$\widehat{V}[\widehat{\beta}_{OLS}] = N^{-1}\widehat{\mathbf{M}}_{xx}^{-1}\widehat{\mathbf{M}}_{x\Omega x}\widehat{\mathbf{M}}_{xx}^{-1}. \quad (4.20)$$

This estimate is called a **sandwich estimate**, with $\widehat{\mathbf{M}}_{x\Omega x}$ sandwiched between $\widehat{\mathbf{M}}_{xx}^{-1}$ and $\widehat{\mathbf{M}}_{xx}^{-1}$.

4.4.5. Heteroskedasticity-Robust Standard Errors for OLS

The obvious choice for $\widehat{\mathbf{M}}_{xx}$ in (4.20) is $N^{-1}\mathbf{X}'\mathbf{X}$. Estimation of $\mathbf{M}_{x\Omega x}$ defined in (4.16) depends on assumptions made about the error term.

In microeconomics applications the model errors are often conditionally heteroskedastic, with $V[u_i|x_i] = E[u_i^2|x_i] = \sigma_i^2$ varying over i . White (1980a) proposed

using $\widehat{\mathbf{M}}_{\mathbf{x}\Omega\mathbf{x}} = N^{-1} \sum_i \widehat{u}_i^2 \mathbf{x}_i \mathbf{x}'_i$. This estimate requires additional assumptions given in Section 4.4.8.

Combining these estimates $\widehat{\mathbf{M}}_{\mathbf{xx}}$ and $\widehat{\mathbf{M}}_{\mathbf{x}\Omega\mathbf{x}}$ and simplifying yields the estimated asymptotic variance matrix estimate

$$\begin{aligned}\widehat{\mathbf{V}}[\widehat{\boldsymbol{\beta}}_{\text{OLS}}] &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \widehat{\Omega} \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \\ &= \left(\sum_{i=1}^N \mathbf{x}_i \mathbf{x}'_i \right)^{-1} \sum_{i=1}^N \widehat{u}_i^2 \mathbf{x}_i \mathbf{x}'_i \left(\sum_{i=1}^N \mathbf{x}_i \mathbf{x}'_i \right)^{-1},\end{aligned}\quad (4.21)$$

where $\widehat{\Omega} = \text{Diag}[\widehat{u}_i^2]$ and $\widehat{u}_i = y_i - \mathbf{x}'_i \widehat{\boldsymbol{\beta}}$ is the OLS residual. This estimate, due to White (1980a), is called the **heteroskedastic-consistent** estimate of the asymptotic variance matrix of the OLS estimator, and it leads to standard errors that are called **heteroskedasticity-robust standard errors**, or even more simply **robust standard errors**. It provides a consistent estimate of $\mathbf{V}[\widehat{\boldsymbol{\beta}}_{\text{OLS}}]$ even though \widehat{u}_i^2 is not consistent for σ_i^2 .

In introductory courses the errors are restricted to be **homoskedastic**. Then $\Omega = \sigma^2 \mathbf{I}$ so that $\mathbf{X}'\Omega\mathbf{X} = \sigma^2 \mathbf{X}'\mathbf{X}$ and hence $\mathbf{M}_{\mathbf{x}\Omega\mathbf{x}} = \sigma^2 \mathbf{M}_{\mathbf{xx}}$. The limit distribution variance matrix in (4.17) simplifies to $\sigma^2 \mathbf{M}_{\mathbf{xx}}^{-1}$, and many computer packages instead use what is sometimes called the **default** OLS variance estimate

$$\widetilde{\mathbf{V}}[\widehat{\boldsymbol{\beta}}_{\text{OLS}}] = s^2 (\mathbf{X}'\mathbf{X})^{-1}, \quad (4.22)$$

where $s^2 = (N - K)^{-1} \sum_i \widehat{u}_i^2$.

Inference based on (4.22) rather than (4.21) is invalid, unless errors are homoskedastic and uncorrelated. In general the erroneous use of (4.22) when errors are heteroskedastic, as is often the case for cross-section data, can lead to either inflation or deflation of the true standard errors.

In practice $\widehat{\mathbf{M}}_{\mathbf{x}\Omega\mathbf{x}}$ is calculated using division by $(N - K)$, rather than by N , to be consistent with the similar division in forming s^2 in the homoskedastic case. Then $\widetilde{\mathbf{V}}[\widehat{\boldsymbol{\beta}}_{\text{OLS}}]$ in (4.21) is multiplied by $N/(N - K)$. With heteroskedastic errors there is no theoretical basis for this adjustment for degrees of freedom, but some simulation studies provide support (see MacKinnon and White, 1985, and Long and Ervin, 2000).

Microeconometric analysis uses robust standard errors wherever possible. Here the errors are robust to heteroskedasticity. Guarding against other misspecifications may also be warranted. In particular, when data are clustered the standard errors should additionally be robust to clustering; see Sections 21.2.3 and 24.5.

4.4.6. Assumptions for Cross-Section Regression

Proposition 4.1 is a quite generic theorem that relies on assumptions about $N^{-1}\mathbf{X}'\mathbf{X}$ and $N^{-1/2}\mathbf{X}'\mathbf{u}$. In practice these assumptions are verified by application of laws of large numbers and central limit theorems to averages of $\mathbf{x}_i \mathbf{x}'_i$ and $\mathbf{x}_i u_i$. These in turn require assumptions about how the observations \mathbf{x}_i and errors u_i are generated, and consequently how y_i defined in (4.7) is generated. The assumptions are referred to collectively as assumptions regarding the **data-generating process** (dgp). A simple pedagogical example is given in Exercise 4.4.

Our objective at this stage is to make assumptions that are appropriate in many applied settings where cross-section data are used. The assumptions, are those in White (1980a), and include three important departures from those in introductory treatments. First, the regressors may be stochastic (Assumptions 1 and 3 that follow), so assumptions on the error are made conditional on regressors. Second, the conditional variance of the error may vary across observations (Assumption 5). Third, the errors are not restricted to be normally distributed.

Here are the assumptions:

1. The data (y_i, \mathbf{x}_i) are independent and not identically distributed (inid) over i .
2. The model is correctly specified so that

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + u_i.$$

3. The regressor vector \mathbf{x}_i is possibly stochastic with finite second moment, additionally $E[|x_{ij}x_{ik}|^{1+\delta}] \leq \infty$ for all $j, k = 1, \dots, K$ for some $\delta > 0$, and the matrix $\mathbf{M}_{\mathbf{xx}}$ defined in (4.15) exists and is a finite positive definite matrix of rank K . Also, \mathbf{X} has rank K in the sample being analyzed.
4. The errors have zero mean, conditional on regressors

$$E[u_i | \mathbf{x}_i] = 0.$$

5. The errors are heteroskedastic, conditional on regressors, with

$$\begin{aligned} \sigma_i^2 &= E[u_i^2 | \mathbf{x}_i], \\ \boldsymbol{\Omega} &= E[\mathbf{u}\mathbf{u}' | \mathbf{X}] = \text{Diag}[\sigma_i^2], \end{aligned} \tag{4.23}$$

where $\boldsymbol{\Omega}$ is an $N \times N$ positive definite matrix. Also, for some $\delta > 0$, $E[|u_i^2|^{1+\delta}] < \infty$.

6. The matrix $\mathbf{M}_{\mathbf{x}\Omega\mathbf{x}}$ defined in (4.16) exists and is a finite positive definite matrix of rank K , where $\mathbf{M}_{\mathbf{x}\Omega\mathbf{x}} = \text{plim } N^{-1} \sum_i u_i^2 \mathbf{x}_i \mathbf{x}'_i$ given independence over i . Also, for some $\delta > 0$, $E[|u_i^2 x_{ij}x_{ik}|^{1+\delta}] < \infty$ for all $j, k = 1, \dots, K$.

4.4.7. Remarks on Assumptions

For completeness we provide a detailed discussion of each assumption, before proving the key results in the following section.

Stratified Random Sampling

Assumption 1 is one that is often implicitly made for cross-section data. Here we make it explicit. It restricts (y_i, \mathbf{x}_i) to be independent over i , but permits the distribution to differ over i . Many microeconomics data sets come from **stratified random sampling** (see Section 3.2). Then the population is partitioned into strata and random draws are made within strata, but some strata are oversampled with the consequence that the sampled (y_i, \mathbf{x}_i) are inid rather than iid. If instead the data come from **simple random sampling** then (y_i, \mathbf{x}_i) are iid, a stronger assumption that is a special case of inid. Many introductory treatments assumed that regressors are **fixed in repeated samples**.

Then (y_i, \mathbf{x}_i) are iid since only y_i is random with a value that depends on the value of \mathbf{x}_i . The fixed regressors assumption is rarely appropriate for microeconomics data, which are usually observational data. It is used instead for experimental data, where \mathbf{x} is the treatment level.

These different assumptions on the distribution of (y_i, \mathbf{x}_i) affect the particular laws of large numbers and central limit theorems used to obtain the asymptotic properties of the OLS estimator. Note that even if (y_i, \mathbf{x}_i) are iid, y_i given \mathbf{x}_i is not iid since, for example, $E[y_i | \mathbf{x}_i] = \mathbf{x}'_i \boldsymbol{\beta}$ varies with \mathbf{x}_i .

Assumption 1 rules out most time-series data since they are dependent over observations. It will also be violated if the sampling scheme involves clustering of observations. The OLS estimator can still be consistent in these cases, provided Assumptions 2–4 hold, but usually it has a variance matrix different from that presented in this chapter.

Correctly Specified Model

Assumption 2 seems very obvious as it is an essential ingredient in the derivation of the OLS estimator. It still needs to be made explicitly, however, since $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is a function of \mathbf{y} and so its properties depend on \mathbf{y} .

If Assumption 2 holds then it is being assumed that the regression model is *linear* in \mathbf{x} , rather than nonlinear, that there are *no omitted variables* in the regression, and that there is *no measurement error* in the regressors, as the regressors \mathbf{x} used to calculate $\widehat{\boldsymbol{\beta}}$ are the same regressors \mathbf{x} that are in the dgp. Also, the parameters $\boldsymbol{\beta}$ are the same across individuals, ruling out random parameter models.

If Assumption 2 fails then OLS can only be interpreted as an optimal linear predictor; see Section 4.2.3.

Stochastic Regressors

Assumption 3 permits regressors to be **stochastic regressors**, as is usually the case when survey data rather than experimental data are used. It is assumed that in the limit the sample second-moment matrix is constant and nonsingular.

If the regressors are iid, as is assumed under simple random sampling, then $\mathbf{M}_{\mathbf{xx}} = E[\mathbf{xx}']$ and Assumption 3 can be reduced to an assumption that the second moment exists. If the regressors are stochastic but not iid, as is the case for stratified random sampling, then we need the stronger Assumption 3, which permits application of the Markov LLN to obtain $\text{plim } N^{-1}\mathbf{X}'\mathbf{X}$. If the regressors are fixed in repeated samples, the common less-satisfactory assumption made in introductory courses, then $\mathbf{M}_{\mathbf{xx}} = \lim N^{-1}\mathbf{X}'\mathbf{X}$ and Assumption 3 becomes assumption that this limit exists.

Weakly Exogenous Regressors

Assumption 4 of **zero conditional mean errors** is crucial because when combined with Assumption 2 it implies that $E[\mathbf{y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta}$, so that the conditional mean is indeed $\mathbf{X}\boldsymbol{\beta}$.

The assumption that $E[u|\mathbf{x}] = 0$ implies that $\text{Cov}[\mathbf{x}, u] = \mathbf{0}$, so that the error is uncorrelated with regressors. This follows as $\text{Cov}[\mathbf{x}, u] = E[\mathbf{x}u] - E[\mathbf{x}]E[u]$ and $E[u|\mathbf{x}] = 0$ implies $E[\mathbf{x}u] = 0$ and $E[u] = 0$ by the law of iterated expectations. The weaker assumption that $\text{Cov}[\mathbf{x}, u] = \mathbf{0}$ can be sufficient for consistency of OLS, whereas the stronger assumption that $E[u|\mathbf{x}] = 0$ is needed for unbiasedness of OLS.

The economic meaning of Assumption 4 is that the error term represents all the excluded factors that are assumed to be uncorrelated with \mathbf{X} and these have, on average, zero impact on y . This is a key assumption that was referred to in Section 2.3 as the *weak exogeneity assumption*. Essentially this means that the knowledge of the data-generating process for \mathbf{X} variables does not contribute useful information for estimating β . When the assumption fails, one or more of the K regressor variables is said to be *jointly dependent* with y , or simply *endogenous*. A general term for correlation of regressors with errors is **endogeneity** or **endogenous regressors**, where the term “endogenous” means caused by factors inside the system. As we will show in Section 4.7, the violation of weak exogeneity may lead to inconsistent estimation. There are many ways in which weak exogeneity can be violated, but one of the most common involves a variable in \mathbf{x} that is a choice or a decision variable that is related to y in a larger model. Ignoring these other relationships, and treating \mathbf{x}_i as if it were randomly assigned to observation i , and hence uncorrelated with u_i , will have non-trivial consequences. **Endogenous sampling** is ruled out by Assumption 4. Instead, if data are collected by stratified random sampling it must be **exogenous stratified sampling**.

Conditionally Heteroskedastic Errors

Independent regression errors uncorrelated with regressors are assumed, a consequence of Assumptions 1, 2, and 4. Introductory courses usually further restrict attention to errors that are homoskedastic with homogeneous or constant variances, in which case $\sigma_i^2 = \sigma^2$ for all i . Then the errors are iid $(0, \sigma^2)$ and are called **spherical errors** since $\Omega = \sigma^2 \mathbf{I}$.

Assumption 5 is instead one of **conditionally heteroskedastic regression errors**, where *heteroskedastic* means heterogeneous variances or different variances. The assumption is stated in terms of the second moment $E[u^2|\mathbf{x}]$, but this equals the variance $V[u|\mathbf{x}]$ since $E[u|\mathbf{x}] = 0$ by Assumption 4. This more general assumption of heteroskedastic errors is made because empirically this is often the case for cross-section regression. Furthermore, relaxing the homoskedasticity assumption is not costly as it is possible to obtain valid standard errors for the OLS estimator even if the functional form for the heteroskedasticity is unknown.

The term *conditionally heteroskedastic* is used for the following reason. Even if (y_i, \mathbf{x}_i) are iid, as is the case for simple random sampling, once we condition on \mathbf{x}_i the conditional mean and conditional variance can vary with \mathbf{x}_i . Similarly, the errors $u_i = y_i - \mathbf{x}'_i \beta$ are iid under simple random sampling, and they are therefore unconditionally homoskedastic. Once we condition on \mathbf{x}_i , and consider the distribution of u_i *conditional* on \mathbf{x}_i , the variance of this conditional distribution is permitted to vary with \mathbf{x}_i .

Limit Variance Matrix of $N^{-1/2}\mathbf{X}'\mathbf{u}$

Assumption 6 is needed to obtain the limit variance matrix of $N^{-1/2}\mathbf{X}'\mathbf{u}$. If regressors are independent of the errors, a stronger assumption than that made in Assumption 4, then Assumption 5 that $E[|u_i^2|^{1+\delta}] < \infty$ and Assumption 3 that $E[|x_{ij}x_{ik}|^{1+\delta}] < \infty$ imply the Assumption 6 condition that $E[|u_i^2x_{ij}x_{ik}|^{1+\delta}] < \infty$.

We have deliberately not made a seventh assumption, that the error \mathbf{u} is normally distributed conditional on \mathbf{X} . An assumption such as normality is needed to obtain the exact small-sample distribution of the OLS estimator. However, we focus on asymptotic methods throughout this book, because exact small-sample distributional results are rarely available for the estimators used in microeconomics, and then the normality assumption is no longer needed.

4.4.8. Derivations for the OLS Estimator

Here we present both small-sample and limit distributions of the OLS estimator and justify White's estimator of the variance matrix of the OLS estimator under Assumptions 1–6.

Small-Sample Distribution

The parameter β is identified under Assumptions 1–4 since then $E[\mathbf{y}|\mathbf{X}] = \mathbf{X}\beta$ and \mathbf{X} has rank K .

In small samples the OLS estimator is unbiased under Assumptions 1–4 and its variance matrix is easily obtained given Assumption 5. These results are obtained by using the law of iterated expectations to first take expectation with respect to \mathbf{u} conditional on \mathbf{X} and then take the unconditional expectation. Then from (4.11)

$$\begin{aligned} E[\widehat{\beta}_{OLS}] &= \beta + E_{\mathbf{X},\mathbf{u}}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}] \\ &= \beta + E_{\mathbf{X}}[E_{\mathbf{u}|\mathbf{X}}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}]] \\ &= \beta + E_{\mathbf{X}}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E_{\mathbf{u}|\mathbf{X}}[\mathbf{u}|\mathbf{X}]] \\ &= \beta, \end{aligned} \tag{4.24}$$

using the law of iterated expectations (Theorem A.23) and given Assumptions 1 and 4, which together imply that $E[\mathbf{u}|\mathbf{X}] = \mathbf{0}$. Similarly, (4.11) yields

$$V[\widehat{\beta}_{OLS}] = E_{\mathbf{X}}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}], \tag{4.25}$$

given Assumption 5, where $E[\mathbf{u}\mathbf{u}'|\mathbf{X}] = \Omega$ and we use Theorem A.23, which tells us that in general

$$V_{\mathbf{X},\mathbf{u}}[\mathbf{g}(\mathbf{X}, \mathbf{u})] = E_{\mathbf{X}}[V_{\mathbf{u}|\mathbf{X}}[\mathbf{g}(\mathbf{X}, \mathbf{u})]] + V_{\mathbf{X}}[E_{\mathbf{u}|\mathbf{X}}[\mathbf{g}(\mathbf{X}, \mathbf{u})]].$$

This simplifies here as the second term is zero since $E_{\mathbf{u}|\mathbf{X}}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}] = \mathbf{0}$.

The OLS estimator is therefore **unbiased** if $E[\mathbf{u}|\mathbf{X}] = \mathbf{0}$. This valuable property generally does not extend to nonlinear estimators. Most nonlinear estimators, such as nonlinear least squares, are biased and even linear estimators such as instrumental

variables estimators can be biased. The OLS estimator is **inefficient**, as its variance is not the smallest possible variance matrix among linear unbiased estimators, unless $\Omega = \sigma^2 \mathbf{I}$. The inefficiency of OLS provides motivation for more efficient estimators such as generalized least squares, though the efficiency loss of OLS is not necessarily great. Under the additional assumption of normality of the errors conditional on \mathbf{X} , an assumption not usually made in microeconomics applications, the OLS estimator is normally distributed conditional on \mathbf{X} .

Consistency

The term $\text{plim} (N^{-1} \mathbf{X}' \mathbf{X})^{-1} = \mathbf{M}_{\mathbf{xx}}^{-1}$ since $\text{plim} N^{-1} \mathbf{X}' \mathbf{X} = \mathbf{M}_{\mathbf{xx}}$ by Assumption 3. Consistency then requires that condition (4.13) holds. This is established using a law of large numbers applied to the average $N^{-1} \mathbf{X}' \mathbf{u} = N^{-1} \sum_i \mathbf{x}_i u_i$, which converges in probability to zero if $E[\mathbf{x}_i u_i] = \mathbf{0}$. Given Assumptions 1 and 2, the $\mathbf{x}_i u_i$ are iid and Assumptions 1–5 permit use of the Markov LLN (Theorem A.9). If Assumption 1 is simplified to (y_i, \mathbf{x}_i) iid then $\mathbf{x}_i u_i$ are iid and Assumptions 1–4 permit simpler use of the Kolmogorov LLN (Theorem A.8).

Limit Distribution

By Assumption 3, $\text{plim} (N^{-1} \mathbf{X}' \mathbf{X})^{-1} = \mathbf{M}_{\mathbf{xx}}^{-1}$. The key is to obtain the limit distribution of $N^{-1/2} \mathbf{X}' \mathbf{u} = N^{-1/2} \sum_i \mathbf{x}_i u_i$ by application of a central limit theorem. Given Assumptions 1 and 2, the $\mathbf{x}_i u_i$ are iid and Assumptions 1–6 permit use of the Liapounov CLT (Theorem A.15). If assumption 1 is strengthened to (y_i, \mathbf{x}_i) iid then $\mathbf{x}_i u_i$ are iid and Assumptions 1–5 permit simpler use of the Lindeberg–Levy CLT (Theorem A.14).

This yields

$$\frac{1}{\sqrt{N}} \mathbf{X}' \mathbf{u} \xrightarrow{d} \mathcal{N} [\mathbf{0}, \mathbf{M}_{\mathbf{x}\Omega\mathbf{x}}], \quad (4.26)$$

where $\mathbf{M}_{\mathbf{x}\Omega\mathbf{x}} = \text{plim} N^{-1} \mathbf{X}' \mathbf{u} \mathbf{u}' \mathbf{X} = \text{plim} N^{-1} \sum_i u_i^2 \mathbf{x}_i \mathbf{x}_i'$ given independence over i . Application of a law of large numbers yields $\mathbf{M}_{\mathbf{x}\Omega\mathbf{x}} = \lim N^{-1} \sum_i E_{\mathbf{x}_i} [\sigma_i^2 \mathbf{x}_i \mathbf{x}_i']$, using $E_{u_i, \mathbf{x}_i} [u_i^2 \mathbf{x}_i \mathbf{x}_i'] = E_{\mathbf{x}_i} [E[u_i^2 | \mathbf{x}_i] \mathbf{x}_i \mathbf{x}_i']$ and $\sigma_i^2 = E[u_i^2 | \mathbf{x}_i]$. It follows that $\mathbf{M}_{\mathbf{x}\Omega\mathbf{x}} = \lim N^{-1} E[\mathbf{X}' \Omega \mathbf{X}]$, where $\Omega = \text{Diag}[\sigma_i^2]$ and the expectation is with respect to only \mathbf{X} , rather than both \mathbf{X} and \mathbf{u} .

The presentation here assumes independence over i . More generally we can permit correlated observations. Then $\mathbf{M}_{\mathbf{x}\Omega\mathbf{x}} = \text{plim} N^{-1} \sum_i \sum_j u_i u_j \mathbf{x}_i \mathbf{x}_j'$ and Ω has i,j th entry $\sigma_{ij} = \text{Cov}[u_i, u_j]$. This complication is deferred to treatment of the nonlinear LS estimator in Section 5.8.

Heteroskedasticity-Robust Standard Errors

We consider the key step of consistent estimation of $\mathbf{M}_{\mathbf{x}\Omega\mathbf{x}}$. Beginning with the original definition of $\mathbf{M}_{\mathbf{x}\Omega\mathbf{x}} = \text{plim} N^{-1} \sum_{i=1}^N u_i^2 \mathbf{x}_i \mathbf{x}_i'$, we replace u_i by $\hat{u}_i = y_i - \mathbf{x}_i' \hat{\beta}$, where

asymptotically $\widehat{u}_i \xrightarrow{p} u_i$ since $\widehat{\beta} \xrightarrow{p} \beta$. This yields the consistent estimate

$$\widehat{\mathbf{M}}_{\mathbf{x}\Omega\mathbf{x}} = \frac{1}{N} \sum_{i=1}^N \widehat{u}_i^2 \mathbf{x}_i \mathbf{x}'_i = N^{-1} \mathbf{X}' \widehat{\Omega} \mathbf{X}, \quad (4.27)$$

where $\widehat{\Omega} = \text{Diag}[\widehat{u}_i^2]$. The additional assumption that $E[|x_{ij}^2 x_{ik} x_{il}|^{1+\delta}] < \Delta$ for positive constants δ and Δ and $j, k, l = 1, \dots, K$ is needed, as $\widehat{u}_i^2 \mathbf{x}_i \mathbf{x}'_i = (u_i - \mathbf{x}'_i (\widehat{\beta} - \beta))^2 \mathbf{x}_i \mathbf{x}'_i$ involves up to the fourth power of \mathbf{x}_i (see White (1980a)).

Note that $\widehat{\Omega}$ does not converge to the $N \times N$ matrix Ω , a seemingly impossible task without additional structure as there are N variances σ_i^2 to be estimated. But all that is needed is that $N^{-1} \mathbf{X}' \widehat{\Omega} \mathbf{X}$ converges to the $K \times K$ matrix $\text{plim } N^{-1} \mathbf{X}' \Omega \mathbf{X} = N^{-1} \text{plim} \sum_i \sigma_i^2 \mathbf{x}_i \mathbf{x}'_i$. This is easier to achieve because the number of regressors K is fixed. To understand White's estimator, consider OLS estimation of the intercept-only model $y_i = \beta + u_i$ with heteroskedastic error. Then in our notation we can show that $\widehat{\beta} = \bar{y}$, $\mathbf{M}_{\mathbf{xx}} = \lim N^{-1} \sum_i 1 = 1$, and $\mathbf{M}_{\mathbf{x}\Omega\mathbf{x}} = \lim N^{-1} \sum_i E[u_i^2]$. An obvious estimator for $\mathbf{M}_{\mathbf{x}\Omega\mathbf{x}}$ is $\widehat{\mathbf{M}}_{\mathbf{x}\Omega\mathbf{x}} = N^{-1} \sum_i \widehat{u}_i^2$, where $\widehat{u}_i = y_i - \widehat{\beta}$. To obtain the probability limit of this estimate, it is enough to consider $N^{-1} \sum_i \widehat{u}_i^2$, since $\widehat{u}_i \xrightarrow{p} 0$ given $\widehat{\beta} \xrightarrow{p} \beta$. If a law of large numbers can be applied this average converges to the limit of its expected value, so $\text{plim } N^{-1} \sum_i \widehat{u}_i^2 = \lim N^{-1} \sum_i E[\widehat{u}_i^2] = \mathbf{M}_{\mathbf{x}\Omega\mathbf{x}}$ as desired. Eicker (1967) gave the formal conditions for this example.

4.5. Weighted Least Squares

If robust standard errors need to be used efficiency gains are usually possible. For example, if heteroskedasticity is present then the feasible generalized least-squares (GLS) estimator is more efficient than the OLS estimator.

In this section we present the feasible GLS estimator, an estimator that makes stronger distributional assumptions about the variance of the error term. It is nonetheless possible to obtain standard errors of the feasible GLS estimator that are robust to misspecification of the error variance, just as in the OLS case.

Many studies in microeconomics do not take advantage of the potential efficiency gains of GLS, for reasons of convenience and because the efficiency gains may be felt to be relatively small. Instead, it is common to use less efficient weighted least-squares estimators, most notably OLS, with robust estimates of the standard errors.

4.5.1. GLS and Feasible GLS

By the Gauss–Markov theorem, presented in introductory texts, the OLS estimator is efficient among linear unbiased estimators if the linear regression model errors are independent and homoskedastic.

Instead, we assume that the error variance matrix $\Omega \neq \sigma^2 \mathbf{I}$. If Ω is known and nonsingular, we can premultiply the linear regression model (4.8) by $\Omega^{-1/2}$, where

$\Omega^{1/2}\Omega^{1/2} = \Omega$, to yield

$$\Omega^{-1/2}\mathbf{y} = \Omega^{-1/2}\mathbf{X}\beta + \Omega^{-1/2}\mathbf{u}.$$

Some algebra yields $V[\Omega^{-1/2}\mathbf{u}] = E[(\Omega^{-1/2}\mathbf{u})(\Omega^{-1/2}\mathbf{u})'|X] = I$. The errors in this transformed model are therefore zero mean, uncorrelated, and homoskedastic. So β can be efficiently estimated by OLS regression of $\Omega^{-1/2}\mathbf{y}$ on $\Omega^{-1/2}\mathbf{X}$.

This argument yields the **generalized least-squares estimator**

$$\hat{\beta}_{GLS} = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}\mathbf{y}. \quad (4.28)$$

The GLS estimator cannot be directly implemented because in practice Ω is not known. Instead, we specify that $\Omega = \Omega(\gamma)$, where γ is a finite-dimensional parameter vector, obtain a consistent estimate $\hat{\gamma}$ of γ , and form $\hat{\Omega} = \Omega(\hat{\gamma})$. For example, if errors are heteroskedastic then specify $V[u|x] = \exp(z'\gamma)$, where z is a subset of x and the exponential function is used to ensure a positive variance. Then $\hat{\gamma}$ can be consistently estimated by nonlinear least-squares regression (see Section 5.8) of the squared OLS residual $\hat{u}_i^2 = (y - \mathbf{x}'\hat{\beta}_{OLS})^2$ on $\exp(z'\gamma)$. This estimate $\hat{\Omega}$ can be used in place of Ω in (4.28). Note that we cannot replace Ω in (4.28) by $\hat{\Omega} = \text{Diag}[\hat{u}_i^2]$ as this yields an inconsistent estimator (see Section 5.8.6).

The **feasible generalized least-squares (FGLS) estimator** is

$$\hat{\beta}_{FGLS} = (\mathbf{X}'\hat{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\Omega}^{-1}\mathbf{y}. \quad (4.29)$$

If Assumptions 1–6 hold and $\Omega(\gamma)$ is correctly specified, a strong assumption that is relaxed in the following, and $\hat{\gamma}$ is consistent for γ , it can be shown that

$$\sqrt{N}(\hat{\beta}_{FGLS} - \beta) \xrightarrow{d} \mathcal{N}\left[\mathbf{0}, \left(\text{plim } N^{-1}\mathbf{X}'\Omega^{-1}\mathbf{X}\right)^{-1}\right]. \quad (4.30)$$

The FGLS estimator has the same limiting variance matrix as the GLS estimator and so is second-moment efficient. For implementation replace Ω by $\hat{\Omega}$ in (4.30).

It can be shown that the GLS estimator minimizes $\mathbf{u}'\Omega^{-1}\mathbf{u}$, see Exercise 4.5, which simplifies to $\sum_i u_i^2/\sigma_i^2$ if errors are heteroskedastic but uncorrelated. The motivation provided for GLS was efficient estimation of β . In terms of the Section 4.2 discussion of loss function and optimal prediction, with heteroskedastic errors the loss function is $L(e) = e^2/\sigma^2$. Compared to OLS with $L(e) = e^2$, the GLS loss function places a relatively smaller penalty on the prediction error for observations with large conditional error variance.

4.5.2. Weighted Least Squares

The result in (4.30) assumes correct specification of the error variance matrix $\Omega(\gamma)$. If instead $\Omega(\gamma)$ is misspecified then the FGLS estimator is still consistent, but (4.30) gives the wrong variance. Fortunately, a robust estimate of the variance of the GLS estimator can be found even if $\Omega(\gamma)$ is misspecified.

Specifically, define $\Sigma = \Sigma(\gamma)$ to be a **working variance matrix** that does not necessarily equal the true variance matrix $\Omega = E[\mathbf{u}\mathbf{u}'|X]$. Form an estimate $\hat{\Sigma} = \Sigma(\hat{\gamma})$,

Table 4.2. Least-Squares Estimators and Their Asymptotic Variance

Estimator ^a	Definition	Estimated Asymptotic Variance
OLS	$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$	$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$
FGLS	$\hat{\beta} = (\mathbf{X}'\hat{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\Omega}^{-1}\mathbf{y}$	$(\mathbf{X}'\hat{\Omega}^{-1}\mathbf{X})^{-1}$
WLS	$\hat{\beta} = (\mathbf{X}'\hat{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\Sigma}^{-1}\mathbf{y}$	$(\mathbf{X}'\hat{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\Sigma}^{-1}\hat{\Omega}\hat{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\hat{\Sigma}^{-1}\mathbf{X})^{-1}$.

^a Estimators are for linear regression model with error conditional variance matrix Ω . For FGLS it is assumed that $\hat{\Omega}$ is consistent for Ω . For OLS and WLS the heteroskedastic robust variance matrix of $\hat{\beta}$ uses $\hat{\Omega}$ equal to a diagonal matrix with squared residuals on the diagonals.

where $\hat{\gamma}$ is an estimate of γ . Then use weighted least squares with weighting matrix $\hat{\Sigma}^{-1}$.

This yields the **weighted least-squares (WLS) estimator**

$$\hat{\beta}_{WLS} = (\mathbf{X}'\hat{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\Sigma}^{-1}\mathbf{y}. \quad (4.31)$$

Statistical inference is then done without the assumption that $\Sigma = \Omega$, the true variance matrix of the error term. In the statistics literature this approach is referred to as a working matrix approach. We call it weighted least squares, but be aware that others instead use weighted least squares to mean GLS or FGLS in the special case that Ω^{-1} is diagonal. Here there is no presumption that the weighting matrix $\Sigma^{-1} = \Omega^{-1}$.

Similar algebra to that for OLS given in Section 4.4.5 yields the estimated asymptotic variance matrix

$$\hat{V}[\hat{\beta}_{WLS}] = (\mathbf{X}'\hat{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\Sigma}^{-1}\hat{\Omega}\hat{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\hat{\Sigma}^{-1}\mathbf{X})^{-1}, \quad (4.32)$$

where $\hat{\Omega}$ is such that

$$\text{plim } N^{-1}\mathbf{X}'\hat{\Sigma}^{-1}\hat{\Omega}\hat{\Sigma}^{-1}\mathbf{X} = \text{plim } N^{-1}\mathbf{X}'\Sigma^{-1}\Omega\Sigma^{-1}\mathbf{X}.$$

In the heteroskedastic case $\hat{\Omega} = \text{Diag}[\hat{u}_i^{*2}]$, where $\hat{u}_i^* = y_i - \mathbf{x}'_i\hat{\beta}_{WLS}$.

For heteroskedastic errors the basic approach is to choose a simple model for heteroskedasticity such as error variance depending on only one or two key regressors. For example, in a linear regression model of the level of wages as a function of schooling and other variables, the heteroskedasticity might be modeled as a function of schooling alone. Suppose this model yields $\hat{\Sigma} = \text{Diag}[\hat{\sigma}_i^2]$. Then OLS regression of $y_i/\hat{\sigma}_i$ on $\mathbf{x}_i/\hat{\sigma}_i$ (with the no-constant option) yields $\hat{\beta}_{WLS}$ and the White robust standard errors from this regression can be shown to equal those based on (4.32).

The weighted least-squares or working matrix approach is especially convenient when there is more than one complication. For example, in the random effects panel data model of Chapter 21 the errors may be viewed as both correlated over time for a given individual and heteroskedastic. One may use the random effects estimator, which controls only for the first complication, but then compute heteroskedastic-consistent standard errors for this estimator.

The various least-squares estimators are summarized in Table 4.2.

Table 4.3. Least Squares: Example with Conditionally Heteroskedastic Errors^a

	OLS	WLS	GLS
Constant	2.213 (0.823) [0.820]	1.060 (0.150) [0.051]	0.996 (0.007) [0.006]
x	0.979 (0.178) [0.275]	0.957 (0.190) [0.232]	0.952 (0.209) [0.208]
R^2	0.236	0.205	0.174

^a Generated data for sample size of 100. OLS, WLS, and GLS are all consistent but OLS and WLS are inefficient. Two different standard errors are given: default standard errors assuming homoskedastic errors in parentheses and heteroskedastic-robust standard errors in square brackets. The data-generating process is given in the text.

4.5.3. Robust Standard Errors for LS Example

As an example of robust standard error estimation, consider estimation of the standard error of least-squares estimates of the slope coefficient for a dgp with multiplicative heteroskedasticity

$$\begin{aligned}y &= 1 + 1 \times x + u, \\u &= x\varepsilon,\end{aligned}$$

where the scalar regressor $x \sim \mathcal{N}[0, 25]$ and $\varepsilon \sim \mathcal{N}[0, 4]$.

The errors are conditionally heteroskedastic, since $V[u|x] = V[x\varepsilon|x] = x^2V[\varepsilon|x] = 4x^2$, which depends on the regressor x . This differs from the unconditional variance, where $V[u] = V[x\varepsilon] = E[(x\varepsilon)^2] - (E[x\varepsilon])^2 = E[x^2]E[\varepsilon^2] = V[x]V[\varepsilon] = 100$, given x and ε independent and the particular dgp here.

Standard errors for the OLS estimator should be calculated using the heteroskedastic-consistent or robust variance estimate (4.21). Since OLS is not fully efficient, WLS may provide efficiency gains. GLS will definitely provide efficiency gains and in this simulated data example we have the advantage of knowing that $V[u|x] = 4x^2$. All estimation methods yield a consistent estimate of the intercept and slope coefficients.

Various least-squares estimates and associated standard errors from a generated data sample of size 100 are given in Table 4.3. We focus on the slope coefficient.

The OLS slope coefficient estimate is 0.979. Two standard error estimates are reported, with the correct heteroskedasticity-robust standard error of 0.275 using (4.21) much larger here than the incorrect estimate of 0.177 that uses $s^2(\mathbf{X}'\mathbf{X})^{-1}$. Such a large difference in standard error estimates could lead to quite different conclusions in statistical inference. In general the direction of bias in the standard errors could be in either direction. For this example it can be shown theoretically that, in the limit, the robust standard errors are $\sqrt{3}$ times larger than the incorrect one. Specifically, for this dgp

and for sample size N the correct and incorrect standard errors of the OLS estimate of the slope coefficient converge to, respectively, $\sqrt{12/N}$ and $\sqrt{4/N}$.

As an example of the WLS estimator, assume that $u = \sqrt{|x|}\varepsilon$ rather than $u = x\varepsilon$, so that it is assumed that $V[u] = \sigma^2|x|$. The WLS estimator can be computed by OLS regression after dividing y , the intercept, and x by $\sqrt{|x|}$. Since this is the wrong model for the heteroskedastic error the correct standard error for the slope coefficient is the robust estimate of 0.232, computed using (4.32).

The GLS estimator for this dgp can be computed by OLS regression after dividing y , the intercept, and x by $|x|$, since the transformed error is then homoskedastic. The usual and robust standard errors for the slope coefficient are similar (0.209 and 0.208). This is expected as both are asymptotically correct because the GLS estimator here uses the correct model for heteroskedasticity. It can be shown theoretically that for this dgp the standard error of the GLS estimate of the slope coefficient converges to $\sqrt{4/N}$.

Both OLS and WLS are less efficient than GLS, as expected, with standard errors for the slope coefficient of, respectively, $0.275 > 0.232 > 0.208$.

The setup in this example is a standard one used in estimation theory for cross-section data. Both y and \mathbf{x} are stochastic random variables. The pair (y_i, \mathbf{x}_i) are independent over i and identically distributed, as is the case under random sampling. The conditional distribution of $y_i|\mathbf{x}_i$ differs over i , however, since the conditional mean and variance of y_i depend on \mathbf{x}_i .

4.6. Median and Quantile Regression

In an intercept-only model, summary statistics for the sample distribution include quantiles, such as the median, lower and upper quartiles, and percentiles, in addition to the sample mean.

In the regression context we might similarly be interested in conditional quantiles. For example, interest may lie in how the percentiles of the earnings distribution for lowly educated workers are much more compressed than those for highly educated workers. In this simple example one can just do separate computations for lowly educated workers and for highly educated workers. However, this approach becomes infeasible if there are several regressors taking several values. Instead, quantile regression methods are needed to estimate the quantiles of the conditional distribution of y given \mathbf{x} .

From Table 4.1, quantile regression corresponds to use of asymmetric absolute loss, whereas the special case of median regression uses absolute error loss. These methods provide an alternative to OLS, which uses squared error loss.

Quantile regression methods have advantages beyond providing a richer characterization of the data. Median regression is more robust to outliers than least-squares regression. Moreover, quantile regression estimators can be consistent under weaker stochastic assumptions than possible with least-squares estimation. Leading examples are the maximum score estimator of Manski (1975) for binary outcome models (see Section 14.6) and the censored least absolute deviations estimator of Powell (1984) for censored models (see Section 16.6).

We begin with a brief explanation of population quantiles before turning to estimation of sample quantiles.

4.6.1. Population Quantiles

For a continuous random variable y , the **population q th quantile** is that value μ_q such that y is less than or equal to μ_q with probability q . Thus

$$q = \Pr[y \leq \mu_q] = F_y(\mu_q),$$

where F_y is the cumulative distribution function (cdf) of y . For example, if $\mu_{0.75} = 3$ then the probability that $y \leq 3$ equals 0.75. It follows that

$$\mu_q = F_y^{-1}(q).$$

Leading examples are the median, $q = 0.5$, the upper quartile, $q = 0.75$, and the lower quartile, $q = 0.25$. For the standard normal distribution $\mu_{0.5} = 0.0$, $\mu_{0.95} = 1.645$, and $\mu_{0.975} = 1.960$. The $100q$ th **percentile** is the q th quantile.

For the regression model, the **population q th quantile** of y conditional on \mathbf{x} is that function $\mu_q(\mathbf{x})$ such that y conditional on \mathbf{x} is less than or equal to $\mu_q(\mathbf{x})$ with probability q , where the probability is evaluated using the conditional distribution of y given \mathbf{x} . It follows that

$$\mu_q(\mathbf{x}) = F_{y|\mathbf{x}}^{-1}(q), \quad (4.33)$$

where $F_{y|\mathbf{x}}$ is the conditional cdf of y given \mathbf{x} and we have suppressed the role of the parameters of this distribution.

It is insightful to derive the quantile function $\mu_q(\mathbf{x})$ if the dgp is assumed to be the linear model with multiplicative heteroskedasticity

$$\begin{aligned} y &= \mathbf{x}'\boldsymbol{\beta} + u, \\ u &= \mathbf{x}'\boldsymbol{\alpha} \times \varepsilon, \\ \varepsilon &\sim \text{iid } [0, \sigma^2], \end{aligned}$$

where it is assumed that $\mathbf{x}'\boldsymbol{\alpha} > 0$. Then the population q th quantile of y conditional on \mathbf{x} is that function $\mu_q(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\alpha})$ such that

$$\begin{aligned} q &= \Pr[y \leq \mu_q(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\alpha})] \\ &= \Pr[u \leq \mu_q(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\alpha}) - \mathbf{x}'\boldsymbol{\beta}] \\ &= \Pr[\varepsilon \leq [\mu_q(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\alpha}) - \mathbf{x}'\boldsymbol{\beta}] / \mathbf{x}'\boldsymbol{\alpha}] \\ &= F_\varepsilon([\mu_q(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\alpha}) - \mathbf{x}'\boldsymbol{\beta}] / \mathbf{x}'\boldsymbol{\alpha}), \end{aligned}$$

where we use $u = y - \mathbf{x}'\boldsymbol{\beta}$ and $\varepsilon = u / \mathbf{x}'\boldsymbol{\alpha}$, and F_ε is the cdf of ε . It follows that $[\mu_q(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\alpha}) - \mathbf{x}'\boldsymbol{\beta}] / \mathbf{x}'\boldsymbol{\alpha} = F_\varepsilon^{-1}(q)$ so that

$$\begin{aligned} \mu_q(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\alpha}) &= \mathbf{x}'\boldsymbol{\beta} + \mathbf{x}'\boldsymbol{\alpha} \times F_\varepsilon^{-1}(q) \\ &= \mathbf{x}'(\boldsymbol{\beta} + \boldsymbol{\alpha} \times F_\varepsilon^{-1}(q)). \end{aligned}$$

Thus for the linear model with multiplicative heteroskedasticity of the form $u = \mathbf{x}'\boldsymbol{\alpha} \times \varepsilon$ the conditional quantiles are linear in \mathbf{x} . In the special case of homoskedasticity, $\mathbf{x}'\boldsymbol{\alpha}$ equals a constant and all conditional quantiles have the same slope and differ only in their intercept, which becomes larger as q increases.

In more general examples the quantile function may be nonlinear in \mathbf{x} , owing to other forms of heteroskedasticity such as $u = h(\mathbf{x}, \boldsymbol{\alpha}) \times \varepsilon$, where $h(\cdot)$ is nonlinear in \mathbf{x} , or because the regression function itself is of nonlinear form $g(\mathbf{x}, \boldsymbol{\beta})$. It is standard to still estimate quantile functions that are linear and interpret them as the best linear predictor under the quantile regression loss function given in (4.34) in the next section.

4.6.2. Sample Quantiles

For univariate random variable y the usual way to obtain the sample quantile estimate is to first order the sample. Then $\hat{\mu}_q$ equals the $[Nq]$ th smallest value, where N is the sample size and $[Nq]$ denotes Nq rounded up to the nearest integer. For example, if $N = 97$, the lower quartile is the 25th observation since $[97 \times 0.25] = [24.25] = 25$.

Koenker and Bassett (1978) observed that the **sample q th quantile** $\hat{\mu}_q$ can equivalently be expressed as the solution to the optimization problem of minimizing with respect to β

$$\sum_{i:y_i \geq \beta}^N q|y_i - \beta| + \sum_{i:y_i < \beta}^N (1-q)|y_i - \beta|.$$

This result is not obvious. To gain some understanding, consider the median, where $q = 0.5$. Then the median is the minimum of $\sum_i |y_i - \beta|$. Suppose in a sample of 99 observations that the 50th smallest observation, the median, equals 10 and the 51st smallest observation equals 12. If we let β equal 12 rather than 10, then $\sum_i |y_i - \beta|$ will increase by 2 for the first 50 ordered observations and decrease by 2 for the remaining 49 observations, leading to an overall net increase of $50 \times 2 - 49 \times 2 = 2$. So the 51st smallest observation is a worse choice than the 50th. Similarly the 49th smallest observation can be shown to be a worse choice than the 50th observation.

This objective function is then readily expanded to the linear regression case, so that the q th **quantile regression estimator** $\hat{\beta}_q$ minimizes over β_q

$$Q_N(\beta_q) = \sum_{i:y_i \geq \mathbf{x}'_i \beta}^N q|y_i - \mathbf{x}'_i \beta_q| + \sum_{i:y_i < \mathbf{x}'_i \beta}^N (1-q)|y_i - \mathbf{x}'_i \beta_q|, \quad (4.34)$$

where we use β_q rather than β to make clear that different choices of q estimate different values of β . Note that this is the asymmetric absolute loss function given in Table 4.1, where \hat{y} is restricted to be linear in \mathbf{x} so that $e = y - \mathbf{x}'\beta_q$. The special case $q = 0.5$ is called the **median regression estimator** or the **least absolute deviations estimator**.

4.6.3. Properties of Quantile Regression Estimators

The objective function (4.34) is not differentiable and so the gradient optimization methods presented in Chapter 10 are not applicable. Fortunately, linear programming methods can be used and these provide relatively fast computation of $\hat{\beta}_q$.

Since there is no explicit solution for $\hat{\beta}_q$ the asymptotic distribution of $\hat{\beta}_q$ cannot be obtained using the approach of Section 4.4 for OLS. The methods of Chapter 5 also require adaptation, as the objective function is nondifferentiable. It can be shown that

$$\sqrt{N}(\hat{\beta}_q - \beta_q) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}], \quad (4.35)$$

(see, for example, Buchinsky, 1998, p. 85), where

$$\begin{aligned} \mathbf{A} &= \text{plim} \frac{1}{N} \sum_{i=1}^N f_{u_q}(0|\mathbf{x}_i) \mathbf{x}_i \mathbf{x}_i' , \\ \mathbf{B} &= \text{plim} \frac{1}{N} \sum_{i=1}^N q(1-q) \mathbf{x}_i \mathbf{x}_i' , \end{aligned} \quad (4.36)$$

and $f_{u_q}(0|\mathbf{x})$ is the conditional density of the error term $u_q = y - \mathbf{x}'\beta_q$ evaluated at $u_q = 0$. Estimation of the variance of $\hat{\beta}_q$ is complicated by the need to estimate $f_{u_q}(0|\mathbf{x})$. It is easier to instead obtain standard errors for $\hat{\beta}_q$ using the bootstrap pairs procedure of Chapter 11.

4.6.4. Quantile Regression Example

In this section we perform conditional quantile estimation and compare it with the usual conditional mean estimation using OLS regression. The application involves Engel curve estimation for household annual medical expenditure. More specifically, we consider the regression relationship between the log of medical expenditure and the log of total household expenditure. This regression yields an estimate of the (constant) elasticity of medical expenditure with respect to total expenditure.

The data are from the World Bank's 1997 Vietnam Living Standards Survey. The sample consists of 5,006 households that have positive level of medical expenditures, after dropping 16.6% of the sample that has zero expenditures to permit taking the natural logarithm. Zero values can be handled using the censored quantile regression methods of Powell (1986a), presented in Section 16.9.2. For simplicity we simply dropped observations with zero expenditures. The largest component of medical expenditure, especially at low levels of income, consists of medications purchased from pharmacies. Although several household characteristic variables are available, for simplicity we only consider one regressor, the log of total household expenditure, to serve as a proxy for household income.

The linear least-squares regression yields an elasticity estimate of 0.57. This estimate would be usually interpreted to mean that medicines are a “necessity” and hence their demand is income inelastic. This estimate is not very surprising, but before accepting it at face value we should acknowledge that there may be considerable heterogeneity in the elasticity across different income groups.

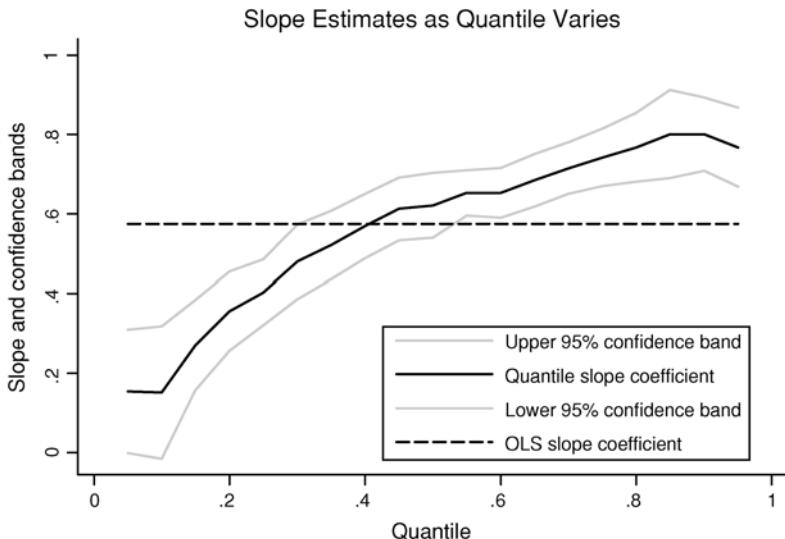


Figure 4.1: Quantile regression estimates of slope coefficient for $q = 0.05, 0.10, \dots, 0.90, 0.95$ and associated 95% confidence bands plotted against q from regression of the natural logarithm of medical expenditure on the natural logarithm of total expenditure.

Quantile regression is a useful tool for studying such heterogeneity, as emphasized by Koenker and Hallock (2001). We minimize the quantity (4.34), where y is log of medical expenditure and $\mathbf{x}'\beta = \beta_1 + \beta_2 x$, where x is log of total household expenditure. This is done for the nineteen quantile values $q = \{0.05, 0.10, \dots, 0.95\}$, where $q = 0.5$ is the median. In each case the standard errors were estimated using the bootstrap method with 50 resamples. The results of this exercise are condensed into Figures 4.1 and 4.2.

Figure 4.1 plots the slope coefficient $\hat{\beta}_{2,q}$ for the different values of q , along with the associated 95% confidence interval. This shows how the quantile estimates of the elasticity varies with quantile value q . The elasticity estimate increases systematically with the level of household income, rising from 0.15 for $q = 0.05$ to a maximum of 0.80 for $q = 0.85$. The least-squares slope estimate of 0.57 is also presented as a horizontal line that does not vary with quantile. The elasticity estimates at lower and higher quantiles are clearly statistically significantly different from each other and from the OLS estimate, which has standard error 0.032. It seems that the aggregate elasticity estimate will vary according to changes in the underlying income distribution. This graph supports the observation of Mosteller and Tukey (1977, p. 236), quoted by Koenker and Hallock (2001), that by focusing only on the conditional mean function the least-squares regression gives an incomplete summary of the joint distribution of dependent and explanatory variables.

Figure 4.2 superimposes three estimated quantile regression lines $\hat{y}_q = \hat{\beta}_{1,q} + \hat{\beta}_{2,q}x$ for $q = 0.1, 0.2, \dots, 0.9$ and the OLS regression line. The OLS regression line, not graphed, is similar to the median ($q = 0.5$) regression line. There is a fanning out of the quantile regression lines in Figure 4.2. This is not surprising given the increase

Regression Lines as Quantile Varies

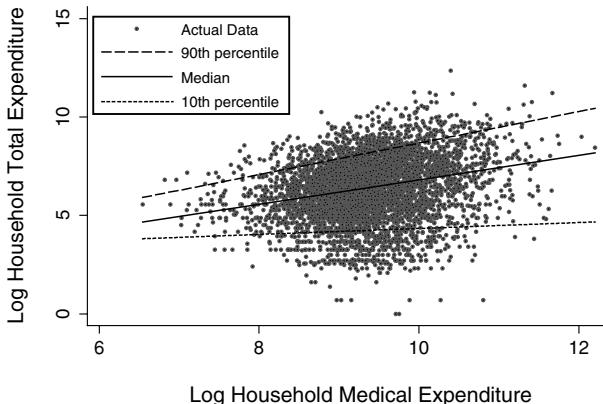


Figure 4.2: Quantile regression estimated lines for $q = 0.1$, $q = 0.5$ and $q = 0.9$ from regression of natural logarithm of medical expenditure on natural logarithm of total expenditure. Data for 5006 Vietnamese households with positive medical expenditures in 1997.

in estimated slopes as q increases as evident in Figure 4.1. Koenker and Bassett (1982) developed quantile regression as a means to test for heteroskedastic errors when the dgp is the linear model. For such a case a fanning out of the quantile regression lines is interpreted as evidence of heteroskedasticity. Another interpretation is that the conditional mean is nonlinear in x with increasing slope and this leads to quantile slope coefficients that increase with quantile q .

More detailed illustrations of quantile regression are given in Buchinsky (1994) and Koenker and Hallock (2001).

4.7. Model Misspecification

The term “model misspecification” in its broadest sense means that one or more of the assumptions made on the data generating process are incorrect. Misspecifications may occur individually or in combination, but analysis is simpler if only the consequences of a single misspecification are considered.

In the following discussion we emphasize misspecifications that lead to inconsistency of the least-squares estimator and loss of identifiability of parameters of interest. The least-squares estimator may nonetheless continue to have a meaningful interpretation, only one different from that intended under the assumption of a correctly specified model. Specifically, the estimator may converge asymptotically to a parameter that differs from the true population value, a concept defined in Section 4.7.5 as the pseudo-true value.

The issues raised here for consistency of OLS are relevant to other estimators in other models. Consistency can then require stronger assumptions than those needed

for consistency of OLS, so that inconsistency resulting from model misspecification is more likely.

4.7.1. Inconsistency of OLS

The most serious consequence of a model misspecification is inconsistent estimation of the regression parameters β . From Section 4.4, the two key conditions needed to demonstrate consistency of the OLS estimator are (1) the dgp is $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$ and (2) the dgp is such that $\text{plim } N^{-1}\mathbf{X}'\mathbf{u} = \mathbf{0}$. Then

$$\begin{aligned}\widehat{\beta}_{\text{OLS}} &= \beta + (N^{-1}\mathbf{X}'\mathbf{X})^{-1}N^{-1}\mathbf{X}'\mathbf{u} \\ &\xrightarrow{P} \beta,\end{aligned}\tag{4.37}$$

where the first equality follows if $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$ (see (4.12)) and the second line uses $\text{plim } N^{-1}\mathbf{X}'\mathbf{u} = \mathbf{0}$.

The OLS estimator is likely to be **inconsistent** if model misspecification leads to either specification of the wrong model for \mathbf{y} , so that condition 1 is violated, or correlation of regressors with the error, so that condition 2 is violated.

4.7.2. Functional Form Misspecification

A linear specification of the conditional mean function is merely an approximation in R^K to the true unknown conditional mean function in parameter space of indeterminate dimension. Even if the correct regressors are chosen, it is possible that the conditional mean is incorrectly specified.

Suppose the dgp is one with a nonlinear regression function

$$y = g(\mathbf{x}) + v,$$

where the dependence of $g(\mathbf{x})$ on unknown parameters is suppressed, and assume $E[v|\mathbf{x}] = 0$. The linear regression model

$$y = \mathbf{x}'\beta + u$$

is erroneously specified. The question is whether the OLS estimator can be given any meaningful interpretation, even though the dgp is in fact nonlinear.

The usual way to interpret regression coefficients is through the true *micro relationship*, which here is

$$E[y_i | \mathbf{x}_i] = g(\mathbf{x}_i).$$

In this case $\widehat{\beta}_{\text{OLS}}$ does not measure the micro response of $E[y_i | \mathbf{x}_i]$ to a change in \mathbf{x}_i , as it does not converge to $\partial g(\mathbf{x}_i)/\partial \mathbf{x}_i$. So the usual interpretation of $\widehat{\beta}_{\text{OLS}}$ is not possible.

White (1980b) showed that the OLS estimator converges to that value of β that minimizes the mean-squared prediction error

$$E_{\mathbf{x}}[(g(\mathbf{x}) - \mathbf{x}'\beta)^2].$$

Hence prediction from OLS is the best linear predictor of the nonlinear regression function if the mean-squared error is used as the loss function. This useful property has already been noted in Section 4.2.3, but it adds little in interpretation of $\hat{\beta}_{OLS}$.

In summary, if the true regression function is nonlinear, OLS is not useful for individual prediction. OLS can still be useful for prediction of aggregate changes, giving the sample average change in $E[y|\mathbf{x}]$ due to change in \mathbf{x} (see Stoker, 1982). However, microeconometric analyses usually seek models that are meaningful at the individual level.

Much of this book presents alternatives to the linear model that are more likely to be correctly specified. For example, Chapter 14 on binary outcomes presents model specifications that ensure that predicted probabilities are restricted to lie between 0 and 1. Also, models and methods that rely on minimal distributional assumptions are preferred because there is then less scope for misspecification.

4.7.3. Endogeneity

Endogeneity is formally defined in Section 2.3. A broad definition is that a regressor is endogenous when it is correlated with the error term. If any one regressor is endogenous then in general OLS estimates of all regression parameters are inconsistent (unless the exogenous regressor is uncorrelated with the endogenous regressor).

Leading examples of endogeneity, dealt with extensively in this book in both linear and nonlinear model settings, include simultaneous equations bias (Section 2.4), omitted variable bias (Section 4.7.4), sample selection bias (Section 16.5), and measurement error bias (Chapter 26). Endogeneity is quite likely to occur when cross-section observational data are used, and economists are very concerned with this complication.

A quite general approach to control for endogeneity is the instrumental variables method, presented in Sections 4.8 and 4.9 and in Sections 6.4 and 6.5. This method cannot always be applied, however, as necessary instruments may not be available.

Other methods to control for endogeneity, reviewed in Section 2.8, include control for confounding variables, differences in differences if repeated cross-section or panel data are available (see Chapter 21), fixed effects if panel data are available and endogeneity arises owing to a time-invariant omitted variable (see Section 21.6), and regression-discontinuity design (see Section 25.6).

4.7.4. Omitted Variables

Omission of a variable in a linear regression equation is often the first example of inconsistency of OLS presented in introductory courses. Such omission may be the consequence of an erroneous exclusion of a variable for which data are available or of exclusion of a variable that is not directly observed. For example, omission of ability in a regression of earnings (or more usually its natural logarithm) on schooling is usually due to unavailability of a comprehensive measure of ability.

Let the true dgp be

$$y = \mathbf{x}'\beta + z\alpha + v, \quad (4.38)$$

4.7. MODEL MISSPECIFICATION

where \mathbf{x} and z are regressors, with z a scalar regressor for simplicity, and v is an error term that is assumed to be uncorrelated with the regressors \mathbf{x} and z . OLS estimation of y on \mathbf{x} and z will yield consistent parameter estimates of β and α .

Suppose instead that y is regressed on \mathbf{x} alone, with z omitted owing to unavailability. Then the term $z\alpha$ is moved into the error term. The estimated model is

$$y = \mathbf{x}'\beta + (z\alpha + v), \quad (4.39)$$

where the error term is now $(z\alpha + v)$. As before v is uncorrelated with \mathbf{x} , but if z is correlated with \mathbf{x} the error term $(z\alpha + v)$ will be correlated with the regressors \mathbf{x} . The OLS estimator will be inconsistent for β if z is correlated with \mathbf{x} .

There is enough structure in this example to determine the direction of the inconsistency. Stacking all observations in an obvious manner gives the dgp $\mathbf{y} = \mathbf{X}\beta + \mathbf{z}\alpha + \mathbf{v}$. Substituting this into $\widehat{\beta}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ yields

$$\widehat{\beta}_{OLS} = \beta + (N^{-1}\mathbf{X}'\mathbf{X})^{-1}(N^{-1}\mathbf{X}'\mathbf{z})\alpha + (N^{-1}\mathbf{X}'\mathbf{X})^{-1}(N^{-1}\mathbf{X}'\mathbf{v}).$$

Under the usual assumption that \mathbf{X} is uncorrelated with \mathbf{v} , the final term has probability limit zero. \mathbf{X} is correlated with \mathbf{z} , however, and

$$\text{plim } \widehat{\beta}_{OLS} = \beta + \delta\alpha, \quad (4.40)$$

where

$$\delta = \text{plim} [(N^{-1}\mathbf{X}'\mathbf{X})^{-1}(N^{-1}\mathbf{X}'\mathbf{z})]$$

is the probability limit of the OLS estimator in regression of the omitted regressor (\mathbf{z}) on the included regressors (\mathbf{X}).

This inconsistency is called **omitted variables bias**, where common terminology states that various misspecifications lead to bias even though formally they lead to inconsistency. The inconsistency exists as long as $\delta \neq 0$, that is, as long as the omitted variable is correlated with the included regressors. In general the inconsistency could be positive or negative and could even lead to a sign reversal of the OLS coefficient.

For the returns to schooling example, the correlation between schooling and ability is expected to be positive, so $\delta > 0$, and the return to ability is expected to be positive, so $\alpha > 0$. It follows that $\delta\alpha > 0$, so the omitted variables bias is positive in this example. OLS of earnings on schooling alone will overstate the effect of education on earnings.

A related form of misspecification is **inclusion of irrelevant regressors**. For example, the regression may be of y on \mathbf{x} and z , even though the dgp is more simply $y = \mathbf{x}'\beta + v$. In this case it is straightforward to show that OLS is consistent, but there is a loss of efficiency.

Controlling for omitted variables bias is necessary if parameter estimates are to be given a causal interpretation. Since too many regressors cause little harm, but too few regressors can lead to inconsistency, microeconomic models estimated from large data sets tend to include many regressors. If omitted variables are still present then one of the methods given at the end of Section 4.7.3 is needed.

4.7.5. Pseudo-True Value

In the omitted variables example the least-squares estimator is subject to *confounding* in the sense that it does not estimate β , but instead estimates a function of β , δ , and α .

The OLS estimate cannot be used as an estimate of β , which, for example, measures the effect of an exogenous change in a regressor \mathbf{x} such as schooling holding all other regressors including ability constant.

From (4.40), however, $\hat{\beta}_{OLS}$ is a consistent estimator of the function $(\beta + \delta\alpha)$ and has a meaningful interpretation. The probability limit of $\hat{\beta}_{OLS}$ of $\beta^* = (\beta + \delta\alpha)$ is referred to as the **pseudo-true value**, see Section 5.7.1 for a formal definition, corresponding to $\hat{\beta}_{OLS}$.

Furthermore, one can obtain the distribution of $\hat{\beta}_{OLS}$ even though it is inconsistent for β . The estimated asymptotic variance of $\hat{\beta}_{OLS}$ measures dispersion around $(\beta + \delta\alpha)$ and is given by the usual estimator, for example by $s^2(\mathbf{X}'\mathbf{X})^{-1}$ if the error in (4.38) is homoskedastic.

4.7.6. Parameter Heterogeneity

The presentation to date has permitted regressors and error terms to vary across individuals but has restricted the regression parameters β to be the same across individuals.

Instead, suppose that the dgp is

$$y_i = \mathbf{x}'_i \beta_i + u_i, \quad (4.41)$$

with subscript i on the parameters. This is an example of **parameter heterogeneity**, where the marginal effect $E[y_i | \mathbf{x}_i] = \beta_i$ is now permitted to differ across individuals.

The **random coefficients model** or **random parameters model** specifies β_i to be independently and identically distributed over i with distribution that does not depend on the observables \mathbf{x}_i . Let the common mean of β_i be denoted β . The dgp can be rewritten as

$$y_i = \mathbf{x}'_i \beta + (u_i + \mathbf{x}'_i (\beta_i - \beta)),$$

and enough assumptions have been made to ensure that the regressors \mathbf{x}_i are uncorrelated with the error term $(u_i + \mathbf{x}'_i (\beta_i - \beta))$. OLS regression of y on \mathbf{x} will therefore consistently estimate β , though note that the error is heteroskedastic even if u_i is homoskedastic.

For panel data a standard model is the random effects model (see Section 21.7) that lets the intercept vary across individuals while the slope coefficients are not random.

For nonlinear models a similar result need not hold, and random parameter models can be preferred as they permit a richer parameterization. Random parameter models are consistent with existence of heterogeneous responses of individuals to changes in \mathbf{x} . A leading example is random parameters logit in Section 15.7.

More serious complications can arise when the regression parameters β_i for an individual are related to observed individual characteristics. Then OLS estimation can lead to inconsistent parameter estimation. An example is the fixed effects model for panel data (see Section 21.6) for which OLS estimation of y on \mathbf{x} is inconsistent. In

this example, but not in all such examples, alternative consistent estimators for a subset of the regression parameters are available.

4.8. Instrumental Variables

A major complication that is emphasized in microeconomics is the possibility of inconsistent parameter estimation caused by endogenous regressors. Then regression estimates measure only the magnitude of association, rather than the magnitude and direction of causation, both of which are needed for policy analysis.

The instrumental variables estimator provides a way to nonetheless obtain consistent parameter estimates. This method, widely used in econometrics and rarely used elsewhere, is conceptually difficult and easily misused.

We provide a lengthy expository treatment that defines an instrumental variable and explains how the instrumental variables method works in a simple setting.

4.8.1. Inconsistency of OLS

Consider the scalar regression model with dependent variable y and single regressor x . The goal of regression analysis is to estimate the conditional mean function $E[y|x]$. A linear conditional mean model, without intercept for notational convenience, specifies

$$E[y|x] = \beta x. \quad (4.42)$$

This model without intercept subsumes the model with intercept if dependent and regressor variables are deviations from their respective means. Interest lies in obtaining a consistent estimate of β as this gives the change in the conditional mean given an *exogenous* change in x . For example, interest may lie in the effect in earnings caused by an increase in schooling attributed to exogenous reasons, such as an increase in the minimum age at which students leave school, that are not a choice of the individual.

The OLS regression model specifies

$$y = \beta x + u, \quad (4.43)$$

where u is an error term. Regression of y on x yields OLS estimate $\hat{\beta}$ of β .

Standard regression results make the assumption that the regressors are uncorrelated with the errors in the model (4.43). Then the only effect of x on y is a direct effect via the term βx . We have the following path analysis diagram:



where there is no association between x and u . So x and u are independent causes of y .

However, in some situations there may be an association between regressors and errors. For example, consider regression of log-earnings (y) on years of schooling (x). The error term u embodies all factors other than schooling that determine earnings,

such as ability. Suppose a person has a high level of u , as a result of high (unobserved) ability. This increases earnings, since $y = \beta x + u$, but it may also lead to higher levels of x , since schooling is likely to be higher for those with high ability. A more appropriate path diagram is then the following:

$$\begin{array}{ccc} x & \longrightarrow & y \\ \uparrow & \nearrow & \\ u & & \end{array}$$

where now there is an association between x and u .

What are the consequences of this correlation between x and u ? Now higher levels of x have two effects on y . From (4.43) there is both a direct effect via βx and an indirect effect via u affecting x , which in turn affects y . The goal of regression is to estimate only the first effect, yielding an estimate of β . The OLS estimate will instead combine these two effects, giving $\hat{\beta} > \beta$ in this example where both effects are positive. Using calculus, we have $y = \beta x + u(x)$ with total derivative

$$\frac{dy}{dx} = \beta + \frac{du}{dx}. \quad (4.44)$$

The data give information on dy/dx , so OLS estimates the total effect $\beta + du/dx$ rather than β alone. The OLS estimator is therefore biased and inconsistent for β , unless there is no association between x and u .

A more formal treatment of the linear regression model with K regressors leads to the same conclusion. From Section 4.7.1 a necessary condition for consistency of OLS is that $\text{plim } N^{-1} \mathbf{X}' \mathbf{u} = \mathbf{0}$. Consistency requires that the regressors are asymptotically uncorrelated with the errors. From (4.37) the magnitude of the inconsistency of OLS is $(\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{u}$, the OLS coefficient from regression of u on \mathbf{x} . This is just the OLS estimate of $du/d\mathbf{x}$, confirming the intuitive result in (4.44).

4.8.2. Instrumental Variable

The inconsistency of OLS is due to endogeneity of x , meaning that changes in x are associated not only with changes in y but also changes in the error u . What is needed is a method to generate only exogenous variation in x . An obvious way is through a randomized experiment, but for most economics applications such experiments are too expensive or even infeasible.

Definition of an Instrument

A crude experimental or treatment approach is still possible using observational data, provided there exists an **instrument** z that has the property that changes in z are associated with changes in x but do not lead to change in y (aside from the indirect route via x). This leads to the following path diagram:

$$\begin{array}{ccc} z & \longrightarrow & x \longrightarrow y \\ & \uparrow & \nearrow \\ & u & \end{array}$$

which introduces a variable z that is causally associated with x but not u . It is still the case that z and y will be correlated, but the only source of such correlation is the indirect path of z being correlated with x , which in turn determines y . The more direct path of z being a regressor in the model for y is ruled out.

More formally, a variable z is called an **instrument** or **instrumental variable** for the regressor x in the scalar regression model $y = \beta x + u$ if (1) z is uncorrelated with the error u and (2) z is correlated with the regressor x .

The first assumption excludes the instrument z from being a regressor in the model for y , since if instead y depended on both x and z , and y is regressed on x alone, then z is being absorbed into the error so that z will then be correlated with the error. The second assumption requires that there is some association between the instrument and the variable being instrumented.

Examples of an Instrument

In many microeconomic applications it is difficult to find legitimate instruments. Here we provide two examples.

Suppose we want to estimate the response of market demand to exogenous changes in market price. Quantity demanded clearly depends on price, but prices are not exogenously given since they are determined in part by market demand. A suitable instrument for price is a variable that is correlated with price but does not directly affect quantity demanded. An obvious candidate is a variable that affects market supply, since this also affect prices, but is not a direct determinant of demand. An example is a measure of favorable growing conditions if an agricultural product is being modeled. The choice of instrument here is uncontroversial, provided favorable growing conditions do not directly affect demand, and is helped greatly by the formal economic model of supply and demand.

Next suppose we want to estimate the returns to exogenous changes in schooling. Most observational data sets lack measures of individual ability, so regression of earnings on schooling has error that includes unobserved ability and hence is correlated with the regressor schooling. We need an instrument z that is correlated with schooling, uncorrelated with ability, and more generally uncorrelated with the error term, which means that it cannot directly determine earnings.

One popular candidate for z is proximity to a college or university (Card, 1995). This clearly satisfies condition 2 because, for example, people whose home is a long way from a community college or state university are less likely to attend college. It most likely satisfies 1, though since it can be argued that people who live a long way from a college are more likely to be in low-wage labor markets one needs to estimate a multiple regression for y that includes additional regressors such as indicators for nonmetropolitan area.

A second candidate for the instrument is month of birth (Angrist and Krueger, 1991). This clearly satisfies condition 1 as there is no reason to believe that month of birth has a direct effect on earnings if the regression includes age in years. Surprisingly condition 2 may also be satisfied, as birth month determines age of first entry

into school in the USA, which in turn may affect years of schooling since laws often specify a minimum school-leaving age. Bound, Jaeger, and Baker (1995) provide a critique of this instrument.

The consequences of choosing poor instruments are considered in detail in Section 4.9.

4.8.3. Instrumental Variables Estimator

For regression with scalar regressor x and scalar instrument z , the **instrumental variables (IV) estimator** is defined as

$$\hat{\beta}_{\text{IV}} = (\mathbf{z}'\mathbf{x})^{-1}\mathbf{z}'\mathbf{y}, \quad (4.45)$$

where, in the scalar regressor case \mathbf{z} , \mathbf{x} and \mathbf{y} are $N \times 1$ vectors. This estimator provides a consistent estimator for the slope coefficient β in the linear model $y = \beta x + u$ if z is correlated with x and uncorrelated with the error term.

There are several ways to derive (4.45). We provide an intuitive derivation, one that differs from derivations usually presented such as that in Section 6.2.5.

Return to the earnings–schooling example. Suppose a one-unit change in the instrument z is associated with 0.2 more years of schooling and with a \$500 increase in annual earnings. This increase in earnings is a consequence of the indirect effect that increase in z led to increase in schooling, which in turn increases income. Then it follows that 0.2 years additional schooling is associated with a \$500 increase in earnings, so that a one-year increase in schooling is associated with a $\$500/0.2 = \$2,500$ increase in earnings. The causal estimate of β is therefore 2,500. In mathematical notation we have estimated the changes dx/dz and dy/dz and calculated the causal estimator as

$$\beta_{\text{IV}} = \frac{dy/dz}{dx/dz}. \quad (4.46)$$

This approach to identification of the causal parameter β is given in Heckman (2000, p. 58); see also the example in Section 2.4.2.

All that remains is consistent estimation of dy/dz and dx/dz . The obvious way to estimate dy/dz is by OLS regression of y on z with slope estimate $(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{y}$. Similarly, estimate dx/dz by OLS regression of x on z with slope estimate $(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{x}$. Then

$$\hat{\beta}_{\text{IV}} = \frac{(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{y}}{(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{x}} = (\mathbf{z}'\mathbf{x})^{-1}\mathbf{z}'\mathbf{y}. \quad (4.47)$$

4.8.4. Wald Estimator

A leading simple example of IV is one where the instrument z is a **binary instrument**. Denote the subsample averages of y and x by \bar{y}_1 and \bar{x}_1 , respectively, when $z = 1$ and by \bar{y}_0 and \bar{x}_0 , respectively, when $z = 0$. Then $\Delta y/\Delta z = (\bar{y}_1 - \bar{y}_0)$ and

4.8. INSTRUMENTAL VARIABLES

$\Delta x / \Delta z = (\bar{x}_1 - \bar{x}_0)$, and (4.46) yields

$$\hat{\beta}_{\text{Wald}} = \frac{(\bar{y}_1 - \bar{y}_0)}{(\bar{x}_1 - \bar{x}_0)}. \quad (4.48)$$

This estimator is called the **Wald estimator**, after Wald (1940), or the **grouping estimator**.

The Wald estimator can also be obtained from the formula (4.45). For the no-intercept model variables are measured in deviations from means, so $\mathbf{z}'\mathbf{y} = \sum_i (z_i - \bar{z})(y_i - \bar{y})$. For binary z this yields $\mathbf{z}'\mathbf{y} = N_1(\bar{y}_1 - \bar{y}) = N_1 N_0 (\bar{y}_1 - \bar{y}_0) / N$, where N_0 and N_1 are the number of observations for which $z = 0$ and $z = 1$. This result uses $\bar{y}_1 - \bar{y} = (N_0 \bar{y}_1 + N_1 \bar{y}_0) / N - (N_0 \bar{y}_0 + N_1 \bar{y}_1) / N = N_0 (\bar{y}_1 - \bar{y}_0) / N$. Similarly, $\mathbf{z}'\mathbf{x} = N_1 N_0 (\bar{x}_1 - \bar{x}_0) / N$. Combining these results, we have that (4.45) yields (4.48).

For the earnings–schooling example it is being assumed that we can define two groups where group membership does not directly determine earnings, though it does affect level of schooling and hence indirectly affects earnings. Then the IV estimate is the difference in average earnings across the two groups divided by the difference in average schooling across the two groups.

4.8.5. Sample Covariance and Correlation Analysis

The IV estimator can also be interpreted in terms of covariances or correlations.

For sample covariances we have directly from (4.45) that

$$\hat{\beta}_{\text{IV}} = \frac{\text{Cov}[z, y]}{\text{Cov}[z, x]}, \quad (4.49)$$

where here $\text{Cov}[\cdot]$ is being used to denote sample covariance.

For sample correlations, note that the OLS estimator for the model (4.43) can be written as $\hat{\beta}_{\text{OLS}} = r_{xy} \sqrt{\mathbf{y}'\mathbf{y}} / \sqrt{\mathbf{x}'\mathbf{x}}$, where $r_{xy} = \mathbf{x}'\mathbf{y} / \sqrt{(\mathbf{x}'\mathbf{x})(\mathbf{y}'\mathbf{y})}$ is the **sample correlation** between x and y . This leads to the interpretation of the OLS estimator as implying that a one standard deviation change in x is associated with an r_{xy} standard deviation change in y . The problem is that the correlation r_{xy} is contaminated by correlation between x and u . An alternative approach is to measure the correlation between x and y indirectly by the correlation between z and y divided by the correlation between z and x . Then

$$\hat{\beta}_{\text{IV}} = \frac{r_{zy}}{r_{zx}} \frac{\sqrt{\mathbf{y}'\mathbf{y}}}{\sqrt{\mathbf{x}'\mathbf{x}}}, \quad (4.50)$$

which can be shown to equal $\hat{\beta}_{\text{IV}}$ in (4.45).

4.8.6. IV Estimation for Multiple Regression

Now consider the multiple regression model with typical observation

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

with K regressor variables, so that \mathbf{x} and $\boldsymbol{\beta}$ are $K \times 1$ vectors.

Instruments

Assume the existence of an $r \times 1$ vector of **instruments** \mathbf{z} , with $r \geq K$, satisfying the following:

1. \mathbf{z} is uncorrelated with the error u .
2. \mathbf{z} is correlated with the regressor vector \mathbf{x} .
3. \mathbf{z} is strongly correlated, rather than weakly correlated, with the regressor vector \mathbf{x} .

The first two properties are necessary for consistency and were presented earlier in the scalar case. The third property, defined in Section 4.9.1, is a strengthening of the second to ensure good finite-sample performance of the IV estimator.

In the multiple regression case \mathbf{z} and \mathbf{x} may share some common components. Some components of \mathbf{x} , called **exogenous regressors**, may be uncorrelated with u . These components are clearly suitable instruments as they satisfy conditions 1 and 2. Other components of \mathbf{x} , called **endogenous regressors**, may be correlated with u . These components lead to inconsistency of OLS and are also clearly unsuitable instruments as they do not satisfy condition 1. Partition \mathbf{x} into $\mathbf{x} = [\mathbf{x}_1' \mathbf{x}_2']'$, where \mathbf{x}_1 contains endogenous regressors and \mathbf{x}_2 contains exogenous regressors. Then a valid instrument is $\mathbf{z} = [\mathbf{z}_1' \mathbf{z}_2']'$, where \mathbf{z}_2 can be an instrument for itself, but we need to find at least as many instruments \mathbf{z}_1 as there are endogenous variables \mathbf{x}_1 .

Identification

Identification in a simultaneous equations model was presented in Section 2.5. Here we have a single equation. The **order condition** requires that the number of instruments must at least equal the number of independent endogenous components, so that $r \geq K$. The model is said to be **just-identified** if $r = K$ and **overidentified** if $r > K$.

In many multiple regression applications there is only one endogenous regressor. For example, the earnings on schooling regression will include many other regressors such as age, geographic location, and family background. Interest lies in the coefficient on schooling, but this is an endogenous variable most likely correlated with the error because ability is unobserved. Possible candidates for the necessary single instrument for schooling have already been given in Section 4.8.2.

If an instrument fails the first condition the instrument is an **invalid instrument**. If an instrument fails the second condition the instrument is an **irrelevant instrument**, and the model may be **unidentified** if too few instruments are relevant. The third condition fails when very low correlation exists between the instrument and the endogenous variable being instrumented. The model is said to be **weakly identified** and the instrument is called a **weak instrument**.

Instrumental Variables Estimator

When the model is just-identified, so that $r = K$, the **instrumental variables estimator** is the obvious matrix generalization of (4.45)

$$\widehat{\boldsymbol{\beta}}_{\text{IV}} = (\mathbf{Z}'\mathbf{X})^{-1} \mathbf{Z}'\mathbf{y}, \quad (4.51)$$

where \mathbf{Z} is an $N \times K$ matrix with i th row \mathbf{z}'_i . Substituting the regression model $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$ for \mathbf{y} in (4.51) yields

$$\begin{aligned}\widehat{\beta}_{\text{IV}} &= (\mathbf{Z}'\mathbf{X})^{-1} \mathbf{Z}'[\mathbf{X}\beta + \mathbf{u}] \\ &= \beta + (\mathbf{Z}'\mathbf{X})^{-1} \mathbf{Z}'\mathbf{u} \\ &= \beta + (N^{-1}\mathbf{Z}'\mathbf{X})^{-1} N^{-1}\mathbf{Z}'\mathbf{u}.\end{aligned}$$

It follows immediately that the IV estimator is consistent if

$$\text{plim } N^{-1}\mathbf{Z}'\mathbf{u} = \mathbf{0}$$

and

$$\text{plim } N^{-1}\mathbf{Z}'\mathbf{X} \neq \mathbf{0}.$$

These are essentially conditions 1 and 2 that \mathbf{z} is uncorrelated with \mathbf{u} and correlated with \mathbf{x} . To ensure that the inverse of $N^{-1}\mathbf{Z}'\mathbf{X}$ exists it is assumed that $\mathbf{Z}'\mathbf{X}$ is of full rank K , a stronger assumption than the order condition that $r = K$.

With heteroskedastic errors the IV estimator is asymptotically normal with mean β and variance matrix consistently estimated by

$$\widehat{\mathbf{V}}[\widehat{\beta}_{\text{IV}}] = (\mathbf{Z}'\mathbf{X})^{-1} \mathbf{Z}'\widehat{\Omega}\mathbf{Z}(\mathbf{Z}'\mathbf{X})^{-1}, \quad (4.52)$$

where $\widehat{\Omega} = \text{Diag}[\widehat{u}_i^2]$. This result is obtained in a manner similar to that for OLS given in Section 4.4.4.

The IV estimator, although consistent, leads to a loss of efficiency that can be very large in practice. Intuitively IV will not work well if the instrument \mathbf{z} has low correlation with the regressor \mathbf{x} (see Section 4.9.3).

4.8.7. Two-Stage Least Squares

The IV estimator in (4.51) requires that the number of instruments equals the number of regressors. For overidentified models the IV estimator can be used, by discarding some of the instruments so that the model is just-identified. However, an asymptotic efficiency loss can occur when discarding these instruments.

Instead, a common procedure is to use the **two-stage least-squares (2SLS) estimator**

$$\widehat{\beta}_{\text{2SLS}} = [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1} [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}], \quad (4.53)$$

presented and motivated in Section 6.4.

The 2SLS estimator is an IV estimator. In a just-identified model it simplifies to the IV estimator given in (4.51) with instruments \mathbf{Z} . In an overidentified model the 2SLS estimator equals the IV estimator given in (4.51) if the instruments are $\widehat{\mathbf{X}}$, where $\widehat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ is the predicted value of \mathbf{x} from OLS regression of \mathbf{x} on \mathbf{z} .

The 2SLS estimator gets its name from the result that it can be obtained by two consecutive OLS regressions: OLS regression of \mathbf{x} on \mathbf{z} to get $\widehat{\mathbf{x}}$ followed by OLS of \mathbf{y} on $\widehat{\mathbf{x}}$, which gives $\widehat{\beta}_{\text{2SLS}}$. This interpretation does not necessarily generalize to nonlinear regressions; see Section 6.5.6.

The 2SLS estimator is often expressed more compactly as

$$\widehat{\beta}_{2SLS} = [\mathbf{X}' \mathbf{P}_Z \mathbf{X}]^{-1} [\mathbf{X}' \mathbf{P}_Z \mathbf{y}], \quad (4.54)$$

where

$$\mathbf{P}_Z = \mathbf{Z}(\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}'$$

is an idempotent **projection matrix** that satisfies $\mathbf{P}_Z = \mathbf{P}_Z'$, $\mathbf{P}_Z \mathbf{P}_Z' = \mathbf{P}_Z$, and $\mathbf{P}_Z \mathbf{Z} = \mathbf{Z}$. The 2SLS estimator can be shown to be asymptotically normal distributed with estimated asymptotic variance

$$\widehat{V}[\widehat{\beta}_{2SLS}] = N [\mathbf{X}' \mathbf{P}_Z \mathbf{X}]^{-1} \left[\mathbf{X}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \widehat{\mathbf{S}} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X} \right] [\mathbf{X}' \mathbf{P}_Z \mathbf{X}]^{-1}, \quad (4.55)$$

where in the usual case of heteroskedastic errors $\widehat{\mathbf{S}} = N^{-1} \sum_i \widehat{u}_i^2 \mathbf{z}_i \mathbf{z}_i'$ and $\widehat{u}_i = y_i - \mathbf{x}_i' \widehat{\beta}_{2SLS}$. A commonly used small-sample adjustment is to divide by $N - K$ rather than N in the formula for $\widehat{\mathbf{S}}$.

In the special case that errors are homoskedastic, simplification occurs and $\widehat{V}[\widehat{\beta}_{2SLS}] = s^2 [\mathbf{X}' \mathbf{P}_Z \mathbf{X}]^{-1}$. This latter result is given in many introductory treatments, but the more general formula (4.55) is preferred as the modern approach is to treat errors as potentially heteroskedastic.

For overidentified models with heteroskedastic errors an estimator that White (1982) calls the **two-stage instrumental variables estimator** is more efficient than 2SLS. Moreover, some commonly used model specification tests require estimation by this estimator rather than 2SLS. For details see Section 6.4.2.

4.8.8. IV Example

As an example of IV estimation, consider estimation of the slope coefficient of x for the dgp

$$\begin{aligned} y &= 0 + 0.5x + u, \\ x &= 0 + z + v, \end{aligned}$$

where $z \sim \mathcal{N}[2, 1]$ and (u, v) are joint normal with means 0, variances 1, and correlation 0.8.

OLS of y on x yields inconsistent estimates as x is correlated with u since by construction x is correlated with v , which in turn is correlated with u . IV estimation yields consistent estimates. The variable z is a valid instrument since by construction is uncorrelated with u but is correlated with x . Transformations of z , such as z^3 , are also valid instruments.

Various estimates and associated standard errors from a generated data sample of size 10,000 are given in Table 4.4. We focus on the slope coefficient.

The OLS estimator is inconsistent, with slope coefficient estimate of 0.902 being more than 50 standard errors from the true value of 0.5. The remaining estimates are consistent and are all within two standard errors of 0.5.

There are several ways to compute the IV estimator. The slope coefficient from OLS regression of y on z is 0.5168 and from OLS regression of x on z it is 1.0124,

Table 4.4. Instrumental Variables Example^a

	OLS	IV	2SLS	IV (z^3)
Constant	−0.804 (0.014)	−0.017 (0.022)	−0.017 (0.032)	−0.014 (0.025)
x	0.902 (0.006)	0.510 (0.010)	0.510 (0.014)	0.509 (0.012)
R^2	0.709	0.576	0.576	0.574

^a Generated data for a sample size of 10,000. OLS is inconsistent and other estimators are consistent. Robust standard errors are reported though they are unnecessary here as errors are homoskedastic. The 2SLS standard errors are incorrect. The data-generating process is given in the text.

yielding an IV estimate of $0.5168/1.0124 = 0.510$ using (4.47). In practice one instead directly computes the IV estimator using (4.45) or (4.51), with z used as the instrument for x and standard errors computed using (4.52). The 2SLS estimator (see (4.54)) can be computed by OLS regression of y on \hat{x} , where \hat{x} is the prediction from OLS regression of x on z . The 2SLS estimates exactly equal the IV estimates in this just-identified model, though the standard errors from this OLS regression of y on \hat{x} are incorrect as will be explained in Section 6.4.5.

The final column uses z^3 rather than z as the instrument for x . This alternative IV estimator is consistent, since z^3 is uncorrelated with u and correlated with x . However, it is less efficient for this particular dgp, and the standard error of the slope coefficient rises from 0.010 to 0.012.

There is an efficiency loss in IV estimation compared to OLS estimation, see (4.61) for a general result for the case of single regressor and single instrument. Here $r_{x,z}^2 = 0.510$, not given in Table 4.4, is high so the loss is not great and the standard error of the slope coefficient increases somewhat from 0.006 to 0.010. In practice the efficiency loss can be much greater than this.

4.9. Instrumental Variables in Practice

Important practical issues include determining whether IV methods are necessary and, if necessary, determining whether the instruments are valid. The relevant specification tests are presented in Section 8.4. Unfortunately, the validity of tests are limited. They require the assumption that in a just-identified model the instruments are valid and test only overidentifying restrictions.

Although IV estimators are consistent given valid instruments, as detailed in the following, IV estimators can be much less efficient than the OLS estimator and can have a finite-sample distribution that for usual finite-sample sizes differs greatly from the asymptotic distribution. These problems are greatly magnified if instruments are weakly correlated with the variables being instrumented. One way that weak instruments can arise is if there are many more instruments than needed. This is simply dealt with by dropping some of the instruments (see also Donald and Newey, 2001). A

more fundamental problem arises when even with the minimal number of instruments one or more of the instruments is weak.

This section focuses on the problem of weak instruments.

4.9.1. Weak Instruments

There is no single definition of a weak instrument. Many authors use the following signals of a **weak instrument**, presented here for progressively more complex models.

- Scalar regressor x and scalar instrument z : A weak instrument is one for which $r_{x,z}^2$ is small.
- Scalar regressor x and vector of instruments \mathbf{z} : The instruments are weak if the R^2 from regression of x on \mathbf{z} , denoted $R_{x,\mathbf{z}}^2$, is small or if the F -statistic for test of overall fit in this regression is small.
- Multiple regressors \mathbf{x} with only one endogenous: A weak instrument is one for which the partial R^2 is low or the partial F -statistic is small, where these partial statistics are defined toward the end of Section 4.9.1.
- Multiple regressors \mathbf{x} with several endogenous: There are several measures.

R^2 Measures

Consider a single equation

$$y = \beta_1 x_1 + \mathbf{x}_2' \boldsymbol{\beta}_2 + u, \quad (4.56)$$

where just one regressor x_1 is endogenous and the remaining regressors in the vector \mathbf{x}_2 are exogenous. Assume that the instrument vector \mathbf{z} includes the exogenous instruments \mathbf{x}_2 , as well as at least one other instrument.

One possible R^2 measure is the usual R^2 from regression of x_1 on \mathbf{z} . However, this could be high only because x_1 is highly correlated with \mathbf{x}_2 whereas intuitively we really need x_1 to be highly correlated with the instrument(s) other than \mathbf{x}_2 .

Bound, Jaeger, and Baker (1995) therefore proposed use of a **partial R^2** , denoted R_p^2 , that purges the effect of \mathbf{x}_2 . R_p^2 is obtained as R^2 from the regression

$$(x_1 - \tilde{x}_1) = (\mathbf{z} - \tilde{\mathbf{z}})' \boldsymbol{\gamma} + v, \quad (4.57)$$

where \tilde{x}_1 and $\tilde{\mathbf{z}}$ are the fitted values from regressions of x_1 on \mathbf{x}_2 and \mathbf{z} on \mathbf{x}_2 . In the just-identified case $\mathbf{z} - \tilde{\mathbf{z}}$ will reduce to $z_1 - \tilde{z}_1$, where z_1 is the single instrument other than \mathbf{x}_2 and \tilde{z}_1 is the fitted value from regression of z_1 on \mathbf{x}_2 .

It is not unusual for R_p^2 to be much lower than $R_{x_1,\mathbf{z}}^2$. The formula for R_p^2 simplifies to $r_{x,z}^2$ when there is only one regressor and it is endogenous. It further simplifies to $\text{Cor}[x, z]$ when there is only one instrument.

When there is more than one endogenous variable, analysis is less straightforward as a number of generalizations of R_p^2 have been proposed.

Consider a single equation with more than one endogenous variable model and focus on estimation of the coefficient of the first endogenous variable. Then in (4.56)

x_1 is endogenous and additionally some of the variables in \mathbf{x}_2 are also endogenous. Several alternative measures replace the right-hand side of (4.57) with a residual that controls for the presence of other endogenous regressors. Shea (1997) proposed a partial R^2 , say R_p^{*2} , that is computed as the squared sample correlation between $(x_1 - \tilde{x}_1)$ and $(\hat{x}_1 - \tilde{\hat{x}}_1)$. Here $(x_1 - \tilde{x}_1)$ is again the residual from regression of x_1 on \mathbf{x}_2 , whereas $(\hat{x}_1 - \tilde{\hat{x}}_1)$ is the residual from regression of \hat{x}_1 (the fitted value from regression of x_1 on \mathbf{z}) on $\hat{\mathbf{x}}_2$ (the fitted value from regression of \mathbf{x}_2 on \mathbf{z}). Poskitt and Skeels (2002) proposed an alternative partial R^2 , which, like Shea's R_p^{*2} , simplifies to R_p^2 when there is only one endogenous regressor. Hall, Rudebusch, and Wilcox (1996) instead proposed use of canonical correlations.

These measures for the coefficient for the first endogenous variable can be repeated for the other endogenous variables. Poskitt and Skeels (2002) additionally consider an R^2 measure that applies jointly to instrumentation of all the endogenous variables.

The problems of inconsistency of estimators and loss of precision are magnified as the partial R^2 measures fall, as detailed in Sections 4.9.2 and 4.9.3. See especially (4.60) and (4.62).

Partial F -Statistics

For poor finite-sample performance, considered in Section 4.9.4, it is common to use a related measure, the F -statistic for whether coefficients are zero in regression of the endogenous regressor on instruments.

For a single regressor that is endogenous we use the usual overall F -statistic, for a test of $\boldsymbol{\pi} = \mathbf{0}$ in the regression $x = \mathbf{z}'\boldsymbol{\pi} + v$ of the endogenous regressor on the instruments. This F -statistic is a function of $R_{x,\mathbf{z}}^2$.

More commonly, some exogenous regressors also appear in the model, and in model (4.56) with single endogenous regressor \mathbf{x}_1 we use the F -statistic for a test of $\boldsymbol{\pi}_1 = \mathbf{0}$ in the regression

$$x = \mathbf{z}'_1 \boldsymbol{\pi}_1 + \mathbf{x}'_2 \boldsymbol{\pi}_2 + v, \quad (4.58)$$

where \mathbf{z}_1 are the instruments other than the exogenous regressors and \mathbf{x}_2 are the exogenous regressors. This is the first-stage regression in the two-stage least-squares interpretation of IV.

This statistic is used as a signal of potential finite-sample bias in the IV estimator. In Section 4.9.4 we explain results of Staiger and Stock (1997) that suggest a value less than 10 is problematic and a value of 5 or less is a sign of extreme finite-sample bias and we consider extension to more than one endogenous regressor.

4.9.2. Inconsistency of IV Estimators

The essential condition for consistency of IV is condition 1 in Section 4.8.6, that the instrument should be uncorrelated with the error term. No test is possible in the just-identified case. In the overidentified case a test of the overidentifying assumptions is possible (see Section 6.4.3). Rejection then could be due to either instrument

endogeneity or model failure. Thus condition 1 is difficult to test directly and determining whether an instrument is exogenous is usually a subjective decision, albeit one often guided by economic theory.

It is always possible to create an exogenous instrument through **functional form restrictions**. For example, suppose there are two regressors so that $y = \beta_1 x_1 + \beta_2 x_2 + u$, with x_1 uncorrelated with u and x_2 correlated with u . Note that throughout this section all variables are assumed to be measured in departures from means, so that without loss of generality the intercept term can be omitted. Then OLS is inconsistent, as x_2 is endogenous. A seemingly good instrument for x_2 is x_1^2 , since x_1^2 is likely to be uncorrelated with u because x_1 is uncorrelated with u . However, the validity of this instrument requires the functional form restriction on the conditional mean that x_1 only enters the model linearly and not quadratically. In practice one should view a linear model as only an approximation, and obtaining instruments in such an artificial way can be easily criticized.

A better way to create a valid instrument is through alternative **exclusion restrictions** that do not rely so heavily on choice of functional form. Some practical examples have been given in Section 4.8.2.

Structural models such as the classical linear simultaneous equations model (see Sections 2.4 and 6.10.6) make such exclusion restrictions very explicit. Even then the restrictions can often be criticized for being too ad hoc, unless compelling economic theory supports the restrictions.

For panel data applications it may be reasonable to assume that only current data may belong in the equation of interest – an exclusion restriction permitting past data to be used as instruments under the assumption that errors are serially uncorrelated (see Section 22.2.4). Similarly, in models of decision making under uncertainty (see Section 6.2.7), lagged variables can be used as instruments as they are part of the information set.

There is no formal test of instrument exogeneity that does not additionally test whether the regression equation is correctly specified. Instrument exogeneity inevitably relies on *a priori* information, such as that from economic or statistical theory. The evaluation by Bound et al. (1995, pp. 446–447) of the validity of the instruments used by Angrist and Krueger (1991) provides an insightful example of the subtleties involved in determining instrument exogeneity.

It is especially important that an instrument be exogenous if an instrument is weak, because with weak instruments even very mild endogeneity of the instrument can lead to IV parameter estimates that are much more inconsistent than the already inconsistent OLS parameter estimates.

For simplicity consider linear regression with one regressor and one instrument; hence $y = \beta x + u$. Then performing some algebra, left as an exercise, yields

$$\frac{\text{plim } \widehat{\beta}_{\text{IV}} - \beta}{\text{plim } \widehat{\beta}_{\text{OLS}} - \beta} = \frac{\text{Cor}[z, u]}{\text{Cor}[x, u]} \times \frac{1}{\text{Cor}[z, x]}. \quad (4.59)$$

Thus with an invalid instrument and low correlation between the instrument and the regressor, the IV estimator can be even more inconsistent than OLS. For example, suppose the correlation between z and x is 0.1, which is not unusual for cross-section

data. Then IV becomes more inconsistent than OLS as soon as the correlation coefficient between z and u exceeds a mere 0.1 times the correlation coefficient between x and u .

Result (4.59) can be extended to the model (4.56) with one endogenous regressor and several exogenous regressors, iid errors, and instruments that include all the exogenous regressors. Then

$$\frac{\text{plim } \widehat{\beta}_{1,2SLS} - \beta_1}{\text{plim } \widehat{\beta}_{1,OLS} - \beta_1} = \frac{\text{Cor}[\widehat{x}, u]}{\text{Cor}[x, u]} \times \frac{1}{R_p^2}, \quad (4.60)$$

where R_p^2 is defined after (4.56). For extension to more than one endogenous regressor see Shea (1997).

These results, emphasized by Bound et al. (1995), have profound implications for the use of IV. If instruments are weak then even mild instrument endogeneity can lead to IV being even more inconsistent than OLS. Perhaps because the conclusion is so negative, the literature has neglected this aspect of weak instruments. A notable recent exception is Hahn and Hausman (2003a).

Most of the literature assumes that condition 1 is satisfied, so that IV is consistent, and focuses on other complications attributable to weak instruments.

4.9.3. Low Precision

Although IV estimation can lead to consistent estimation when OLS is inconsistent, it also leads to a loss in precision. Intuitively, from Section 4.8.2 the instrument z is a treatment that leads to exogenous movement in x but does so with considerable noise.

The loss in precision increases, and standard errors increase, with weaker instruments. This is easily seen in the simplest case of a single endogenous regressor and single instrument with iid errors. Then the asymptotic variance is

$$\begin{aligned} V[\widehat{\beta}_{IV}] &= \sigma^2 (\mathbf{x}' \mathbf{z})^{-1} \mathbf{z}' (\mathbf{z}' \mathbf{x})^{-1} \\ &= [\sigma^2 / \mathbf{x}' \mathbf{x}] / [(\mathbf{z}' \mathbf{x})^2 / (\mathbf{z}' \mathbf{z})(\mathbf{x}' \mathbf{x})] \\ &= V[\widehat{\beta}_{OLS}] / r_{xz}^2. \end{aligned} \quad (4.61)$$

For example, if the squared sample correlation coefficient between z and x equals 0.1, then IV standard errors are 10 times those of OLS. Moreover, the IV estimator has larger variance than the OLS estimator unless $\text{Cor}[z, x] = 1$.

Result (4.61) can be extended to the model (4.56) with one endogenous regressor and several exogenous regressors, iid errors, and instruments that include all the exogenous regressors. Then

$$\text{se}[\widehat{\beta}_{1,2SLS}] = \text{se}[\widehat{\beta}_{1,OLS}] / R_p, \quad (4.62)$$

where $\text{se}[\cdot]$ denotes asymptotic standard error and R_p^2 is defined after (4.56). For extension to more than one endogenous regressor this R_p^2 is replaced by the R_p^{*2} proposed by Shea (1997). This provided the motivation for Shea's test statistic.

The poor precision is concentrated on the coefficients for endogenous variables. For exogenous variables the standard errors for 2SLS coefficient estimates are similar to

those for OLS. Intuitively, exogenous variables are being instrumented by themselves, so they have a very strong instrument.

For the coefficients of an endogenous regressor it is a low *partial R*², rather than *R*², that leads to a loss of estimator precision. This explains why 2SLS standard errors can be much higher than OLS standard errors despite the high raw correlation between the endogenous variable and the instruments. Going the other way, 2SLS standard errors for coefficients of endogenous variables that are much larger than OLS standard errors provide a clear signal that instruments are weak.

Statistics used to detect low precision of IV caused by weak instruments are called measures of **instrument relevance**. To some extent they are unnecessary as the problem is easily detected if IV standard errors are much larger than OLS standard errors.

4.9.4. Finite-Sample Bias

This section summarizes a relatively challenging and as yet unfinished literature on “weak instruments” that focuses on the practical problem that even in “large” samples asymptotic theory can provide a poor approximation to the distribution of the IV estimator. In particular the IV estimator is biased in finite samples even if asymptotically consistent. The bias can be especially pronounced when instruments are weak.

This bias of IV, which is toward the inconsistent OLS estimator, can be remarkably large, as demonstrated in a simple Monte Carlo experiment by Nelson and Startz (1990), and by a real data application involving several hundred thousand observations but very weak instruments by Bound et al. (1995). Moreover, the standard errors can also be very biased, as also demonstrated by Nelson and Startz (1990).

The theoretical literature entails quite specialized and advanced econometric theory, as it is actually difficult to obtain the sample mean of the IV estimator. To see this, consider adapting to the IV estimator the usual proof of unbiasedness of the OLS estimator given in Section 4.4.8. For $\widehat{\beta}_{IV}$ defined in (4.51) for the just-identified case this yields

$$\begin{aligned} E[\widehat{\beta}_{IV}] &= \beta + E_{Z,X,u}[(Z'X)^{-1}Z'u] \\ &= \beta + E_{Z,X}[(Z'X)^{-1}Z' \times [E[u|Z, X]]], \end{aligned}$$

where the unconditional expectation with respect to all stochastic variables, Z , X , and u , is obtained by first taking expectation with respect to u conditional on Z and X , using the law of Iterated Expectations (see Section A.8.). An obvious sufficient condition for the IV estimator to have mean β is that $E[u|Z, X] = \mathbf{0}$. This assumption is too strong, however, because it implies $E[u|X] = \mathbf{0}$, in which case there would be no need to instrument in the first place. So there is no simple way to obtain $E[\widehat{\beta}_{IV}]$. A similar problem does not arise in establishing consistency. Then $\widehat{\beta}_{IV} = \beta + (N^{-1}Z'X)^{-1}N^{-1}Z'u$, where the term $N^{-1}Z'u$ can be considered in isolation of X and the assumption $E[u|Z] = \mathbf{0}$ leads to $\text{plim } N^{-1}Z'u = \mathbf{0}$.

Therefore we need to use alternative methods to obtain the mean of the IV estimator. Here we merely summarize key results.

Initial research made the strong assumption of joint normality of variables and homoskedastic errors. Then the IV estimator has a Wishart distribution (defined in Chapter 13). Surprisingly, the mean of the IV estimator does not even exist in the just-identified case, a signal that there may be finite-sample problems. The mean does exist if there is at least one overidentifying restriction, and the variance exists if there are at least two overidentifying restrictions. Even when the mean exists the IV estimator is biased, with bias in the direction of OLS. With more overidentifying restrictions the bias increases, eventually equaling the bias of the OLS estimator. A detailed discussion is given in Davidson and MacKinnon (1993, pp. 221–224). Approximations based on power-series expansions have also been used.

What determines the size of the finite-sample bias? For regression with a single regressor x that is endogenous and is related to the instruments \mathbf{z} by the reduced form model $x = \boldsymbol{\pi} + v$, the **concentration parameter** τ^2 is defined as $\tau^2 = \boldsymbol{\pi}'\mathbf{Z}\mathbf{Z}'\boldsymbol{\pi}/\sigma_v^2$. The bias of IV can be shown to be an increasing function of τ^2 . The quantity τ^2/K , where K is the number of instruments, is the population analogue of the F -statistic for a test of whether $\boldsymbol{\pi} = \mathbf{0}$. The statistic $F - 1$, where F is the actual F -statistic in the first-stage reduced form model, can be shown to be an approximately unbiased estimate of τ^2/K . This leads to tests for finite-sample bias being based on the F -statistic given in Section 4.9.2.

Staiger and Stock (1997) obtained results under weaker distributional assumptions. In particular, normality is no longer needed. Their approach uses weak instrument asymptotics that find the limit distribution of IV estimators for a sequence of models with τ^2/K held constant as $N \rightarrow \infty$. In a simple model $1/F$ provides an approximate estimate of the finite-sample bias of the IV estimator relative to OLS. More generally, the extent of the bias for given F varies with the number of endogenous regressors and the number of instruments. Simulations show that to ensure that the maximal bias in IV is no more than 10% that of OLS we need $F > 10$. This threshold is widely cited but falls to around 6.5, for example, if one is comfortable with bias in IV of 20% of that for OLS. So a less strict rule of thumb is $F > 5$. Shea (1997) demonstrated that low partial R^2 is also associated with finite-sample bias but there is no similar rule of thumb for use of partial R^2 as a diagnostic for finite-sample bias.

For models with more than one endogenous regressor, separate F -statistics can be computed for each endogenous regressor. For a joint statistic Stock, Wright and Yogo (2002) propose using the minimum eigenvalue of a matrix analogue of the first-stage test F -statistic. Stock and Yogo (2003) present relevant critical values for this eigenvalue as the desired degree of bias, the number of endogenous variables, and the number of overidentifying restrictions vary. These tables include the single endogenous regressor as a special case and presume at least two overidentifying restrictions, so they do not apply to just-identified models.

Finite-sample bias problems arise not only for the IV estimate but also for IV standard errors and test statistics. Stock et al. (2002) present a similar approach to Wald tests whereby a test of $\beta = \beta_0$ at a nominal level of 5% is to have actual size of, say, no more than 15%. Stock and Yogo (2003) also present detailed tables taking this size distortion approach that include just-identified models.

4.9.5. Responses to Weak Instruments

What can the practitioner do in the face of weak instruments?

As already noted one approach is to limit the number of instruments used. This can be done by dropping instruments or by combining instruments.

If finite-sample bias is a concern then alternative estimators may have better small-sample properties than 2SLS. A number of alternatives, many variants of IV, are presented in Section 6.4.4.

Despite the emphasis on finite-sample bias the other problems created by weak instruments may be of greater importance in applications. It is possible with a large enough sample for the first-stage reduced form F -statistic to be large enough that finite-sample bias is not a problem. Meanwhile, the partial R^2 may be very small, leading to fragility to even slight correlation between the model error and instrument. This is difficult to test for and to overcome.

There also can be great loss in estimator precision, as detailed in Sections 4.9.3 and 4.9.4. In such cases either larger samples are needed or alternative approaches to estimating causal marginal effects must be used. These methods are summarized in Section 2.8 and presented elsewhere in this book.

4.9.6. IV Application

Kling (2001) analyzed in detail the use of college proximity as an instrument for schooling. Here we use the same data from the NLS young men's cohort on 3,010 males aged 24 to 34 years old in 1976 as used to produce Table 1 of Kling (2001) and originally used by Card (1995). The model estimated is

$$\ln w_i = \alpha + \beta_1 s_i + \beta_2 e_i + \beta_3 e_i^2 + \mathbf{x}'_{2i} \gamma + u_i,$$

where s denotes years of schooling, e denotes years of work experience, e^2 denotes experience squared, and \mathbf{x}_2 is a vector of 26 control variables that are mainly geographic indicators and measure of parental education.

The schooling variable is considered endogenous, owing to lack of data on ability. Additionally, the two work experience variables are endogenous, since work experience is calculated as age minus years of schooling minus six, as is common in this literature, and schooling is endogenous. At least three instruments are needed.

Here exactly three instruments are used, so the model is just-identified. The first instrument is *col4*, an indicator for whether a four-year college is nearby. This instrument has already been discussed in Section 4.8.2. The other two instruments are age and age squared. These are highly correlated with experience and experience squared, yet it is believed they can be omitted from the model for log-wage since it is work experience that matters. The remaining regressor vector \mathbf{x}_2 is used as an instrument for itself.

Although age is clearly exogenous, some unobservables such as social skills may be correlated with both age and wage. Then the use of age and age squared as instruments can be questioned. This illustrates the general point that there can be disagreement on assumptions of instrument validity.

Table 4.5. Returns to Schooling: Instrumental Variables Estimates^a

	OLS	IV
Schooling (s)	0.073 (0.004)	0.132 (0.049)
R^2	0.304	0.207
Shea's partial R^2	—	0.006
First-stage F -statistic for s	—	8.07

^a Sample of 3,010 young males. Dependent variable is log hourly wage. Coefficient and standard error for schooling given; estimates for experience, experience squared, 26 control variables, and an intercept are not reported. For the three endogenous regressors – schooling (s), experience (e), and experience squared (e^2) – the three instruments are an indicator for whether a four-year college (col) is nearby, age, and age squared. The partial R^2 and first-stage F -statistic are weak instruments diagnostics explained in the test.

Results are given in Table 4.5. The OLS estimate of β_1 is 0.073, so that wages rise by 7.6% ($= 100 \times (e^{0.073} - 1)$) on average with each extra year of schooling. This estimate is an inconsistent estimate of β_1 given omitted ability. The IV estimate, or equivalently the 2SLS estimate since the model is just-identified, is 0.132. An extra year of schooling is estimated to lead to a 14.1% ($= 100 \times (e^{0.132} - 1)$) increase in wage.

The IV estimator is much less efficient than OLS. A formal test does not reject homoskedasticity and we follow Kling (2001) and use the usual standard errors, which are very close to the heteroskedastic-robust standard errors. The standard error of $\hat{\beta}_{1,OLS}$ is 0.004 whereas that for $\hat{\beta}_{1,IV}$ is 0.049, over 10 times larger. The standard errors for the other two endogenous regressors are about 4 times larger and the standard errors for the exogenous regressors are about 1.2 times larger. The R^2 falls from 0.304 to 0.207.

R^2 measures confirm that the instruments are not very relevant for schooling. A simple test is to note that the regression (4.58) of schooling on all of the instruments yields $R^2 = 0.297$, which only falls a little to $R^2 = 0.291$ if the three additional instruments are dropped. More formally, Shea's partial R^2 here equals $0.0064 = 0.08^2$, which from (4.62) predicts that the standard error of $\hat{\beta}_{1,IV}$ will be inflated by a multiple $12.5 = 1/0.08$, very close to the inflation observed here. This reduces the t -statistic on schooling from 19.64 to 2.68. In many applications such a reduction would lead to statistical insignificance. In addition, from Section 4.9.2 even slight correlation between the instrument $col4_i$ and the error term u_i will lead to inconsistency of IV.

To see whether finite-sample bias may also be a problem we run the regression (4.58) of schooling on all of the instruments. Testing the joint significance of the three additional instruments yields an F -statistic of 8.07, suggesting that the bias of IV may be 10 or 20% that of OLS. A similar regression for the other two endogenous variables yields much higher F -statistics since, for example, age is a good additional instrument

for experience. Given that there are three endogenous regressors it is actually better to use the method of Stock et al. (2002) discussed in Section 4.9.4, though here the problem is restricted to schooling since for experience and experience squared, respectively, Shea's partial R^2 equals 0.0876 and 0.0138, whereas the first-stage F -statistics are 1,772 and 1,542.

If additional instruments are available then the model becomes overidentified and standard procedure is to additionally perform a test of overidentifying restrictions (see Section 8.4.4).

4.10. Practical Considerations

The estimation procedures in this chapter are implemented in all standard econometrics packages for cross-section data, except that not all packages implement quantile regression. Most provide robust standard errors as an option rather than the default.

The most difficult estimator to apply can be the instrumental variables estimator, as in many potential applications it can be difficult to obtain instruments that are uncorrelated with the error yet reasonably correlated with the regressor or regressors being instrumented. Such instruments can be obtained through specification of a complete structural model, such as a simultaneous equations system. Current applied research emphasizes alternative approaches such as natural experiments.

4.11. Bibliographic Notes

The results in this chapter are presented in many first-year graduate texts, such as those by Davidson and MacKinnon (2004), Greene (2003), Hayashi (2000), Johnston and diNardo (1997), Mittelhammer, Judge, and Miller (2000), and Ruud (2000). We have emphasized regression with stochastic regressors, robust standard errors, quantile regression, endogeneity, and instrumental variables.

- 4.2** Manski (1991) has a nice discussion of regression in a general setting that includes discussion of the loss functions given in Section 4.2.
- 4.3** The returns to schooling example is well studied. Angrist and Krueger (1999) and Card (1999) provide recent surveys.
- 4.4** For a history of least squares see Stigler (1986). The method was introduced by Legendre in 1805. Gauss in 1810 applied least squares to the linear model with normally distributed error and proposed the elimination method for computation, and in later work he proposed the theorem now called the Gauss–Markov theorem. Galton introduced the concept of regression, meaning mean-reversion in the context of inheritance of family traits, in 1887. For an early “modern” treatment with application to pauperism and welfare availability see Yule (1897). Statistical inference based on least-squares estimates of the linear regression model was developed most notably by Fisher. The heteroskedastic-consistent estimate of the variance matrix of the OLS estimator, due to White (1980a) building on earlier work by Eicker (1963), has had a profound impact on statistical inference in microeconomics and has been extended to many settings.
- 4.6** Boscovich in 1757 proposed a least absolute deviations estimator that predates least squares; see Stigler (1986). A review of quantile regression, introduced by Koenker and

- Bassett (1978), is given in Buchinsky (1994). A more elementary exposition is given in Koenker and Hallock (2001).
- 4.7** The earliest known use of instrumental variables estimation to secure identification in a simultaneous equations setting was by Wright (1928). Another oft-cited early reference is Reiersol (1941), who used instrumental variables methods to control for measurement error in the regressors. Sargan (1958) gives a classic early treatment of IV estimation. Stock and Trebbi (2003) provide additional early references.
- 4.8** Instrumental variables estimation is presented in econometrics texts, with emphasis on algebra but not necessarily intuition. The method is widely used in econometrics because of the desirability of obtaining estimates with a causal interpretation.
- 4.9** The problem of weak instruments was drawn to the attention of applied researchers by Nelson and Startz (1990) and Bound et al. (1995). There are a number of theoretical antecedents, most notably the work of Nagar (1959). The problem has dampened enthusiasm for IV estimation, and small-sample bias owing to weak instruments is currently a very active research topic. Results often assume iid normal errors and restrict analysis to one endogenous regressor. The survey by Stock et al. (2002) provides many references with emphasis on weak instrument asymptotics. It also briefly considers extensions to nonlinear models. The survey by Hahn and Hausman (2003b) presents additional methods and results that we have not reviewed here. For recent work on bias in standard errors see Bond and Windmeijer (2002). For a careful application see C.-I. Lee (2001).

Exercises

- 4–1** Consider the linear regression model $y_i = \mathbf{x}_i'\beta + u_i$ with nonstochastic regressors \mathbf{x}_i and error u_i that has mean zero but is correlated as follows: $E[u_i u_j] = \sigma^2$ if $i = j$, $E[u_i u_j] = \rho\sigma^2$ if $|i - j| = 1$, and $E[u_i u_j] = 0$ if $|i - j| > 1$. Thus errors for immediately adjacent observations are correlated whereas errors are otherwise uncorrelated. In matrix notation we have $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$, where $\Omega = E[\mathbf{u}\mathbf{u}']$. For this model answer each of the following questions using results given in Section 4.4.
- (a) Verify that Ω is a band matrix with nonzero terms only on the diagonal and on the first off-diagonal; and give these nonzero terms.
 - (b) Obtain the asymptotic distribution of $\hat{\beta}_{OLS}$ using (4.19).
 - (c) State how to obtain a consistent estimate of $V[\hat{\beta}_{OLS}]$ that does not depend on unknown parameters.
 - (d) Is the usual OLS output estimate $s^2(\mathbf{X}'\mathbf{X})^{-1}$ a consistent estimate of $V[\hat{\beta}_{OLS}]$?
 - (e) Is White's heteroskedasticity robust estimate of $V[\hat{\beta}_{OLS}]$ consistent here?
- 4–2** Suppose we estimate the model $y_i = \mu + u_i$, where $u_i \sim \mathcal{N}[0, \sigma_i^2]$.
- (a) Show that the OLS estimator of μ simplifies to $\hat{\mu} = \bar{y}$.
 - (b) Hence directly obtain the variance of \bar{y} . Show that this equals White's heteroskedastic consistent estimate of the variance given in (4.21).
- 4–3** Suppose the dgp is $y_i = \beta_0 x_i + u_i$, $u_i = x_i \varepsilon_i$, $x_i \sim \mathcal{N}[0, 1]$, and $\varepsilon_i \sim \mathcal{N}[0, 1]$. Assume that data are independent over i and that x_i is independent of ε_i . Note that the first four central moments of $\mathcal{N}[0, \sigma^2]$ are 0, σ^2 , 0, and $3\sigma^4$.
- (a) Show that the error term u_i is conditionally heteroskedastic.
 - (b) Obtain $\text{plim } N^{-1}\mathbf{X}'\mathbf{X}$. [Hint: Obtain $E[x_i^2]$ and apply a law of large numbers.]

- (c) Obtain $\sigma_0^2 = V[u_i]$, where the expectation is with respect to all stochastic variables in the model.
- (d) Obtain $\text{plim } N^{-1} \mathbf{X}' \Omega_0 \mathbf{X} = \lim N^{-1} E[\mathbf{X}' \Omega_0 \mathbf{X}]$, where $\Omega_0 = \text{Diag}[V[u_i | x_i]]$.
- (e) Using answers to the preceding parts give the default OLS result (4.22) for the variance matrix in the limit distribution of $\sqrt{N}(\hat{\beta}_{\text{OLS}} - \beta_0)$, ignoring potential heteroskedasticity. Your ultimate answer should be numerical.
- (f) Now give the variance in the limit distribution of $\sqrt{N}(\hat{\beta}_{\text{OLS}} - \beta_0)$, taking account of any heteroskedasticity. Your ultimate answer should be numerical.
- (g) Do any differences between answers to parts (e) and (f) accord with your prior beliefs?

4–4 Consider the linear regression model with scalar regressor $y_i = \beta x_i + u_i$ with data (y_i, x_i) iid over i though the error may be conditionally heteroskedastic.

- (a) Show that $(\hat{\beta}_{\text{OLS}} - \beta) = (N^{-1} \sum_i x_i^2)^{-1} N^{-1} \sum_i x_i u_i$.
- (b) Apply Kolmogorov law of large numbers (Theorem A.8) to the averages of x_i^2 and $x_i u_i$ to show that $\hat{\beta}_{\text{OLS}} \xrightarrow{P} \beta$. State any additional assumptions made on the dgp for x_i and u_i .
- (c) Apply the Lindeberg-Levy central limit theorem (Theorem A.14) to the averages of $x_i u_i$ to show that $N^{-1} \sum_i x_i u_i / N^{-2} \sum_i E[u_i^2 x_i^2] \xrightarrow{P} \mathcal{N}[0, 1]$. State any additional assumptions made on the dgp for x_i and u_i .
- (d) Use the product limit normal rule (Theorem A.17) to show that part (c) implies $N^{-1/2} \sum_i x_i u_i \xrightarrow{D} \mathcal{N}[0, \lim N^{-1} \sum_i E[u_i^2 x_i^2]]$. State any assumptions made on the dgp for x_i and u_i .
- (e) Combine results using (2.14) and the product limit normal rule (Theorem A.17) to obtain the limit distribution of β .

4–5 Consider the linear regression model $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$.

- (a) Obtain the formula for $\hat{\beta}$ that minimizes $Q(\beta) = \mathbf{u}' \mathbf{W} \mathbf{u}$, where \mathbf{W} is of full rank.
[Hint: The chain rule for matrix differentiation for column vectors \mathbf{x} and \mathbf{z} is $\partial f(\mathbf{x}) / \partial \mathbf{x} = (\partial z' / \partial \mathbf{x}) \times (\partial f(\mathbf{z}) / \partial \mathbf{z})$, for $f(\mathbf{x}) = f(g(\mathbf{x})) = f(\mathbf{z})$ where $\mathbf{z} = g(\mathbf{x})$].
- (b) Show that this simplifies to the OLS estimator if $\mathbf{W} = \mathbf{I}$.
- (c) Show that this gives the GLS estimator if $\mathbf{W} = \Omega^{-1}$.
- (d) Show that this gives the 2SLS estimator if $\mathbf{W} = \mathbf{Z}(\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}'$.

4–6 Consider IV estimation (Section 4.8) of the model $y = \mathbf{x}'\beta + u$ using instruments \mathbf{z} in the just-identified case with \mathbf{Z} an $N \times K$ matrix of full rank.

- (a) What essential assumptions must \mathbf{z} satisfy for the IV estimator to be consistent for β ? Explain.
- (b) Show that given just identification the 2SLS estimator defined in (4.53) reduces to the IV estimator given in (4.51).
- (c) Give a real-world example of a situation where IV estimation is needed because of inconsistency of OLS, and specify suitable instruments.

4–7 (Adapted from Nelson and Startz, 1990.) Consider the three-equation model, $y = \beta x + u$; $x = \lambda u + \varepsilon$; $z = \gamma \varepsilon + v$, where the mutually independent errors u , ε , and v are iid normal with mean 0 and variances, respectively, σ_u^2 , σ_ε^2 , and σ_v^2 .

- (a) Show that $\text{plim}(\hat{\beta}_{\text{OLS}} - \beta) = \lambda \sigma_u^2 / (\lambda^2 \sigma_u^2 + \sigma_\varepsilon^2)$.
- (b) Show that $\rho_{XZ}^2 = \gamma \sigma_\varepsilon^2 / (\lambda^2 \sigma_u^2 + \sigma_\varepsilon^2) (\gamma^2 \sigma_\varepsilon^2 + \sigma_v^2)$.
- (c) Show that $\hat{\beta}_{\text{IV}} = m_{zy} / m_{zx} = \beta + m_{zu} / (\lambda m_{zu} + m_{zv})$, where, for example, $m_{zy} = \sum_i z_i y_i$.

4.11. BIBLIOGRAPHIC NOTES

- (d) Show that $\widehat{\beta}_{IV} - \beta \rightarrow 1/\lambda$ as γ (or ρ_{xz}) $\rightarrow 0$.
- (e) Show that $\widehat{\beta}_{IV} - \beta \rightarrow \infty$ as $m_{zu} \rightarrow -\gamma\sigma_e^2/\lambda$.
- (f) What do the last two results imply regarding finite-sample biases and the moments of $\widehat{\beta}_{IV} - \beta$ when the instruments are poor?

4–8 Select a 50% random subsample of the Section 4.6.4 data on log health expenditure (y) and log total expenditure (x).

- (a) Obtain OLS estimates and contrast usual and White standard errors for the slope coefficient.
- (b) Obtain median regression estimates and compare these to the OLS estimates.
- (c) Obtain quantile regression estimates for $q = 0.25$ and $q = 0.75$.
- (d) Reproduce Figure 4.2 using your answers from parts (a)–(c).

4–9 Select a 50% random subsample of the Section 4.9.6 data on earnings and education, and reproduce as much of Table 4.5 as possible and provide appropriate interpretation.

CHAPTER 5

Maximum Likelihood and Nonlinear Least-Squares Estimation

5.1. Introduction

A nonlinear estimator is one that is a nonlinear function of the dependent variable. Most estimators used in microeconomics, aside from the OLS and IV estimators in the linear regression model presented in Chapter 4, are nonlinear estimators. Nonlinearity can arise in many ways. The conditional mean may be nonlinear in parameters. The loss function may lead to a nonlinear estimator even if the conditional mean is linear in parameters. Censoring and truncation also lead to nonlinear estimators even if the original model has conditional mean that is linear in parameters.

Here we present the essential statistical inference results for nonlinear estimation. Very limited small-sample results are available for nonlinear estimators. Statistical inference is instead based on asymptotic theory that is applicable for large samples. The estimators commonly used in microeconomics are consistent and asymptotically normal.

The asymptotic theory entails two major departures from the treatment of the linear regression model given in an introductory graduate course. First, alternative methods of proof are needed since there is no direct formula for most nonlinear estimators. Second, the asymptotic distribution is generally obtained under the weakest distributional assumptions possible. This departure was introduced in Section 4.4 to permit heteroskedasticity-robust inference for the OLS estimator. Under such weaker assumptions the default standard errors reported by a simple regression program are invalid. Some care is needed, however, as these weaker assumptions can lead to inconsistency of the estimator itself, a much more fundamental problem.

As much as possible the presentation here is expository. Definitions of convergence in probability and distribution, laws of large numbers (LLN), and central limit theorems (CLT) are presented in many texts, and here these topics are relegated to Appendix A. Applied researchers rarely aim to formally prove consistency and asymptotic normality. It is not unusual, however, to encounter data applications with estimation problems sufficiently recent or complex as to demand reading recent econometric journal articles. Then familiarity with proofs of consistency and asymptotic normality

is very helpful, especially to obtain a good idea in advance of the likely form of the variance matrix of the estimator.

Section 5.2 provides an overview of key results. A more formal treatment of extremum estimators that maximize or minimize any objective function is given in Section 5.3. Estimators based on estimating equations are defined and presented in Section 5.4. Statistical inference based on robust standard errors is presented briefly in Section 5.5, with complete treatment deferred to Chapter 7. Maximum likelihood estimation and quasi-maximum likelihood estimation are presented in Sections 5.6 and 5.7. Nonlinear least-squares estimation is given in Section 5.8. Section 5.9 presents a detailed example.

The remaining leading parametric estimation procedures – generalized method of moments and nonlinear instrumental variables – are given separate treatment in Chapter 6.

5.2. Overview of Nonlinear Estimators

This section provides a summary of asymptotic properties of nonlinear estimators, given more rigorously in Section 5.3, and presents ways to interpret regression coefficients in nonlinear models. The material is essential for understanding use of the cross-section and panel data models presented in later chapters.

5.2.1. Poisson Regression Example

It is helpful to introduce a specific example of nonlinear estimation. Here we consider Poisson regression, analyzed in more detail in Chapter 20.

The Poisson distribution is appropriate for a dependent variable y that takes only nonnegative integer values $0, 1, 2, \dots$. It can be used to model the number of occurrences of an event, such as number of patent applications by a firm and number of doctor visits by an individual.

The Poisson density, or more formally the Poisson probability mass function, with rate parameter λ is

$$f(y|\lambda) = e^{-\lambda} \lambda^y / y!, \quad y = 0, 1, 2, \dots,$$

where it can be shown that $E[y] = \lambda$ and $V[y] = \lambda$.

A regression model specifies the parameter λ to vary across individuals according to a specific function of regressor vector \mathbf{x} and parameter vector β . The usual Poisson specification is

$$\lambda = \exp(\mathbf{x}'\beta),$$

which has the advantage of ensuring that the mean $\lambda > 0$. The density of the **Poisson regression model** for a single observation is therefore

$$f(y|\mathbf{x}, \beta) = e^{-\exp(\mathbf{x}'\beta)} \exp(\mathbf{x}'\beta)^y / y!. \quad (5.1)$$

Consider maximum likelihood estimation based on the sample $\{(y_i, \mathbf{x}_i), i = 1, \dots, N\}$. The **maximum likelihood (ML) estimator** maximizes the log-likelihood function (see Section 5.6). The likelihood function is the joint density, which given independent observations is the product $\prod_i f(y_i | \mathbf{x}_i, \boldsymbol{\beta})$ of the individual densities, where we have conditioned on the regressors. The log-likelihood function is then the log of a product, which equals the sum of logs, or $\sum_i \ln f(y_i | \mathbf{x}_i, \boldsymbol{\beta})$.

For the Poisson density (5.1), the log-density for the i th observation is

$$\ln f(y_i | \mathbf{x}_i, \boldsymbol{\beta}) = -\exp(\mathbf{x}'_i \boldsymbol{\beta}) + y_i \mathbf{x}'_i \boldsymbol{\beta} - \ln y_i!.$$

So the Poisson ML estimator $\hat{\boldsymbol{\beta}}$ maximizes

$$Q_N(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^N \{-\exp(\mathbf{x}'_i \boldsymbol{\beta}) + y_i \mathbf{x}'_i \boldsymbol{\beta} - \ln y_i!\}, \quad (5.2)$$

where the scale factor $1/N$ is included so that $Q_N(\boldsymbol{\beta})$ remains finite as $N \rightarrow \infty$. The Poisson ML estimator is the solution to the first-order conditions $\partial Q_N(\boldsymbol{\beta})/\partial \boldsymbol{\beta}|_{\hat{\boldsymbol{\beta}}} = \mathbf{0}$, or

$$\frac{1}{N} \sum_{i=1}^N (y_i - \exp(\mathbf{x}'_i \boldsymbol{\beta})) \mathbf{x}_i|_{\hat{\boldsymbol{\beta}}} = \mathbf{0}. \quad (5.3)$$

There is no explicit solution for $\hat{\boldsymbol{\beta}}$ in (5.3). Numerical methods to compute $\hat{\boldsymbol{\beta}}$ are given in Chapter 10. In this chapter we instead focus on the statistical properties of the resulting estimate $\hat{\boldsymbol{\beta}}$.

5.2.2. m-Estimators

More generally, we define an **m-estimator** $\hat{\boldsymbol{\theta}}$ of the $q \times 1$ parameter vector $\boldsymbol{\theta}$ as an estimator that maximizes an objective function that is a sum or average of N subfunctions

$$Q_N(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^N q(y_i, \mathbf{x}_i, \boldsymbol{\theta}), \quad (5.4)$$

where $q(\cdot)$ is a scalar function, y_i is the dependent variable, \mathbf{x}_i is a regressor vector, and the results in this section assume independence over i .

For simplicity y_i is written as a scalar, but the results extend to vector \mathbf{y}_i and so cover multivariate and panel data and systems of equations. The objective function is subscripted by N to denote that it depends on the sample data. Throughout the book q is used to denote the dimension of $\boldsymbol{\theta}$. Note that here q is additionally being used to denote the subfunction $q(\cdot)$ in (5.4).

Many econometrics estimators and models are m-estimators, corresponding to specific functional forms for $q(y, \mathbf{x}, \boldsymbol{\theta})$. Leading examples are maximum likelihood (see (5.39) later) and nonlinear least squares (NLS) (see (5.67) later). The Poisson ML estimator that maximizes (5.2) is an example of (5.4) with $\boldsymbol{\theta} = \boldsymbol{\beta}$ and $q(y, \mathbf{x}, \boldsymbol{\beta}) = -\exp(\mathbf{x}' \boldsymbol{\beta}) + y \mathbf{x}' \boldsymbol{\beta} - \ln y!$.

We focus attention on the estimator $\hat{\boldsymbol{\theta}}$ that is computed as the solution to the associated first-order conditions $\partial Q_N(\boldsymbol{\theta})/\partial \boldsymbol{\theta}|_{\hat{\boldsymbol{\theta}}} = \mathbf{0}$, or equivalently

$$\frac{1}{N} \sum_{i=1}^N \frac{\partial q(y_i, \mathbf{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \Big|_{\hat{\boldsymbol{\theta}}} = \mathbf{0}. \quad (5.5)$$

This is a system of q equations in q unknowns that generally has no explicit solution for $\hat{\theta}$.

The term m-estimator, attributed to Huber (1967), is interpreted as an abbreviation for **maximum-likelihood-like estimator**. Many econometrics authors, including Amemiya (1985, p. 105), Greene (2003, p. 461), and Wooldridge (2002, p. 344), define an m-estimator as optimizing over a sum of terms, as in (5.4). Other authors, including Serfling (1980), define an m-estimator as solutions of equations such as (5.5). Huber (1967) considered both cases and Huber (1981, p. 43) explicitly defined an m-estimator in both ways. In this book we call the former type of estimator an m-estimator and the latter an estimating equations estimator (which will be treated separately in Section 5.4).

5.2.3. Asymptotic Properties of m-Estimators

The key desirable asymptotic properties of an estimator are that it be consistent and that it have an asymptotic distribution to permit statistical inference at least in large samples.

Consistency

The first step in determining the properties of $\hat{\theta}$ is to define exactly what $\hat{\theta}$ is intended to estimate. We suppose that there is a unique value of θ , denoted θ_0 and called the **true parameter value**, that generates the data. This identification condition (see Section 2.5) requires both correct specification of the component of the dgp of interest and uniqueness of this representation. Thus for the Poisson example it may be assumed that the dgp is one with Poisson parameter $\exp(\mathbf{x}'\beta_0)$ and \mathbf{x} is such that $\mathbf{x}'\beta^{(1)} = \mathbf{x}'\beta^{(2)}$ if and only if $\beta^{(1)} = \beta^{(2)}$.

The formal notation with subscript 0 for the true parameter value is used extensively in Chapters 5 to 8. The motivation is that θ can take many different values, but interest lies in two particular values – the true value θ_0 and the estimated value $\hat{\theta}$.

The estimate $\hat{\theta}$ will never exactly equal θ_0 , even in large samples, because of the intrinsic randomness of a sample. Instead, we require $\hat{\theta}$ to be **consistent** for θ_0 (see Definition A.2 in Appendix A), meaning that $\hat{\theta}$ must **converge in probability** to θ_0 , denoted $\hat{\theta} \xrightarrow{P} \theta_0$.

Rigorously establishing consistency of m-estimators is difficult. Formal results are given in Section 5.3.2 and a useful informal condition is given in Section 5.3.7. Specializations to ML and NLS estimators are given in later sections.

Limit Normal Distribution

Given consistency, as $N \rightarrow \infty$ the estimator $\hat{\theta}$ has a distribution with all mass at θ_0 . As for OLS, we magnify or rescale $\hat{\theta}$ by multiplication by \sqrt{N} to obtain a random variable that has nondegenerate distribution as $N \rightarrow \infty$. Statistical inference is then conducted assuming N is large enough for asymptotic theory to provide a good approximation, but not so large that $\hat{\theta}$ collapses on θ_0 .

We therefore consider the behavior of $\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. For most estimators this has a finite-sample distribution that is too complicated to use for inference. Instead, asymptotic theory is used to obtain the limit of this distribution as $N \rightarrow \infty$. For most microeconometrics estimators this limit is the multivariate normal distribution. More formally $\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ **converges in distribution** to the multivariate normal, where convergence in distribution is defined in Appendix A.

Recall from Section 4.4 that the OLS estimator can be expressed as

$$\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \left(\frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \mathbf{x}'_i \right)^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^N \mathbf{x}_i u_i,$$

and the limit distribution was derived by obtaining the probability limit of the first term on the right-hand side and the limit normal distribution of the second term. The limit distribution of an m-estimator is obtained in a similar way. In Section 5.3.3 we show that for an estimator that solves (5.5) we can always write

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = - \left(\frac{1}{N} \sum_{i=1}^N \frac{\partial^2 q_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \Big|_{\boldsymbol{\theta}^+} \right)^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{\partial q_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}_0}, \quad (5.6)$$

where $q_i(\boldsymbol{\theta}) = q(y_i, \mathbf{x}_i, \boldsymbol{\theta})$, for some $\boldsymbol{\theta}^+$ between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$, provided second derivatives and the inverse exist. This result is obtained by a Taylor series expansion.

Under appropriate assumptions this yields the following **limit distribution** of an m-estimator:

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{A}_0^{-1} \mathbf{B}_0 \mathbf{A}_0^{-1}], \quad (5.7)$$

where \mathbf{A}_0^{-1} is the probability limit of the first term in the right-hand side of (5.6), and the second term is assumed to converge to the $\mathcal{N}[\mathbf{0}, \mathbf{B}_0]$ distribution. The expressions for \mathbf{A}_0 and \mathbf{B}_0 are given in Table 5.1.

Asymptotic Normality

To obtain the distribution of $\widehat{\boldsymbol{\theta}}$ from the limit distribution result (5.7), divide the left-hand side of (5.7) by \sqrt{N} and hence divide the variance by N . Then

$$\widehat{\boldsymbol{\theta}} \stackrel{a}{\sim} \mathcal{N}[\boldsymbol{\theta}_0, V[\widehat{\boldsymbol{\theta}}]], \quad (5.8)$$

where $\stackrel{a}{\sim}$ means “is **asymptotically distributed** as,” and $V[\widehat{\boldsymbol{\theta}}]$ denotes the **asymptotic variance** of $\widehat{\boldsymbol{\theta}}$ with

$$V[\widehat{\boldsymbol{\theta}}] = N^{-1} \mathbf{A}_0^{-1} \mathbf{B}_0 \mathbf{A}_0^{-1}. \quad (5.9)$$

A complete discussion of the term asymptotic distribution has already been given in Section 4.4.4, and is also given in Section A.6.4.

The result (5.9) depends on the unknown true parameter $\boldsymbol{\theta}_0$. It is implemented by computing the **estimated asymptotic variance**

$$\widehat{V}[\widehat{\boldsymbol{\theta}}] = N^{-1} \widehat{\mathbf{A}}^{-1} \widehat{\mathbf{B}} \widehat{\mathbf{A}}^{-1}, \quad (5.10)$$

where $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$ are consistent estimates of \mathbf{A}_0 and \mathbf{B}_0 .

Table 5.1. Asymptotic Properties of m -Estimators

Property ^a	Algebraic Formula
Objective function	$Q_N(\theta) = N^{-1} \sum_i q(y_i, \mathbf{x}_i, \theta)$ is maximized wrt θ
Examples	ML: $q_i = \ln f(y_i \mathbf{x}_i, \theta)$ is the log-density NLS: $q_i = -(y_i - g(\mathbf{x}_i, \theta))^2$ is minus the squared error
First-order conditions	$\partial Q_N(\theta)/\partial\theta = N^{-1} \sum_{i=1}^N \partial q(y_i, \mathbf{x}_i, \theta)/\partial\theta _{\hat{\theta}} = \mathbf{0}$.
Consistency	Is $\text{plim } Q_N(\theta)$ maximized at $\theta = \theta_0$?
Consistency (informal)	Does $E[\partial q(y_i, \mathbf{x}_i, \theta)/\partial\theta _{\theta_0}] = \mathbf{0}$?
Limit distribution	$\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{A}_0^{-1} \mathbf{B}_0 \mathbf{A}_0^{-1}]$ $\mathbf{A}_0 = \text{plim } N^{-1} \sum_{i=1}^N \partial^2 q_i(\theta)/\partial\theta\partial\theta' _{\theta_0}$ $\mathbf{B}_0 = \text{plim } N^{-1} \sum_{i=1}^N \partial q_i/\partial\theta \times \partial q_i/\partial\theta' _{\theta_0}$. $\hat{\theta} \stackrel{a}{\sim} \mathcal{N}[\theta_0, N^{-1} \hat{\mathbf{A}}^{-1} \hat{\mathbf{B}} \hat{\mathbf{A}}^{-1}]$
Asymptotic distribution	$\hat{\mathbf{A}} = N^{-1} \sum_{i=1}^N \partial^2 q_i(\theta)/\partial\theta\partial\theta' _{\hat{\theta}}$ $\hat{\mathbf{B}} = N^{-1} \sum_{i=1}^N \partial q_i/\partial\theta \times \partial q_i/\partial\theta' _{\hat{\theta}}$

^a The limit distribution variance and asymptotic variance estimate are robust sandwich forms that assume independence over i . See Section 5.5.2 for other variance estimates.

The default output for many econometrics packages instead often uses a simpler estimate $\hat{\mathbf{V}}[\hat{\theta}] = -N^{-1}\hat{\mathbf{A}}^{-1}$ that is only valid in some special cases. See Section 5.5 for further discussion, including various ways to estimate \mathbf{A}_0 and \mathbf{B}_0 and then perform hypothesis tests.

The two leading examples of m-estimators are the ML and the NLS estimators. Formal results for these estimators are given in, respectively, Propositions 5.5 and 5.6. Simpler representations of the asymptotic distributions of these estimators are given in, respectively, (5.48) and (5.77).

Poisson ML Example

Like other ML estimators, the Poisson ML estimator is consistent if the density is correctly specified. However, applying (5.25) from Section 5.3.7 to (5.3) reveals that the essential condition for consistency is actually the weaker condition that $E[y|\mathbf{x}] = \exp(\mathbf{x}'\beta_0)$, that is, correct specification of the mean. Similar robustness of the ML estimator to partial misspecification of the distribution holds for some other special cases detailed in Section 5.7.

For the Poisson ML estimator $\partial q(\beta)/\partial\beta = (y - \exp(\mathbf{x}'\beta_0))\mathbf{x}$, leading to

$$\mathbf{A}_0 = -\text{plim } N^{-1} \sum_i \exp(\mathbf{x}'_i \beta_0) \mathbf{x}_i \mathbf{x}'_i$$

and

$$\mathbf{B}_0 = \text{plim } N^{-1} \sum_i V[y_i | \mathbf{x}_i] \mathbf{x}_i \mathbf{x}'_i.$$

Then $\hat{\beta} \stackrel{a}{\sim} \mathcal{N}[\theta_0, N^{-1} \hat{\mathbf{A}}^{-1} \hat{\mathbf{B}} \hat{\mathbf{A}}^{-1}]$, where $\hat{\mathbf{A}} = -N^{-1} \sum_i \exp(\mathbf{x}'_i \hat{\beta}) \mathbf{x}_i \mathbf{x}'_i$ and $\hat{\mathbf{B}} = N^{-1} \sum_i (y_i - \exp(\mathbf{x}'_i \hat{\beta}))^2 \mathbf{x}_i \mathbf{x}'_i$.

Table 5.2. Marginal Effect: Three Different Estimates

Formula	Description
$N^{-1} \sum_i \partial E[y_i \mathbf{x}_i] / \partial \mathbf{x}_i$	Average response of all individuals
$\partial E[y \mathbf{x}] / \partial \mathbf{x} _{\bar{\mathbf{x}}}$	Response of the average individual
$\partial E[y \mathbf{x}] / \partial \mathbf{x} _{\mathbf{x}^*}$	Response of a representative individual with $\mathbf{x} = \mathbf{x}^*$

If the data are actually Poisson distributed, then $V[y|\mathbf{x}] = E[y|\mathbf{x}] = \exp(\mathbf{x}'\beta_0)$, leading to possible simplification since $\mathbf{A}_0 = -\mathbf{B}_0$ so that $\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1} = -\mathbf{A}_0^{-1}$. However, in most applications with count data $V[y|\mathbf{x}] > E[y|\mathbf{x}]$, so it is best not to impose this restriction.

5.2.4. Coefficient Interpretation in Nonlinear Regression

An important goal of estimation is often prediction, rather than testing the statistical significance of regressors.

Marginal Effects

Interest often lies in measuring **marginal effects**, the change in the conditional mean of y when regressors \mathbf{x} change by one unit.

For the linear regression model, $E[y|\mathbf{x}] = \mathbf{x}'\beta$ implies $\partial E[y|\mathbf{x}] / \partial \mathbf{x} = \beta$ so that the coefficient has a direct interpretation as the marginal effect. For nonlinear regression models, this interpretation is no longer possible. For example, if $E[y|\mathbf{x}] = \exp(\mathbf{x}'\beta)$, then $\partial E[y|\mathbf{x}] / \partial \mathbf{x} = \exp(\mathbf{x}'\beta)\beta$ is a function of both parameters and regressors, and the size of the marginal effect depends on \mathbf{x} in addition to β .

General Regression Function

For a *general regression function*

$$E[y|\mathbf{x}] = g(\mathbf{x}, \beta),$$

the marginal effect varies with the evaluation value of \mathbf{x} .

It is customary to present one of the estimates of the marginal effect given in Table 5.2. The first estimate averages the marginal effects for all individuals. The second estimate evaluates the marginal effect at $\mathbf{x} = \bar{\mathbf{x}}$. The third estimate evaluates at specific characteristics $\mathbf{x} = \mathbf{x}^*$. For example, \mathbf{x}^* may represent a person who is female with 12 years of schooling and so on. More than one representative individual might be considered.

These three measures differ in nonlinear models, whereas in the linear model they all equal β . Even the sign of the effect may be unrelated to the sign of the parameter, with $\partial E[y|\mathbf{x}] / \partial x_j$ positive for some values of \mathbf{x} and negative for other values of \mathbf{x} . Considerable care must be taken in interpreting coefficients in nonlinear models.

Computer programs and applied studies often report the second of these measures. This can be useful in getting a sense for the magnitude of the marginal effect, but policy interest usually lies in the overall effect, the first measure, or the effect on a representative individual or group, the third measure. The first measure tends to change relatively little across different choices of functional form $g(\cdot)$, whereas the other two measures can change considerably. One can also present the full distribution of the marginal effects using a histogram or nonparametric density estimate.

Single-Index Models

Direct interpretation of regression coefficients is possible for **single-index models** that specify

$$E[y|\mathbf{x}] = g(\mathbf{x}'\boldsymbol{\beta}), \quad (5.11)$$

so that the data and parameters enter the nonlinear mean function $g(\cdot)$ by way of the single index $\mathbf{x}'\boldsymbol{\beta}$. Then nonlinearity is of the mild form that the mean is a nonlinear function of a linear combination of the regressors and parameters. For single-index models the effect on the conditional mean of a change in the j th regressor using **calculus methods** is

$$\frac{\partial E[y|\mathbf{x}]}{\partial x_j} = g'(\mathbf{x}'\boldsymbol{\beta})\beta_j,$$

where $g'(z) = \partial g(z)/\partial z$. It follows that the **relative effects** of changes in regressors are given by the ratio of the coefficients since

$$\frac{\partial E[y|\mathbf{x}]/\partial x_j}{\partial E[y|\mathbf{x}]/\partial x_k} = \frac{\beta_j}{\beta_k},$$

because the common factor $g'(\mathbf{x}'\boldsymbol{\beta})$ cancels. Thus if β_j is two times β_k then a one-unit change in x_j has twice the effect as a one-unit change in x_k . If $g(\cdot)$ is additionally **monotonic** then it follows that the **signs** of the coefficients give the signs of the effects, for all possible \mathbf{x} .

Single-index models are advantageous owing to their simple interpretation. Many standard nonlinear models such as logit, probit, and Tobit are of single-index form. Moreover, some choices of function $g(\cdot)$ permit additional interpretation, notably the exponential function considered later in this section and the logistic cdf analyzed in Section 14.3.4.

Finite-Difference Method

We have emphasized the use of calculus methods. The **finite-difference method** instead computes the marginal effect by comparing the conditional mean when x_j is increased by one unit with the value before the increase. Thus

$$\frac{\Delta E[y|\mathbf{x}]}{\Delta x_j} = g(\mathbf{x} + \mathbf{e}_j, \boldsymbol{\beta}) - g(\mathbf{x}, \boldsymbol{\beta}),$$

where \mathbf{e}_j is a vector with j th entry one and other entries zero.

For the linear model finite-difference and calculus methods lead to the same estimated effects, since $\Delta E[y|\mathbf{x}]/\Delta x_j = (\mathbf{x}'\boldsymbol{\beta} + \beta_j) - \mathbf{x}'\boldsymbol{\beta} = \beta_j$. For nonlinear models, however, the two approaches give different estimates of the marginal effect, unless the change in x_j is infinitesimally small.

Often calculus methods are used for continuous regressors and finite-difference methods are used for integer-valued regressors, such as a $(0, 1)$ indicator variable.

Exponential Conditional Mean

As an example, consider coefficient interpretation for an exponential conditional mean function, so that $E[y|\mathbf{x}] = \exp(\mathbf{x}'\boldsymbol{\beta})$. Many count and duration models use the exponential form.

A little algebra yields $\partial E[y|\mathbf{x}]/\partial x_j = E[y|\mathbf{x}] \times \beta_j$. So the parameters can be interpreted as **semi-elasticities**, with a one-unit change in x_j increasing the conditional mean by the multiple β_j . For example, if $\beta_j = 0.2$ then a one-unit change in x_j is predicted to lead to a 0.2 times proportionate increase in $E[y|\mathbf{x}]$, or an increase of 20%.

If instead the finite-difference method is used, the marginal effect is computed as $\Delta E[y|\mathbf{x}]/\Delta x_j = \exp(\mathbf{x}'\boldsymbol{\beta} + \beta_j) - \exp(\mathbf{x}'\boldsymbol{\beta}) = \exp(\mathbf{x}'\boldsymbol{\beta})(e^{\beta_j} - 1)$. This differs from the calculus result, unless β_j is small so that $e^{\beta_j} \approx 1 + \beta_j$. For example, if $\beta_j = 0.2$ the increase is 22.14% rather than 20%.

5.3. Extremum Estimators

This section is intended for use in an advanced graduate course in microeconomics. It presents the key results on consistency and asymptotic normality of extremum estimators, a very general class of estimators that minimize or maximize an objective function. The presentation is very condensed. A more complete understanding requires an advanced treatment such as that in Amemiya (1985), the basis of the treatment here, or in Newey and McFadden (1994).

5.3.1. Extremum Estimators

For cross-section analysis of a single dependent variable the sample is one of N observations, $\{(y_i, \mathbf{x}_i), i = 1, \dots, N\}$, on a dependent variable y_i , and a column vector \mathbf{x}_i of regressors. In matrix notation the sample is (\mathbf{y}, \mathbf{X}) , where \mathbf{y} is an $N \times 1$ vector with i th entry y_i and \mathbf{X} is a matrix with i th row \mathbf{x}_i' , as defined more completely in Section 1.6.

Interest lies in estimating the $q \times 1$ parameter vector $\boldsymbol{\theta} = [\theta_1, \dots, \theta_q]'$. The value $\boldsymbol{\theta}_0$, termed the **true parameter value**, is the particular value of $\boldsymbol{\theta}$ in the process that generated the data, called the **data-generating process**.

We consider estimators $\hat{\boldsymbol{\theta}}$ that maximize over $\boldsymbol{\theta} \in \Theta$ the stochastic objective function $Q_N(\boldsymbol{\theta}) = Q_N(\mathbf{y}, \mathbf{X}, \boldsymbol{\theta})$, where for notational simplicity the dependence of $Q_N(\boldsymbol{\theta})$

on the data is indicated only via the subscript N . Such estimators are called **extremum estimators**, since they solve a maximization or minimization problem.

The extremum estimator may be a **global maximum**, so

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta} \in \Theta} Q_N(\boldsymbol{\theta}). \quad (5.12)$$

Usually the extremum estimator is a **local maximum**, computed as the solution to the associated first-order conditions

$$\frac{\partial Q_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \Big|_{\hat{\boldsymbol{\theta}}} = \mathbf{0}, \quad (5.13)$$

where $\partial Q_N(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ is a $q \times 1$ column vector with k th entry $\partial Q_N(\boldsymbol{\theta})/\partial \theta_k$. The local maximum is emphasized because it is the local maximum that may be asymptotic normal distributed. The local and global maxima coincide if $Q_N(\boldsymbol{\theta})$ is globally concave.

There are two leading examples of extremum estimators. For m-estimators considered in this chapter, notably ML and NLS estimators, $Q_N(\boldsymbol{\theta})$ is a sample average such as average of squared residuals. For the generalized method of moments estimator (see Section 6.3) $Q_N(\boldsymbol{\theta})$ is a quadratic form in sample averages.

For concreteness the discussion focuses on single-equation cross-section regression. But the results are quite general and apply to any estimator based on optimization that satisfies properties given in this section. In particular there is no restriction to a scalar dependent variable and several authors use the notation \mathbf{z}_i in place of (y_i, \mathbf{x}_i) . Then $Q_N(\boldsymbol{\theta})$ equals $Q_N(\mathbf{Z}, \boldsymbol{\theta})$ rather than $Q_N(\mathbf{y}, \mathbf{X}, \boldsymbol{\theta})$.

5.3.2. Formal Consistency Theorems

We first consider parameter identification, introduced in Section 2.5. Intuitively the parameter $\boldsymbol{\theta}_0$ is identified if the distribution of the data, or feature of the distribution of interest, is determined by $\boldsymbol{\theta}_0$ whereas any other value of $\boldsymbol{\theta}$ leads to a different distribution. For example, in linear regression we required $E[\mathbf{y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta}_0$ and $\mathbf{X}\boldsymbol{\beta}^{(1)} = \mathbf{X}\boldsymbol{\beta}^{(2)}$ if and only if $\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(2)}$.

An estimation procedure may not identify $\boldsymbol{\theta}_0$. For example, this is the case if the estimation procedure omits some relevant regressors. We say that an estimation method identifies $\boldsymbol{\theta}_0$ if the probability limit of the objective function, taken with respect to the dgp with parameter $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, is maximized uniquely at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. This identification condition is an asymptotic one. Practical estimation problems that can arise in a finite sample are discussed in Chapter 10.

Consistency is established in the following manner. As $N \rightarrow \infty$ the stochastic objective function $Q_N(\boldsymbol{\theta})$, an average in the case of m-estimation, may converge in probability to a limit function, denoted $Q_0(\boldsymbol{\theta})$, that in the simplest case is nonstochastic. The corresponding maxima (global or local) of $Q_N(\boldsymbol{\theta})$ and $Q_0(\boldsymbol{\theta})$ should then occur for values of $\boldsymbol{\theta}$ close to each other. Since the maximum of $Q_N(\boldsymbol{\theta})$ is $\hat{\boldsymbol{\theta}}$ by definition, it follows that $\hat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}_0$ provided $\boldsymbol{\theta}_0$ maximizes $Q_0(\boldsymbol{\theta})$.

Clearly, consistency and identification are closely related, and Amemiya (1985, p. 230) states that a simple approach is to view identification to mean existence of a consistent estimator. For further discussion see Newey and McFadden (1994, p. 2124) and Deistler and Seifert (1978).

Key applications of this approach include Jennrich (1969) and Amemiya (1973). Amemiya (1985) and Newey and McFadden (1994) present quite general theorems. These theorems require several assumptions, including smoothness (continuity) and existence of necessary derivatives of the objective function, assumptions on the dgp to ensure convergence of $Q_N(\theta)$ to $Q_0(\theta)$, and maximization of $Q_0(\theta)$ at $\theta = \theta_0$. Different consistency theorems use slightly different assumptions.

We present two consistency theorems due to Amemiya (1985), one for a global maximum and one for a local maximum. The notation in Amemiya's theorems has been modified as Amemiya (1985) defines the objective function without the normalization $1/N$ present in, for example, (5.4).

Theorem 5.1 (Consistency of Global Maximum) (Amemiya, 1985, Theorem 4.1.1): *Make the following assumptions:*

- (i) *The parameter space Θ is a compact subset of R^q .*
- (ii) *The objective function $Q_N(\theta)$ is a measurable function of the data for all $\theta \in \Theta$, and $Q_N(\theta)$ is continuous in $\theta \in \Theta$.*
- (iii) *$Q_N(\theta)$ converges uniformly in probability to a nonstochastic function $Q_0(\theta)$, and $Q_0(\theta)$ attains a unique global maximum at θ_0 .*

Then the estimator $\hat{\theta} = \arg \max_{\theta \in \Theta} Q_N(\theta)$ is **consistent** for θ_0 , that is, $\hat{\theta} \xrightarrow{P} \theta_0$.

Uniform convergence in probability of $Q_N(\theta)$ to

$$Q_0(\theta) = \text{plim } Q_N(\theta) \quad (5.14)$$

in condition (iii) means that $\sup_{\theta \in \Theta} |Q_N(\theta) - Q_0(\theta)| \xrightarrow{P} 0$.

For a local maximum, first derivatives need to exist, but one need then only consider the behavior of $Q_N(\theta)$ and its derivative in the neighborhood of θ_0 .

Theorem 5.2 (Consistency of Local Maximum) (Amemiya, 1985, Theorem 4.1.2): *Make the following assumptions:*

- (i) *The parameter space Θ is an open subset of R^q .*
- (ii) *$Q_N(\theta)$ is a measurable function of the data for all $\theta \in \Theta$, and $\partial Q_N(\theta)/\partial \theta$ exists and is continuous in an open neighborhood of θ_0 .*
- (iii) *The objective function $Q_N(\theta)$ converges uniformly in probability to $Q_0(\theta)$ in an open neighborhood of θ_0 , and $Q_0(\theta)$ attains a unique local maximum at θ_0 .*

Then one of the solutions to $\partial Q_N(\theta)/\partial \theta = \mathbf{0}$ is **consistent** for θ_0 .

An example of use of Theorem 5.2 is given later in Section 5.3.4.

Condition (i) in Theorem 5.1 permits a global maximum to be at the boundary of the parameter space, whereas in Theorem 5.2 a local maximum has to be in the interior of the parameter space. Condition (ii) in Theorem 5.2 also implies continuity of $Q_N(\theta)$ in the open neighborhood of θ_0 , where a neighborhood $N(\theta_0)$ of θ_0 is open if and only if there exists a ball with center θ_0 entirely contained in $N(\theta_0)$. In both theorems condition (iii) is the essential condition. The maximum, global or local, of $Q_0(\theta)$ must occur at $\theta = \theta_0$. The second part of (iii) provides the identification condition that θ_0 has a meaningful interpretation and is unique.

For a local maximum, analysis is straightforward if there is only one local maximum. Then $\hat{\theta}$ is uniquely defined by $\partial Q_N(\theta)/\partial\theta|_{\hat{\theta}} = \mathbf{0}$. When there is more than one local maximum, the theorem simply says that one of the local maxima is consistent, but no guidance is given as to which one is consistent. It is best in such cases to consider the global maximum and apply Theorem 5.1. See Newey and McFadden (1994, p. 2117) for a discussion.

An important distinction is made between model specification, reflected in the choice of objective function $Q_N(\theta)$, and the actual dgp of (\mathbf{y}, \mathbf{X}) used in obtaining $Q_0(\theta)$ in (5.14). For some dgps an estimator may be consistent, whereas for other dgps an estimator may be inconsistent. In some cases, such as the Poisson ML and OLS estimators, consistency arises under a wide range of dgps provided the conditional mean is correctly specified. In other cases consistency requires stronger assumptions on the dgp such as correct specification of the density.

5.3.3. Asymptotic Normality

Results on asymptotic normality are usually restricted to the local maximum of $Q_N(\theta)$. Then $\hat{\theta}$ solves (5.13), which in general is nonlinear in $\hat{\theta}$ and has no explicit solution for $\hat{\theta}$. Instead, we replace the left-hand side of this equation by a linear function of $\hat{\theta}$, by use of a Taylor series expansion, and then solve for $\hat{\theta}$.

The most often used version of Taylor's theorem is an approximation with a remainder term. Here we instead consider an **exact first-order Taylor expansion**. For the differentiable function $f(\cdot)$ there always exists a point x^+ between x and x_0 such that

$$f(x) = f(x_0) + f'(x^+)(x - x_0),$$

where $f'(x) = \partial f(x)/\partial x$ is the derivative of $f(x)$. This result is also known as the **mean value theorem**.

Application to the current setting requires several changes. The scalar function $f(\cdot)$ is replaced by a vector function $\mathbf{f}(\cdot)$ and the scalar arguments x , x_0 , and x^+ are replaced by the vectors $\hat{\theta}$, θ_0 , and θ^+ . Then

$$\mathbf{f}(\hat{\theta}) = \mathbf{f}(\theta_0) + \left. \frac{\partial \mathbf{f}(\theta)}{\partial \theta'} \right|_{\theta^+} (\hat{\theta} - \theta_0), \quad (5.15)$$

where $\partial \mathbf{f}(\theta)/\partial \theta$ is a matrix, for some unknown θ^+ between $\hat{\theta}$ and θ_0 , and formally θ^+ differs for each row of this matrix (see Newey and McFadden, 1994, p. 2141). For the local extremum estimator the function $\mathbf{f}(\theta) = \partial Q_N(\theta)/\partial \theta$ is already a first

derivative. Then an exact first-order Taylor series expansion around θ_0 yields

$$\frac{\partial Q_N(\theta)}{\partial \theta} \Big|_{\hat{\theta}} = \frac{\partial Q_N(\theta)}{\partial \theta} \Big|_{\theta_0} + \frac{\partial^2 Q_N(\theta)}{\partial \theta \partial \theta'} \Big|_{\theta^+} (\hat{\theta} - \theta_0), \quad (5.16)$$

where $\partial^2 Q_N(\theta)/\partial \theta \partial \theta'$ is a $q \times q$ matrix with (j, k) th entry $\partial^2 Q_N(\theta)/\partial \theta_j \partial \theta_k$, and θ^+ is a point between $\hat{\theta}$ and θ_0 .

The first-order conditions set the left-hand side of (5.16) to zero. Setting the right-hand side to $\mathbf{0}$ and solving for $(\hat{\theta} - \theta_0)$ yields

$$\sqrt{N}(\hat{\theta} - \theta_0) = - \left(\frac{\partial^2 Q_N(\theta)}{\partial \theta \partial \theta'} \Big|_{\theta^+} \right)^{-1} \sqrt{N} \frac{\partial Q_N(\theta)}{\partial \theta} \Big|_{\theta_0}, \quad (5.17)$$

where we rescale by \sqrt{N} to ensure a nondegenerate limit distribution (discussed further in the following).

Result (5.17) provides a solution for $\hat{\theta}$. It is of no use for numerical computation of $\hat{\theta}$, since it depends on θ_0 and θ^+ , both of which are unknown, but it is fine for theoretical analysis. In particular, if it has been established that $\hat{\theta}$ is consistent for θ_0 then the unknown θ^+ converges in probability to θ_0 , because it lies between $\hat{\theta}$ and θ_0 and by consistency $\hat{\theta}$ converges in probability to θ_0 .

The result (5.17) expresses $\sqrt{N}(\hat{\theta} - \theta_0)$ in a form similar to that used to obtain the limit distribution of the OLS estimator (see Section 5.2.3). All we need do is assume a probability limit for the first term on the right-hand side of (5.17) and a limit normal distribution for the second term.

This leads to the following theorem, from Amemiya (1985), for an extremum estimator satisfying a local maximum. Again note that Amemiya (1985) defines the objective function without the normalization $1/N$. Also, Amemiya defines \mathbf{A}_0 and \mathbf{B}_0 in terms of limE rather than plim .

Theorem 5.3 (Limit Distribution of Local Maximum) (Amemiya, 1985, Theorem 4.1.3): *In addition to the assumptions of the preceding theorem for consistency of the local maximum make the following assumptions:*

- (i) $\partial^2 Q_N(\theta)/\partial \theta \partial \theta'$ exists and is continuous in an open convex neighborhood of θ_0 .
- (ii) $\partial^2 Q_N(\theta)/\partial \theta \partial \theta'|_{\theta^+}$ converges in probability to the finite nonsingular matrix

$$\mathbf{A}_0 = \text{plim } \partial^2 Q_N(\theta)/\partial \theta \partial \theta' \Big|_{\theta_0} \quad (5.18)$$

for any sequence θ^+ such that $\theta^+ \xrightarrow{P} \theta_0$.

- (iii) $\sqrt{N} \partial Q_N(\theta)/\partial \theta|_{\theta_0} \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{B}_0]$, where

$$\mathbf{B}_0 = \text{plim} \left[N \partial Q_N(\theta)/\partial \theta \times \partial Q_N(\theta)/\partial \theta' \Big|_{\theta_0} \right]. \quad (5.19)$$

Then the limit distribution of the extremum estimator is

$$\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{A}_0^{-1} \mathbf{B}_0 \mathbf{A}_0^{-1}], \quad (5.20)$$

where the estimator $\hat{\theta}$ is the consistent solution to $\partial Q_N(\theta)/\partial \theta = \mathbf{0}$.

The proof follows directly from the Limit Normal Product Rule (Theorem A.17) applied to (5.17). Note that the proof assumes that consistency of $\hat{\boldsymbol{\theta}}$ has already been established. The expressions for \mathbf{A}_0 and \mathbf{B}_0 given in Table 5.1 are specializations to the case $Q_N(\boldsymbol{\theta}) = N^{-1} \sum_i q_i(\boldsymbol{\theta})$ with independence over i .

The probability limits in (5.18) and (5.19) are obtained with respect to the dgp for (\mathbf{y}, \mathbf{X}) . In some applications the regressors are assumed to be nonstochastic and the expectation is with respect to \mathbf{y} only. In other cases the regressors are treated as stochastic and the expectations are then with respect to both \mathbf{y} and \mathbf{X} .

5.3.4. Poisson ML Estimator Asymptotic Properties Example

We formally prove consistency and asymptotic normality of the Poisson ML estimator, under exogenous stratified sampling with stochastic regressors so that (y_i, \mathbf{x}_i) are iid, without necessarily assuming that y_i is Poisson distributed.

The key step to prove *consistency* is to obtain $Q_0(\boldsymbol{\beta}) = \text{plim } Q_N(\boldsymbol{\beta})$ and verify that $Q_0(\boldsymbol{\beta})$ attains a maximum at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. For $Q_N(\boldsymbol{\beta})$ defined in (5.1), we have

$$\begin{aligned} Q_0(\boldsymbol{\beta}) &= \text{plim } N^{-1} \sum_i \left\{ -e^{\mathbf{x}'_i \boldsymbol{\beta}} + y_i \mathbf{x}'_i \boldsymbol{\beta} - \ln y_i! \right\} \\ &= \lim N^{-1} \sum_i \left\{ -E[e^{\mathbf{x}'_i \boldsymbol{\beta}}] + E[y_i \mathbf{x}'_i \boldsymbol{\beta}] - E[\ln y_i!] \right\} \\ &= \lim N^{-1} \sum_i \left\{ -E[e^{\mathbf{x}'_i \boldsymbol{\beta}}] + E[e^{\mathbf{x}'_i \boldsymbol{\beta}_0} \mathbf{x}'_i \boldsymbol{\beta}] - E[\ln y_i!] \right\}. \end{aligned}$$

The second equality assumes a law of large numbers can be applied to each term. Since (y_i, \mathbf{x}_i) are iid, the Markov LLN (Theorem A.8) can be applied if each of the expected values given in the second line exists and additionally the corresponding $(1 + \delta)$ th absolute moment exists for some $\delta > 0$ and the side condition given in Theorem A.8 is satisfied. For example, set $\delta = 1$ so that second moments are used. The third line requires the assumption that the dgp is such that $E[y|\mathbf{x}] = \exp(\mathbf{x}' \boldsymbol{\beta}_0)$. The first two expectations in the third line are with respect to \mathbf{x} , which is stochastic. Note that $Q_0(\boldsymbol{\beta})$ depends on both $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$. Differentiating with respect to $\boldsymbol{\beta}$, and assuming that limits, derivatives, and expectations can be interchanged, we get

$$\frac{\partial Q_0(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\lim N^{-1} \sum_i E[e^{\mathbf{x}'_i \boldsymbol{\beta}} \mathbf{x}_i] + \lim N^{-1} \sum_i E[e^{\mathbf{x}'_i \boldsymbol{\beta}_0} \mathbf{x}_i],$$

where the derivative of $E[\ln y!]$ with respect to $\boldsymbol{\beta}$ is zero since $E[\ln y!]$ will depend on $\boldsymbol{\beta}_0$, the true parameter value in the dgp, but not on $\boldsymbol{\beta}$. Clearly, $\partial Q_0(\boldsymbol{\beta})/\partial \boldsymbol{\beta} = \mathbf{0}$ at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $\partial^2 Q_0(\boldsymbol{\beta})/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}' = -\lim N^{-1} \sum_i E[\exp(\mathbf{x}'_i \boldsymbol{\beta}) \mathbf{x}_i \mathbf{x}'_i]$ is negative definite, so $Q_0(\boldsymbol{\beta})$ attains a local maximum at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and the Poisson ML estimator is consistent by Theorem 5.2. Since here $Q_N(\boldsymbol{\beta})$ is globally concave the local maximum equals the global maximum and consistency can also be established using Theorem 5.1.

For *asymptotic normality* of the Poisson ML estimator, the exact first-order Taylor series expansion of the Poisson ML estimator first-order conditions (5.3) yields

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = - \left[-N^{-1} \sum_i e^{\mathbf{x}'_i \boldsymbol{\beta}^+} \mathbf{x}_i \mathbf{x}'_i \right]^{-1} N^{-1/2} \sum_i (y_i - e^{\mathbf{x}'_i \boldsymbol{\beta}_0}) \mathbf{x}_i, \quad (5.21)$$

for some unknown β^+ between $\widehat{\beta}$ and β_0 . Making sufficient assumptions on regressors \mathbf{x} so that the Markov LLN can be applied to the first term, and using $\beta^+ \xrightarrow{P} \beta_0$ since $\widehat{\beta} \xrightarrow{P} \beta_0$, we have

$$-N^{-1} \sum_i e^{\mathbf{x}'_i \beta^+} \mathbf{x}_i \mathbf{x}'_i \xrightarrow{P} \mathbf{A}_0 = -\lim N^{-1} \sum_i E[e^{\mathbf{x}'_i \beta_0} \mathbf{x}_i \mathbf{x}'_i]. \quad (5.22)$$

For the second term in (5.21) begin by assuming scalar regressor x . Then $X = (y - \exp(x\beta_0))x$ has mean $E[X] = 0$, as $E[y|x] = \exp(x\beta_0)$ has already been assumed for consistency, and variance $V[X] = E[V[y|x]x^2]$. The Liapounov CLT (Theorem A.15) can be applied if the side condition involving a $(2 + \delta)$ th absolute moment of $y - \exp(x\beta_0)x$ is satisfied. For this example with $y \geq 0$ it is sufficient to assume that the third moment of y exists, that is, $\delta = 1$, and x is bounded. Applying the CLT gives

$$Z_N = \frac{\sum_i (y_i - e^{\beta_0 x_i})x_i}{\sqrt{\sum_i E[V[y_i|x_i]x_i^2]}} \xrightarrow{d} \mathcal{N}[0, 1],$$

so

$$N^{-1/2} \sum_i (y_i - e^{\beta_0 x_i})x_i \xrightarrow{d} \mathcal{N}\left[0, \lim N^{-1} \sum_i E[V[y_i|x_i]x_i^2]\right],$$

assuming the limit in the expression for the asymptotic variance exists. This result can be extended to the vector regressor case using the Cramer–Wold device (see Theorem A.16). Then

$$N^{-1/2} \sum_i (y_i - e^{\mathbf{x}'_i \beta_0}) \mathbf{x}_i \xrightarrow{d} \mathcal{N}\left[\mathbf{0}, \mathbf{B}_0 = \lim N^{-1} \sum_i E[V[y_i|x_i] \mathbf{x}_i \mathbf{x}'_i]\right]. \quad (5.23)$$

Thus (5.21) yields $\sqrt{N}(\widehat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{A}_0^{-1} \mathbf{B}_0 \mathbf{A}_0^{-1}]$, where \mathbf{A}_0 is defined in (5.22) and \mathbf{B}_0 is defined in (5.23).

Note that for this particular example $y|\mathbf{x}$ need not be Poisson distributed for the Poisson ML estimator to be consistent and asymptotically normal. The essential assumption for consistency of the Poisson ML estimator is that the dgp is such that $E[y|\mathbf{x}] = \exp(\mathbf{x}' \beta_0)$.

For asymptotic normality the essential assumption is that $V[y|\mathbf{x}]$ exists, though additional assumptions on existence of higher moments are needed to permit use of LLN and CLT. If in fact $V[y|\mathbf{x}] = \exp(\mathbf{x}' \beta_0)$ then $\mathbf{A}_0 = -\mathbf{B}_0$ and more simply $\sqrt{N}(\widehat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, -\mathbf{A}_0^{-1}]$. The results for this ML example extend to the LEF class of densities defined in Section 5.7.3.

5.3.5. Proofs of Consistency and Asymptotic Normality

The assumptions made in Theorems 5.1–5.3 are quite general and need not hold in every application. These assumptions need to be verified on a case-by-case basis, in a manner similar to the preceding Poisson ML estimator example. Here we sketch out details for m-estimators.

For consistency, the key step is to obtain the probability limit of $Q_N(\theta)$. This is done by application of an LLN because for an m-estimator $Q_N(\theta)$ is the average

$N^{-1} \sum_i q_i(\theta)$. Different assumptions on the dgp lead to the use of different LLNs and more substantively to different expressions for $Q_0(\theta)$.

Asymptotic normality requires assumptions on the dgp in addition to those required for consistency. Specifically, we need assumptions on the dgp to enable application of an LLN to obtain \mathbf{A}_0 and to enable application of a CLT to obtain \mathbf{B}_0 .

For an m-estimator an LLN is likely to verify condition (ii) of Theorem 5.3 as each entry in the matrix $\partial^2 Q_N(\theta)/\partial\theta\partial\theta'$ is an average since $Q_N(\theta)$ is an average. A CLT is likely to yield condition (iii) of Theorem 5.3, since $\sqrt{N} \partial Q_N(\theta)/\partial\theta|_{\theta_0}$ has mean $\mathbf{0}$ from the informal consistency condition (5.24) in Section 5.3.7 and finite variance $E[N \partial Q_N(\theta)/\partial\theta \times \partial Q_N(\theta)/\partial\theta'|_{\theta_0}]$.

The particular CLT and LLN used to obtain the limit distribution of the estimator vary with assumptions about the dgp for (y, \mathbf{x}) . In all cases the dependent variable is stochastic. However, the regressors may be fixed or stochastic, and in the latter case they may exhibit time-series dependence. These issues have already been considered for OLS in Section 4.4.7.

The common microeconomics assumption is that regressors are stochastic with independence across observations, which is reasonable for cross-section data from national surveys. For simple random sampling, the data (y_i, \mathbf{x}_i) are iid and Kolmogorov LLN and Lindeberg–Levy CLT (Theorems A.8 and A.14) can be used. Furthermore, under simple random sampling (5.18) and (5.19) then simplify to

$$\mathbf{A}_0 = E \left[\frac{\partial^2 q(y, \mathbf{x}, \theta)}{\partial\theta\partial\theta'} \Big|_{\theta_0} \right]$$

and

$$\mathbf{B}_0 = E \left[\frac{\partial q(y, \mathbf{x}, \theta)}{\partial\theta} \frac{\partial q(y, \mathbf{x}, \theta)}{\partial\theta'} \Big|_{\theta_0} \right],$$

where (y, \mathbf{x}) denotes a single observation and expectations are with respect to the joint distribution of (y, \mathbf{x}) . This simpler notation is used in several texts.

For stratified random sampling and for fixed regressors the data (y_i, \mathbf{x}_i) are iid and Markov LLN and Liapounov CLT (Theorems A.9 and A.15) need to be used. These require moment assumptions additional to those made in the iid case. In the stochastic regressors case, expectations are with respect to the joint distribution of (y, \mathbf{x}) , whereas in the fixed regressors case, such as in a controlled experiment where the level of \mathbf{x} can be set, the expectations in (5.18) and (5.19) are with respect to y only.

For time-series data the regressors are assumed to be stochastic, but they are also assumed to be dependent across observations, a necessary framework to accommodate lagged dependent variables. Hamilton (1994) focuses on this case, which is also studied extensively by White (2001a). The simplest treatments restrict the random variables (y, \mathbf{x}) to have stationary distribution. If instead the data are nonstationary with unit roots then rates of convergence may no longer be \sqrt{N} and the limit distributions may be nonnormal.

Despite these important conceptual and theoretical differences about the stochastic nature of (y, \mathbf{x}) , however, for cross-section regression the eventual limit theorem is usually of the general form given in Theorem 5.3.

5.3.6. Discussion

The form of the variance matrix in (5.20) is called the **sandwich form**, with \mathbf{B}_0 sandwiched between \mathbf{A}_0^{-1} and \mathbf{A}_0^{-1} . The sandwich form, introduced in Section 4.4.4, will be discussed in more detail in Section 5.5.2.

The asymptotic results can be extended to inconsistent estimators. Then θ_0 is replaced by the **pseudo-true value** θ^* , defined to be that value of θ that yields the local maximum of $Q_0(\theta)$. This is considered in further detail for quasi-ML estimation in Section 5.7.1. In most cases, however, the estimator is consistent and in later chapters the subscript 0 is often dropped to simplify notation.

In the preceding results the objective function $Q_N(\theta)$ is initially defined with normalization by $1/N$, the first derivative of $Q_N(\theta)$ is then normalized by \sqrt{N} , and the second derivative is not normalized, leading to a \sqrt{N} -consistent estimator. In some cases alternative normalizations may be needed, most notably time series with nonstationary trend.

The results assume that $Q_N(\theta)$ is a continuous differentiable function. This excludes some estimators such as least absolute deviations, for which $Q_N(\theta) = N^{-1} \sum_i |y_i - \mathbf{x}'_i \beta|$. One way to proceed in this case is to obtain a differentiable approximating function $Q_N^*(\theta)$ such that $Q_N^*(\theta) - Q_N(\theta) \xrightarrow{P} 0$ and apply the preceding theorem to $Q_N^*(\theta)$.

The key component to obtaining the limit distribution is linearization using a Taylor series expansion. Taylor series expansions can be a poor global approximation to a function. They work well in the statistical application here as the approximation is asymptotically a local one, since consistency implies that for large sample sizes $\hat{\theta}$ is close to the point of expansion θ_0 . More refined asymptotic theory is possible using the Edgeworth expansion (see Section 11.4.3). The bootstrap (see Chapter 11) is a method to empirically implement an Edgeworth expansion.

5.3.7. Informal Approach to Consistency of an m-Estimator

For the practitioner the limit normal result of Theorem 5.3 is much easier to prove than formal proof of consistency using Theorem 5.1 or 5.2. Here we present an informal approach to determining the nature and strength of distributional assumptions needed for an m-estimator to be consistent.

For an m-estimator that is a local maximum, the first-order conditions (5.4) imply that $\hat{\theta}$ is chosen so that the average of $\partial q_i(\theta)/\partial\theta|_{\hat{\theta}}$ equals zero. Intuitively, a necessary condition for this to yield a consistent estimator for θ_0 is that in the limit the average of $\partial q(\theta)/\partial\theta|_{\theta_0}$ goes to $\mathbf{0}$, or that

$$\text{plim } \frac{\partial Q_N(\theta)}{\partial\theta} \Big|_{\theta_0} = \lim \frac{1}{N} \sum_{i=1}^N E \left[\frac{\partial q_i(\theta)}{\partial\theta} \Big|_{\theta_0} \right] = \mathbf{0}, \quad (5.24)$$

where the first equality requires the assumption that a law of large numbers can be applied and expectation in (5.24) is taken with respect to the population dgp for (\mathbf{y}, \mathbf{X}) . The limit is used as the equality need not be exact, provided any departure from zero disappears as $N \rightarrow \infty$. For example, consistency should hold if the expectation equals

$1/N$. The condition (5.24) provides a very useful check for the practitioner. An **informal approach to consistency** is to look at the first-order conditions for the estimator $\hat{\theta}$ and determine whether in the limit these have expectation zero when evaluated at $\theta = \theta_0$.

Even less formally, if we consider the components in the sum, the **essential condition** for consistency is whether for the typical observation

$$E[\partial q(\theta)/\partial\theta|_{\theta_0}] = \mathbf{0}. \quad (5.25)$$

This condition can provide a very useful guide to the practitioner. However, it is neither a necessary nor a sufficient condition. If the expectation in (5.25) equals $1/N$ then it is still likely that the probability limit in (5.24) equals zero, so the condition (5.25) is not necessary. To see that it is not sufficient, consider y iid with mean μ_0 estimated using just one observation, say the first observation y_1 . Then $\hat{\mu}$ solves $y_1 - \mu = 0$ and (5.25) is satisfied. But clearly $y_1 \xrightarrow{P} \mu_0$ as the single observation y_1 has a variance that does not go to zero. The problem is that here the plim in (5.24) does not equal $\lim E$. Formal proof of consistency requires use of theorems such as Theorem 5.1 or 5.2.

For Poisson regression use of (5.25) reveals that the essential condition for consistency is correct specification of the conditional mean of $y|\mathbf{x}$ (see Section 5.2.3). Similarly, the OLS estimator solves $N^{-1} \sum_i \mathbf{x}_i(y_i - \mathbf{x}'_i \beta) = \mathbf{0}$, so from (5.25) consistency essentially requires that $E[\mathbf{x}(y - \mathbf{x}' \beta_0)] = \mathbf{0}$. This condition fails if $E[y|\mathbf{x}] \neq \mathbf{x}' \beta_0$, which can happen for many reasons, as given in Section 4.7. In other examples use of (5.25) can indicate that consistency will require considerably more parametric assumptions than correct specification of the conditional mean.

To link use of (5.24) to condition (iii) in Theorem 5.2, note the following:

$$\begin{aligned} \partial Q_0(\theta)/\partial\theta &= \mathbf{0} && \text{(condition (iii) in Theorem 5.2)} \\ \Rightarrow \partial(\text{plim } Q_N(\theta))/\partial\theta &= \mathbf{0} && \text{(from definition of } Q_0(\theta)) \\ \Rightarrow \partial(\lim E[Q_N(\theta)])/\partial\theta &= \mathbf{0} && \text{(as an LLN } \Rightarrow Q_0 = \text{plim } Q_N = \lim E[Q_N]) \\ \Rightarrow \lim \partial E[Q_N(\theta)]/\partial\theta &= \mathbf{0} && \text{(interchanging limits and differentiation), and} \\ \Rightarrow \lim E[\partial Q_N(\theta)/\partial\theta] &= \mathbf{0} && \text{(interchanging differentiation and expectation).} \end{aligned}$$

The last line is the informal condition (5.24). However, obtaining this result requires additional assumptions, including restriction to local maximum, application of a law of large numbers, interchangeability of limits and differentiation, and interchangeability of differentiation and expectation (i.e., integration). In the scalar case a sufficient condition for interchanging differentiation and limits is $\lim_{h \rightarrow 0} (E[Q_N(\theta + h)] - E[Q_N(\theta)]) / h = dE[Q_N(\theta)]/d\theta$ uniformly in θ .

5.4. Estimating Equations

The derivation of the limit distribution given in Section 5.3.3 can be extended from a local extremum estimator to estimators defined as being the solution of an estimating equation that sets an average to zero. Several examples are given in Chapter 6.

5.4.1. Estimating Equations Estimator

Let $\widehat{\boldsymbol{\theta}}$ be defined as the solution to the system of q **estimating equations**

$$\mathbf{h}_N(\widehat{\boldsymbol{\theta}}) = \frac{1}{N} \sum_{i=1}^N \mathbf{h}(y_i, \mathbf{x}_i, \widehat{\boldsymbol{\theta}}) = \mathbf{0}, \quad (5.26)$$

where $\mathbf{h}(\cdot)$ is a $q \times 1$ vector, and independence over i is assumed. Examples of $\mathbf{h}(\cdot)$ are given later in Section 5.4.2.

Since $\widehat{\boldsymbol{\theta}}$ is chosen so that the sample average of $\mathbf{h}(y, \mathbf{x}, \widehat{\boldsymbol{\theta}})$ equals zero, we expect that $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ if in the limit the average of $\mathbf{h}(y, \mathbf{x}, \boldsymbol{\theta}_0)$ goes to zero, that is, if $\text{plim } \mathbf{h}_N(\boldsymbol{\theta}_0) = \mathbf{0}$. If an LLN can be applied this requires that $\lim E[\mathbf{h}_N(\boldsymbol{\theta}_0)] = \mathbf{0}$, or more loosely that for the i th observation

$$E[\mathbf{h}(y_i, \mathbf{x}_i, \boldsymbol{\theta}_0)] = \mathbf{0}. \quad (5.27)$$

The easiest way to formally establish consistency is actually to derive (5.26) as the first-order conditions for an m-estimator.

Assuming consistency, the limit distribution of the **estimating equations estimator** can be obtained in the same manner as in Section 5.3.3 for the extremum estimator. Take an exact first-order Taylor series expansion of $\mathbf{h}_N(\boldsymbol{\theta})$ around $\boldsymbol{\theta}_0$, as in (5.15) with $\mathbf{f}(\boldsymbol{\theta}) = \mathbf{h}_N(\boldsymbol{\theta})$, and set the right-hand side to $\mathbf{0}$ and solve. Then

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = - \left(\frac{\partial \mathbf{h}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \Big|_{\boldsymbol{\theta}^+} \right)^{-1} \sqrt{N} \mathbf{h}_N(\boldsymbol{\theta}_0). \quad (5.28)$$

This leads to the following theorem.

Theorem 5.4 (Limit Distribution of Estimating Equations Estimator):

Assume that the estimating equations estimator that solves (5.26) is consistent for $\boldsymbol{\theta}_0$ and make the following assumptions:

- (i) $\partial \mathbf{h}_N(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}'$ exists and is continuous in an open convex neighborhood of $\boldsymbol{\theta}_0$.
- (ii) $\partial \mathbf{h}_N(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}' \Big|_{\boldsymbol{\theta}^+}$ converges in probability to the finite nonsingular matrix

$$\mathbf{A}_0 = \text{plim} \frac{\partial \mathbf{h}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \Big|_{\boldsymbol{\theta}_0} = \text{plim} \frac{1}{N} \sum_{i=1}^N \frac{\partial \mathbf{h}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \Big|_{\boldsymbol{\theta}_0}, \quad (5.29)$$

for any sequence $\boldsymbol{\theta}^+$ such that $\boldsymbol{\theta}^+ \xrightarrow{p} \boldsymbol{\theta}_0$.

- (iii) $\sqrt{N} \mathbf{h}_N(\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{B}_0]$, where

$$\mathbf{B}_0 = \text{plim} N \mathbf{h}_N(\boldsymbol{\theta}_0) \mathbf{h}_N(\boldsymbol{\theta}_0)' = \text{plim} \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \mathbf{h}_i(\boldsymbol{\theta}_0) \mathbf{h}_j(\boldsymbol{\theta}_0)'. \quad (5.30)$$

Then the limit distribution of the estimating equations estimator is

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{A}_0^{-1} \mathbf{B}_0 \mathbf{A}_0'^{-1}], \quad (5.31)$$

where, unlike for the extremum estimator, the matrix \mathbf{A}_0 may not be symmetric since it is no longer necessarily a Hessian matrix.

This theorem can be proved by adaptation of Amemiya's proof of Theorem 5.3. Note that Theorem 5.4 assumes that consistency has already been established.

Godambe (1960) showed that for analysis conditional on regressors the most efficient estimating equations estimator sets $\mathbf{h}_i(\boldsymbol{\theta}) = \partial \ln f(y_i | \mathbf{x}_i, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}$. Then (5.26) are the first-order conditions for the ML estimator.

5.4.2. Analogy Principle

The analogy principle uses population conditions to motivate estimators. The book by Manski (1988a) emphasizes the importance of the analogy principle as a unifying theme for estimation. Manski (1988a, p. xi) provides the following quote from Goldberger (1968, p. 4):

The **analogy principle** of estimation . . . proposes that population parameters be estimated by sample statistics which have the same property in the sample as the parameters do in the population.

Analogue estimators are estimators obtained by application of the analogy principle. **Population moment conditions** suggest as estimator the solution to the corresponding **sample moment condition**.

Extremum estimator examples of application of the analogy principle have been given in Section 4.2. For instance, if the goal of prediction is to minimize expected loss in the population and squared error loss is used, then the regression parameters β are estimated by minimizing the sample sum of squared errors.

Method of moments estimators are also examples. For instance, in the iid case if $E[y_i - \mu] = 0$ in the population then we use as estimator $\hat{\mu}$ that solves the corresponding sample moment conditions $N^{-1} \sum_i (y_i - \mu) = 0$, leading to $\hat{\mu} = \bar{y}$, the sample mean.

An estimating equations estimator may be motivated as an analogue estimator. If (5.27) holds in the population then estimate $\boldsymbol{\theta}$ by solving the corresponding sample moment condition (5.26).

Estimating equations estimators are extensively used in microeconomics. The relevant theory can be subsumed within that for **generalized method of moments**, presented in the next chapter, which is an extension that permits there to be more moment conditions than parameters. In applied statistics the approach is used in the context of **generalized estimating equations**.

5.5. Statistical Inference

A detailed treatment of hypothesis tests and confidence intervals is given in Chapter 7. Here we outline how to test linear restrictions, including exclusion restrictions, using the most common method, the Wald test for estimators that may be nonlinear. Asymptotic theory is used, so formal results lead to chi-square and normal distributions rather than the small sample F - and t -distributions from linear regression under normality. Moreover, there are several ways to consistently estimate the variance matrix of an

extremum estimator, leading to alternative estimates of standard errors and associated test statistics and p -values.

5.5.1. Wald Hypothesis Tests of Linear Restrictions

Consider testing h linearly independent restrictions, say H_0 against H_a , where

$$\begin{aligned} H_0 &: \mathbf{R}\theta_0 - \mathbf{r} = \mathbf{0}, \\ H_a &: \mathbf{R}\theta_0 - \mathbf{r} \neq \mathbf{0}, \end{aligned}$$

with \mathbf{R} an $h \times q$ matrix of constants and \mathbf{r} an $h \times 1$ vector of constants. For example, if $\theta = [\theta_1, \theta_2, \theta_3]$ then to test whether $\theta_{10} - \theta_{20} = 2$, $\mathbf{R} = [1, -1, 0]$ and $\mathbf{r} = -2$.

The Wald test rejects H_0 if $\widehat{\mathbf{R}\theta} - \mathbf{r}$, the sample estimate of $\mathbf{R}\theta_0 - \mathbf{r}$, is significantly different from $\mathbf{0}$. This requires knowledge of the distribution of $\widehat{\mathbf{R}\theta} - \mathbf{r}$. Suppose $\sqrt{N}(\widehat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{C}_0]$, where $\mathbf{C}_0 = \mathbf{A}_0^{-1} \mathbf{B}_0 \mathbf{A}_0^{-1}$ from (5.20). Then

$$\widehat{\theta} \xrightarrow{a} \mathcal{N}[\theta_0, N^{-1} \mathbf{C}_0],$$

so that under H_0 the linear combination

$$\widehat{\mathbf{R}\theta} - \mathbf{r} \xrightarrow{a} \mathcal{N}[\mathbf{0}, \mathbf{R}(N^{-1} \mathbf{C}_0) \mathbf{R}'],$$

where the mean is zero since $\mathbf{R}\theta_0 - \mathbf{r} = \mathbf{0}$ under H_0 .

Chi-Square Tests

It is convenient to move from the multivariate normal distribution to the chi-square distribution by taking the quadratic form. This yields the **Wald statistic**

$$W = (\widehat{\mathbf{R}\theta} - \mathbf{r})' (\mathbf{R}(N^{-1} \widehat{\mathbf{C}}) \mathbf{R}')^{-1} (\widehat{\mathbf{R}\theta} - \mathbf{r}) \xrightarrow{d} \chi^2(h) \quad (5.32)$$

under H_0 , where $\mathbf{R}(N^{-1} \mathbf{C}_0) \mathbf{R}'$ is of full rank h under the assumption of linearly independent restrictions, and $\widehat{\mathbf{C}}$ is a consistent estimator of \mathbf{C}_0 . Large values of W lead to rejection, and H_0 is rejected at level α if $W > \chi_\alpha^2(h)$ and is not rejected otherwise.

Practitioners frequently instead use the F -statistic $F = W/h$. Inference is then based on the $F(h, N - q)$ distribution in the hope that this might provide a better finite sample approximation. Note that h times the $F(h, N)$ distribution converges to the $\chi^2(h)$ distribution as $N \rightarrow \infty$.

The replacement of \mathbf{C}_0 by $\widehat{\mathbf{C}}$ in obtaining (5.32) makes no difference asymptotically, but in finite samples different $\widehat{\mathbf{C}}$ will lead to different values of W . In the case of classical linear regression this step corresponds to replacing σ^2 by s^2 . Then W/h is exactly F distributed if the errors are normally distributed (see Section 7.2.1).

Tests of a Single Coefficient

Often attention is focused on testing difference from zero of a single coefficient, say the j th coefficient. Then $\mathbf{R}\theta - \mathbf{r} = \theta_j$ and $W = \widehat{\theta}_j^2 / (N^{-1} \widehat{c}_{jj})$, where \widehat{c}_{jj} is the j th diagonal

element in $\widehat{\mathbf{C}}$. Taking the square root of \mathbf{W} yields

$$t = \frac{\widehat{\theta}_j}{\text{se}[\widehat{\theta}_j]} \xrightarrow{d} \mathcal{N}[0, 1] \quad (5.33)$$

under H_0 , where $\text{se}[\widehat{\theta}_j] = \sqrt{N^{-1}\widehat{c}_{jj}}$ is the asymptotic standard error of $\widehat{\theta}_j$. Large values of t lead to rejection, and unlike \mathbf{W} the statistic t can be used for one-sided tests.

Formally $\sqrt{N}\widehat{\theta}$ is an asymptotic z -statistic, but we use the notation t as it yields the usual “ t -statistic,” the estimate divided by its standard error. In finite samples, some statistical packages use the standard normal distribution whereas others use the t -distribution to compute critical values, p -values, and confidence intervals. Neither is exactly correct in finite samples, except in the very special case of linear regression with errors assumed to be normally distributed, in which case the t -distribution is exact. Both lead to the same results in infinitely large samples as the t -distribution then collapses to the standard normal.

5.5.2. Variance Matrix Estimation

There are many possible ways to estimate $\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0'^{-1}$, because there are many ways to consistently estimate \mathbf{A}_0 and \mathbf{B}_0 . Thus different econometrics programs should give the same coefficient estimates but, quite reasonably, can give standard errors, t -statistics, and p -values that differ in finite samples. It is up to the practitioner to determine the method used and the strength of the associated distributional assumptions on the dgp.

Sandwich Estimate of the Variance Matrix

The limit distribution of $\sqrt{N}(\widehat{\theta} - \theta_0)$ has variance matrix $\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0'^{-1}$. It follows that $\widehat{\theta}$ has asymptotic variance matrix $N^{-1}\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0'^{-1}$, where division by N arises because we are considering $\widehat{\theta}$ rather than $\sqrt{N}(\widehat{\theta} - \theta_0)$.

A **sandwich** estimate of the asymptotic variance of $\widehat{\theta}$ is any estimate of the form

$$\widehat{\mathbf{V}}[\widehat{\theta}] = N^{-1}\widehat{\mathbf{A}}^{-1}\widehat{\mathbf{B}}\widehat{\mathbf{A}}'^{-1}, \quad (5.34)$$

where $\widehat{\mathbf{A}}$ is consistent for \mathbf{A}_0 and $\widehat{\mathbf{B}}$ is consistent for \mathbf{B}_0 . This is called the *sandwich form* since $\widehat{\mathbf{B}}$ is sandwiched between $\widehat{\mathbf{A}}^{-1}$ and $\widehat{\mathbf{A}}'^{-1}$. For many estimators \mathbf{A} is a Hessian matrix so $\widehat{\mathbf{A}}^{-1}$ is symmetric, but this need not always be the case.

A **robust sandwich** estimate is a sandwich estimate where the estimate $\widehat{\mathbf{B}}$ is consistent for \mathbf{B}_0 under relatively weak assumptions. It leads to what are termed **robust standard errors**. A leading example is White’s heteroskedastic-consistent estimate of the variance matrix of the OLS estimator (see Section 4.4.5). In various specific contexts, detailed in later sections, robust sandwich estimates are called **Huber** estimates, after Huber (1967); **Eicker–White** estimates, after Eicker (1967) and White (1980a,b, 1982); and in stationary time-series applications **Newey–West** estimates, after Newey and West (1987b).

Estimation of A and B

Here we present different estimators for \mathbf{A}_0 and \mathbf{B}_0 for both the estimating equations estimator that solves $\mathbf{h}_N(\hat{\theta}) = \mathbf{0}$ and the local extremum estimator that solves $\partial Q_N(\theta)/\partial\theta|_{\hat{\theta}} = \mathbf{0}$.

Two standard estimates of \mathbf{A}_0 in (5.29) and (5.18) are the **Hessian** estimate

$$\widehat{\mathbf{A}}_H = \left. \frac{\partial \mathbf{h}_N(\theta)}{\partial \theta'} \right|_{\hat{\theta}} = \left. \frac{\partial^2 Q_N(\theta)}{\partial \theta \partial \theta'} \right|_{\hat{\theta}}, \quad (5.35)$$

where the second equality explains the use of the term Hessian, and the **expected Hessian** estimate

$$\widehat{\mathbf{A}}_{EH} = E \left[\left. \frac{\partial \mathbf{h}_N(\theta)}{\partial \theta'} \right|_{\hat{\theta}} \right] = E \left[\left. \frac{\partial^2 Q_N(\theta)}{\partial \theta \partial \theta'} \right|_{\hat{\theta}} \right]. \quad (5.36)$$

The first is analytically simpler and potentially relies on fewer distributional assumptions; the latter is more likely to be negative definite and invertible.

For \mathbf{B}_0 in (5.30) or (5.19) it is not possible to use the obvious estimate $N\mathbf{h}_N(\hat{\theta})\mathbf{h}_N(\hat{\theta})'$, since this equals zero as $\hat{\theta}$ is defined to satisfy $\mathbf{h}_N(\hat{\theta}) = \mathbf{0}$. One estimate is to make potentially strong distributional assumptions to get

$$\widehat{\mathbf{B}}_E = E \left[N \mathbf{h}_N(\theta) \mathbf{h}_N(\theta)' \right] \Big|_{\hat{\theta}} = E \left[N \left. \frac{\partial Q_N(\theta)}{\partial \theta} \frac{\partial Q_N(\theta)}{\partial \theta'} \right|_{\hat{\theta}} \right]. \quad (5.37)$$

Weaker assumptions are possible for m-estimators and estimating equations estimators with data independent over i . Then (5.30) simplifies to

$$\mathbf{B}_0 = E \left[\frac{1}{N} \sum_{i=1}^N \mathbf{h}_i(\theta) \mathbf{h}_i(\theta)' \right],$$

since independence implies that, for $i \neq j$, $E[\mathbf{h}_i \mathbf{h}_j'] = E[\mathbf{h}_i] E[\mathbf{h}_j']$, which in turn equals zero given $E[\mathbf{h}_i(\theta)] = \mathbf{0}$. This leads to the **outer product (OP)** estimate or **BHHH** estimate (after Berndt, Hall, Hall, and Hausman, 1974)

$$\widehat{\mathbf{B}}_{OP} = \frac{1}{N} \sum_{i=1}^N \mathbf{h}_i(\hat{\theta}) \mathbf{h}_i(\hat{\theta})' = \frac{1}{N} \sum_{i=1}^N \left. \frac{\partial q_i(\theta)}{\partial \theta} \right|_{\hat{\theta}} \left. \frac{\partial q_i(\theta)}{\partial \theta'} \right|_{\hat{\theta}}. \quad (5.38)$$

$\widehat{\mathbf{B}}_{OP}$ requires fewer assumptions than $\widehat{\mathbf{B}}_E$.

In practice a **degrees of freedom adjustment** is often used in estimating \mathbf{B}_0 , with division in (5.38) for $\widehat{\mathbf{B}}_{OP}$ by $(N - q)$ rather than N , and similar multiplication of $\widehat{\mathbf{B}}_E$ in (5.37) by $N/(N - q)$. There is no theoretical justification for this adjustment in nonlinear models, but in some simulation studies this adjustment leads to better finite-sample performance and it does coincide with the degrees of freedom adjustment made for OLS with homoskedastic errors. No similar adjustment is made for $\widehat{\mathbf{A}}_H$ or $\widehat{\mathbf{A}}_{EH}$.

Simplification occurs in some special cases with $\mathbf{A}_0 = -\mathbf{B}_0$. Leading examples are OLS or NLS with homoskedastic errors (see Section 5.8.3) and maximum likelihood with correctly specified distribution (see Section 5.6.4). Then either $-\widehat{\mathbf{A}}^{-1}$ or $\widehat{\mathbf{B}}^{-1}$ may be used to estimate the variance of $\sqrt{N}(\hat{\theta} - \theta_0)$. These estimates are less robust to misspecification of the dgp than those using the sandwich form. Misspecification of

the dgp, however, may additionally lead to inconsistency of $\hat{\theta}$, in which case even inference based on the robust sandwich estimate will be invalid.

For the Poisson example of Section 5.2, $\hat{\mathbf{A}}_H = \hat{\mathbf{A}}_{EH} = -N^{-1} \sum_i \exp(\mathbf{x}'_i \hat{\beta}) \mathbf{x}_i \mathbf{x}'_i$ and $\hat{\mathbf{B}}_{OP} = (N - q)^{-1} \sum_i (y_i - \exp(\mathbf{x}'_i \hat{\beta}))^2 \mathbf{x}_i \mathbf{x}'_i$. If $V[y|\mathbf{x}] = \exp(\mathbf{x}' \beta_0)$, the case if $y|\mathbf{x}$ is actually Poisson distributed, then $\hat{\mathbf{B}}_E = -[N/(N - q)] \hat{\mathbf{A}}_{EH}$ and simplification occurs.

5.6. Maximum Likelihood

The ML estimator holds special place among estimators. It is the most efficient estimator among consistent asymptotically normal estimators. It is also important pedagogically, as many methods for nonlinear regression such as m-estimation can be viewed as extensions and adaptations of results first obtained for ML estimation.

5.6.1. Likelihood Function

The Likelihood Principle

The **likelihood principle**, due to R. A. Fisher (1922), is to choose as estimator of the parameter vector θ_0 that value of θ that maximizes the likelihood of observing the actual sample. In the discrete case this likelihood is the probability obtained from the probability mass function; in the continuous case this is the density. Consider the discrete case. If one value of θ implies that the probability of the observed data occurring is .0012, whereas a second value of θ gives a higher probability of .0014, then the second value of θ is a better estimator.

The joint probability mass function or density $f(\mathbf{y}, \mathbf{X}|\theta)$ is viewed here as a function of θ given the data (\mathbf{y}, \mathbf{X}) . This is called the **likelihood function** and is denoted by $L_N(\theta|\mathbf{y}, \mathbf{X})$. Maximizing $L_N(\theta)$ is equivalent to maximizing the **log-likelihood function**

$$\mathcal{L}_N(\theta) = \ln L_N(\theta).$$

We take the natural logarithm because in application this leads to an objective function that is the sum rather than the product of N terms.

Conditional Likelihood

The likelihood function $L_N(\theta) = f(\mathbf{y}, \mathbf{X}|\theta) = f(\mathbf{y}|\mathbf{X}, \theta) f(\mathbf{X}|\theta)$ requires specification of both the conditional density of \mathbf{y} given \mathbf{X} and the marginal density of \mathbf{X} .

Instead, estimation is usually based on the **conditional likelihood function** $L_N(\theta) = f(\mathbf{y}|\mathbf{X}, \theta)$, since the goal of regression is to model the behavior of \mathbf{y} given \mathbf{X} . This is not a restriction if $f(\mathbf{y}|\mathbf{X})$ and $f(\mathbf{X})$ depend on mutually exclusive sets of parameters. When this is the case it is common terminology to drop the adjective conditional. For rare exceptions such as endogenous sampling (see Chapters 3 and 24) consistent estimation requires that estimation is based on the full joint density $f(\mathbf{y}, \mathbf{X}|\theta)$ rather than the conditional density $f(\mathbf{y}|\mathbf{X}, \theta)$.

Table 5.3. Maximum Likelihood: Commonly Used Densities

Model	Range of y	Density $f(y)$	Common Parameterization
Normal	$(-\infty, \infty)$	$[2\pi\sigma^2]^{-1/2}e^{-(y-\mu)^2/2\sigma^2}$	$\mu = \mathbf{x}'\beta, \sigma^2 = \sigma^2$
Bernoulli	0 or 1	$p^y(1-p)^{1-y}$	Logit $p = e^{\mathbf{x}'\beta}/(1 + e^{\mathbf{x}'\beta})$
Exponential	$(0, \infty)$	$\lambda e^{-\lambda y}$	$\lambda = e^{\mathbf{x}'\beta}$ or $1/\lambda = e^{\mathbf{x}'\beta}$
Poisson	$0, 1, 2, \dots$	$e^{-\lambda}\lambda^y/y!$	$\lambda = e^{\mathbf{x}'\beta}$

For cross-section data the observations (y_i, \mathbf{x}_i) are independent over i with conditional density function $f(y_i|\mathbf{x}_i, \theta)$. Then by independence the joint conditional density $f(\mathbf{y}|\mathbf{X}, \theta) = \prod_{i=1}^N f(y_i|\mathbf{x}_i, \theta)$, leading to the (conditional) log-likelihood function

$$Q_N(\theta) = N^{-1} \mathcal{L}_N(\theta) = \frac{1}{N} \sum_{i=1}^N \ln f(y_i|\mathbf{x}_i, \theta), \quad (5.39)$$

where we divide by N so that the objective function is an average.

Results extend to multivariate data, systems of equations, and panel data by replacing the scalar y_i by vector \mathbf{y}_i and letting $f(\mathbf{y}_i|\mathbf{x}_i, \theta)$ be the joint density of \mathbf{y}_i conditional on \mathbf{x}_i . See also Section 5.7.5.

Examples

Across a wide range of data types the following method is used to generate fully parametric cross-section regression models. First choose the one-parameter or two-parameter (or in some rare cases three-parameter) distribution that would be used for the dependent variable y in the iid case studied in a basic statistics course. Then parameterize the one or two underlying parameters in terms of regressors \mathbf{x} and parameters θ .

Some commonly used distributions and parameterizations are given in Table 5.3. Additional distributions are given in Appendix B, which also presents methods to draw pseudo-random variates.

For *continuous data on $(-\infty, \infty)$* , the normal is the standard distribution. The classical linear regression model sets $\mu = \mathbf{x}'\beta$ and assumes σ^2 is constant.

For *discrete binary data* taking values 0 or 1, the density is always the Bernoulli, a special case of the binomial with one trial. The usual parameterizations for the Bernoulli probability lead to the logit model, given in Table 5.3, and the probit model with $p = \Phi(\mathbf{x}'\beta)$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function. These models are analyzed in Chapter 14.

For *positive continuous data on $(0, \infty)$* , notably duration data considered in Chapters 17–19, the richer Weibull, gamma, and log-normal models are often used in addition to the exponential given in Table 5.3.

For *integer-valued count data* taking values $0, 1, 2, \dots$ (see Chapter 20) the richer negative binomial is often used in addition to the Poisson presented in Section 5.2.1. Setting $\lambda = \exp(\mathbf{x}'\beta)$ ensures a positive conditional mean.

For *incompletely observed data*, censored or truncated variants of these distributions may be used. The most common example is the censored normal, which is called the Tobit model and is presented in Section 16.3.

Standard likelihood-based models are rarely specified by making assumptions on the distribution of an error term. They are instead defined directly in terms of the distribution of the dependent variable. In the special case that $y \sim \mathcal{N}[\mathbf{x}'\boldsymbol{\beta}, \sigma^2]$ we can equivalently define $y = \mathbf{x}'\boldsymbol{\beta} + u$, where the error term $u \sim \mathcal{N}[0, \sigma^2]$. However, this relies on an additive property of the normal shared by few other distributions. For example, if y is Poisson distributed with mean $\exp(\mathbf{x}'\boldsymbol{\beta})$ we can always write $y = \exp(\mathbf{x}'\boldsymbol{\beta}) + u$, but the error u no longer has a familiar distribution.

5.6.2. Maximum Likelihood Estimator

The **maximum likelihood estimator** (MLE) is the estimator that maximizes the (conditional) log-likelihood function and is clearly an extremum estimator. Usually the MLE is the local maximum that solves the first-order conditions

$$\frac{1}{N} \frac{\partial \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{1}{N} \sum_{i=1}^N \frac{\partial \ln f(y_i | \mathbf{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}. \quad (5.40)$$

More formally this estimator is the **conditional MLE**, as it is based on the conditional density of y given \mathbf{x} , but it is common practice to use the simpler term MLE.

The gradient vector $\partial \mathcal{L}_N(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ is called the **score vector**, as it sums the first derivatives of the log density, and when evaluated at $\boldsymbol{\theta}_0$ it is called the **efficient score**.

5.6.3. Information Matrix Equality

The results of Section 5.3 simplify for the MLE, provided the density is correctly specified and is one for which the range of y does not depend on $\boldsymbol{\theta}$.

Regularity Conditions

The ML **regularity conditions** are that

$$E_f \left[\frac{\partial \ln f(y|\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right] = \int \frac{\partial \ln f(y|\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(y|\mathbf{x}, \boldsymbol{\theta}) = \mathbf{0} \quad (5.41)$$

and

$$-E_f \left[\frac{\partial^2 \ln f(y|\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right] = E_f \left[\frac{\partial \ln f(y|\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln f(y|\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right], \quad (5.42)$$

where the notation $E_f [\cdot]$ is used to make explicit that the expectation is with respect to the specified density $f(y|\mathbf{x}, \boldsymbol{\theta})$. Result (5.41) implies that the score vector has expected value zero, and (5.42) yields (5.44).

Derivation given in Section 5.6.7 requires that the range of y does not depend on $\boldsymbol{\theta}$ so that integration and differentiation can be interchanged.

Information Matrix Equality

The **information matrix** is the expectation of the **outer product of the score vector**,

$$\mathcal{I} = E \left[\frac{\partial \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right]. \quad (5.43)$$

The terminology information matrix is used as \mathcal{I} is the variance of $\partial \mathcal{L}_N(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$, since by (5.41) $\partial \mathcal{L}_N(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ has mean zero. Then large values of \mathcal{I} mean that small changes in $\boldsymbol{\theta}$ lead to large changes in the log-likelihood, which accordingly contains considerable information about $\boldsymbol{\theta}$. The quantity \mathcal{I} is more precisely called **Fisher Information**, as there are alternative information measures.

For log-likelihood function (5.39), the regularity condition (5.42) implies that

$$-E_f \left[\frac{\partial^2 \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \Big|_{\boldsymbol{\theta}_0} \right] = E_f \left[\frac{\partial \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \Big|_{\boldsymbol{\theta}_0} \right], \quad (5.44)$$

if the expectation is with respect to $f(y|\mathbf{x}, \boldsymbol{\theta}_0)$. The relationship (5.44) is called the **information matrix (IM) equality** and implies that the information matrix also equals $-E[\partial^2 \mathcal{L}_N(\boldsymbol{\theta})/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}']$. The IM equality (5.44) implies that $-\mathbf{A}_0 = \mathbf{B}_0$, where \mathbf{A}_0 and \mathbf{B}_0 are defined in (5.18) and (5.19). Theorem 5.3 then simplifies since $\mathbf{A}_0^{-1} \mathbf{B}_0 \mathbf{A}_0^{-1} = -\mathbf{A}_0^{-1} = \mathbf{B}_0^{-1}$.

The equality (5.42) is in turn a special case of the **generalized information matrix equality**

$$E_f \left[\frac{\partial \mathbf{m}(y, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right] = -E_f \left[\mathbf{m}(y, \boldsymbol{\theta}) \frac{\partial \ln f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right], \quad (5.45)$$

where $\mathbf{m}(\cdot)$ is a vector moment function with $E_f [\mathbf{m}(y, \boldsymbol{\theta})] = \mathbf{0}$ and expectations are with respect to the density $f(y|\boldsymbol{\theta})$. This result, also obtained in Section 5.6.7, is used in Chapters 7 and 8 to obtain simpler forms of some test statistics.

5.6.4. Distribution of the ML Estimator

The regularity conditions (5.41) and (5.42) lead to simplification of the general results of Section 5.3.

The essential consistency condition (5.25) is that $E[\partial \ln f(y|\mathbf{x}, \boldsymbol{\theta})/\partial \boldsymbol{\theta}|_{\boldsymbol{\theta}_0}] = \mathbf{0}$. This holds by the regularity condition (5.41), provided *the expectation is with respect to $f(y|\mathbf{x}, \boldsymbol{\theta}_0)$* . Thus *if the dgp is $f(y|\mathbf{x}, \boldsymbol{\theta}_0)$* , that is, the density has been correctly specified, the MLE is consistent for $\boldsymbol{\theta}_0$.

For the asymptotic distribution, simplification occurs since $-\mathbf{A}_0 = \mathbf{B}_0$ by the IM equality, which again assumes that the density is correctly specified.

These results can be collected into the following proposition.

Proposition 5.5 (Distribution of ML Estimator): *Make the following assumptions:*

- (i) *The dgp is the conditional density $f(y_i|\mathbf{x}_i, \boldsymbol{\theta}_0)$ used to define the likelihood function.*

(ii) The density function $f(\cdot)$ satisfies $f(y, \boldsymbol{\theta}^{(1)}) = f(y, \boldsymbol{\theta}^{(2)})$ iff $\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(2)}$

(iii) The matrix

$$\mathbf{A}_0 = \text{plim} \frac{1}{N} \left. \frac{\partial^2 \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right|_{\boldsymbol{\theta}_0} \quad (5.46)$$

exists and is finite nonsingular.

(iv) The order of differentiation and integration of the log-likelihood can be reversed.

Then the **ML estimator** $\widehat{\boldsymbol{\theta}}_{ML}$, defined to be a solution of the first-order conditions $\partial N^{-1} \mathcal{L}_N(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} = \mathbf{0}$, is consistent for $\boldsymbol{\theta}_0$, and

$$\sqrt{N} (\widehat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N} [\mathbf{0}, -\mathbf{A}_0^{-1}]. \quad (5.47)$$

Condition (i) states that the conditional density is correctly specified; conditions (i) and (ii) ensure that $\boldsymbol{\theta}_0$ is identified; condition (iii) is analogous to the assumption on $\text{plim } N^{-1} \mathbf{X}' \mathbf{X}$ in the case of OLS estimation; and condition (iv) is necessary for the regularity conditions to hold. As in the general case probability limits and expectations are with respect to the dgp for (\mathbf{y}, \mathbf{X}) , or with respect to just \mathbf{y} if regressors are assumed to be nonstochastic or analysis is conditional on \mathbf{X} .

Relaxation of condition (i) is considered in detail in Section 5.7. Most ML examples satisfy condition (iv), but it does rule out some models such as y uniformly distributed on the interval $[0, \theta]$ since in this case the range of y varies with θ . Then not only does $\mathbf{A}_0 \neq -\mathbf{B}_0$ but the global MLE converges at a rate other than \sqrt{N} and has limit distribution that is nonnormal. See, for example, Hirano and Porter (2003).

Given Proposition 5.5, the resulting *asymptotic distribution* of the MLE is often expressed as

$$\widehat{\boldsymbol{\theta}}_{ML} \xrightarrow{a} \mathcal{N} \left[\boldsymbol{\theta}, - \left(E \left[\frac{\partial^2 \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right] \right)^{-1} \right], \quad (5.48)$$

where for notational simplicity the evaluation at $\boldsymbol{\theta}_0$ is suppressed and we assume that an LLN applies so that the plim operator in the definition of \mathbf{A}_0 is replaced by $\lim E$ and then drop the limit. This notation is often used in later chapters.

The right-hand side of (5.48) is the Cramer–Rao lower bound (CRLB), which from basic statistics courses is the lower bound of the variance of unbiased estimators in small samples. For large samples, considered here, the CRLB is the lower bound for the variance matrix of consistent asymptotically normal (CAN) estimators with convergence to normality of $\sqrt{N} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ uniform in compact intervals of $\boldsymbol{\theta}_0$ (see Rao, 1973, pp. 344–351). Loosely speaking the MLE has the strong attraction of having the smallest asymptotic variance among root- N consistent estimators. This result requires the strong assumption of correct specification of the conditional density.

5.6.5. Weibull Regression Example

As an example, consider regression based on the Weibull distribution, which is used to model duration data such as length of unemployment spell (see Chapter 17).

The density for the Weibull distribution is $f(y) = \gamma\alpha y^{\alpha-1} \exp(-\gamma y^\alpha)$, where $y > 0$ and the parameters $\alpha > 0$ and $\gamma > 0$. It can be shown that $E[y] = \gamma^{-1/\alpha}\Gamma(\alpha^{-1} + 1)$, where $\Gamma(\cdot)$ is the gamma function. The standard Weibull regression model is obtained by specifying $\gamma = \exp(\mathbf{x}'\beta)$, in which case $E[y|\mathbf{x}] = \exp(-\mathbf{x}'\beta/\alpha)\Gamma(\alpha^{-1} + 1)$. Given independence over i the log-likelihood function is

$$N^{-1}\mathcal{L}_N(\theta) = N^{-1} \sum_i \{\mathbf{x}'_i\beta + \ln \alpha + (\alpha - 1) \ln y_i - \exp(\mathbf{x}'_i\beta)y_i^\alpha\}.$$

Differentiation with respect to β and α leads to the first-order conditions

$$\begin{aligned} N^{-1} \sum_i \{1 - \exp(\mathbf{x}'_i\beta)y_i^\alpha\} \mathbf{x}_i &= \mathbf{0}, \\ N^{-1} \sum_i \left\{ \frac{1}{\alpha} + \ln y_i - \exp(\mathbf{x}'_i\beta)y_i^\alpha \ln y_i \right\} &= 0. \end{aligned}$$

Unlike the Poisson example, consistency essentially requires correct specification of the distribution. To see this, consider the first-order conditions for β . The informal condition (5.25) that $E[\{1 - \exp(\mathbf{x}'\beta)y^\alpha\}\mathbf{x}] = \mathbf{0}$ requires that $E[y^\alpha|\mathbf{x}] = \exp(-\mathbf{x}'\beta)$, where the power α is not restricted to be an integer. The first-order conditions for α lead to an even more esoteric moment condition on y .

So we need to proceed on the assumption that the density is indeed Weibull with $\gamma = \exp(\mathbf{x}'\beta_0)$ and $\alpha = \alpha_0$. Theorem 5.5 can be applied as the range of y does not depend on the parameters. Then, from (5.48), the Weibull MLE is asymptotically normal with asymptotic variance

$$\mathbf{V} \begin{bmatrix} \widehat{\beta} \\ \widehat{\alpha} \end{bmatrix} = \left(-E \begin{bmatrix} \sum_i -e^{\mathbf{x}'_i\beta_0} y_i^{\alpha_0} \mathbf{x}_i \mathbf{x}'_i & \sum_i -e^{\mathbf{x}'_i\beta_0} y_i^{\alpha_0} \ln(y_i) \mathbf{x}_i \\ \sum_i -e^{\mathbf{x}'_i\beta_0} y_i^{\alpha_0} \ln(y_i) \mathbf{x}'_i & \sum_i d_i \end{bmatrix} \right)^{-1}, \quad (5.49)$$

where $d_i = -(1/\alpha_0^2) - e^{\mathbf{x}'_i\beta_0} y_i^{\alpha_0} (\ln y_i)^2$. The matrix inverse in (5.49) needs to be obtained by partitioned inversion because the off-diagonal term $\partial^2 \mathcal{L}_N(\beta, \alpha)/\partial\beta\partial\alpha$ does not have expected value zero. Simplification occurs in models with zero expected cross-derivative $E[\partial^2 \mathcal{L}_N(\beta, \alpha)/\partial\beta\partial\alpha'] = \mathbf{0}$, such as regression with normally distributed errors, in which case the information matrix is said to be **block diagonal** in β and α .

5.6.6. Variance Matrix Estimation for MLE

There are several ways to consistently estimate the variance matrix of an extremum estimator, as already noted in Section 5.5.2. For the MLE additional possibilities arise if the information matrix equality is assumed to hold. Then $\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}$, $-\mathbf{A}_0^{-1}$, and \mathbf{B}_0^{-1} are all asymptotically equivalent, as are the corresponding consistent estimates of these quantities. A detailed discussion for the MLE is given in Davidson and MacKinnon (1993, chapter 18).

The sandwich estimate $\widehat{\mathbf{A}}^{-1}\widehat{\mathbf{B}}\widehat{\mathbf{A}}^{-1}$ is called the Huber estimate, after Huber (1967), or White estimate, after White (1982), who considered the distribution of the MLE without imposing the information matrix equality. The sandwich estimate is in theory more robust than $-\widehat{\mathbf{A}}^{-1}$ or $\widehat{\mathbf{B}}^{-1}$. It is important to note, however, that the cause of failure of the information matrix equality may additionally lead to the more fundamental complication of inconsistency of $\widehat{\theta}_{ML}$. This is the subject of Section 5.7.

5.6.7. Derivation of ML Regularity Conditions

We now formally derive the regularity conditions stated in Section 5.6.3. For notational simplicity the subscript i and the regressor vector are suppressed.

Begin by deriving the first condition (5.41). The density integrates to one, that is,

$$\int f(y|\theta)dy = 1.$$

Differentiating both sides with respect to θ yields $\frac{\partial}{\partial \theta} \int f(y|\theta)dy = \mathbf{0}$. If the range of integration (the range of y) does not depend on θ this implies

$$\int \frac{\partial f(y|\theta)}{\partial \theta} dy = \mathbf{0}. \quad (5.50)$$

Now $\partial \ln f(y|\theta)/\partial \theta = [\partial f(y|\theta)/\partial \theta]/[f(y|\theta)]$, which implies

$$\frac{\partial f(y|\theta)}{\partial \theta} = \frac{\partial \ln f(y|\theta)}{\partial \theta} f(y|\theta). \quad (5.51)$$

Substituting (5.51) in (5.50) yields

$$\int \frac{\partial \ln f(y|\theta)}{\partial \theta} f(y|\theta) dy = \mathbf{0}, \quad (5.52)$$

which is (5.41) provided the expectation is with respect to the density $f(y|\theta)$.

Now consider the second condition (5.42), initially deriving a more general result. Suppose

$$E[\mathbf{m}(y, \theta)] = \mathbf{0},$$

for some (possibly vector) function $\mathbf{m}(\cdot)$. Then when the expectation is taken with respect to the density $f(y|\theta)$

$$\int \mathbf{m}(y, \theta) f(y|\theta) dy = \mathbf{0}. \quad (5.53)$$

Differentiating both sides with respect to θ' and assuming differentiation and integration are interchangeable yields

$$\int \left(\frac{\partial \mathbf{m}(y, \theta)}{\partial \theta'} f(y|\theta) + \mathbf{m}(y, \theta) \frac{\partial f(y|\theta)}{\partial \theta'} \right) dy = \mathbf{0}. \quad (5.54)$$

Substituting (5.51) in (5.54) yields

$$\int \left(\frac{\partial \mathbf{m}(y, \theta)}{\partial \theta'} f(y|\theta) + \mathbf{m}(y, \theta) \frac{\partial \ln f(y|\theta)}{\partial \theta'} f(y|\theta) \right) dy = \mathbf{0}, \quad (5.55)$$

or

$$E \left[\frac{\partial \mathbf{m}(y, \theta)}{\partial \theta'} \right] = -E \left[\mathbf{m}(y, \theta) \frac{\partial \ln f(y|\theta)}{\partial \theta'} \right], \quad (5.56)$$

when the expectation is taken with respect to the density $f(y|\theta)$. The regularity condition (5.42) is the special case $\mathbf{m}(y, \theta) = \partial \ln f(y|\theta)/\partial \theta$ and leads to the IM equality (5.44). The more general result (5.56) leads to the generalized IM equality (5.45).

What happens when integration and differentiation cannot be interchanged? The starting point (5.50) no longer holds, as by the fundamental theorem of calculus the derivative with respect to θ of $\int f(y|\theta)dy$ includes an additional term reflecting the presence of a function θ in the range of the integral. Then $E[\partial \ln f(y|\theta)/\partial \theta] \neq \mathbf{0}$.

What happens when the density is misspecified? Then (5.52) still holds, but it does not necessarily imply (5.41), since in (5.41) the expectation will no longer be with respect to the specified density $f(y|\theta)$.

5.7. Quasi-Maximum Likelihood

The **quasi-MLE** $\hat{\theta}_{QML}$ is defined to be the estimator that maximizes a log-likelihood function that is misspecified, as the result of specification of the wrong density. Generally such misspecification leads to inconsistent estimation.

In this section general properties of the quasi-MLE are presented, followed by some special cases where the quasi-MLE retains consistency.

5.7.1. Psuedo-True Value

In principle any misspecification of the density may lead to inconsistency, as then the expectation in evaluation of $E[\partial \ln f(y|\mathbf{x}, \theta)/\partial \theta|_{\theta_0}]$ (see Section 5.6.4) is no longer with respect to $f(y|\mathbf{x}, \theta_0)$.

By adaptation of the general consistency proof in Section 5.3.2, the quasi-MLE $\hat{\theta}_{QML}$ converges in probability to the **pseudo-true value** θ^* defined as

$$\theta^* = \arg \max_{\theta \in \Theta} (\text{plim } N^{-1} \mathcal{L}_N(\theta)). \quad (5.57)$$

The probability limit is taken with respect to the true dgp. If the true dgp differs from the assumed density $f(y|\mathbf{x}, \theta)$ used to form $\mathcal{L}_N(\theta)$, then usually $\theta^* \neq \theta_0$ and the quasi-MLE is inconsistent.

Huber (1967) and White (1982) showed that the asymptotic distribution of the quasi-MLE is similar to that for the MLE, except that it is centered around θ^* and the IM equality no longer holds. Then

$$\sqrt{N}(\hat{\theta}_{QML} - \theta^*) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{A}^{*-1} \mathbf{B}^* \mathbf{A}^{*-1}], \quad (5.58)$$

where \mathbf{A}^* and \mathbf{B}^* are as defined in (5.18) and (5.19) except that probability limits are taken with respect to the unknown true dgp and are evaluated at θ^* . Consistent estimates $\hat{\mathbf{A}}^*$ and $\hat{\mathbf{B}}^*$ can be obtained as in Section 5.5.2, with evaluation at $\hat{\theta}_{QML}$.

This distributional result is used for statistical inference if the quasi-MLE retains consistency. If the quasi-MLE is inconsistent then usually θ^* has no simple interpretation, aside from that given in the next section. However, (5.58) may still be useful if nonetheless there is interest in knowing the precision of estimation. The result (5.58) also provides motivation for White's **information matrix test** (see Section 8.2.8) and for **Vuong's test** for discriminating between parametric models (see Section 8.5.3).

5.7.2. Kullback–Liebler Distance

Recall from Section 4.2.3 that if $E[y|\mathbf{x}] \neq \mathbf{x}'\boldsymbol{\beta}_0$ then the OLS estimator can still be interpreted as the best linear predictor of $E[y|\mathbf{x}]$ under squared error loss. White (1982) proposed a qualitatively similar interpretation for the quasi-MLE.

Let $f(\mathbf{y}|\boldsymbol{\theta})$ denote the assumed joint density of y_1, \dots, y_N and let $h(\mathbf{y})$ denote the true density, which is unknown, where for simplicity dependence on regressors is suppressed. Define the **Kullback–Leibler information criterion (KLIC)**

$$\text{KLIC} = E \left[\ln \left(\frac{h(\mathbf{y})}{f(\mathbf{y}|\boldsymbol{\theta})} \right) \right], \quad (5.59)$$

where expectation is with respect to $h(\mathbf{y})$. KLIC takes a minimum value of 0 when there is a $\boldsymbol{\theta}_0$ such that $h(\mathbf{y}) = f(\mathbf{y}|\boldsymbol{\theta}_0)$, that is, the density is correctly specified, and larger values of KLIC indicate greater ignorance about the true density.

Then the quasi-MLE $\widehat{\boldsymbol{\theta}}_{\text{QML}}$ minimizes the distance between $f(\mathbf{y}|\boldsymbol{\theta})$ and $h(\mathbf{y})$, where distance is measured using KLIC. To obtain this result, note that under suitable assumptions $\text{plim } N^{-1}\mathcal{L}_N(\boldsymbol{\theta}) = E[\ln f(\mathbf{y}|\boldsymbol{\theta})]$, so $\widehat{\boldsymbol{\theta}}_{\text{QML}}$ converges to $\boldsymbol{\theta}^*$ that maximizes $E[\ln f(\mathbf{y}|\boldsymbol{\theta})]$. However, this is equivalent to minimizing KLIC, since $\text{KLIC} = E[\ln h(\mathbf{y})] - E[\ln f(\mathbf{y}|\boldsymbol{\theta})]$ and the first term does not depend on $\boldsymbol{\theta}$ as the expectation is with respect to $h(\mathbf{y})$.

5.7.3. Linear Exponential Family

In some special cases the quasi-MLE is consistent even when the density is partially misspecified. One well-known example is that the quasi-MLE for the linear regression model with normality is consistent even if the errors are nonnormal, provided $E[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}_0$. The Poisson MLE provides a second example (see Section 5.3.4).

Similar robustness to misspecification is enjoyed by other models based on densities in the **linear exponential family (LEF)**. An LEF density can be expressed as

$$f(y|\mu) = \exp\{a(\mu) + b(y) + c(\mu)y\}, \quad (5.60)$$

where we have given the **mean parameterization** of the LEF, so that $\mu = E[y]$. It can be shown that for this density $E[y] = -[c'(\mu)]^{-1}a'(\mu)$ and $V[y] = [c'(\mu)]^{-1}$, where $c'(\mu) = \partial c(\mu)/\partial \mu$ and $a'(\mu) = \partial a(\mu)/\partial \mu$. Different functions $a(\cdot)$ and $c(\cdot)$ lead to different densities in the family. The term $b(y)$ in (5.60) is a normalizing constant that ensures probabilities sum or integrate to one. The remainder of the density $\exp\{a(\mu) + c(\mu)y\}$ is an exponential function that is linear in y , hence explaining the term linear exponential.

Most densities cannot be expressed in this form. Several important densities are LEF densities, however, including those given in Table 5.4. These densities, already presented in Table 5.3, are reexpressed in Table 5.4 in the form (5.60). Other LEF densities are the binomial with number of trials known (the Bernoulli being a special case), some negative binomials models (the geometric and the Poisson being special cases), and the one-parameter gamma (the exponential being a special case).

Table 5.4. Linear Exponential Family Densities: Leading Examples

Distribution	$f(y) = \exp\{a(\cdot) + b(y) + c(\cdot)y\}$	$E[y]$	$V[y] = [c'(\mu)]^{-1}$
Normal (σ^2 known)	$\exp\left\{-\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2) - \frac{y^2}{2\sigma^2} + \frac{\mu}{\sigma^2}y\right\}$	μ	σ^2
Bernoulli	$\exp\{\ln(1-p) + \ln[p/(1-p)]y\}$	$\mu = p$	$\mu(1-\mu)$
Exponential	$\exp\{\ln\lambda - \lambda y\}$	$\mu = 1/\lambda$	μ^2
Poisson	$\exp\{-\lambda - \ln y! + y \ln \lambda\}$	$\mu = \lambda$	μ

For regression the parameter $\mu = E[y|\mathbf{x}]$ is modeled as

$$\mu = g(\mathbf{x}, \boldsymbol{\beta}), \quad (5.61)$$

for specified function $g(\cdot)$ that varies across models (see Section 5.7.4) depending in part on restrictions on the range of y and hence μ . The LEF log-likelihood is then

$$\mathcal{L}_N(\boldsymbol{\beta}) = \sum_{i=1}^N \{a(g(\mathbf{x}_i, \boldsymbol{\beta})) + b(y_i) + c(g(\mathbf{x}_i, \boldsymbol{\beta}))y_i\}, \quad (5.62)$$

with first-order conditions that can be reexpressed, using the aforementioned information on the first-two moments of y , as

$$\frac{\partial \mathcal{L}_N(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^N \frac{y_i - g(\mathbf{x}_i, \boldsymbol{\beta})}{\sigma_i^2} \times \frac{\partial g(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{0}, \quad (5.63)$$

where $\sigma_i^2 = [c'(g(\mathbf{x}_i, \boldsymbol{\beta}))]^{-1}$ is the assumed variance function corresponding to the particular LEF density. For example, for Bernoulli, exponential, and Poisson, σ_i^2 equals, respectively, $g_i(1-g_i)$, $1/g_i^2$, and g_i , where $g_i = g(\mathbf{x}_i, \boldsymbol{\beta})$.

The quasi-MLE solves these equations, but it is no longer assumed that the LEF density is correctly specified. Gouriéroux, Monfort, and Trognon (1984a) proved that the quasi-MLE $\hat{\boldsymbol{\beta}}_{QML}$ is consistent provided $E[y|\mathbf{x}] = g(\mathbf{x}, \boldsymbol{\beta}_0)$. This is clear from taking the expected value of the first-order conditions (5.63), which evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ are a weighted sum of errors $y - g(\mathbf{x}, \boldsymbol{\beta}_0)$ with expected value equal to zero if $E[y|\mathbf{x}] = g(\mathbf{x}, \boldsymbol{\beta}_0)$.

Thus the quasi-MLE based on an LEF density is consistent provided only that the conditional mean of y given \mathbf{x} is correctly specified. Note that the actual dgp for y need not be LEF. It is the specified density, potentially incorrectly specified, that is LEF.

Even with correct conditional mean, however, adjustment of default ML output for variance, standard errors, and t -statistics based on $-\mathbf{A}_0^{-1}$ is warranted. In general the sandwich form $\mathbf{A}_0^{-1} \mathbf{B}_0 \mathbf{A}_0^{-1}$ should be used, unless the conditional variance of y given \mathbf{x} is also correctly specified, in which case $\mathbf{A}_0 = -\mathbf{B}_0$. For Bernoulli models, however, $\mathbf{A}_0 = -\mathbf{B}_0$ always. Consistent standard errors can be obtained using (5.36) and (5.38).

The LEF is a very special case. In general, misspecification of any aspect of the density leads to inconsistency of the MLE. Even in the LEF case the quasi-MLE can

be used only to predict the conditional mean whereas with a correctly specified density one can predict the conditional distribution.

5.7.4. Generalized Linear Models

Models based on an assumed LEF density are called **generalized linear models** (GLMs) in the statistics literature (see the book with this title by McCullagh and Nelder, 1989). The class of generalized linear models is the most widely used framework in applied statistics for nonlinear cross-section regression, as from Table 5.3 it includes nonlinear least squares, Poisson, geometric, probit, logit, binomial (known number of trials), gamma, and exponential regression models. We provide a short overview that introduces standard GLM terminology.

Standard GLMs specify the conditional mean $g(\mathbf{x}, \beta)$ in (5.61) to be of the simpler single-index form, so that $\mu = g(\mathbf{x}'\beta)$. Then $g^{-1}(\mu) = \mathbf{x}'\beta$, and the function $g^{-1}(\cdot)$ is called the **link function**. For example, the usual specification for the Poisson model corresponds to the log-link function since if $\mu = \exp(\mathbf{x}'\beta)$ then $\ln \mu = \mathbf{x}'\beta$.

The first-order conditions (5.63) become $\sum_i [(y_i - g_i)/c'(g_i)] g'_i \mathbf{x}_i = \mathbf{0}$, where $g_i = g(\mathbf{x}'_i \beta)$ and $g'_i = g'(\mathbf{x}'_i \beta)$. There are computational advantages in choosing the link function so that $c'(g(\mu)) = g'(\mu)$, since then these first-order conditions reduce to $\sum_i (y_i - g_i) \mathbf{x}_i = \mathbf{0}$, or the error $(y_i - g_i)$ is orthogonal to the regressors. The **canonical link function** is defined to be that function $g^{-1}(\cdot)$ which leads to $c'(g(\mu)) = g'(\mu)$ and varies with $c(\mu)$ and hence the GLM. The canonical link function leads to $\mu = \mathbf{x}'\beta$ for normal, $\mu = \exp(\mathbf{x}'\beta)$ for Poisson, and $\mu = \exp(\mathbf{x}'\beta)/[1 + \exp(\mathbf{x}'\beta)]$ for binary data. The last of these is the logit form given earlier in Table 5.3.

Two times the difference between the maximum achievable log-likelihood and the fitted log-likelihood is called the **deviance**, a measure that generalizes the residual sum of squares in linear regression to other LEF regression models.

Models based on the LEF are very restrictive as all moments depend on just one underlying parameter, $\mu = g(\mathbf{x}'\beta)$. The GLM literature places some additional structure by making the convenient assumption that the LEF variance is potentially misspecified by a scalar multiple α , so that $V[y|\mathbf{x}] = \alpha \times [c'(g(\mathbf{x}, \beta))]^{-1}$, where $\alpha \neq 1$ necessarily. For example, for the Poisson model let $V[y|\mathbf{x}] = \alpha g(\mathbf{x}, \beta)$ rather than $g(\mathbf{x}, \beta)$. Given such variance misspecification it can be shown that $\mathbf{B}_0 = -\alpha \mathbf{A}_0$, so the variance matrix of the quasi-MLE is $-\alpha \mathbf{A}_0^{-1}$, which requires only a rescaling of the nonsandwich ML variance matrix $-\mathbf{A}_0^{-1}$ by multiplication by α . A commonly used consistent estimate for α is $\hat{\alpha} = (N - K)^{-1} \sum_i (y_i - \hat{g}_i)^2 / \hat{\sigma}_i^2$, where $\hat{g}_i = g(\mathbf{x}_i, \hat{\beta}_{QML})$, $\hat{\sigma}_i^2 = [c'(\hat{g}_i)]^{-1}$, and division is by $(N - K)$ rather than N is felt to provide a better estimate in small samples. See the preceding references and Cameron and Trivedi (1986, 1998) for further details.

Many statistical packages include a GLM module that as a default gives standard errors that are correct provided $V[y|\mathbf{x}] = \alpha [c'(g(\mathbf{x}, \beta))]^{-1}$. Alternatively, one can estimate using ML, with standard errors obtained using the robust sandwich formula $\mathbf{A}_0^{-1} \mathbf{B}_0 \mathbf{A}_0^{-1}$. In practice the sandwich standard errors are similar to those obtained using the simple GLM correction. Yet another way to estimate a GLM is by weighted nonlinear least squares, as detailed at the end of Section 5.8.6.

5.7.5. Quasi-MLE for Multivariate Dependent Variables

This chapter has focused on scalar dependent variables, but the theory applies also to the multivariate case. Suppose the dependent variable \mathbf{y} is an $m \times 1$ vector, and the data $(\mathbf{y}_i, \mathbf{x}_i)$, $i = 1, \dots, N$, are independent over i . Examples given in later chapters include seemingly unrelated equations, panel data with m observations for the i th individual on the same dependent variable, and clustered data where data for the ij th observation are correlated over m possible values of j .

Given specification of $f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$, the joint density of $\mathbf{y} = (y_1, \dots, y_m)$ conditional on \mathbf{x} , the fully efficient MLE maximizes $N^{-1} \sum_i \ln f(\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\theta})$ as noted after (5.39). However, in multivariate applications the joint density of \mathbf{y} can be complicated. A simpler estimator is possible given knowledge only of the m univariate densities $f_j(y_j | \mathbf{x}, \boldsymbol{\theta})$, $j = 1, \dots, m$, where y_j is the j th component of \mathbf{y} . For example, for multivariate count data one might work with m independent univariate negative binomial densities for each count rather than a richer multivariate count model that permits correlation.

Consider then the quasi-MLE $\hat{\boldsymbol{\theta}}_{QML}$ based on the product of the univariate densities, $\prod_j f_j(y_j | \mathbf{x}, \boldsymbol{\theta})$, that maximizes

$$Q_N(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^m \ln f(y_{ij} | \mathbf{x}_i, \boldsymbol{\theta}). \quad (5.64)$$

Wooldridge (2002) calls this estimator the **partial MLE**, since the density has been only partially specified.

The partial MLE is an m -estimator with $q_i = \sum_j \ln f(y_{ij} | \mathbf{x}_i, \boldsymbol{\theta})$. The essential consistency condition (5.25) requires that $E[\sum_j \partial f(y_{ij} | \mathbf{x}_i, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}]|_{\boldsymbol{\theta}_0} = \mathbf{0}$. This condition holds if the marginal densities $f(y_{ij} | \mathbf{x}_i, \boldsymbol{\theta}_0)$ are correctly specified, since then $E[\partial f(y_{ij} | \mathbf{x}_i, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}]|_{\boldsymbol{\theta}_0} = \mathbf{0}$ by the regularity condition (5.41).

Thus the partial MLE is consistent provided the univariate densities $f_j(y_j | \mathbf{x}, \boldsymbol{\theta})$ are correctly specified. Consistency does not require that $f(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta}) = \prod_j f_j(y_j | \mathbf{x}, \boldsymbol{\theta})$. Dependence of y_1, \dots, y_m will lead to failure of the information matrix equality, however, so standard errors should be computed using the sandwich form for the variance matrix with

$$\begin{aligned} \mathbf{A}_0 &= \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^m \left. \frac{\partial^2 \ln f_{ij}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right|_{\boldsymbol{\theta}_0}, \\ \mathbf{B}_0 &= \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^m \sum_{k=1}^m \left. \frac{\partial \ln f_{ij}}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}_0} \left. \frac{\partial \ln f_{ik}}{\partial \boldsymbol{\theta}'} \right|_{\boldsymbol{\theta}_0} \end{aligned} \quad (5.65)$$

where $f_{ij} = f(y_{ij} | \mathbf{x}_i, \boldsymbol{\theta})$. Furthermore, the partial MLE is inefficient compared to the MLE based on the joint density. Further discussion is given in Sections 6.9 and 6.10.

5.8. Nonlinear Least Squares

The NLS estimator is the natural extension of LS estimation for the linear model to the nonlinear model with $E[y | \mathbf{x}] = g(\mathbf{x}, \boldsymbol{\beta})$, where $g(\cdot)$ is nonlinear in $\boldsymbol{\beta}$. The analysis and results are essentially the same as for linear least squares, with the single change that in

Table 5.5. Nonlinear Least Squares: Common Examples

Model	Regression Function $g(\mathbf{x}, \beta)$
Exponential	$\exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)$
Regressor raised to power	$\beta_1 x_1 + \beta_2 x_2^{\beta_3}$
Cobb–Douglas production	$\beta_1 x_1^{\beta_2} x_2^{\beta_3}$
CES production	$[\beta_1 x_1^{\beta_3} + \beta_2 x_2^{\beta_3}]^{1/\beta_3}$
Nonlinear restrictions	$\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, where $\beta_3 = -\beta_2 \beta_1$

the formulas for variance matrices the regressor vector \mathbf{x} is replaced by $\partial g(\mathbf{x}, \beta)/\partial\beta|_{\widehat{\beta}}$, the derivative of the conditional mean function evaluated at $\beta = \widehat{\beta}$.

For microeconometric analysis, controlling for heteroskedastic errors may be necessary, as in the linear case. The NLS estimator and extensions that model heteroskedastic errors are generally less efficient than the MLE, but they are widely used in microeconomics because they rely on weaker distributional assumptions.

5.8.1. Nonlinear Regression Model

The **nonlinear regression model** defines the scalar dependent variable y to have conditional mean

$$E[y_i|\mathbf{x}_i] = g(\mathbf{x}_i, \beta), \quad (5.66)$$

where $g(\cdot)$ is a specified function, \mathbf{x} is a vector of explanatory variables, and β is a $K \times 1$ vector of parameters. The linear regression model of Chapter 4 is the special case $g(\mathbf{x}, \beta) = \mathbf{x}'\beta$.

Common reasons for specifying a nonlinear function for $E[y|\mathbf{x}]$ include range restriction (e.g., to ensure that $E[y|\mathbf{x}] > 0$) and specification of supply or demand or cost or expenditure models that satisfy restrictions from producer or consumer theory. Some commonly used nonlinear regression models are given in Table 5.5.

5.8.2. NLS Estimator

The error term is defined to be the difference between the dependent variable and its conditional mean, $y_i - g(\mathbf{x}_i, \beta)$. The **nonlinear least-squares estimator** $\widehat{\beta}_{NLS}$ minimizes the sum of squared residuals, $\sum_i(y_i - g(\mathbf{x}_i, \beta))^2$, or equivalently maximizes

$$Q_N(\beta) = -\frac{1}{2N} \sum_{i=1}^N (y_i - g(\mathbf{x}_i, \beta))^2, \quad (5.67)$$

where the scale factor 1/2 simplifies the subsequent analysis.

Differentiation leads to the NLS first-order conditions

$$\frac{\partial Q_N(\beta)}{\partial \beta} = \frac{1}{N} \sum_{i=1}^N \frac{\partial g_i}{\partial \beta} (y_i - g_i) = \mathbf{0}, \quad (5.68)$$

where $g_i = g(\mathbf{x}_i, \beta)$. These conditions restrict the residual $(y - g)$ to be orthogonal to $\partial g / \partial \beta$, rather than to \mathbf{x} as in the linear case. There is no explicit solution for $\widehat{\beta}_{\text{NLS}}$, which instead is computed using iterative methods (given in Chapter 10).

The nonlinear regression model can be more compactly represented in matrix notation. Stacking observations yields

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} g_1 \\ \vdots \\ g_N \end{bmatrix} + \begin{bmatrix} u_1 \\ \vdots \\ u_N \end{bmatrix}, \quad (5.69)$$

where $g_i = g(\mathbf{x}_i, \beta)$, or equivalently

$$\mathbf{y} = \mathbf{g} + \mathbf{u}, \quad (5.70)$$

where \mathbf{y} , \mathbf{g} , and \mathbf{u} are $N \times 1$ vectors with i th entries of, respectively, y_i , g_i , and u_i . Then

$$Q_N(\beta) = -\frac{1}{2N}(\mathbf{y} - \mathbf{g})'(\mathbf{y} - \mathbf{g})$$

and

$$\frac{\partial Q_N(\beta)}{\partial \beta} = \frac{1}{N} \frac{\partial \mathbf{g}'}{\partial \beta} (\mathbf{y} - \mathbf{g}), \quad (5.71)$$

where

$$\frac{\partial \mathbf{g}'}{\partial \beta} = \begin{bmatrix} \frac{\partial g_1}{\partial \beta_1} & \cdots & \frac{\partial g_N}{\partial \beta_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_1}{\partial \beta_K} & \cdots & \frac{\partial g_N}{\partial \beta_K} \end{bmatrix} \quad (5.72)$$

is the $K \times N$ matrix of partial derivatives of $\mathbf{g}(\mathbf{x}, \beta)'$ with respect to β .

5.8.3. Distribution of the NLS Estimator

The distribution of the NLS estimator will vary with the dgp. The dgp can always be written as

$$y_i = g(\mathbf{x}_i, \beta_0) + u_i, \quad (5.73)$$

a nonlinear regression model with additive error u . The conditional mean is correctly specified if $E[y|\mathbf{x}] = g(\mathbf{x}, \beta_0)$ in the dgp. Then the error must satisfy $E[u|\mathbf{x}] = 0$.

Given the NLS first-order conditions (5.68), the essential consistency condition (5.25) becomes

$$E[\partial g(\mathbf{x}, \beta)/\partial \beta|_{\beta_0} \times (y - g(\mathbf{x}_i, \beta_0))] = \mathbf{0}.$$

Equivalently, given (5.73), we need $E[\partial g(\mathbf{x}, \beta)/\partial \beta|_{\beta_0} \times u] = \mathbf{0}$. This holds if $E[u|\mathbf{x}] = 0$, so consistency requires correct specification of the conditional mean as in the linear case. If instead $E[u|\mathbf{x}] \neq 0$ then consistent estimation requires nonlinear instrumental methods (which are presented in Section 6.5).

The limit distribution of $\sqrt{N}(\widehat{\beta}_{\text{NLS}} - \beta_0)$ is obtained using an exact first-order Taylor series expansion of the first-order conditions (5.68). This yields

$$\begin{aligned}\sqrt{N}(\widehat{\beta}_{\text{NLS}} - \beta_0) &= - \left(\frac{-1}{N} \sum_{i=1}^N \frac{\partial g_i}{\partial \beta} \frac{\partial g_i}{\partial \beta'} + \frac{1}{N} \sum_{i=1}^N \frac{\partial^2 g_i}{\partial \beta \partial \beta'} (y_i - g_i) \Big|_{\beta^+} \right)^{-1} \\ &\quad \times \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{\partial g_i}{\partial \beta} u_i \Big|_{\beta_0},\end{aligned}$$

for some β^+ between $\widehat{\beta}_{\text{NLS}}$ and β_0 . For \mathbf{A}_0 in (5.18) simplification occurs because the term involving $(\partial^2 g / \partial \beta \partial \beta')$ drops out since $E[u|\mathbf{x}] = 0$. Thus asymptotically we need consider only

$$\sqrt{N}(\widehat{\beta}_{\text{NLS}} - \beta_0) = \left(\frac{1}{N} \sum_{i=1}^N \frac{\partial g_i}{\partial \beta} \frac{\partial g_i}{\partial \beta'} \Big|_{\beta_0} \right)^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{\partial g_i}{\partial \beta} u_i \Big|_{\beta_0},$$

which is exactly the same as OLS, see Section 4.4.4, except \mathbf{x}_i is replaced by $\partial g_i / \partial \beta' \Big|_{\beta_0}$.

This yields the following proposition, analogous to Proposition 4.1 for the OLS estimator.

Proposition 5.6 (Distribution of NLS Estimator): *Make the following assumptions:*

- (i) *The model is (5.73); that is, $y_i = g(\mathbf{x}_i, \beta_0) + u_i$.*
- (ii) *In the dgp $E[u_i|\mathbf{x}_i] = 0$ and $E[\mathbf{u}\mathbf{u}'|\mathbf{X}] = \Omega_0$, where $\Omega_{0,ij} = \sigma_{ij}$.*
- (iii) *The mean function $g(\cdot)$ satisfies $g(\mathbf{x}, \beta^{(1)}) = g(\mathbf{x}, \beta^{(2)})$ iff $\beta^{(1)} = \beta^{(2)}$.*
- (iv) *The matrix*

$$\mathbf{A}_0 = \text{plim} \frac{1}{N} \sum_{i=1}^N \frac{\partial g_i}{\partial \beta} \frac{\partial g_i}{\partial \beta'} \Big|_{\beta_0} = \text{plim} \frac{1}{N} \frac{\partial \mathbf{g}}{\partial \beta} \frac{\partial \mathbf{g}'}{\partial \beta'} \Big|_{\beta_0} \quad (5.74)$$

exists and is finite nonsingular.

- (v) $N^{-1/2} \sum_{i=1}^N \partial g_i / \partial \beta \times u_i \Big|_{\beta_0} \xrightarrow{d} \mathcal{N}[0, \mathbf{B}_0]$, where

$$\mathbf{B}_0 = \text{plim} \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \sigma_{ij} \frac{\partial g_i}{\partial \beta} \frac{\partial g_j}{\partial \beta'} \Big|_{\beta_0} = \text{plim} \frac{1}{N} \frac{\partial \mathbf{g}'}{\partial \beta} \Omega_0 \frac{\partial \mathbf{g}}{\partial \beta'} \Big|_{\beta_0}. \quad (5.75)$$

Then the NLS estimator $\widehat{\beta}_{\text{NLS}}$, defined to be a root of the first-order conditions $\partial N^{-1} Q_N(\beta) / \partial \beta = \mathbf{0}$, is consistent for β_0 and

$$\sqrt{N}(\widehat{\beta}_{\text{NLS}} - \beta_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{A}_0^{-1} \mathbf{B}_0 \mathbf{A}_0^{-1}]. \quad (5.76)$$

Conditions (i) to (iii) imply that the regression function is correctly specified and the regressors are uncorrelated with the errors and that β_0 is identified. The errors can be heteroskedastic and correlated over i . Conditions (iv) and (v) assume the relevant limit results necessary for application of Theorem 5.3. For condition (v) to be satisfied some restrictions will need to be placed on the error correlation over i . The probability limits in (5.74) and (5.75) are with respect to the dgp for \mathbf{X} ; they become regular limits if \mathbf{X} is nonstochastic.

The matrices \mathbf{A}_0 and \mathbf{B}_0 in Proposition 5.6 are the same as the matrices $\mathbf{M}_{\mathbf{xx}}$ and $\mathbf{M}_{\mathbf{x}\Omega\mathbf{x}}$ in Section 4.4.4 for the OLS estimator with \mathbf{x}_i replaced by $\partial g_i / \partial \beta|_{\beta_0}$. The asymptotic theory for NLS is the same as that for OLS, with this single change.

In the special case of spherical errors, $\Omega_0 = \sigma_0^2 \mathbf{I}$, so $\mathbf{B}_0 = \sigma_0^2 \mathbf{A}_0$ and $V[\hat{\beta}_{NLS}] = \sigma_0^2 \mathbf{A}_0^{-1}$. Nonlinear least squares is then asymptotically efficient among LS estimators. However, cross-section data errors are not necessarily heteroskedastic.

Given Proposition 5.6, the resulting *asymptotic distribution* of the NLS estimator can be expressed as

$$\hat{\beta}_{NLS} \xrightarrow{a} \mathcal{N} \left[\beta, (\mathbf{D}' \mathbf{D})^{-1} \mathbf{D}' \Omega_0 \mathbf{D} (\mathbf{D}' \mathbf{D})^{-1} \right], \quad (5.77)$$

where the derivative matrix $\mathbf{D} = \partial \mathbf{g} / \partial \beta'|_{\beta_0}$ has i th row $\partial g_i / \partial \beta'|_{\beta_0}$ (see (5.72)), for notational simplicity the evaluation at β_0 is suppressed, and we assume that an LLN applies, so that the plim operator in the definitions of \mathbf{A}_0 and \mathbf{B}_0 are replaced by $\lim E$, and then drop the limit. This notation is often used in later chapters.

5.8.4. Variance Matrix Estimation for NLS

We consider statistical inference for the usual microeconomics situation of independent errors with **heteroskedasticity of unknown functional form**. This requires a consistent estimate of $\mathbf{A}_0^{-1} \mathbf{B}_0 \mathbf{A}_0^{-1}$ defined in Proposition 5.6.

For \mathbf{A}_0 defined in (5.74) it is straightforward to use the obvious estimator

$$\hat{\mathbf{A}} = \frac{1}{N} \sum_{i=1}^N \left. \frac{\partial g_i}{\partial \beta} \right|_{\hat{\beta}} \left. \frac{\partial g_i}{\partial \beta'} \right|_{\hat{\beta}}, \quad (5.78)$$

as \mathbf{A}_0 does not involve moments of the errors.

Given independence over i the double sum in \mathbf{B}_0 defined in (5.75) simplifies to the single sum

$$\mathbf{B}_0 = \text{plim} \frac{1}{N} \sum_{i=1}^N \sigma_i^2 \left. \frac{\partial g_i}{\partial \beta} \frac{\partial g_i}{\partial \beta'} \right|_{\beta_0}.$$

As for the OLS estimator (see Section 4.4.5) it is only necessary to consistently estimate the $K \times K$ matrix sum \mathbf{B}_0 . This does not require consistent estimation of σ_i^2 , the N individual components in the sum.

White (1980b) gave conditions under which

$$\widehat{\mathbf{B}} = \frac{1}{N} \sum_{i=1}^N \widehat{u}_i^2 \frac{\partial g_i}{\partial \beta} \frac{\partial g_i}{\partial \beta'} \Big|_{\widehat{\beta}} = \frac{1}{N} \frac{\partial \mathbf{g}'}{\partial \beta} \Big|_{\widehat{\beta}} \widehat{\Omega} \frac{\partial \mathbf{g}}{\partial \beta'} \Big|_{\widehat{\beta}} \quad (5.79)$$

is consistent for \mathbf{B}_0 , where $\widehat{u}_i = y_i - g(\mathbf{x}_i, \widehat{\beta})$, $\widehat{\beta}$ is consistent for β_0 , and

$$\widehat{\Omega} = \text{Diag}[\widehat{u}_i^2]. \quad (5.80)$$

This leads to the following **heteroskedastic-consistent** estimate of the asymptotic variance matrix of the NLS estimator:

$$\widehat{\mathbf{V}}[\widehat{\beta}_{\text{NLS}}] = (\widehat{\mathbf{D}}' \widehat{\mathbf{D}})^{-1} \widehat{\mathbf{D}}' \widehat{\Omega} \widehat{\mathbf{D}} (\widehat{\mathbf{D}}' \widehat{\mathbf{D}})^{-1}, \quad (5.81)$$

where $\widehat{\mathbf{D}} = \frac{\partial \mathbf{g}}{\partial \beta'} \Big|_{\widehat{\beta}}$. This equation is the same as the OLS result in Section 4.4.5, with the regressor matrix \mathbf{X} replaced by $\widehat{\mathbf{D}}$. In practice, a degrees of freedom correction may be used, so that $\widehat{\mathbf{B}}$ in (5.79) is computed using division by $(N - K)$ rather than by N . Then the right-hand side in (5.81) should be multiplied by $N/(N - K)$.

Generalization to errors correlated over i is given in Section 5.8.7.

5.8.5. Exponential Regression Example

As an example, suppose that y given \mathbf{x} has exponential conditional mean, so that $E[y|\mathbf{x}] = \exp(\mathbf{x}'\beta)$. The model can be expressed as a nonlinear regression with

$$y = \exp(\mathbf{x}'\beta) + u,$$

where the error term u has $E[u|\mathbf{x}] = 0$ and the error is potentially heteroskedastic.

The NLS estimator has first-order conditions

$$N^{-1} \sum_i (y_i - \exp(\mathbf{x}'\beta)) \exp(\mathbf{x}'\beta) \mathbf{x}_i = \mathbf{0}, \quad (5.82)$$

so consistency of $\widehat{\beta}_{\text{NLS}}$ requires only that the conditional mean be correctly specified with $E[y|\mathbf{x}] = \exp(\mathbf{x}'\beta_0)$. Here $\partial g/\partial \beta = \exp(\mathbf{x}'\beta)\mathbf{x}$, so the general NLS result (5.81) yields the heteroskedastic-robust estimate

$$\widehat{\mathbf{V}}[\widehat{\beta}_{\text{NLS}}] = \left(\sum_i e^{2\mathbf{x}'\beta} \mathbf{x}_i \mathbf{x}'_i \right)^{-1} \sum_i \widehat{u}_i^2 e^{2\mathbf{x}'\beta} \mathbf{x}_i \mathbf{x}'_i \left(\sum_i e^{2\mathbf{x}'\beta} \mathbf{x}_i \mathbf{x}'_i \right)^{-1}, \quad (5.83)$$

where $\widehat{u}_i = y_i - \exp(\mathbf{x}'\widehat{\beta}_{\text{NLS}})$.

5.8.6. Weighted NLS and FGNLS

For cross-section data the errors are often heteroskedastic. Then feasible generalized NLS that controls for the heteroskedasticity is more efficient than NLS.

Feasible generalized nonlinear least squares (FGNLS) is still generally less efficient than ML. The notable exception is that FGNLS is asymptotically equivalent to the MLE when the conditional density for y is an LEF density. A special case is that FGLS is asymptotically equivalent to the MLE in the linear regression under normality.

Table 5.6. Nonlinear Least-Squares Estimators and Their Asymptotic Variance^a

Estimator	Objective Function	Estimated Asymptotic Variance
NLS	$Q_N(\beta) = \frac{1}{2N} \mathbf{u}' \mathbf{u}$	$(\widehat{\mathbf{D}}' \widehat{\mathbf{D}})^{-1} \widehat{\mathbf{D}}' \widehat{\Omega} \widehat{\mathbf{D}} (\widehat{\mathbf{D}}' \widehat{\mathbf{D}})^{-1}$
FGNLS	$Q_N(\beta) = \frac{1}{2N} \mathbf{u}' \Omega(\widehat{\gamma})^{-1} \mathbf{u}$	$(\widehat{\mathbf{D}}' \widehat{\Omega}^{-1} \widehat{\mathbf{D}})^{-1}$
WNLS	$Q_N(\beta) = \frac{1}{2N} \mathbf{u}' \widehat{\Sigma}^{-1} \mathbf{u}$	$(\widehat{\mathbf{D}}' \widehat{\Sigma}^{-1} \widehat{\mathbf{D}})^{-1} \widehat{\mathbf{D}}' \widehat{\Sigma}^{-1} \widehat{\Omega} \widehat{\Sigma}^{-1} \widehat{\mathbf{D}} (\widehat{\mathbf{D}}' \widehat{\Sigma}^{-1} \widehat{\mathbf{D}})^{-1}$

^a Functions are for a nonlinear regression model with error $\mathbf{u} = \mathbf{y} - \mathbf{g}$ defined in (5.70) and error conditional variance matrix Ω . $\widehat{\mathbf{D}}$ is the derivative of the conditional mean vector with respect to β' evaluated at $\widehat{\beta}$. For FGNLS it is assumed that $\widehat{\Omega}$ is consistent for Ω . For NLS and WNLS the heteroskedastic robust variance matrix uses $\widehat{\Omega}$ equal to a diagonal matrix with squared residuals on the diagonals, an estimate that need not be consistent for Ω .

If heteroskedasticity is incorrectly modeled then the FGNLS estimator retains consistency but one should then obtain standard errors that are robust to misspecification of the model for heteroskedasticity. The analysis is very similar to that for the linear model given in Section 4.5.

Feasible Generalized Nonlinear Least Squares

The **feasible generalized nonlinear least-squares estimator** $\widehat{\beta}_{\text{FGNLS}}$ maximizes

$$Q_N(\beta) = -\frac{1}{2N} (\mathbf{y} - \mathbf{g})' \Omega(\widehat{\gamma})^{-1} (\mathbf{y} - \mathbf{g}), \quad (5.84)$$

where it is assumed that $E[\mathbf{u}\mathbf{u}'|\mathbf{X}] = \Omega(\gamma_0)$ and $\widehat{\gamma}$ is a consistent estimate $\widehat{\gamma}$ of γ_0 .

If the assumptions made for the NLS estimator are satisfied and in fact $\Omega_0 = \Omega(\gamma_0)$, then the FGNLS estimator is consistent and asymptotically normal with estimated asymptotic variance matrix given in Table 5.6. The variance matrix estimate is similar to that for linear FGLS, $[\mathbf{X}' \Omega(\widehat{\gamma})^{-1} \mathbf{X}]^{-1}$, except that \mathbf{X} is replaced by $\widehat{\mathbf{D}} = \partial \mathbf{g} / \partial \beta'|_{\widehat{\beta}}$.

The FGNLS estimator is the most efficient consistent estimator that minimizes quadratic loss functions of the form $(\mathbf{y} - \mathbf{g})' \mathbf{V} (\mathbf{y} - \mathbf{g})$, where \mathbf{V} is a weighting matrix.

In general, implementation of FGNLS requires inversion of the $N \times N$ matrix $\Omega(\widehat{\gamma})$. This may be computationally impossible for large N , but in practice $\Omega(\widehat{\gamma})$ usually has a structure, such as diagonality, that leads to an analytical solution for the inverse.

Weighted NLS

The FGNLS approach is fully efficient but leads to invalid standard error estimates if the model for Ω_0 is misspecified. Here we consider an approach between NLS and FGNLS that specifies a model for the variance matrix of the errors but then obtains robust standard errors. The discussion mirrors that in Section 4.5.2.

The **weighted nonlinear least squares (WNLS) estimator** $\widehat{\beta}_{\text{WNLS}}$ maximizes

$$Q_N(\beta) = -\frac{1}{2N} (\mathbf{y} - \mathbf{g})' \widehat{\Sigma}^{-1} (\mathbf{y} - \mathbf{g}), \quad (5.85)$$

where $\Sigma = \Sigma(\gamma)$ is a **working error variance matrix**, $\widehat{\Sigma} = \Sigma(\widehat{\gamma})$, where $\widehat{\gamma}$ is an estimate of γ , and, in a departure from FGNLS, $\Sigma \neq \Omega_0$.

Under assumptions similar to those for the NLS estimator and assuming that $\Sigma_0 = \text{plim } \widehat{\Sigma}$, the WNLS estimator is consistent and asymptotically normal with estimated asymptotic variance matrix given in Table 5.6.

This estimator is called WNLS to distinguish it from FGNLS, which assumed that $\Sigma = \Omega_0$. The WNLS estimator hopefully lies between NLS and FGNLS in terms of efficiency, though it may be less efficient than NLS if a poor model of the error variance matrix is chosen. The NLS and OLS estimators are special cases of WNLS with $\Sigma = \sigma^2 \mathbf{I}$.

Heteroskedastic Errors

An obvious working model for heteroskedasticity is $\sigma_i^2 = E[u_i^2 | \mathbf{x}_i] = \exp(\mathbf{z}'_i \boldsymbol{\gamma}_0)$, where the vector \mathbf{z} is a specified function of \mathbf{x} (such as selected subcomponents of \mathbf{x}) and using the exponential ensures a positive variance.

Then $\Sigma = \text{Diag}[\exp(\mathbf{z}'_i \boldsymbol{\gamma})]$ and $\widehat{\Sigma} = \text{Diag}[\exp(\mathbf{z}'_i \widehat{\boldsymbol{\gamma}})]$, where $\widehat{\boldsymbol{\gamma}}$ can be obtained by nonlinear regression of squared NLS residuals $(y_i - g(\mathbf{x}_i, \widehat{\beta}_{\text{NLS}}))^2$ on $\exp(\mathbf{z}'_i \boldsymbol{\gamma})$. Since Σ is diagonal, $\Sigma^{-1} = \text{Diag}[1/\sigma_i^2]$. Then (5.84) simplifies and the WNLS estimator maximizes

$$Q_N(\boldsymbol{\beta}) = -\frac{1}{2N} \sum_{i=1}^N \frac{(y_i - g(\mathbf{x}_i, \boldsymbol{\beta}))^2}{\widehat{\sigma}_i^2}. \quad (5.86)$$

The variance matrix of the WNLS estimator given in Table 5.6 yields

$$\widehat{V}[\widehat{\boldsymbol{\beta}}_{\text{WNLS}}] = \left(\sum_{i=1}^N \frac{1}{\widehat{\sigma}_i^2} \widehat{\mathbf{d}}_i \widehat{\mathbf{d}}_i' \right)^{-1} \left(\sum_{i=1}^N \widehat{u}_i^2 \frac{1}{\widehat{\sigma}_i^4} \widehat{\mathbf{d}}_i \widehat{\mathbf{d}}_i' \right) \left(\sum_{i=1}^N \frac{1}{\widehat{\sigma}_i^2} \widehat{\mathbf{d}}_i \widehat{\mathbf{d}}_i' \right)^{-1}, \quad (5.87)$$

where $\widehat{\mathbf{d}}_i = \partial g(\mathbf{x}_i, \boldsymbol{\beta}) / \partial \boldsymbol{\beta}|_{\widehat{\boldsymbol{\beta}}}$ and $\widehat{u}_i = y_i - g(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_{\text{WNLS}})$ is the residual. In practice a degrees of freedom correction may be used, so that the right-hand side of (5.87) is multiplied by $N/(N - K)$. If the stronger assumption is made that $\Sigma = \Omega_0$, then WNLS becomes FGNLS and

$$\widehat{V}[\widehat{\boldsymbol{\beta}}_{\text{FGNLS}}] = \left(\sum_{i=1}^N \frac{1}{\widehat{\sigma}_i^2} \widehat{\mathbf{d}}_i \widehat{\mathbf{d}}_i' \right)^{-1}. \quad (5.88)$$

The WNLS and FGNLS estimators can be implemented using an NLS program. First, do NLS regression of y_i on $g(\mathbf{x}_i, \boldsymbol{\beta})$. Second, obtain $\widehat{\boldsymbol{\gamma}}$ by, for example, NLS regression of $(y_i - g(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_{\text{NLS}}))^2$ on $\exp(\mathbf{z}'_i \boldsymbol{\gamma})$ if $\sigma_i^2 = \exp(\mathbf{z}'_i \boldsymbol{\gamma})$. Third, perform an NLS regression of $y_i / \widehat{\sigma}_i$ on $g(\mathbf{x}_i, \boldsymbol{\beta}) / \widehat{\sigma}_i$, where $\widehat{\sigma}_i^2 = \exp(\mathbf{z}'_i \widehat{\boldsymbol{\gamma}})$. This is equivalent to maximizing (5.86). White robust sandwich standard errors from this transformed regression give robust WNLS standard errors based on (5.87). The usual nonrobust standard errors from this transformed regression give FGNLS standard errors based on (5.88).

With heteroskedastic errors it is very tempting to go one step further and attempt FGNLS using $\widehat{\Omega} = \text{Diag}[\widehat{u}_i^2]$. This will give inconsistent parameter estimates of $\boldsymbol{\beta}_0$, however, as FGNLS regression of y_i on $g(\mathbf{x}_i, \boldsymbol{\beta})$ then reduces to NLS regression of $y_i / |\widehat{u}_i|$ on $g(\mathbf{x}_i, \boldsymbol{\beta}) / |\widehat{u}_i|$. The technique suffers from the fundamental problem of

correlation between regressors and error term. Alternative semiparametric methods that enable an estimator as efficient as feasible GLS, without specifying a functional form for Ω_0 , are presented in Section 9.7.6.

Generalized Linear Models

Implementation of the weighted NLS approach requires a reasonable specification for the working matrix. A somewhat ad-hoc approach, already presented, is to let $\sigma_i^2 = \exp(\mathbf{z}'_i \boldsymbol{\gamma})$, where \mathbf{z} is often a subset of \mathbf{x} . For example, in regression of earnings on schooling and other control variables we might model heteroskedasticity more simply as being a function of just a few of the regressors, most notably schooling.

Some types of cross-section data provide a natural model for heteroskedasticity that is very parsimonious. For example, for count data the Poisson density specifies that the variance equals the mean, so $\sigma_i^2 = g(\mathbf{x}_i, \boldsymbol{\beta})$. This provides a working model for heteroskedasticity that introduces no further parameters than those already used in modeling the conditional mean.

This approach of letting the working model for the variance be a function of the mean arises naturally for generalized linear models, introduced in Sections 5.7.3 and 5.7.4. From (5.63) the first-order conditions for the quasi-MLE based on an LEF density are of the form

$$\sum_{i=1}^N \frac{y_i - g(\mathbf{x}_i, \boldsymbol{\beta})}{\sigma_i^2} \times \frac{\partial g(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{0},$$

where $\sigma_i^2 = [c'(g(\mathbf{x}_i, \boldsymbol{\beta}))]^{-1}$ is the assumed variance function corresponding to the particular GLM (see (5.60)). For example, for Poisson, Bernoulli, and exponential distributions σ_i^2 equals, respectively, g_i , $g_i(1 - g_i)$, and $1/g_i^2$, where $g_i = g(\mathbf{x}_i, \boldsymbol{\beta})$.

These first-order conditions can be solved for $\boldsymbol{\beta}$ in one step that allows for dependence of σ_i^2 on $\boldsymbol{\beta}$. In a simpler two-step method one computes $\widehat{\sigma}_i^2 = c'(g(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}))$ given an initial NLS estimate of $\widehat{\boldsymbol{\beta}}$ and then does a weighted NLS regression of $y_i/\widehat{\sigma}_i$ on $g(\mathbf{x}_i, \boldsymbol{\beta})/\widehat{\sigma}_i$. The resulting estimator of $\boldsymbol{\beta}$ is asymptotically equivalent to the quasi-MLE that directly solves (5.63) (see Gouriéroux, Monfort, and Trognan 1984a, or Cameron and Trivedi, 1986). Thus FGNLS is asymptotically equivalent to ML estimation when the density is an LEF density. To guard against misspecification of σ_i^2 inference is based on robust sandwich standard errors, or one lets $\widehat{\sigma}_i^2 = \widehat{\alpha}[c'(g(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}))]^{-1}$, where the estimate $\widehat{\alpha}$ is given in Section 5.7.4.

5.8.7. Time Series

The general NLS result in Proposition 5.6 applies to all types of data, including time-series data. The subsequent results on variance matrix estimation focused on the cross-section case of heteroskedastic errors, but they are easily adapted to the case of time-series data with serially correlated errors. Indeed, results on robust variance matrix estimation using spectral methods for the time-series case preceded those for the cross-section case.

The time-series nonlinear regression model is

$$y_t = g(\mathbf{x}_t, \boldsymbol{\beta}) + u_t, \quad t = 1, \dots, T.$$

If the error u_t is serially correlated it is common to use the **autoregressive moving average** or ARMA(p, q) model

$$u_t = \rho_1 u_{t-1} + \dots + \rho_p u_{t-p} + \varepsilon_t + \alpha_1 \varepsilon_{t-1} + \dots + \alpha_q \varepsilon_{t-q},$$

where ε_t is iid with mean 0 and variance σ^2 , and restrictions may be placed on ARMA model parameters to ensure stationarity and invertibility. The ARMA error model implies a particular structure to the error variance matrix $\Omega_0 = \Omega(\rho, \alpha)$.

The ARMA model provides a good model for Ω_0 in the time-series case. In contrast, in the cross-section case, it is more difficult to correctly model heteroskedasticity, leading to greater emphasis on robust inference that does not require specification of a model for Ω_0 .

What if errors are both heteroskedastic and serially correlated? The NLS estimator is consistent though inefficient if errors are serially correlated, provided \mathbf{x}_t does not include lagged dependent variables in which case it becomes inconsistent. White and Domowitz (1984) generalized (5.79) to obtain a robust estimate of the variance matrix of the NLS estimator given heteroskedasticity and serial correlation of unknown functional form, assuming serial correlation of no more than say, l , lags. In practice a minor refinement due to Newey and West (1987b) is used. This refinement is a rescaling that ensures that the variance matrix estimate is semi-positive definite. Several other refinements have also been proposed and the assumption of fixed lag length has been relaxed so that it is possible for $l \rightarrow \infty$ at a sufficiently slower rate than $N \rightarrow \infty$. This permits an AR component for the error.

5.9. Example: ML and NLS Estimation

Maximum likelihood and NLS estimation, standard error calculation, and coefficient interpretation are illustrated using simulation data.

5.9.1. Model and Estimators

The exponential distribution is used for continuous positive data, notably duration data studied in Chapter 17. The exponential density is

$$f(y) = \lambda e^{-\lambda y}, \quad y > 0, \quad \lambda > 0,$$

with mean $1/\lambda$ and variance $1/\lambda^2$. We introduce regressors into this model by setting

$$\lambda = \exp(\mathbf{x}' \boldsymbol{\beta}),$$

which ensures $\lambda > 0$. Note that this implies that

$$E[y|\mathbf{x}] = \exp(-\mathbf{x}' \boldsymbol{\beta}).$$

An alternative parameterization instead specifies $E[y|\mathbf{x}] = \exp(\mathbf{x}'\boldsymbol{\beta})$, so that $\lambda = \exp(-\mathbf{x}'\boldsymbol{\beta})$. Note that the exponential is used in two different ways: for the density and for the conditional mean.

The OLS estimator from regression of y on \mathbf{x} is inconsistent, since it fits a straight line when the regression function is in fact an exponential curve.

The MLE is easily obtained. The log-density is $\ln f(y|\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta} - y \exp(\mathbf{x}'\boldsymbol{\beta})$, leading to ML first-order conditions $N^{-1} \sum_i (1 - y_i \exp(\mathbf{x}_i'\boldsymbol{\beta})) \mathbf{x}_i = \mathbf{0}$, or

$$N^{-1} \sum_i \frac{y_i - \exp(-\mathbf{x}'\boldsymbol{\beta})}{\exp(-\mathbf{x}'\boldsymbol{\beta})} \mathbf{x}_i = \mathbf{0}.$$

To perform NLS regression, note that the model can also be expressed as a nonlinear regression with

$$y = \exp(-\mathbf{x}'\boldsymbol{\beta}) + u,$$

where the error term u has $E[u|\mathbf{x}] = 0$, though it is heteroskedastic. The first-order conditions for an exponential conditional mean for this model, aside from a sign reversal, have already been given in (5.82) and clearly lead to an estimator that differs from the MLE.

As an example of weighted NLS we suppose that the error variance is proportional to the mean. Then the working variance is $V[y] = E[y]$ and weighted least squares can be implemented by NLS regression of $y_i/\hat{\sigma}_i$ on $\exp(-\mathbf{x}_i'\boldsymbol{\beta})/\hat{\sigma}_i$, where $\hat{\sigma}_i^2 = \exp(-\mathbf{x}_i'\hat{\boldsymbol{\beta}}_{\text{NLS}})$. This estimator is less efficient than the MLE and may or may not be more efficient than NLS.

Feasible generalized NLS can be implemented here, since we know the dgp. Since $V[y] = 1/\lambda^2$ for the exponential density, so the variance equals the mean squared, it follows that $V[u|\mathbf{x}] = [\exp(-\mathbf{x}'\boldsymbol{\beta})]^2$. The FGNLS estimator estimates σ_i^2 by $\hat{\sigma}_i^2 = [\exp(-\mathbf{x}_i'\hat{\boldsymbol{\beta}}_{\text{NLS}})]^2$ and can be implemented by NLS regression of $y_i/\hat{\sigma}_i$ on $\exp(-\mathbf{x}_i'\boldsymbol{\beta})/\hat{\sigma}_i$. In general FGNLS is less efficient than the MLE. In this example it is actually fully efficient as the exponential density is an LEF density (see the discussion at the end of Section 5.8.6).

5.9.2. Simulation and Results

For simplicity we consider regression on an intercept and a regressor. The data-generating process is

$$\begin{aligned} y|\mathbf{x} &\sim \text{exponential}[\lambda], \\ \lambda &= \exp(\beta_1 + \beta_2 x), \end{aligned}$$

where $x \sim \mathcal{N}[1, 1^2]$ and $(\beta_1, \beta_2) = (2, -1)$. A large sample of size 10,000 was drawn to minimize differences in estimates, particularly standard errors, arising from sampling variability. For the particular sample of 10,000 drawn here the sample mean of y is 0.62 and the sample standard deviation of y is 1.29.

Table 5.7 presents OLS, ML, NLS, WNLS, and FGNLS estimates. Up to three different standard error estimates are also given. The default regression output yields nonrobust standard errors, given in parentheses. For OLS and NLS estimators these

Table 5.7. Exponential Example: Least-Squares and ML Estimates^a

Variable	Estimator				
	OLS	ML	NLS	WNLS	FGNLS
Constant	−0.0093 (0.0161) [0.0172]	1.9829 (0.0141) [0.0144]	1.8876 (0.0307) [0.1421] {0.2110}	1.9906 (0.0225) [0.0359]	1.9840 (0.0148) [0.0146]
x	0.6198 (0.0113) [0.0254]	−0.9896 (0.0099) [0.0099]	−0.9575 (0.0097) [0.0612] {0.0880}	−0.9961 (0.0098) [0.0224]	−0.9907 (0.0100) [0.0101]
InL	—	−208.71	−232.98	−208.93	−208.72
R^2	0.2326	0.3906	0.3913	0.3902	0.3906

^a All estimators are consistent, aside from OLS. Up to three alternative standard error estimates are given: nonrobust in parentheses, robust outer product in square brackets, and an alternative robust estimate for NLS in braces. The conditional dgp is an exponential distribution with intercept 2 and slope parameter −1. Sample size $N = 10,000$.

assume iid errors, an erroneous assumption here, and for the MLE these impose the IM equality, a valid assumption here since the assumed density is the dgp. The robust standard errors, given in square brackets, use the robust sandwich variance estimate $N^{-1}\widehat{\mathbf{A}}_H^{-1}\widehat{\mathbf{B}}_{OP}\widehat{\mathbf{A}}_H^{-1}$, where $\widehat{\mathbf{B}}_{OP}$ is the outer product estimated given in (5.38). These estimates are heteroskedastic consistent. For standard errors of the NLS estimator an alternative better estimate is given in braces (and is explained in the next section). The standard error estimates presented here use numerical rather than analytical derivatives in computing $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$.

5.9.3. Comparison of Estimates and Standard Errors

The OLS estimator is inconsistent, yielding estimates unrelated to (β_1, β_2) in the exponential dgp.

The remaining estimators are consistent, and the ML, NLS, WNLS, and FGNLS estimators are within two standard errors of the true parameter values of $(2, -1)$, where the robust standard errors need to be used for NLS. The FGNLS estimates are quite close to the ML estimates, a consequence of using a dgp in the LEF.

For the MLE the nonrobust and robust ML standard errors are quite similar. This is expected as they are asymptotically equivalent (since the information matrix equality holds if the MLE is based on the true density) and the sample size here is large.

For NLS the nonrobust standard errors are invalid, because the dgp has heteroskedastic errors, and greatly overstate the precision of the NLS estimates. The formula for the robust variance matrix estimate for NLS is given in (5.81), where $\widehat{\Omega} = \text{Diag}[\widehat{u}_i^2]$. An alternative that uses $\widehat{\Omega} = \text{Diag}[\widehat{\mathbf{E}}[u_i^2]]$, where $\widehat{\mathbf{E}}[u_i^2] = [\exp(-\mathbf{x}'\widehat{\beta})]^2$, is given in braces. The two estimates differ: 0.0612 compared to 0.0880 for the slope coefficient. The difference arises because $\widehat{u}_i^2 = (y_i - \exp(\mathbf{x}'\widehat{\beta}))^2$ differs from

$[\exp(-\mathbf{x}'_i \hat{\beta})]^2$. More generally standard errors estimated using the outer product (see Section 5.5.2) can be biased even in quite large samples. NLS is considerably less efficient than MLE, with standard errors many times those of the MLE using the preferred estimates in braces.

The WNLS estimator does not use the correct model for heteroskedasticity, so the nonrobust and robust standard errors again differ. Using the robust standard errors the WNLS estimator is more efficient than NLS and less efficient than the MLE.

In this example the FGNLS estimator is as efficient as the MLE, a consequence of the known dgp being in the LEF. The results indicate this, with coefficients and standard errors very close to those for the MLE. The robust and nonrobust standard errors for the FGNLS estimator are essentially the same, as expected since here the model for heteroskedasticity is correctly specified.

Table 5.7 also reports the estimated log-likelihood, $\ln L = \sum_i [\mathbf{x}'_i \hat{\beta} - \exp(-\mathbf{x}'_i \hat{\beta}) y_i]$, and an R-squared measure, $R^2 = 1 - \sum_i (y_i - \hat{y}_i)^2 / \sum_i (y_i - \bar{y})^2$, where $\hat{y}_i = \exp(-\mathbf{x}'_i \hat{\beta})$, evaluated at the ML, NLS, WNLS, and FGNLS estimates. The R^2 differs little across models and is lowest for the NLS estimator, as expected since NLS minimizes $\sum_i (y_i - \hat{y}_i)^2$. The log-likelihood is maximized by the MLE, as expected, and is considerably lower for the NLS estimator.

5.9.4. Coefficient Interpretation

Interest lies in changes in $E[y|x]$ when x changes. We consider the ML estimates of $\hat{\beta}_2 = -0.99$ given in Table 5.7.

The conditional mean $\exp(-\beta_1 - \beta_2 x)$ is of single-index form, so that if an additional regressor z with coefficient β_3 were included, then the marginal effect of a one-unit change in z would be $\hat{\beta}_3/\hat{\beta}_2$ times that of a one-unit change in x (see Section 5.2.4).

The conditional mean is monotonically decreasing in x , so the sign of $\hat{\beta}_2$ is the reverse of the marginal effect (see Section 5.2.4). Here the marginal effect of an increase in x is an increase in the conditional mean, since $\hat{\beta}_2$ is negative.

We now consider the magnitude of the marginal effect of changes in x using calculus methods. Here $\partial E[y|x]/\partial x = -\beta_2 \exp(-\mathbf{x}' \beta)$ varies with the evaluation point x and ranges from 0.01 to 19.09 in the sample. The sample-average response is $0.99N^{-1} \sum_i \exp(\mathbf{x}'_i \hat{\beta}) = 0.61$. The response evaluated at the sample mean of x , $0.99 \exp(\bar{\mathbf{x}}' \hat{\beta}) = 0.37$, is considerably smaller. Since $\partial E[y|x]/\partial x = -\beta_2 E[y|x]$, yet another estimate of the marginal effect is $0.99\bar{y} = 0.61$.

Finite-difference methods lead to a different estimated marginal effect. For $\Delta x = 1$ we obtain $\Delta E[y|x] = (\exp(\beta_2 + 1) - \exp(\beta_2)) \exp(-\mathbf{x}' \beta)$ (see Section 5.2.4). This yields an average response over the sample of 1.04, rather than 0.61. The finite-difference and calculus methods coincide, however, if Δx is small.

The preceding marginal effects are additive. For the exponential conditional mean we can also consider multiplicative or proportionate marginal effects (see Section 5.2.4). For example, a 0.1-unit change in x is predicted to lead to a proportionate increase in $E[y|x]$ of 0.1×0.99 or a 9.9% increase. Again a finite-difference approach will yield a different estimate.

Which of these measures is most useful? The restriction to single-index form is very useful as the relative impact of regressors can be immediately calculated. For the magnitude of the response it is most accurate to compute the average response across the sample, using noncalculus methods, of a c -unit change in the regressor, where the magnitude of c is a meaningful amount such as a one standard deviation change in x .

Similar calculations can be done for the NLS, WNLS, and FGNLS estimates, with similar results. For the OLS estimator, note that the coefficient of x can be interpreted as giving the sample-average marginal effect of a change in x (see Section 4.7.2). Here the OLS estimate $\hat{\beta}_2 = 0.61$ equals to two decimal places the sample-average response computed earlier using the exponential MLE. Here OLS provides a good estimate of the sample-average marginal response, even though it can provide a very poor estimate of the marginal response for any particular value of x .

5.10. Practical Considerations

Most econometrics packages provide simple commands to obtain the maximum likelihood estimators for the standard models introduced in Section 5.6.1. For other densities many packages provide an ML routine to which the user provides the equation for the density and possibly first derivatives or even second derivatives. Similarly, for NLS one provides the equation for the conditional mean to an NLS routine. For some nonlinear models and data sets the ML and NLS routines provided in packages can encounter computational difficulties in obtaining estimates. In such circumstances it may be necessary to use more robust optimization routines provided as add-on modules to Gauss, Matlab and OX. Gauss, Matlab and OX are better tools for nonlinear modeling, but require a higher initial learning investment.

For cross-section data it is becoming standard to use standard errors based on the sandwich form of the variance matrix. These are often provided as a command option. For LS estimators this gives heteroskedastic-consistent standard errors. For maximum likelihood one should be aware that misspecification of the density can lead to inconsistency in addition to requiring the use of sandwich errors.

The parameters of nonlinear models are usually not directly interpretable, and it is good practice to additionally compute the implied marginal effects caused by changes in regressors (see Section 5.2.4). Some packages do this automatically; for others several lines of postestimation code using saved regression coefficients may be needed.

5.11. Bibliographic Notes

A brief history of the development of asymptotic theory results for extremum estimators is given in Newey and McFadden (1994, p. 2115). A major econometrics advance was made by Amemiya (1973), who developed quite general theorems that were applied to the Tobit model MLE. Useful book-length treatments include those by Gallant (1987), Gallant and White (1987), Bierens (1993), and White (1994, 2001a). Statistical foundations are given in many books, including Amemiya (1985, Chapter 3), Davidson and MacKinnon (1993, Chapter 4),

Greene (2003, appendix D), Davidson (1994), and Zaman (1996).

- 5.3** The presentation of general extremum estimation results draws heavily on Amemiya (1985, Chapter 4), and to a lesser extent on Newey and McFadden (1994). The latter reference is very comprehensive.
- 5.4** The estimating equations approach is used in the generalized linear models literature (see McCullagh and Nelder, 1989). Econometricians subsume this in generalized method of moments (see Chapter 6).
- 5.5** Statistical inference is presented in detail in Chapter 7.
- 5.6** See the pioneering article by Fisher (1922) for general results for ML estimation, including efficiency, and for comparison of the likelihood approach with the inverse-probability or Bayesian approach and with method of moments estimation.
- 5.7** Modern applications frequently use the quasi-ML framework and sandwich estimates of the variance matrix (see White, 1982, 1994). In statistics the approach is called generalized linear models, with McCullagh and Nelder (1989) a standard reference.
- 5.8** Similarly for NLS estimation, sandwich estimates of the variance matrix are used that require relatively weak assumptions on the error process. The papers by White (1980a,c) had a big impact on statistical inference in econometrics. Generalization and a detailed review of the asymptotic theory is given in White and Domowitz (1984). Amemiya (1983) has extensively surveyed methods for nonlinear regression.

Exercises

- 5–1** Suppose we obtain model estimates that yield predicted conditional mean $\hat{E}[y|x] = \exp(1 + 0.01x)/[1 + \exp(1 + 0.01x)]$. Suppose the sample is of size 100 and x takes integer values 1, 2, ..., 100. Obtain the following estimates of the estimated marginal effect $\partial\hat{E}[y|x]/\partial x$.
- The average marginal effect over all observations.
 - The marginal effect of the average observation.
 - The marginal effect when $x = 90$.
 - The marginal effect of a one-unit change when $x = 90$, computed using the finite-difference method.
- 5–2** Consider the following special one-parameter case of the gamma distribution, $f(y) = (y/\lambda^2) \exp(-y/\lambda)$, $y > 0$, $\lambda > 0$. For this distribution it can be shown that $E[y] = 2\lambda$ and $V[y] = 2\lambda^2$. Here we introduce regressors and suppose that in the true model the parameter λ depends on regressors according to $\lambda_i = \exp(\mathbf{x}_i'\beta)/2$. Thus $E[y_i|\mathbf{x}_i] = \exp(\mathbf{x}_i'\beta)$ and $V[y_i|\mathbf{x}_i] = [\exp(\mathbf{x}_i'\beta)]^2/2$. Assume the data are independent over i and \mathbf{x}_i is nonstochastic and $\beta = \beta_0$ in the dgp.
- Show that the log-likelihood function (scaled by N^{-1}) for this gamma model is $Q_N(\beta) = N^{-1} \sum_i \{\ln y_i - 2\mathbf{x}_i'\beta + 2 \ln 2 - 2y_i \exp(-\mathbf{x}_i'\beta)\}$.
 - Obtain $\text{plim } Q_N(\beta)$. You can assume that assumptions for any LLN used are satisfied. [Hint: $E[\ln y_i]$ depends on β_0 but not β .]
 - Prove that $\hat{\beta}$ that is the local maximum of $Q_N(\beta)$ is consistent for β_0 . State any assumptions made.
 - Now state what LLN you would use to verify part (b) and what additional information, if any, is needed to apply this law. A brief answer will do. There is no need for a formal proof.

5–3 Continue with the gamma model of Exercise 5–2.

- (a) Show that $\partial Q_N(\beta)/\partial\beta = N^{-1} \sum_i 2[(y_i - \exp(\mathbf{x}_i'\beta))/\exp(\mathbf{x}_i'\beta)]\mathbf{x}_i$.
- (b) What essential condition indicated by the first-order conditions needs to be satisfied for $\hat{\beta}$ to be consistent?
- (c) Apply a central limit theorem to obtain the limit distribution of $\sqrt{N}\partial Q_N/\partial\beta|_{\beta_0}$. Here you can assume that the assumptions necessary for a CLT are satisfied.
- (d) State what CLT you would use to verify part (c) and what additional information, if any, is needed to apply this law. A brief answer will do. There is no need for a formal proof.
- (e) Obtain the probability limit of $\partial^2 Q_N/\partial\beta\partial\beta'|_{\beta_0}$.
- (f) Combine the previous results to obtain the limit distribution of $\sqrt{N}(\hat{\beta} - \beta_0)$.
- (g) Given part (f), state how to test $H_0 : \beta_{0j} \geq \beta_j^*$ against $H_a : \beta_{0j} < \beta_j^*$ at level 0.05, where β_j is the j th component of β .

5–4 A nonnegative integer variable y that is geometric distributed has density (or more formally probability mass function) $f(y) = (y+1)(2\lambda)^y(1+2\lambda)^{-(y+0.5)}$, $y = 0, 1, 2, \dots$, $\lambda > 0$. Then $E[y] = \lambda$ and $V[y] = \lambda(1+2\lambda)$. Introduce regressors and suppose $y_i = \exp(\mathbf{x}_i'\beta)$. Assume the data are independent over i and \mathbf{x}_i is non-stochastic and $\beta = \beta_0$ in the dgp.

- (a) Repeat Exercise 5–2 for this model.
- (b) Repeat Exercise 5–3 for this model.

5–5 Suppose a sample yields estimates $\hat{\theta}_1 = 5$, $\hat{\theta}_2 = 3$, $se[\hat{\theta}_1] = 2$, and $se[\hat{\theta}_2] = 1$ and the correlation coefficient between $\hat{\theta}_1$ and $\hat{\theta}_2$ equals 0.5. Perform the following tests at level 0.05, assuming asymptotic normality of the parameter estimates.

- (a) Test $H_0 : \theta_1 = 0$ against $H_a : \theta_1 \neq 0$.
- (b) Test $H_0 : \theta_1 = 2\theta_2$ against $H_a : \theta_1 \neq 2\theta_2$.
- (c) Test $H_0 : \theta_1 = 0, \theta_2 = 0$ against H_a : at least one of $\theta_1, \theta_2 \neq 0$.

5–6 Consider the nonlinear regression model $y = \exp(\mathbf{x}'\beta)/[1 + \exp(\mathbf{x}'\beta)] + u$, where the error term is possibly heteroskedastic.

- (a) Within what range does this restrict $E[y|\mathbf{x}]$ to lie?
- (b) Give the first-order conditions for the NLS estimator.
- (c) Obtain the asymptotic distribution of the NLS estimator using result (5.77).

5–7 This question presumes access to software that allows NLS and ML estimation. Consider the gamma regression model of Exercise 5–2. An appropriate gamma variate can be generated using $y = -\lambda \ln r_1 - \lambda \ln r_2$, where $\lambda = \exp(\mathbf{x}'\beta)/2$ and r_1 and r_2 are random draws from Uniform[0, 1]. Let $\mathbf{x}'\beta = \beta_1 + \beta_2 x$. Generate a sample of size 1,000 when $\beta_1 = -1.0$ and $\beta_2 = 1$ and $\mathbf{x} \sim \mathcal{N}[0, 1]$.

- (a) Obtain estimates of β_1 and β_2 from NLS regression of y on $\exp(\beta_1 + \beta_2 x)$.
- (b) Should sandwich standard errors be used here?
- (c) Obtain ML estimates of β_1 and β_2 from NLS regression of y on $\exp(\beta_1 + \beta_2 x)$.
- (d) Should sandwich standard errors be used here?

CHAPTER 6

Generalized Method of Moments and Systems Estimation

6.1. Introduction

The previous chapter focused on m-estimation, including ML and NLS estimation. Now we consider a much broader class of extremum estimators, those based on method of moments (MM) and generalized method of moments (GMM).

The basis of MM and GMM is specification of a set of population moment conditions involving data and unknown parameters. The MM estimator solves the sample moment conditions that correspond to the population moment conditions. For example, the sample mean is the MM estimator of the population mean. In some cases there may be no explicit analytical solution for the MM estimator, but numerical solution may still be possible. Then the estimator is an example of the estimating equations estimator introduced briefly in Section 5.4.

In some situations, however, MM estimation may be infeasible because there are more moment conditions and hence equations to solve than there are parameters. A leading example is IV estimation in an overidentified model. The GMM estimator, due to Hansen (1982), extends the MM approach to accommodate this case.

The GMM estimator defines a class of estimators, with different GMM estimators obtained by using different population moment conditions, just as different specified densities lead to different ML estimators. We emphasize this moment-based approach to estimation, even in cases where alternative presentations are possible, as it provides a unified approach to estimation and can provide an obvious way to extend methods from linear to nonlinear models.

The basics of GMM estimation are given in Sections 6.2 and 6.3, which present, respectively, expository examples and asymptotic results for statistical inference. The remainder of the chapter details more specialized estimators. Instrumental variables estimators are presented in Sections 6.4 and 6.5. For linear models the treatment in Sections 4.8 and 4.9 may be sufficient, but extension to nonlinear models uses the GMM approach. Section 6.6 covers methods to compute standard errors of sequential two-step m-estimators. Sections 6.7 and 6.8 present the minimum distance estimator, a variant of GMM, and the empirical likelihood estimator, an alternative estimator to

GMM. Systems estimation methods, used in a relatively small fraction of microeconomics studies, are discussed in Sections 6.9 and 6.10.

This chapter reviews many estimation methods from a GMM perspective. Applications of these methods to actual data include a linear IV application in Section 4.9.6 and a linear panel GMM application in Section 22.3.

6.2. Examples

GMM estimators are based on the analogy principle (see Section 5.4.2) that population moment conditions lead to sample moment conditions that can be used to estimate parameters. This section provides several leading applications of this principle, with properties of the resulting estimator deferred to Section 6.3.

6.2.1. Linear Regression

A classic example of **method of moments** is estimation of the population mean when y is iid with mean μ . In the population

$$E[y - \mu] = 0.$$

Replacing the expectations operator $E[\cdot]$ for the population by the average operator $N^{-1} \sum_{i=1}^N (\cdot)$ for the sample yields the corresponding sample moment

$$\frac{1}{N} \sum_{i=1}^N (y_i - \mu) = 0.$$

Solving for μ leads to the estimator $\hat{\mu}_{\text{MM}} = N^{-1} \sum_i y_i = \bar{y}$. The MM estimate of the population mean is the sample mean.

This approach can be extended to the *linear regression model* $y = \mathbf{x}'\beta + u$, where \mathbf{x} and β are $K \times 1$ vectors. Suppose the error term u has zero mean conditional on regressors. The single conditional moment restriction $E[u|\mathbf{x}] = 0$ leads to K unconditional moment conditions $E[\mathbf{x}u] = \mathbf{0}$, since

$$E[\mathbf{x}u] = E_{\mathbf{x}}[E[\mathbf{x}u|\mathbf{x}]] = E_{\mathbf{x}}[\mathbf{x}E[u|\mathbf{x}]] = E_{\mathbf{x}}[\mathbf{x}\cdot 0] = \mathbf{0}, \quad (6.1)$$

using the **law of iterated expectations** (see Section A.8) and the assumption that $E[u|\mathbf{x}] = 0$. Thus

$$E[\mathbf{x}(y - \mathbf{x}'\beta)] = \mathbf{0},$$

if the error has conditional mean zero. The MM estimator is the solution to the corresponding sample moment condition

$$\frac{1}{N} \sum_{i=1}^N \mathbf{x}_i(y_i - \mathbf{x}'_i\beta) = \mathbf{0}.$$

This yields $\hat{\beta}_{\text{MM}} = (\sum_i \mathbf{x}_i \mathbf{x}'_i)^{-1} \sum_i \mathbf{x}_i y_i$.

The OLS estimator is therefore a special case of MM estimation. The MM derivation of the OLS estimator, however, differs significantly from the usual one of minimization of a sum of squared residuals.

6.2.2. Nonlinear Regression

For nonlinear regression the method of moments approach reduces to NLS if regression errors are additive. For more general nonlinear regression with nonadditive errors (defined in the following) method of moments yields a consistent estimator whereas NLS is inconsistent.

From Section 5.8.3 the nonlinear regression model with **additive error** is a model that specifies

$$y = g(\mathbf{x}, \boldsymbol{\beta}) + u.$$

A moment approach similar to that for the linear model yields that $E[u|\mathbf{x}] = 0$ implies that $E[\mathbf{h}(\mathbf{x})(y - \mathbf{x}'\boldsymbol{\beta})] = \mathbf{0}$, where $\mathbf{h}(\mathbf{x})$ is any function of \mathbf{x} . The particular choice $\mathbf{h}(\mathbf{x}) = \partial g(\mathbf{x}, \boldsymbol{\beta})/\partial \boldsymbol{\beta}$, motivated in Section 6.3.7, leads to corresponding sample moment condition that equals the first-order conditions for the NLS estimator given in Section 5.8.2.

The more general nonlinear regression model with **nonadditive error** specifies

$$u = r(y, \mathbf{x}, \boldsymbol{\beta}),$$

where again $E[u|\mathbf{x}] = 0$ but now y is no longer restricted to being an additive function of u . For example, in Poisson regression one may define the standardized error $u = [y - \exp(\mathbf{x}'\boldsymbol{\beta})]/[\exp(\mathbf{x}'\boldsymbol{\beta})]^{1/2}$ that has $E[u|\mathbf{x}] = 0$ and $V[u|\mathbf{x}] = 1$ since y has conditional mean and variance equal to $\exp(\mathbf{x}'\boldsymbol{\beta})$.

The NLS estimator is inconsistent given nonadditive error. Minimizing $N^{-1} \sum_i u_i^2 = N^{-1} \sum_i r(y_i, \mathbf{x}_i, \boldsymbol{\beta})^2$ leads to first-order conditions

$$\frac{1}{N} \sum_{i=1}^N \frac{\partial r(y_i, \mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} r(y_i, \mathbf{x}_i, \boldsymbol{\beta}) = \mathbf{0}.$$

Here y_i appears in both terms in the product and there is no guarantee that this product has expected value of zero even if $E[r(\cdot)|\mathbf{x}] = 0$. This inconsistency did not arise with additive errors $r(\cdot) = y - g(\mathbf{x}, \boldsymbol{\beta})$, as then $\partial r(\cdot)/\partial \boldsymbol{\beta} = -\partial g(\mathbf{x}, \boldsymbol{\beta})/\partial \boldsymbol{\beta}$, so only the second term in the product depended on y .

A moment-based approach yields a consistent estimator. The assumption that $E[u|\mathbf{x}] = 0$ implies

$$E[\mathbf{h}(\mathbf{x})r(y, \mathbf{x}, \boldsymbol{\beta})] = \mathbf{0},$$

where $\mathbf{h}(\mathbf{x})$ is a function of \mathbf{x} . If $\text{dim}[\mathbf{h}(\mathbf{x})] = K$ then the corresponding sample moment

$$\frac{1}{N} \sum_{i=1}^N \mathbf{h}(\mathbf{x}_i)r(y_i, \mathbf{x}_i, \boldsymbol{\beta}) = \mathbf{0}$$

yields a consistent estimate of $\boldsymbol{\beta}$, where solution is by numerical methods.

6.2.3. Maximum Likelihood

The Kullback–Leibler information criterion was defined in Section 5.7.2. From this definition, a local maximum of KLIC occurs if $E[\mathbf{s}(\boldsymbol{\theta})] = \mathbf{0}$, where $\mathbf{s}(\boldsymbol{\theta}) = \partial \ln f(y|\mathbf{x}, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ and $f(y|\mathbf{x}, \boldsymbol{\theta})$ is the conditional density.

Replacing population moments by sample moments yields an estimator $\hat{\boldsymbol{\theta}}$ that solves $N^{-1} \sum_i \mathbf{s}_i(\boldsymbol{\theta}) = \mathbf{0}$. These are the ML first-order conditions, so the MLE can be motivated as an MM estimator.

6.2.4. Additional Moment Restrictions

Using additional moments can improve the efficiency of estimation but requires adaptation of regular method of moments if there are more moment conditions than parameters to estimate.

A simple example of an inefficient estimator is the sample mean. This is an inefficient estimator of the population mean unless the data are a random sample from the normal distribution or some other member of the exponential family of distributions. One way to improve efficiency is to use alternative estimators. The sample median, consistent for μ if the distribution is symmetric, may be more efficient. Obviously the MLE could be used if the distribution is fully specified, but here we instead improve efficiency by using additional moment restrictions.

Consider estimation of $\boldsymbol{\beta}$ in the linear regression model. The OLS estimator is inefficient even assuming homoskedastic errors, unless errors are normally distributed. From Section 6.2.1, the OLS estimator is an MM estimator based on $E[\mathbf{x}u] = \mathbf{0}$. Now make the additional moment assumption that errors are conditionally symmetric, so that $E[u^3|\mathbf{x}] = 0$ and hence $E[\mathbf{x}u^3] = \mathbf{0}$. Then estimation of $\boldsymbol{\beta}$ may be based on the $2K$ moment conditions

$$\begin{bmatrix} E[\mathbf{x}(y - \mathbf{x}'\boldsymbol{\beta})] \\ E[\mathbf{x}(y - \mathbf{x}'\boldsymbol{\beta})^3] \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}.$$

The MM estimator would attempt to estimate $\boldsymbol{\beta}$ as the solution to the corresponding sample moment conditions $N^{-1} \sum_i \mathbf{x}_i(y_i - \mathbf{x}'_i \boldsymbol{\beta}) = \mathbf{0}$ and $N^{-1} \sum_i \mathbf{x}_i(y_i - \mathbf{x}'_i \boldsymbol{\beta})^3 = \mathbf{0}$. However, with $2K$ equations and only K unknown parameters $\boldsymbol{\beta}$, it is not possible for all of these sample moment conditions to be satisfied.

The GMM estimator instead sets the sample moments as close to zero as possible using quadratic loss. Then $\hat{\boldsymbol{\beta}}_{\text{GMM}}$ minimizes

$$Q_N(\boldsymbol{\beta}) = \begin{bmatrix} \frac{1}{N} \sum_i \mathbf{x}_i u_i \\ \frac{1}{N} \sum_i \mathbf{x}_i u_i^3 \end{bmatrix}' \mathbf{W}_N \begin{bmatrix} \frac{1}{N} \sum_i \mathbf{x}_i u_i \\ \frac{1}{N} \sum_i \mathbf{x}_i u_i^3 \end{bmatrix}, \quad (6.2)$$

where $u_i = y_i - \mathbf{x}'_i \boldsymbol{\beta}$ and \mathbf{W}_N is a $2K \times 2K$ weighting matrix. For some choices of \mathbf{W}_N this estimator is more efficient than OLS. This example is analyzed in Section 6.3.6.

6.2.5. Instrumental Variables Regression

Instrumental variables estimation is a leading example of generalized method of moments estimation.

Consider the linear regression model $y = \mathbf{x}'\beta + u$, with the complication that some components of \mathbf{x} are correlated with the error term so that OLS is inconsistent for β . Assume the existence of **instruments** \mathbf{z} (introduced in Section 4.8) that are correlated with \mathbf{x} but satisfy $E[u|\mathbf{z}] = 0$. Then $E[y - \mathbf{x}'\beta|\mathbf{z}] = 0$. Using algebra similar to that used to obtain (6.1) for the OLS example, we multiply by \mathbf{z} to get the K unconditional population moment conditions

$$E[\mathbf{z}(y - \mathbf{x}'\beta)] = \mathbf{0}. \quad (6.3)$$

The method of moments estimator solves the corresponding sample moment condition

$$\frac{1}{N} \sum_{i=1}^N \mathbf{z}_i(y_i - \mathbf{x}'_i\beta) = \mathbf{0}.$$

If $\dim(\mathbf{z}) = K$ this yields $\widehat{\beta}_{MM} = (\sum_i \mathbf{z}_i \mathbf{x}'_i)^{-1} \sum_i \mathbf{z}_i y_i$, which is the linear IV estimator introduced in Section 4.8.6.

No unique solution exists if there are more potential instruments than regressors, since then $\dim(\mathbf{z}) > K$ and there are more equations than unknowns. One possibility is to use just K instruments, but there is then an efficiency loss. The GMM estimator instead chooses $\widehat{\beta}$ to make the vector $N^{-1} \sum_i \mathbf{z}_i(y_i - \mathbf{x}'_i\beta)$ as small as possible using quadratic loss, so that $\widehat{\beta}_{GMM}$ minimizes

$$Q_N(\beta) = \left[\frac{1}{N} \sum_{i=1}^N \mathbf{z}_i(y_i - \mathbf{x}'_i\beta) \right]' \mathbf{W}_N \left[\frac{1}{N} \sum_{i=1}^N \mathbf{z}_i(y_i - \mathbf{x}'_i\beta) \right], \quad (6.4)$$

where \mathbf{W}_N is a $\dim(\mathbf{z}) \times \dim(\mathbf{z})$ weighting matrix. The 2SLS estimator (see Section 4.8.6) corresponds to a particular choice of \mathbf{W}_N .

Instrumental variables methods for linear models are presented in considerable detail in Section 6.4. An advantage of the GMM approach is that it provides a way to specify the optimal choice of weighting matrix \mathbf{W}_N , leading to an estimator more efficient than 2SLS.

Section 6.5 covers IV methods for nonlinear models. One advantage of the GMM approach is that generalization to nonlinear regression is straightforward. Then we simply replace $y - \mathbf{x}'\beta$ in the preceding expression for $Q_N(\beta)$ by the nonlinear model error $u = y - g(\mathbf{x}'\beta)$ or $u = r(y, \mathbf{x}, \beta)$.

6.2.6. Panel Data

Another leading application of GMM and related estimation methods is to panel data regression.

As an example, suppose $y_{it} = \mathbf{x}'_{it}\beta + u_{it}$, where i denotes individual and t denotes time. From Section 6.2.1, pooled OLS regression of y_{it} on \mathbf{x}_{it} is an MM estimator based on the condition $E[\mathbf{x}_{it}u_{it}] = \mathbf{0}$. Suppose it is additionally assumed that the error u_{it} is uncorrelated with regressors in periods other than the current period. Then

6.2. EXAMPLES

$E[\mathbf{x}_{is}u_{it}] = \mathbf{0}$ for $s \neq t$ provides additional moment conditions that can be used to obtain more efficient estimators.

Chapters 22 and 23 provide many applications of GMM methods to panel data.

6.2.7. Moment Conditions from Economic Theory

Economic theory can generate moment conditions that can be used as the basis for estimation.

Begin with the model

$$y_t = E[y_t | \mathbf{x}_t, \beta] + u_t,$$

where the first term on the right-hand side measures the “anticipated” component of y conditional on \mathbf{x} and the second component measures the “unanticipated” component. As examples, y may denote return on an asset or the rate of inflation. Under the twin assumptions of rational expectations and market clearing or market efficiency, we may obtain the result that the unanticipated component is unpredictable using any information that was available at time t for determining $E[y|\mathbf{x}]$. Then

$$E[(y_t - E[y_t | \mathbf{x}_t, \beta]) | \mathcal{I}_t] = 0,$$

where \mathcal{I}_t denotes information available at time t .

By the law of iterated expectations, $E[\mathbf{z}_t(y_t - E[y_t | \mathbf{x}_t, \beta])] = 0$, where \mathbf{z}_t is formed from any subset of \mathcal{I}_t . Since any part of the information set can be used as an instrument, this provides many moment conditions that can be the basis of estimation. If time-series data are available then GMM minimizes the quadratic form

$$Q_T(\beta) = \left[\frac{1}{T} \sum_{t=1}^T \mathbf{z}_t u_t \right]' \mathbf{W}_T \left[\frac{1}{T} \sum_{t=1}^T \mathbf{z}_t u_t \right],$$

where $u_t = y_t - E[y_t | \mathbf{x}_t, \beta]$. If cross-section data are available at a single time point t then GMM minimizes the quadratic form

$$Q_N(\beta) = \left[\frac{1}{N} \sum_{i=1}^N \mathbf{z}_i u_i \right]' \mathbf{W}_N \left[\frac{1}{N} \sum_{i=1}^N \mathbf{z}_i u_i \right],$$

where $u_i = y_i - E[y_i | \mathbf{x}_i, \beta]$ and the subscript t can be dropped as only one time period is analyzed.

This approach is not restricted to the additive structure used in motivation. All that is needed is an error u_t with the property that $E[u_t | \mathcal{I}_t] = 0$. Such conditions arise from the Euler conditions from intertemporal models of decision making under certainty. For example, Hansen and Singleton (1982) present a model of maximization of expected lifetime utility that leads to the Euler condition $E[u_t | \mathcal{I}_t] = 0$, where $u_t = \beta g_{t+1}^\alpha r_{t+1} - 1$, $g_{t+1} = c_{t+1}/c_t$ is the ratio of consumption in two periods, and r_{t+1} is asset return. The parameters β and α , the intertemporal discount rate and the coefficient of relative risk aversion, respectively, can be estimated by GMM using either time-series or cross-section data as was done previously, with this new definition of u_t . Hansen (1982) and Hansen and Singleton (1982) consider time-series data; MacCurdy (1983) modeled both consumption and labor supply using panel data.

Table 6.1. Generalized Method of Moments: Examples

Moment Function $\mathbf{h}(\cdot)$	Estimation Method
$y - \mu$	Method of moments for population mean
$\mathbf{x}(y - \mathbf{x}'\boldsymbol{\beta})$	Ordinary least-squares regression
$\mathbf{z}(y - \mathbf{x}'\boldsymbol{\beta})$	Instrumental variables regression
$\partial \ln f(y \mathbf{x}, \boldsymbol{\theta})/\partial \boldsymbol{\theta}$	Maximum likelihood estimation

6.3. Generalized Method of Moments

This section presents the general theory of GMM estimation. Generalized method of moments defines a class of estimators. Different choice of moment condition and weighting matrix lead to different GMM estimators, just as different choices of distribution lead to different ML estimators. We address these issues, in addition to presenting the usual properties of consistency and asymptotic normality and methods to estimate the variance matrix of the GMM estimator.

6.3.1. Method of Moments Estimator

The starting point is to assume the existence of r moment conditions for q parameters,

$$\mathbb{E}[\mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}_0)] = \mathbf{0}, \quad (6.5)$$

where $\boldsymbol{\theta}$ is a $q \times 1$ vector, $\mathbf{h}(\cdot)$ is an $r \times 1$ vector function with $r \geq q$, and $\boldsymbol{\theta}_0$ denotes the value of $\boldsymbol{\theta}$ in the dgp. The vector \mathbf{w} includes all observables including, where relevant, a dependent variable \mathbf{y} , potentially endogenous regressors \mathbf{x} , and instrumental variables \mathbf{z} . The dependent variable \mathbf{y} may be a vector, so that applications with systems of equations or with panel data are subsumed. The expectation is with respect to all stochastic components of \mathbf{w} and hence \mathbf{y} , \mathbf{x} , and \mathbf{z} .

The choice of functional form for $\mathbf{h}(\cdot)$ is qualitatively similar to the choice of model and will vary with application. Table 6.1 summarizes some single-equation examples of $\mathbf{h}(\mathbf{w}) = \mathbf{h}(y, \mathbf{x}, \mathbf{z}, \boldsymbol{\theta})$ already presented in Section 6.2.

If $r = q$ then method of moments can be applied. Equality to zero of the population moment is replaced by equality to zero of the corresponding sample moment, and the **method of moments estimator** $\widehat{\boldsymbol{\theta}}_{MM}$ is defined to be the solution to

$$\frac{1}{N} \sum_{i=1}^N \mathbf{h}(\mathbf{w}_i, \widehat{\boldsymbol{\theta}}) = \mathbf{0}. \quad (6.6)$$

This is an estimating equations estimator that equivalently minimizes

$$\mathcal{Q}_N(\boldsymbol{\theta}) = \left[\frac{1}{N} \sum_{i=1}^N \mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}) \right]' \left[\frac{1}{N} \sum_{i=1}^N \mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}) \right],$$

with asymptotic distribution presented in Section 5.4 and reproduced in (6.13) in Section 6.3.3.

6.3.2. GMM Estimator

The GMM estimator is based on r independent moment conditions (6.5) while q parameters are estimated.

If $r = q$ the model is said to be **just-identified** and the MM estimator in (6.6) can be used. More formally $r = q$ is only a necessary condition for just-identification and we additionally require that \mathbf{G}_0 in Proposition 5.1 is of rank q . Identification is addressed in Section 6.3.9.

If $r > q$ the model is said to be **overidentified** and (6.6) has no solution for $\hat{\theta}$ as there are more equations (r) than unknowns (q). Instead, $\hat{\theta}$ is chosen so that a quadratic form in $N^{-1} \sum_i \mathbf{h}(\mathbf{w}_i, \hat{\theta})$ is as close to zero as possible. Specifically, the **generalized methods of moments estimator** $\hat{\theta}_{\text{GMM}}$ minimizes the objective function

$$Q_N(\theta) = \left[\frac{1}{N} \sum_{i=1}^N \mathbf{h}(\mathbf{w}_i, \theta) \right]' \mathbf{W}_N \left[\frac{1}{N} \sum_{i=1}^N \mathbf{h}(\mathbf{w}_i, \theta) \right], \quad (6.7)$$

where the $r \times r$ weighting matrix \mathbf{W}_N is symmetric positive definite, possibly stochastic with finite probability limit, and does not depend on θ . The subscript N on \mathbf{W}_N is used to indicate that its value may depend on the sample. The dimension r of \mathbf{W}_N , however, is fixed as $N \rightarrow \infty$. The objective function can also be expressed in matrix notation as $Q_N(\theta) = N^{-1} \mathbf{l}' \mathbf{H}(\theta) \times \mathbf{W}_N \times N^{-1} \mathbf{H}(\theta)' \mathbf{l}$, where \mathbf{l} is an $N \times 1$ vector of ones and $\mathbf{H}(\theta)$ is an $N \times r$ matrix with i th row $\mathbf{h}(y_i, \mathbf{x}_i, \theta)'$.

Different choices of weighting matrix \mathbf{W}_N lead to different estimators that, although consistent, have different variances if $r > q$. A simple choice, though often a poor choice, is to let \mathbf{W}_N be the identity matrix. Then $Q_N(\theta) = \bar{h}_1^2 + \bar{h}_2^2 + \cdots + \bar{h}_r^2$ is the sum of r squared sample averages, where $\bar{h}_j = N^{-1} \sum_i h_j(\mathbf{w}_i, \theta)$ and $h_j(\cdot)$ is the j th component of $\mathbf{h}(\cdot)$. The optimal choice of \mathbf{W}_N is given in Section 6.3.5.

Differentiating $Q_N(\theta)$ in (6.7) with respect to θ yields the GMM first-order conditions

$$\left[\frac{1}{N} \sum_{i=1}^N \frac{\partial \mathbf{h}_i(\hat{\theta})'}{\partial \theta} \Bigg|_{\hat{\theta}} \right] \times \mathbf{W}_N \times \left[\frac{1}{N} \sum_{i=1}^N \mathbf{h}_i(\hat{\theta}) \right] = \mathbf{0}, \quad (6.8)$$

where $\mathbf{h}_i(\theta) = \mathbf{h}_i(\mathbf{w}_i, \theta)$ and we have multiplied by the scaling factor $1/2$. These equations will generally be nonlinear in $\hat{\theta}$ and can be quite complicated to solve as $\hat{\theta}$ may appear in both the first and third terms. Numerical solution methods are presented in Chapter 10.

6.3.3. Distribution of GMM Estimator

The asymptotic distribution of the GMM estimator is given in the following proposition, derived in Section 6.3.9.

Proposition 6.1 (Distribution of GMM Estimator): *Make the following assumptions:*

- (i) *The dgp imposes the moment condition (6.5); that is, $E[\mathbf{h}(\mathbf{w}, \theta_0)] = \mathbf{0}$.*
- (ii) *The $r \times 1$ vector function $\mathbf{h}(\cdot)$ satisfies $\mathbf{h}(\mathbf{w}, \theta^{(1)}) = \mathbf{h}(\mathbf{w}, \theta^{(2)})$ iff $\theta^{(1)} = \theta^{(2)}$.*

(iii) The following $r \times q$ matrix exists and is finite with rank q :

$$\mathbf{G}_0 = \text{plim} \frac{1}{N} \sum_{i=1}^N \left[\frac{\partial \mathbf{h}_i}{\partial \boldsymbol{\theta}'} \Big|_{\boldsymbol{\theta}_0} \right]. \quad (6.9)$$

(iv) $\mathbf{W}_N \xrightarrow{P} \mathbf{W}_0$, where \mathbf{W}_0 is finite symmetric positive definite.

(v) $N^{-1/2} \sum_{i=1}^N \mathbf{h}_i |_{\boldsymbol{\theta}_0} \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{S}(\boldsymbol{\theta}_0)]$, where

$$\mathbf{S}_0 = \text{plim} N^{-1} \sum_{i=1}^N \sum_{j=1}^N \left[\mathbf{h}_i \mathbf{h}'_j \Big|_{\boldsymbol{\theta}_0} \right]. \quad (6.10)$$

Then the **GMM estimator** $\hat{\boldsymbol{\theta}}_{\text{GMM}}$, defined to be a root of the first-order conditions $\partial Q_N(\boldsymbol{\theta})/\partial \boldsymbol{\theta} = \mathbf{0}$ given in (6.8), is consistent for $\boldsymbol{\theta}_0$ and

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{\text{GMM}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, (\mathbf{G}'_0 \mathbf{W}_0 \mathbf{G}_0)^{-1} (\mathbf{G}'_0 \mathbf{W}_0 \mathbf{S}_0 \mathbf{W}_0 \mathbf{G}_0) (\mathbf{G}'_0 \mathbf{W}_0 \mathbf{G}_0)^{-1}]. \quad (6.11)$$

Some leading specializations are the following.

First, in microeconomic analysis data are usually assumed to be independent over i , so (6.10) simplifies to

$$\mathbf{S}_0 = \text{plim} \frac{1}{N} \sum_{i=1}^N \left[\mathbf{h}_i \mathbf{h}'_i \Big|_{\boldsymbol{\theta}_0} \right]. \quad (6.12)$$

If additionally the data are assumed to be identically distributed then (6.9) and (6.10) simplify to $\mathbf{G}_0 = E[\partial \mathbf{h}/\partial \boldsymbol{\theta}' \Big|_{\boldsymbol{\theta}_0}]$ and $\mathbf{S}_0 = E[\mathbf{h} \mathbf{h}' \Big|_{\boldsymbol{\theta}_0}]$, a notation used by many authors.

Second, in the just-identified case that $r = q$, the situation for many estimators including ML and LS, the results simplify to those already presented in Section 5.4 for the estimating equations estimator. To see this note that when $r = q$ the matrices \mathbf{G}_0 , \mathbf{W}_0 , and \mathbf{S}_0 are square matrices that are invertible, so $(\mathbf{G}'_0 \mathbf{W}_0 \mathbf{G}_0)^{-1} = \mathbf{G}_0^{-1} \mathbf{W}_0^{-1} (\mathbf{G}'_0)^{-1}$ and the variance matrix in (6.11) simplifies. It follows that, for the MM estimator in (6.6),

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{\text{MM}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{G}_0^{-1} \mathbf{S}_0 (\mathbf{G}'_0)^{-1}]. \quad (6.13)$$

An MM estimator can always be computed as a GMM estimator and will be invariant to the choice of full rank weighting matrix.

Third, the best choice of matrix \mathbf{W}_N is one such that $\mathbf{W}_0 = \mathbf{S}_0^{-1}$. Then the variance matrix in (6.11) simplifies to $(\mathbf{G}'_0 \mathbf{S}_0^{-1} \mathbf{G}_0)^{-1}$. This is expanded on in Section 6.3.5.

6.3.4. Variance Matrix Estimation

Statistical inference for the GMM estimator is possible given consistent estimates $\widehat{\mathbf{G}}$ of \mathbf{G}_0 , $\widehat{\mathbf{W}}$ of \mathbf{W}_0 , and $\widehat{\mathbf{S}}$ of \mathbf{S}_0 in (6.11). Consistent estimates are easily obtained under relatively weak distributional assumptions.

For \mathbf{G}_0 the obvious estimator is

$$\widehat{\mathbf{G}} = \frac{1}{N} \sum_{i=1}^N \left. \frac{\partial \mathbf{h}_i}{\partial \boldsymbol{\theta}'} \right|_{\widehat{\boldsymbol{\theta}}} . \quad (6.14)$$

For \mathbf{W}_0 the sample weighting matrix \mathbf{W}_N is used. The estimator for the $r \times r$ matrix \mathbf{S}_0 varies with the stochastic assumptions made about the dgp. Microeconometric analysis usually assumes independence over i , so that \mathbf{S}_0 is of the simpler form (6.12). An obvious estimator is then

$$\widehat{\mathbf{S}} = \frac{1}{N} \sum_{i=1}^N \mathbf{h}_i(\widehat{\boldsymbol{\theta}}) \mathbf{h}_i(\widehat{\boldsymbol{\theta}})' . \quad (6.15)$$

Since $\mathbf{h}(\cdot)$ is $r \times 1$, there are at most a finite number of $r(r + 1)/2$ unique entries in \mathbf{S}_0 to be estimated. So $\widehat{\mathbf{S}}$ is consistent as $N \rightarrow \infty$ without need to parameterize the variance $E[\mathbf{h}_i \mathbf{h}_i']$, assumed to exist, to depend on fewer parameters. All that is required are some mild additional assumptions to ensure that $\text{plim } N^{-1} \sum_i \widehat{\mathbf{h}}_i \widehat{\mathbf{h}}_i' = \text{plim } N^{-1} \sum_i \mathbf{h}_i \mathbf{h}_i'$. For example, if $\widehat{\mathbf{h}}_i = \mathbf{x}_i \widehat{u}_i$, where \widehat{u}_i is the OLS residual, we know from Section 4.4 that existence of fourth moments of the regressors needs to be assumed.

Combining these results, we have that the GMM estimator is asymptotically normally distributed with mean $\boldsymbol{\theta}_0$ and estimated asymptotic variance

$$\widehat{V}[\widehat{\boldsymbol{\theta}}_{\text{GMM}}] = \frac{1}{N} (\widehat{\mathbf{G}}' \mathbf{W}_N \widehat{\mathbf{G}})^{-1} \widehat{\mathbf{G}}' \mathbf{W}_N \widehat{\mathbf{S}} \mathbf{W}_N \widehat{\mathbf{G}} (\widehat{\mathbf{G}}' \mathbf{W}_N \widehat{\mathbf{G}})^{-1} . \quad (6.16)$$

This variance matrix estimator is a robust estimator that is an extension of the Eicker–White heteroskedastic-consistent estimator for least-squares estimators.

One can also take expectations and use $\widehat{\mathbf{G}}_E = N^{-1} \sum_i E[\partial \mathbf{h}_i / \partial \boldsymbol{\theta}']|_{\widehat{\boldsymbol{\theta}}}$ for \mathbf{G}_0 and $\widehat{\mathbf{S}}_E = N^{-1} \sum_i E[\mathbf{h}_i \mathbf{h}_i']|_{\widehat{\boldsymbol{\theta}}}$ for \mathbf{S}_0 . However, this usually requires additional distributional assumptions to take the expectation, and the variance matrix estimate will not be as robust to distributional misspecification.

In the time-series case \mathbf{h}_t is subscripted by time t , and asymptotic theory is based on the number of time periods $T \rightarrow \infty$. For time-series data, with \mathbf{h}_t a vector MA(q) process, the usual estimator of $V[\widehat{\boldsymbol{\theta}}_{\text{GMM}}]$ is one proposed by Newey and West (1987b) that uses (6.16) with $\widehat{\mathbf{S}} = \widehat{\Omega}_0 + \sum_{j=1}^q (1 - \frac{j}{q+1}) (\widehat{\Omega}_j + \widehat{\Omega}'_j)$, where $\widehat{\Omega}_j = T^{-1} \sum_{t=j+1}^T \widehat{\mathbf{h}}_t \widehat{\mathbf{h}}_{t-j}'$. This permits time-series correlation in \mathbf{h}_t in addition to contemporaneous correlation. Further details on covariance matrix estimation, including improvements in the time-series case, are given in Davidson and MacKinnon (1993, Section 17.5), Hamilton (1994), and Haan and Levin (1997).

6.3.5. Optimal Weighting Matrix

Application of GMM requires specification of moment function $\mathbf{h}(\cdot)$ and weighting matrix \mathbf{W}_N in (6.7).

The easy part is choosing \mathbf{W}_N to obtain the GMM estimator with the smallest asymptotic variance given a specified function $\mathbf{h}(\cdot)$. This is often called **optimal GMM**

even though it is a limited form of optimality since a poor choice of $\mathbf{h}(\cdot)$ could still lead to a very inefficient estimator.

For just-identified models the same estimator (the MM estimator) is obtained for any full rank weighting matrix, so one might just as well set $\mathbf{W}_N = \mathbf{I}_q$.

For overidentified models with $r > q$, and \mathbf{S}_0 known, the most efficient GMM estimator is obtained by choosing the weighting matrix $\mathbf{W}_N = \mathbf{S}_0^{-1}$. Then the variance matrix given in the proposition simplifies and

$$\sqrt{N}(\widehat{\boldsymbol{\theta}}_{\text{GMM}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}\left[\mathbf{0}, (\mathbf{G}'_0 \mathbf{S}_0^{-1} \mathbf{G}_0)^{-1}\right], \quad (6.17)$$

a result due to Hansen (1982).

This result can be obtained using matrix arguments similar to those that establish that GLS is the most efficient WLS estimator in the linear model. Even more simply, one can work directly with the objective function. For LS estimators that minimize the quadratic form $\mathbf{u}'\mathbf{W}\mathbf{u}$ the most efficient estimator is GLS that sets $\mathbf{W} = \Sigma^{-1} = \mathbf{V}[\mathbf{u}]^{-1}$. The GMM objective function in (6.7) is of this quadratic form with $\mathbf{u} = N^{-1} \sum_i \mathbf{h}_i(\boldsymbol{\theta})$ and so the optimal $\mathbf{W} = (\mathbf{V}[N^{-1} \sum_i \mathbf{h}_i(\boldsymbol{\theta})])^{-1} = \mathbf{S}_0^{-1}$. The optimal GMM estimator weights by the inverse of the variance matrix of the sample moment conditions.

Optimal GMM

In practice \mathbf{S}_0 is unknown and we let $\mathbf{W}_N = \widehat{\mathbf{S}}^{-1}$, where $\widehat{\mathbf{S}}$ is consistent for \mathbf{S}_0 . The optimal GMM estimator can be obtained using a two-step procedure. At the first step a GMM estimator is obtained using a suboptimal choice of \mathbf{W}_N , such as $\mathbf{W}_N = \mathbf{I}_r$ for simplicity. From this first step, form estimate $\widehat{\mathbf{S}}$ using (6.15). At the second step perform an optimal GMM estimator with **optimal weighting matrix** $\mathbf{W}_N = \widehat{\mathbf{S}}^{-1}$.

Then the **optimal GMM estimator** or **two-step GMM estimator** $\widehat{\boldsymbol{\theta}}_{\text{OGMM}}$ based on $\mathbf{h}_i(\boldsymbol{\theta})$ minimizes

$$Q_N(\boldsymbol{\theta}) = \left[\frac{1}{N} \sum_{i=1}^N \mathbf{h}_i(\boldsymbol{\theta}) \right]' \widehat{\mathbf{S}}^{-1} \left[\frac{1}{N} \sum_{i=1}^N \mathbf{h}_i(\boldsymbol{\theta}) \right]. \quad (6.18)$$

The limit distribution is given in (6.17). The optimal GMM estimator is asymptotically normally distributed with mean $\boldsymbol{\theta}_0$ and estimated asymptotic variance with the relatively simple formula

$$\mathbf{V}[\widehat{\boldsymbol{\theta}}_{\text{OGMM}}] = N^{-1}(\widehat{\mathbf{G}}' \widetilde{\mathbf{S}}^{-1} \widehat{\mathbf{G}})^{-1}. \quad (6.19)$$

Usually evaluation of $\widehat{\mathbf{G}}$ and $\widetilde{\mathbf{S}}$ is at $\widehat{\boldsymbol{\theta}}_{\text{OGMM}}$, so $\widetilde{\mathbf{S}}$ uses the same formula as $\widehat{\mathbf{S}}$ except that evaluation is at $\widehat{\boldsymbol{\theta}}_{\text{OGMM}}$. An alternative is to continue to evaluate (6.19) at the first-step estimator, as any consistent estimate of $\boldsymbol{\theta}_0$ can be used.

Remarkably, the optimal GMM estimator in (6.18) requires no additional stochastic assumptions beyond those needed to permit use of (6.16) to estimate the variance matrix of suboptimal GMM. In both cases $\widehat{\mathbf{S}}$ needs to be consistent for \mathbf{S}_0 and from the discussion after (6.15) this requires few additional assumptions. This stands in stark contrast to the additional assumptions needed for GLS to be more efficient than OLS when errors are heteroskedastic. Heteroskedasticity in the errors will affect the optimal choice of $\mathbf{h}_i(\boldsymbol{\theta})$, however (see Section 6.3.7).

Small-Sample Bias of Two-Step GMM

Theory suggests that for overidentified models it is best to use optimal GMM. In implementation, however, the theoretical optimal weighting matrix $\mathbf{W}_N = \mathbf{S}_0^{-1}$ needs to be replaced by a consistent estimate $\widehat{\mathbf{S}}^{-1}$. This replacement makes no difference asymptotically, but it will make a difference in finite samples. In particular, individual observations that increase $\mathbf{h}_i(\theta)$ in (6.18) are likely to increase $\widehat{\mathbf{S}} = N^{-1} \sum_i \widehat{\mathbf{h}}_i \widehat{\mathbf{h}}_i'$ in (6.18), leading to correlation between $N^{-1} \sum_i \mathbf{h}_i(\theta)$ and $\widehat{\mathbf{S}}$. Note that $\mathbf{S}_0 = \text{plim } N^{-1} \sum_i \mathbf{h}_i \mathbf{h}_i'$ is not similarly affected because the probability limit is taken.

Altonji and Segal (1996) demonstrated this problem in estimation of covariance structure models using panel data (see Section 22.5). They used the related minimum distance estimator (see Section 6.7) but in the literature their results are interpreted as being relevant to GMM estimation with cross-section data or short panels. In simulations the optimal estimator was more efficient than a one-step estimator, as expected. However, the optimal estimator had finite-sample bias so large that its root mean-squared error was much larger than that for the one-step estimator.

Altonji and Segal (1996) also proposed a variant, an **independently weighted optimal** estimator that forms the weighting matrix using observations other than used to construct the sample moments. They split the sample into G groups, with $G = 2$ an obvious choice, and minimize

$$Q_N(\theta) = \frac{1}{G} \sum_g \mathbf{h}_g(\theta) \widehat{\mathbf{S}}_{(-g)}^{-1} \mathbf{h}_g(\theta), \quad (6.20)$$

where $\mathbf{h}_g(\theta)$ is computed for the g th group and $\widehat{\mathbf{S}}_{(-g)}$ is computed using all but the g th group. This estimator is less biased, since the weighting matrix $\widehat{\mathbf{S}}_{(-g)}^{-1}$ is by construction independent of $\mathbf{h}_g(\theta)$. However, splitting the sample leads to efficiency loss. Horowitz (1998a) instead used the bootstrap (see Section 11.6.4).

In the Altonji and Segal (1996) example \mathbf{h}_i involves second moments, so $\widehat{\mathbf{S}}$ involves fourth moments. Finite-sample problems for the optimal estimator may not be as significant in other examples where \mathbf{h}_i involves only first moments. Nonetheless, Altonji and Segal's results do suggest caution in using optimal GMM and that differences between one-step GMM and optimal GMM estimates may indicate problems of finite-sample bias in optimal GMM.

Number of Moment Restrictions

In general adding further moment restrictions improves asymptotic efficiency, as it reduces the limit variance $(\mathbf{G}_0' \mathbf{S}_0^{-1} \mathbf{G}_0)^{-1}$ of the optimal GMM estimator or at worst leaves it unchanged.

The benefits of adding further moment conditions vary with the application. For example, if the estimator is the MLE then there is no gain since the MLE is already fully efficient. The literature has focused on IV estimation where gains may be considerable because the variable being instrumented may be much more highly correlated with a combination of many instruments than with a single instrument.

There is a limit, however, as the number of moment restrictions cannot exceed the number of observations. Moreover, adding more moment conditions increases the

likelihood of finite-sample bias and related problems similar to those of weak instruments in linear models (see Section 4.9). Stock et al. (2002) briefly consider weak instruments in nonlinear models.

6.3.6. Regression with Symmetric Error Example

To demonstrate the GMM asymptotic results we return to the additional moment restrictions example introduced in Section 6.2.4. For this example the objective function for $\widehat{\beta}_{\text{GMM}}$ has already been given in (6.2). All that is required is specification of \mathbf{W}_N , such as $\mathbf{W}_N = \mathbf{I}$.

To obtain the distribution of this estimator we use the general notation of Section 6.3. The function $\mathbf{h}(\cdot)$ in (6.5) specializes to

$$\mathbf{h}(\mathbf{y}, \mathbf{x}, \boldsymbol{\beta}) = \begin{bmatrix} \mathbf{x}(\mathbf{y} - \mathbf{x}'\boldsymbol{\beta}) \\ \mathbf{x}(\mathbf{y} - \mathbf{x}'\boldsymbol{\beta})^3 \end{bmatrix} \Rightarrow \frac{\partial \mathbf{h}(\mathbf{y}, \mathbf{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} = \begin{bmatrix} -\mathbf{x}\mathbf{x}' \\ -3\mathbf{x}\mathbf{x}'(\mathbf{y} - \mathbf{x}'\boldsymbol{\beta})^2 \end{bmatrix}.$$

These expressions lead directly to expressions for \mathbf{G}_0 and \mathbf{S}_0 using (6.9) and (6.12), so that (6.14) and (6.15) then yield consistent estimates

$$\widehat{\mathbf{G}} = \begin{bmatrix} -\frac{1}{N} \sum_i \mathbf{x}_i \mathbf{x}_i' \\ -\frac{1}{N} \sum_i 3\widehat{u}_i^2 \mathbf{x}_i \mathbf{x}_i' \end{bmatrix} \quad (6.21)$$

and

$$\widehat{\mathbf{S}} = \begin{bmatrix} \frac{1}{N} \sum_i \widehat{u}_i^2 \mathbf{x}_i \mathbf{x}_i' & \frac{1}{N} \sum_i \widehat{u}_i^4 \mathbf{x}_i \mathbf{x}_i' \\ \frac{1}{N} \sum_i \widehat{u}_i^4 \mathbf{x}_i \mathbf{x}_i' & \frac{1}{N} \sum_i \widehat{u}_i^6 \mathbf{x}_i \mathbf{x}_i' \end{bmatrix}, \quad (6.22)$$

where $\widehat{u}_i = y - \mathbf{x}_i' \widehat{\boldsymbol{\beta}}$. Alternative estimates can be obtained by first evaluating the expectations in \mathbf{G}_0 and \mathbf{S}_0 , but this will require assumptions on $E[u^2|\mathbf{x}]$, $E[u^4|\mathbf{x}]$, and $E[u^6|\mathbf{x}]$. Substituting $\widehat{\mathbf{G}}$, $\widehat{\mathbf{S}}$, and \mathbf{W}_N into (6.16) gives the estimated asymptotic variance matrix for $\widehat{\boldsymbol{\beta}}_{\text{GMM}}$.

Now consider GMM with an optimal weighting matrix. This again minimizes (6.2), but from (6.18) now $\mathbf{W}_N = \widehat{\mathbf{S}}^{-1}$, where $\widehat{\mathbf{S}}$ is defined in (6.22). Computation of $\widehat{\mathbf{S}}$ requires first-step consistent estimates $\widehat{\boldsymbol{\beta}}$. An obvious choice is GMM with $\mathbf{W}_N = \mathbf{I}$. In this example the OLS estimator is also consistent and could instead be used. Using (6.19) gives this two-step estimator an estimated asymptotic variance matrix $\widehat{V}[\widehat{\boldsymbol{\beta}}_{\text{OGMM}}]$ equal to

$$\left(\begin{bmatrix} \sum_i \widetilde{u}_i \mathbf{x}_i \mathbf{x}_i' \\ \sum_i \widetilde{u}_i^3 \mathbf{x}_i \mathbf{x}_i' \end{bmatrix}' \begin{bmatrix} \sum_i \widetilde{u}_i^2 \mathbf{x}_i \mathbf{x}_i' & \sum_i \widetilde{u}_i^4 \mathbf{x}_i \mathbf{x}_i' \\ \sum_i \widetilde{u}_i^4 \mathbf{x}_i \mathbf{x}_i' & \sum_i \widetilde{u}_i^6 \mathbf{x}_i \mathbf{x}_i' \end{bmatrix}^{-1} \begin{bmatrix} \sum_i \widetilde{u}_i \mathbf{x}_i \mathbf{x}_i' \\ \sum_i \widetilde{u}_i^3 \mathbf{x}_i \mathbf{x}_i' \end{bmatrix} \right)^{-1},$$

where $\widetilde{u}_i = y_i - \mathbf{x}_i' \widehat{\boldsymbol{\beta}}_{\text{OGMM}}$ and the various divisions by N have canceled out.

Analytical results for the efficiency gain of optimal GMM in this example are easily obtained by specialization to the nonregression case where y is iid with mean μ . Furthermore, assume that y is Laplace distributed with scale parameter equal to unity, in which case the density is $f(y) = (1/2) \times \exp\{-|y - \mu|\}$ with $E[y] = \mu$, $V[y] = 2$, and higher central moments $E[(y - \mu)^r]$ equal to zero for r odd and equal to $r!$ for r even. The sample median is fully efficient as it is the MLE, and it can be shown to

have asymptotic variance $1/N$. The sample mean \bar{y} is inefficient with variance $V[\bar{y}] = V[y]/N = 2/N$. The optimal GMM estimator $\hat{\mu}^{\text{opt}}$ based on the two moment conditions $E[(y - \mu)] = 0$ and $E[(y - \mu)^3] = 0$ has weighting matrix that places much less weight on the second moment condition, because it has relatively high variance, and has negative off-diagonal entries. The optimal GMM estimator $\hat{\mu}_{\text{OGMM}}$ can be shown to have asymptotic variance $1.7143/N$ (see Exercise 6.3). It is therefore more efficient than the sample mean (variance $2/N$), though is still considerably less efficient than the sample median.

For this example the identity matrix is an exceptionally poor choice of weighting matrix. It places too much weight on the second moment condition, yielding a sub-optimal GMM estimator of μ with asymptotic variance $19.14/N$ that is many times greater than even $V[\bar{y}] = 2/N$. For details see Exercise 6.3.

6.3.7. Optimal Moment Condition

Section 6.3.5 gives the surprising result that optimal GMM requires essentially no more assumptions than does GMM without an optimal weighting matrix. However, this optimality is very limited as it is conditional on the choice of moment function $\mathbf{h}(\cdot)$ in (6.5) or (6.18).

The GMM defines a class of estimators, with different choice of $\mathbf{h}(\cdot)$ corresponding to different members of the class. Some choices of $\mathbf{h}(\cdot)$ are better than others, depending on additional stochastic assumptions. For example, $\mathbf{h}_i = \mathbf{x}_i u_i$ yields the OLS estimator whereas $\mathbf{h}_i = \mathbf{x}_i u_i / V[u_i | \mathbf{x}_i]$ yields the GLS estimator when errors are heteroskedastic. This multitude of potential choices for $\mathbf{h}(\cdot)$ can make any particular GMM estimator appear ad hoc. However, qualitatively similar decisions have to be made in m-estimation in choosing, for example, to minimize the sum of squared errors rather than the weighted sum of squared errors or the sum of absolute deviations of errors.

If complete distributional assumptions are made the most efficient estimator is the MLE. Thus the optimal choice of $\mathbf{h}(\cdot)$ in (6.5) is

$$\mathbf{h}(\mathbf{w}, \boldsymbol{\theta}) = \frac{\partial \ln f(\mathbf{w}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}},$$

where $f(\mathbf{w}, \boldsymbol{\theta})$ is the joint density of \mathbf{w} . For regression with dependent variable(s) \mathbf{y} and regressors \mathbf{x} this is the unconditional MLE based on the unconditional joint density $f(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})$ of \mathbf{y} and \mathbf{x} . In many applications $f(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}) = f(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta})g(\mathbf{x})$, where the (suppressed) parameters of the marginal density of \mathbf{x} do not depend on the parameters of interest $\boldsymbol{\theta}$. Then it is just as efficient to use the conditional MLE based on the conditional density $f(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta})$. This can be used as the basis for MM estimation, or GMM estimation with weighting matrix $\mathbf{W}_N = \mathbf{I}_q$, though any full-rank matrix \mathbf{W}_N will also give the MLE. This result is of limited practical use, however, as the purpose of GMM estimation is to avoid making a full set of distributional assumptions.

When incomplete distributional assumptions are made, a common starting point is specification of a **conditional moment condition**, where conditioning is on exogenous variables. This is usually a low-order moment condition for the model error such

as $E[u|\mathbf{x}] = 0$ or $E[u|\mathbf{z}] = 0$. This conditional moment condition can lead to many **unconditional moment conditions** that might be the basis for GMM estimation, such as $E[\mathbf{zu}] = \mathbf{0}$. Newey (1990a, 1993) obtained results on the optimal choice of unconditional moment condition for data independent over i .

Specifically, begin with s conditional moment condition restrictions

$$E[\mathbf{r}(y, \mathbf{x}, \boldsymbol{\theta}_0)|\mathbf{z}] = \mathbf{0}, \quad (6.23)$$

where $\mathbf{r}(\cdot)$ is a residual-type $s \times 1$ vector function introduced in Section 6.2.2. A scalar example is $E[y - \mathbf{x}'\boldsymbol{\theta}_0|\mathbf{z}] = 0$. The instrumental variables notation is being used where \mathbf{x} are regressors, some potentially endogenous, and \mathbf{z} are instruments that include the exogenous components of \mathbf{x} . In simpler models without endogeneity $\mathbf{z} = \mathbf{x}$.

GMM estimation of the q parameters $\boldsymbol{\theta}$ based on (6.23) is not possible, as typically there are only a few conditional moment restrictions, and often just one, so $s \leq q$. Instead, we introduce an $r \times s$ matrix function of the instruments $\mathbf{D}(\mathbf{z})$, where $r \geq q$, and note that by the law of iterated expectations $E[\mathbf{D}(\mathbf{z})\mathbf{r}(y, \mathbf{x}, \boldsymbol{\theta}_0)] = \mathbf{0}$, which can be used as the basis for GMM estimation. The **optimal instruments** or optimal choice of matrix function $\mathbf{D}(\mathbf{z})$ can be shown to be the $q \times s$ matrix

$$\mathbf{D}^*(\mathbf{z}, \boldsymbol{\theta}_0) = E \left[\frac{\partial \mathbf{r}(y, \mathbf{x}, \boldsymbol{\theta}_0)'}{\partial \boldsymbol{\theta}} | \mathbf{z} \right] \{V[\mathbf{r}(y, \mathbf{x}, \boldsymbol{\theta}_0)|\mathbf{z}]\}^{-1}. \quad (6.24)$$

A derivation is given in, for example, Davidson and MacKinnon (1993, p. 604). The optimal instrument matrix $\mathbf{D}^*(\mathbf{z})$ is a $q \times s$ matrix, so the unconditional moment condition $E[\mathbf{D}^*(\mathbf{z})\mathbf{r}(y, \mathbf{x}, \boldsymbol{\theta}_0)] = \mathbf{0}$ yields exactly as many moment conditions as parameters. The optimal GMM estimator simply solves the corresponding sample moment conditions

$$\frac{1}{N} \sum_{i=1}^N \mathbf{D}^*(\mathbf{z}_i, \boldsymbol{\theta}) \mathbf{r}(y_i, \mathbf{x}_i, \boldsymbol{\theta}) = \mathbf{0}. \quad (6.25)$$

The optimal estimator requires additional assumptions, namely the expectations used in forming $\mathbf{D}^*(\mathbf{z}, \boldsymbol{\theta}_0)$ in (6.24), and implementation requires replacing unknown parameters by known parameters so that generated regressors $\widehat{\mathbf{D}}$ are used.

For example, if $r(y, \mathbf{x}, \boldsymbol{\theta}) = y - \exp(\mathbf{x}'\boldsymbol{\theta})$ then $\partial r / \partial \boldsymbol{\theta} = -\exp(\mathbf{x}'\boldsymbol{\theta})\mathbf{x}$ and (6.24) requires specification of $E[\exp(\mathbf{x}'\boldsymbol{\theta}_0)\mathbf{x}|\mathbf{z}]$ and $V[y - \exp(\mathbf{x}'\boldsymbol{\theta})|\mathbf{z}]$. One possibility is to assume $E[\exp(\mathbf{x}'\boldsymbol{\theta}_0)\mathbf{x}|\mathbf{z}]$ is a low-order polynomial in \mathbf{z} , in which case there will be more moment conditions than parameters and so estimation is by GMM rather than simply by solving (6.25), and to assume errors are homoskedastic. If these additional assumptions are wrong then the estimator is still consistent, provided (6.23) is valid, and consistent standard errors can be obtained using the robust form of the variance matrix in (6.16). It is common to more simply use \mathbf{z} rather than $\mathbf{D}^*(\mathbf{z}, \boldsymbol{\theta})$ as the instrument.

Optimal Moment Condition for Nonlinear Regression Example

The result (6.24) is useful in some cases, especially those where $\mathbf{z} = \mathbf{x}$. Here we confirm that GLS is the most efficient GMM estimator based on $E[u|\mathbf{x}] = 0$.

Consider the nonlinear regression model $y = g(\mathbf{x}, \boldsymbol{\beta}) + u$. If the starting point is the conditional moment restriction $E[u|\mathbf{x}] = 0$, or $E[y - g(\mathbf{x}, \boldsymbol{\beta})|\mathbf{x}] = 0$, then $\mathbf{z} = \mathbf{x}$ in (6.23), and (6.24) yields

$$\begin{aligned}\mathbf{D}^*(\mathbf{x}, \boldsymbol{\beta}) &= E\left[\frac{\partial}{\partial \boldsymbol{\beta}}(y - g(\mathbf{x}, \boldsymbol{\beta}_0))|\mathbf{x}\right]\{V[y - g(\mathbf{x}, \boldsymbol{\beta}_0)|\mathbf{x}]\}^{-1} \\ &= -\frac{\partial g(\mathbf{x}, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \times \frac{1}{V[u|\mathbf{x}]},\end{aligned}$$

which requires only specification of $V[u|\mathbf{x}]$. From (6.25) the optimal GMM estimator directly solves the corresponding sample moment conditions

$$\frac{1}{N} \sum_{i=1}^N -\frac{\partial g(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \times \frac{(y_i - g(\mathbf{x}_i, \boldsymbol{\beta}))}{\sigma_i^2} = \mathbf{0},$$

where $\sigma_i^2 = V[u_i|\mathbf{x}_i]$ is functionally independent of $\boldsymbol{\beta}$. These are the first-order conditions for generalized NLS when the error is heteroskedastic. Implementation is possible using a consistent estimate $\hat{\sigma}_i^2$ of σ_i^2 , in which case GMM estimation is the same as FGNLS. One can obtain standard errors robust to misspecification of σ_i^2 as detailed in Section 5.8.

Specializing to the linear model, $g(\mathbf{x}, \boldsymbol{\beta}) = \mathbf{x}'\boldsymbol{\beta}$ and the optimal GMM estimator based on $E[u|\mathbf{x}] = 0$ is GLS, and specializing further to the case of homoskedastic errors, the optimal GMM estimator based on $E[u|\mathbf{x}] = 0$ is OLS. As already seen in the example in Section 6.3.6, more efficient estimation may be possible if additional conditional moment conditions are used.

6.3.8. Tests of Overidentifying Restrictions

Hypothesis tests on $\boldsymbol{\theta}$ can be performed using the Wald test (see Section 5.5), or with other methods given in Section 7.5.

In addition there is a quite general model specification test that can be used for over-identified models with more moment conditions (r) than parameters (q). The test is one of the closeness of $N^{-1} \sum_i \hat{\mathbf{h}}_i$ to $\mathbf{0}$, where $\hat{\mathbf{h}}_i = \mathbf{h}(\mathbf{w}_i, \hat{\boldsymbol{\theta}})$. This is an obvious test of $H_0: E[\mathbf{h}(\mathbf{w}, \boldsymbol{\theta}_0)] = \mathbf{0}$, the initial population moment conditions. For just-identified models, estimation imposes $N^{-1} \sum_i \hat{\mathbf{h}}_i = \mathbf{0}$ and the test is not possible. For over-identified models, however, the first-order conditions (6.8) set a $q \times r$ matrix times $N^{-1} \sum_i \hat{\mathbf{h}}_i$ to zero, where $q < r$, so $\sum_i \hat{\mathbf{h}}_i \neq \mathbf{0}$.

In the special case that $\boldsymbol{\theta}$ is estimated by $\hat{\boldsymbol{\theta}}_{OGMM}$ defined in (6.18), Hansen (1982) showed that the **overidentifying restrictions (OIR) test statistic**

$$OIR = \left(N^{-1} \sum_i \hat{\mathbf{h}}_i\right)' \hat{\mathbf{S}}^{-1} \left(N^{-1} \sum_i \hat{\mathbf{h}}_i\right) \quad (6.26)$$

is asymptotically distributed as $\chi^2(r - q)$ under $H_0 : E[\mathbf{h}(\mathbf{w}, \boldsymbol{\theta}_0)] = \mathbf{0}$. Note that OIR equals the GMM objective function (6.18) evaluated at $\hat{\boldsymbol{\theta}}_{OGMM}$. If OIR is large then the population moment conditions are rejected and the GMM estimator is inconsistent for $\boldsymbol{\theta}$.

It is not obvious a priori that the particular quadratic form in $N^{-1} \sum_i \widehat{\mathbf{h}}_i$ given in (6.26) is $\chi^2(r - q)$ distributed under H_0 . A formal derivation is given in the next section and an intuitive explanation in the case of linear IV estimation is provided in Section 8.4.4.

A classic application is to life-cycle models of consumption (see Section 6.2.7), in which case the orthogonality conditions are Euler conditions. A large chi-square test statistic is then often stated to mean rejection of the life-cycle hypothesis. However, it should instead be more narrowly interpreted as rejection of the particular specification of utility function and set of stochastic assumptions used in the study.

6.3.9. Derivations for the GMM Estimator

The algebra is simplified by introducing a more compact notation. The GMM estimator minimizes

$$Q_N(\boldsymbol{\theta}) = \mathbf{g}_N(\boldsymbol{\theta})' \mathbf{W}_N \mathbf{g}_N(\boldsymbol{\theta}), \quad (6.27)$$

where $\mathbf{g}_N(\boldsymbol{\theta}) = N^{-1} \sum_i \mathbf{h}_i(\boldsymbol{\theta})$. Then the GMM first-order conditions (6.8) are

$$\mathbf{G}_N(\widehat{\boldsymbol{\theta}})' \mathbf{W}_N \mathbf{g}_N(\widehat{\boldsymbol{\theta}}) = \mathbf{0}, \quad (6.28)$$

where $\mathbf{G}_N(\boldsymbol{\theta}) = \partial \mathbf{g}_N(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}' = N^{-1} \sum_i \partial \mathbf{h}_i(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}'$.

For *consistency* we consider the informal condition that the probability limit of $\partial Q_N(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}|_{\boldsymbol{\theta}_0}$ equals zero. From (6.28) this will be the case as $\mathbf{G}_N(\boldsymbol{\theta}_0)$ and \mathbf{W}_N have finite probability limits, by assumptions (iii) and (iv) of Proposition 6.1, and $\text{plim } \mathbf{g}_N(\boldsymbol{\theta}_0) = \mathbf{0}$ as a consequence of assumption (v). More intuitively, $\mathbf{g}_N(\boldsymbol{\theta}_0) = N^{-1} \sum_i \mathbf{h}_i(\boldsymbol{\theta}_0)$ has probability limit zero if a law of large numbers can be applied and $E[\mathbf{h}_i(\boldsymbol{\theta}_0)] = \mathbf{0}$, which was assumed at the outset in (6.5).

The parameter $\boldsymbol{\theta}_0$ is *identified* by the key assumption (ii) and additionally assumptions (iii) and (iv), which restrict the probability limits of $\mathbf{G}_N(\boldsymbol{\theta}_0)$ and \mathbf{W}_N to be full-rank matrices. The assumption that $\mathbf{G}_0 = \text{plim } \mathbf{G}_N(\boldsymbol{\theta}_0)$ is a full-rank matrix is called the **rank condition for identification**. A weaker necessary condition for identification is the **order condition** that $r \geq q$.

For *asymptotic normality*, a more general theory is needed than that for an m-estimator based on an objective function $Q_N(\boldsymbol{\beta}) = N^{-1} \sum_i q(\mathbf{w}_i, \boldsymbol{\theta})$ that involves just one sum. We rescale (6.28) by multiplication by \sqrt{N} , so that

$$\mathbf{G}_N(\widehat{\boldsymbol{\theta}})' \mathbf{W}_N \sqrt{N} \mathbf{g}_N(\widehat{\boldsymbol{\theta}}) = \mathbf{0}. \quad (6.29)$$

The approach of the general Theorem 5.3 is to take a Taylor series expansion around $\boldsymbol{\theta}_0$ of the entire left-hand side of (6.28). Since $\widehat{\boldsymbol{\theta}}$ appears in both the first and third terms this is complicated and requires existence of first derivatives of $\mathbf{G}_N(\boldsymbol{\theta})$ and hence second derivatives of $\mathbf{g}_N(\boldsymbol{\theta})$. Since $\mathbf{G}_N(\widehat{\boldsymbol{\theta}})$ and \mathbf{W}_N have finite probability limits it is sufficient to more simply take an exact Taylor series expansion of only $\sqrt{N} \mathbf{g}_N(\widehat{\boldsymbol{\theta}})$. This yields an expression similar to that in the Chapter 5 discussion of m-estimation, with

$$\sqrt{N} \mathbf{g}_N(\widehat{\boldsymbol{\theta}}) = \sqrt{N} \mathbf{g}_N(\boldsymbol{\theta}_0) + \mathbf{G}_N(\boldsymbol{\theta}^+) \sqrt{N} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \quad (6.30)$$

6.4. LINEAR INSTRUMENTAL VARIABLES

recalling that $\mathbf{G}_N(\theta) = \partial \mathbf{g}_N(\theta) / \partial \theta'$, where θ^+ is a point between θ_0 and $\widehat{\theta}$. Substituting (6.30) back into (6.29) yields

$$\mathbf{G}_N(\widehat{\theta})' \mathbf{W}_N \left[\sqrt{N} \mathbf{g}_N(\theta_0) + \mathbf{G}_N(\theta^+) \sqrt{N} (\widehat{\theta} - \theta_0) \right] = \mathbf{0}.$$

Solving for $\sqrt{N}(\widehat{\theta} - \theta_0)$ yields

$$\sqrt{N}(\widehat{\theta} - \theta_0) = -[\mathbf{G}_N(\widehat{\theta})' \mathbf{W}_N \mathbf{G}_N(\theta^+)]^{-1} \mathbf{G}_N(\widehat{\theta})' \mathbf{W}_N \sqrt{N} \mathbf{g}_N(\theta_0). \quad (6.31)$$

Equation (6.31) is the key result for obtaining the limit distribution of the GMM estimator. We obtain the probability limits of each of the first five terms using $\widehat{\theta} \xrightarrow{p} \theta_0$, given consistency, in which case $\theta^+ \xrightarrow{p} \theta_0$. The last term on the right-hand side of (6.31) has a limit normal distribution by assumption (v). Thus

$$\sqrt{N}(\widehat{\theta} - \theta_0) \xrightarrow{d} -(\mathbf{G}'_0 \mathbf{W}_0 \mathbf{G}_0)^{-1} \mathbf{G}'_0 \mathbf{W}_0 \times \mathcal{N}[0, \mathbf{S}_0],$$

where \mathbf{G}_0 , \mathbf{W}_0 , and \mathbf{S}_0 have been defined in Proposition 6.1. Applying the limit normal product rule (Theorem A.17) yields (6.11).

This derivation treats the GMM first-order conditions as being q linear combinations of the r sample moments $\mathbf{g}_N(\widehat{\theta})$, since $\mathbf{G}_N(\widehat{\theta})' \mathbf{W}_N$ is a $q \times r$ matrix. The MM estimator is the special case $q = r$, since then $\mathbf{G}_N(\widehat{\theta})' \mathbf{W}_N$ is a full-rank square matrix, so $\mathbf{G}_N(\widehat{\theta})' \mathbf{W}_N \mathbf{g}_N(\widehat{\theta}) = \mathbf{0}$ implies that $\mathbf{g}_N(\widehat{\theta}) = \mathbf{0}$.

To derive the distribution of the OIR test statistic in (6.26), begin with a first-order Taylor series expansion of $\sqrt{N} \mathbf{g}_N(\widehat{\theta})$ around θ_0 to obtain

$$\begin{aligned} \sqrt{N} \mathbf{g}_N(\widehat{\theta}_{\text{OGMM}}) &= \sqrt{N} \mathbf{g}_N(\theta_0) + \mathbf{G}_N(\theta^+) \sqrt{N} (\widehat{\theta}_{\text{OGMM}} - \theta_0) \\ &= \sqrt{N} \mathbf{g}_N(\theta_0) - \mathbf{G}_0 (\mathbf{G}'_0 \mathbf{S}_0^{-1} \mathbf{G}_0)^{-1} \mathbf{G}'_0 \mathbf{S}_0^{-1} \sqrt{N} \mathbf{g}_N(\theta_0) + o_p(1) \\ &= [\mathbf{I} - \mathbf{M}_0 \mathbf{S}_0^{-1}] \sqrt{N} \mathbf{g}_N(\theta_0) + o_p(1), \end{aligned}$$

where the second equality uses (6.31) with \mathbf{W}_N consistent for \mathbf{S}_0^{-1} , $\mathbf{M}_0 = \mathbf{G}_0 (\mathbf{G}'_0 \mathbf{S}_0^{-1} \mathbf{G}_0)^{-1} \mathbf{G}'_0$, and $o_p(1)$ is defined in Definition A.22. It follows that

$$\begin{aligned} \mathbf{S}_0^{-1/2} \sqrt{N} \mathbf{g}_N(\widehat{\theta}_{\text{OGMM}}) &= \mathbf{S}_0^{-1/2} [\mathbf{I} - \mathbf{M}_0 \mathbf{S}_0^{-1}] \sqrt{N} \mathbf{g}_N(\theta_0) + o_p(1) \\ &= [\mathbf{I} - \mathbf{S}_0^{-1/2} \mathbf{M}_0 \mathbf{S}_0^{-1/2}] \mathbf{S}_0^{-1/2} \sqrt{N} \mathbf{g}_N(\theta_0) + o_p(1). \end{aligned} \quad (6.32)$$

Now $[\mathbf{I} - \mathbf{S}_0^{-1/2} \mathbf{M}_0 \mathbf{S}_0^{-1/2}] = [\mathbf{I} - \mathbf{S}_0^{-1/2} \mathbf{G}_0 (\mathbf{G}'_0 \mathbf{S}_0^{-1} \mathbf{G}_0)^{-1} \mathbf{G}'_0 \mathbf{S}_0^{-1/2}]$ is an idempotent matrix of rank $(r - q)$, and $\mathbf{S}_0^{-1/2} \sqrt{N} \mathbf{g}_N(\theta_0) \xrightarrow{d} \mathcal{N}[0, \mathbf{I}]$ given $\sqrt{N} \mathbf{g}_N(\theta_0) \xrightarrow{d} \mathcal{N}[0, \mathbf{S}_0]$. From standard results for quadratic forms of normal variables it follows that the inner product

$$\tau_N = (\mathbf{S}_0^{-1/2} \sqrt{N} \mathbf{g}_N(\widehat{\theta}_{\text{OGMM}}))' (\mathbf{S}_0^{-1/2} \sqrt{N} \mathbf{g}_N(\widehat{\theta}_{\text{OGMM}}))$$

converges to the $\chi^2(r - q)$ distribution.

6.4. Linear Instrumental Variables

Correlation of regressors with the error term leads to inconsistency of least-squares methods. Examples of such failure include omitted variables, simultaneity,

measurement error in the regressors, and sample selection bias. Instrumental variables methods provide a general approach that can handle any of these problems, provided suitable instruments exist.

Instrumental variables methods fall naturally into the GMM framework as a surplus of instruments leads to an excess of moment conditions that can be used for estimation. Many IV results are most easily obtained using the GMM framework.

Linear IV is important enough to appear in many places in this book. An introduction was given in Sections 4.8 and 4.9. This section presents single-equation linear IV as a particular application of GMM. For completeness the section also presents the earlier literature on a special case, the two-stage least-squares estimator. Systems linear IV estimation is summarized in Section 6.9.5. Tests of endogeneity and tests of overidentifying restrictions for linear models are detailed in Section 8.4. Chapter 22 presents linear IV estimation with panel data.

6.4.1. Linear GMM with Instruments

Consider the linear regression model

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + u_i, \quad (6.33)$$

where each component of \mathbf{x} is viewed as being an **exogenous regressor** if it is uncorrelated with the error in model (6.33) or an **endogenous regressor** if it is correlated. If all regressors are exogenous then LS estimators can be used, but if any components of \mathbf{x} are endogenous then LS estimators are inconsistent for $\boldsymbol{\beta}$.

From Section 4.8, consistent estimates can be obtained by IV estimation. The key assumption is the existence of an $r \times 1$ vector of **instruments** \mathbf{z} that satisfies

$$\mathbb{E}[u_i | \mathbf{z}_i] = \mathbf{0}. \quad (6.34)$$

Exogenous regressors can be instrumented by themselves. As there must be at least as many instruments as regressors, the challenge is to find additional instruments that at least equal the number of endogenous variables in the model. Some examples of such instruments have been given in Section 4.8.2.

Linear GMM Estimator

From Section 6.2.5, the conditional moment restriction (6.34) and model (6.33) imply the unconditional moment restriction

$$\mathbb{E}[\mathbf{z}_i(y_i - \mathbf{x}'_i \boldsymbol{\beta})] = \mathbf{0}, \quad (6.35)$$

where for notational simplicity the following analysis uses $\boldsymbol{\beta}$ rather than the more formal $\boldsymbol{\beta}_0$ to denote the true parameter value. A quadratic form in the corresponding sample moments leads to the GMM objective function $Q_N(\boldsymbol{\beta})$ given in (6.4).

In matrix notation define $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ as usual and let \mathbf{Z} denote the $N \times r$ matrix of instruments with i th row \mathbf{z}'_i . Then $\sum_i \mathbf{z}_i(y_i - \mathbf{x}'_i \boldsymbol{\beta}) = \mathbf{Z}'\mathbf{u}$ and (6.4) becomes

$$Q_N(\boldsymbol{\beta}) = \left[\frac{1}{N} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{Z} \right] \mathbf{W}_N \left[\frac{1}{N} \mathbf{Z}' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right], \quad (6.36)$$

6.4. LINEAR INSTRUMENTAL VARIABLES

where \mathbf{W}_N is an $r \times r$ full-rank symmetric weighting matrix with leading examples given at the end of this section. The first-order conditions

$$\frac{\partial Q_N(\beta)}{\partial \beta} = -2 \left[\frac{1}{N} \mathbf{X}' \mathbf{Z} \right] \mathbf{W}_N \left[\frac{1}{N} \mathbf{Z}' (\mathbf{y} - \mathbf{X}\beta) \right] = \mathbf{0}$$

can actually be solved for β in this special case of GMM, leading to the **GMM estimator in the linear IV model**

$$\hat{\beta}_{\text{GMM}} = [\mathbf{X}' \mathbf{Z} \mathbf{W}_N \mathbf{Z}' \mathbf{X}]^{-1} \mathbf{X}' \mathbf{Z} \mathbf{W}_N \mathbf{Z}' \mathbf{y}, \quad (6.37)$$

where the divisions by N have canceled out.

Distribution of Linear GMM Estimator

The general results of Section 6.3 can be used to derive the asymptotic distribution. Alternatively, since an explicit solution for $\hat{\beta}_{\text{GMM}}$ exists the analysis for OLS given in Section 4.4. can be adapted. Substituting $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$ into (6.37) yields

$$\hat{\beta}_{\text{GMM}} = \beta + [(N^{-1} \mathbf{X}' \mathbf{Z}) \mathbf{W}_N (N^{-1} \mathbf{Z}' \mathbf{X})]^{-1} (N^{-1} \mathbf{X}' \mathbf{Z}) \mathbf{W}_N (N^{-1} \mathbf{Z}' \mathbf{u}). \quad (6.38)$$

From the last term, consistency of the GMM estimator essentially requires that $\text{plim } N^{-1} \mathbf{Z}' \mathbf{u} = \mathbf{0}$. Under pure random sampling this requires that (6.35) holds, whereas under other common sampling schemes (see Section 24.3) the stronger assumption (6.34) is needed.

Additionally, the **rank condition for identification** of β that $\text{plim } N^{-1} \mathbf{Z}' \mathbf{X}$ is of rank K ensures that the inverse in the right-hand side exists, provided \mathbf{W}_N is of full rank. A weaker **order condition** is that $r \geq K$.

The limit distribution is based on the expression for $\sqrt{N}(\hat{\beta}_{\text{GMM}} - \beta)$ obtained by simple manipulation of (6.38). This yields an asymptotic normal distribution for $\hat{\beta}_{\text{GMM}}$ with mean β and estimated asymptotic variance

$$\hat{V}[\hat{\beta}_{\text{GMM}}] = N [\mathbf{X}' \mathbf{Z} \mathbf{W}_N \mathbf{Z}' \mathbf{X}]^{-1} [\mathbf{X}' \mathbf{Z} \mathbf{W}_N \hat{\mathbf{S}} \mathbf{W}_N \mathbf{Z}' \mathbf{X}] [\mathbf{X}' \mathbf{Z} \mathbf{W}_N \mathbf{Z}' \mathbf{X}]^{-1}, \quad (6.39)$$

where $\hat{\mathbf{S}}$ is a consistent estimate of

$$\mathbf{S} = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \mathbf{E}[u_i^2 \mathbf{z}_i \mathbf{z}_i'],$$

given the usual cross-section assumption of independence over i . The essential additional assumption needed for (6.39) is that $N^{-1/2} \mathbf{Z}' \mathbf{u} \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{S}]$. Result (6.39) also follows from Proposition 6.1 with $\mathbf{h}(\cdot) = \mathbf{z}(\mathbf{y} - \mathbf{x}'\beta)$ and hence $\partial \mathbf{h}/\partial \beta' = -\mathbf{z}\mathbf{x}'$.

For cross-section data with heteroskedastic errors, \mathbf{S} is consistently estimated by

$$\hat{\mathbf{S}} = \frac{1}{N} \sum_{i=1}^N \hat{u}_i^2 \mathbf{z}_i \mathbf{z}_i' = \mathbf{Z}' \mathbf{D} \mathbf{Z} / N, \quad (6.40)$$

where $\hat{u}_i = y_i - \mathbf{x}'_i \hat{\beta}_{\text{GMM}}$ is the GMM residual and \mathbf{D} is an $N \times N$ diagonal matrix with entries \hat{u}_i^2 . A commonly used small-sample adjustment is to divide by $N - K$

Table 6.2. GMM Estimators in Linear IV Model and Their Asymptotic Variance^a

Estimator	Definition and Asymptotic Variance
GMM (general \mathbf{W}_N)	$\hat{\beta}_{\text{GMM}} = [\mathbf{X}' \mathbf{Z} \mathbf{W}_N \mathbf{Z}' \mathbf{X}]^{-1} \mathbf{X}' \mathbf{Z} \mathbf{W}_N \mathbf{Z}' \mathbf{y}$ $\hat{V}[\hat{\beta}] = N[\mathbf{X}' \mathbf{Z} \mathbf{W}_N \mathbf{Z}' \mathbf{X}]^{-1} [\mathbf{X}' \mathbf{Z} \mathbf{W}_N \hat{\mathbf{S}} \mathbf{W}_N \mathbf{Z}' \mathbf{X}] [\mathbf{X}' \mathbf{Z} \mathbf{W}_N \mathbf{Z}' \mathbf{X}]^{-1}$
Optimal GMM ($\mathbf{W}_N = \hat{\mathbf{S}}^{-1}$)	$\hat{\beta}_{\text{OGMM}} = [\mathbf{X}' \mathbf{Z} \hat{\mathbf{S}}^{-1} \mathbf{Z}' \mathbf{X}]^{-1} \mathbf{X}' \mathbf{Z} \hat{\mathbf{S}}^{-1} \mathbf{Z}' \mathbf{y}$ $\hat{V}[\hat{\beta}] = N[\mathbf{X}' \mathbf{Z} \hat{\mathbf{S}}^{-1} \mathbf{Z}' \mathbf{X}]^{-1}$
2SLS ($\mathbf{W}_N = [N^{-1} \mathbf{Z}' \mathbf{Z}]^{-1}$)	$\hat{\beta}_{\text{2SLS}} = [\mathbf{X}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X}]^{-1} \mathbf{X}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{y}$ $\hat{V}[\hat{\beta}] = N[\mathbf{X}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X}]^{-1} [\mathbf{X}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \hat{\mathbf{S}} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X}]$ $\times [\mathbf{X}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X}]^{-1}$ $\hat{V}[\hat{\beta}] = s^2 [\mathbf{X}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X}]^{-1}$ if homoskedastic errors
IV (just-identified)	$\hat{\beta}_{\text{IV}} = [\mathbf{Z}' \mathbf{X}]^{-1} \mathbf{Z}' \mathbf{y}$ $\hat{V}[\hat{\beta}] = N(\mathbf{Z}' \mathbf{X})^{-1} \hat{\mathbf{S}} (\mathbf{Z}' \mathbf{X})^{-1}$

^a Equations are based on a linear regression model with dependent variable \mathbf{y} , regressors \mathbf{X} , and instruments \mathbf{Z} . $\hat{\mathbf{S}}$ is defined in (6.40) and s^2 is defined after (6.41). All variance matrix estimates assume errors that are independent across observations and heteroskedastic, aside from the simplification for homoskedastic errors given for the 2SLS estimator. Optimal GMM uses the optimal weighting matrix.

rather than N in the formula for $\hat{\mathbf{S}}$. In the more restrictive case of homoskedastic errors, $E[u_i^2 | \mathbf{z}_i] = \sigma^2$ and so $\mathbf{S} = \lim N^{-1} \sum_i \sigma^2 E[\mathbf{z}_i \mathbf{z}_i']$, leading to estimate

$$\hat{\mathbf{S}} = s^2 \mathbf{Z}' \mathbf{Z} / N, \quad (6.41)$$

where $s^2 = (N - K)^{-1} \sum_{i=1}^N \hat{u}_i^2$ is consistent for σ^2 . These results mimic similar results for OLS presented in Section 4.4.5.

6.4.2. Different Linear GMM Estimators

Implementation of the results of Section 6.4.1 requires specification of the weighting matrix \mathbf{W}_N . For just-identified models all choices of \mathbf{W}_N lead to the same estimator. For overidentified models there are two common choices of \mathbf{W}_N , given in the following.

Table 6.2 summarizes these estimators and gives the appropriate specialization of the estimated variance matrix formula given in (6.39), assuming independent heteroskedastic errors.

Instrumental Variables Estimator

In the just-identified case $r = K$ and $\mathbf{X}' \mathbf{Z}$ is a square matrix that is invertible. Then $[\mathbf{X}' \mathbf{Z} \mathbf{W}_N \mathbf{Z}' \mathbf{X}]^{-1} = (\mathbf{Z}' \mathbf{X})^{-1} \mathbf{W}_N^{-1} (\mathbf{X}' \mathbf{Z})^{-1}$ and (6.37) simplifies to the **instrumental variables** estimator

$$\hat{\beta}_{\text{IV}} = (\mathbf{Z}' \mathbf{X})^{-1} \mathbf{Z}' \mathbf{y}, \quad (6.42)$$

introduced in Section 4.8.6. For just-identified models the GMM estimator for any choice of \mathbf{W}_N equals the IV estimator.

The simple IV estimator can also be used in overidentified models, by discarding some of the instruments so that the model is just-identified, but this results in an efficiency loss compared to using all the instruments.

Optimal-Weighted GMM

From Section 6.3.5, for overidentified models the most efficient GMM estimator, meaning GMM with optimal choice of weighting matrix, sets $\mathbf{W}_N = \widehat{\mathbf{S}}^{-1}$ in (6.37).

The **optimal GMM estimator or two-step GMM estimator** in the linear IV model is

$$\widehat{\boldsymbol{\beta}}_{\text{OGMM}} = [(\mathbf{X}'\mathbf{Z})\widehat{\mathbf{S}}^{-1}(\mathbf{Z}'\mathbf{X})]^{-1} (\mathbf{X}'\mathbf{Z})\widehat{\mathbf{S}}^{-1}(\mathbf{Z}'\mathbf{y}). \quad (6.43)$$

For heteroskedastic errors, $\widehat{\mathbf{S}}$ is computed using (6.40) based on a consistent first-step estimate $\widehat{\boldsymbol{\beta}}$ such as the 2SLS estimator defined in (6.44). White (1982) called this estimator a **two-stage IV estimator**, since both steps entail IV estimation.

The estimated asymptotic variance matrix for optimal GMM given in Table 6.2 is of relatively simple form as (6.39) simplifies when $\mathbf{W}_N = \widehat{\mathbf{S}}^{-1}$. In computing the estimated variance one can use $\widehat{\mathbf{S}}$ as presented in Table 6.2, but it is more common to instead use an estimator $\widetilde{\mathbf{S}}$, say, that is also computed using (6.40) but evaluates the residual at the optimal GMM estimator rather than the first-step estimate used to form $\widehat{\mathbf{S}}$ in (6.43).

Two-Stage Least Squares

If errors are homoskedastic rather than heteroskedastic, $\widehat{\mathbf{S}}^{-1} = [s^2 N^{-1} \mathbf{Z}' \mathbf{Z}]^{-1}$ from (6.41). Then $\mathbf{W}_N = (N^{-1} \mathbf{Z}' \mathbf{Z})^{-1}$ in (6.37), leading to the **two-stage least-squares estimator**, introduced in Section 4.8.7, that can be expressed compactly as

$$\widehat{\boldsymbol{\beta}}_{\text{2SLS}} = [\mathbf{X}' \mathbf{P}_{\mathbf{Z}} \mathbf{X}]^{-1} [\mathbf{X}' \mathbf{P}_{\mathbf{Z}} \mathbf{y}], \quad (6.44)$$

where $\mathbf{P}_{\mathbf{Z}} = \mathbf{Z}(\mathbf{Z}\mathbf{Z}')^{-1}\mathbf{Z}'$. The basis of the term two-stage least-squares is presented in the next section. The 2SLS estimator is also called the **generalized instrumental variables (GIV) estimator** as it generalizes the IV estimator to the overidentified case of more instruments than regressors. It is also called the **one-step GMM** because (6.44) can be calculated in one step, whereas optimal GMM requires two steps.

The 2SLS estimator is asymptotically normal distributed with estimated asymptotic variance given in Table 6.2. The general form should be used if one wishes to guard against heteroskedastic errors whereas the simpler form, presented in many introductory textbooks, is consistent only if errors are indeed homoskedastic.

Optimal GMM versus 2SLS

Both the optimal GMM and the 2SLS estimator lead to efficiency gains in overidentified models. Optimal GMM has the advantage of being more efficient than 2SLS, if errors are heteroskedastic, though the efficiency gain need not be great. Some of the GMM testing procedures given in Section 7.5 and Chapter 8 assume estimation

using the optimal weighting matrix. Optimal GMM has the disadvantage of requiring additional computation compared to 2SLS. Moreover, as discussed in Section 6.3.5, asymptotic theory may provide a poor small-sample approximation to the distribution of the optimal GMM estimator.

In cross-section applications it is common to use the less efficient 2SLS, though with inference based on heteroskedastic robust standard errors.

Even More Efficient GMM Estimation

The estimator $\widehat{\beta}_{OGMM}$ is the most efficient estimator based on the unconditional moment condition $E[\mathbf{z}_i u_i] = \mathbf{0}$, where $u_i = y_i - \mathbf{x}'_i \beta$. However, this is not the best moment condition to use if the starting point is the conditional moment condition $E[u_i | \mathbf{z}_i] = \mathbf{0}$ and errors are heteroskedastic, meaning $V[u_i | \mathbf{z}_i]$ varies with \mathbf{z}_i .

Applying the general results of Section 6.3.7, we can write the optimal moment condition for GMM estimation based on $E[u_i | \mathbf{z}_i] = \mathbf{0}$ as

$$E [E [\mathbf{x}_i | \mathbf{z}_i] u_i / V[u_i | \mathbf{z}_i]] = \mathbf{0}. \quad (6.45)$$

As with the LS regression example in Section 6.3.7, one should divide by the error variance $V[u | \mathbf{z}]$. Implementation is more difficult than in the LS case, however, as a model for $E[\mathbf{x} | \mathbf{z}]$ needs to be specified in addition to one for $V[u | \mathbf{z}]$. This may be possible with additional structure. In particular, for a linear simultaneous equations system $E[\mathbf{x}_i | \mathbf{z}_i]$ is linear in \mathbf{z} so that estimation is based on $E[\mathbf{x}_i u_i / V[u_i | \mathbf{z}_i]] = 0$.

For linear models the GMM estimator is usually based on the simpler condition $E[\mathbf{z}_i u_i] = \mathbf{0}$. Given this condition, the optimal GMM estimator defined in (6.43) is the most efficient GMM estimator.

6.4.3. Alternative Derivations of Two-Stage Least Squares

The 2SLS estimator, the standard IV estimator for overidentified models, was derived in Section 6.4.2 as a GMM estimator.

Here we present three other derivations of the 2SLS estimator. One of these derivations, due to Theil, provided the original motivation for 2SLS, which predates GMM. Theil's interpretation is emphasized in introductory treatments. However, it does not generalize to nonlinear models, whereas the GMM interpretation does.

We consider the linear model

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}, \quad (6.46)$$

with $E[\mathbf{u} | \mathbf{Z}] = \mathbf{0}$ and additionally $V[\mathbf{u} | \mathbf{Z}] = \sigma^2 \mathbf{I}$.

GLS in a Transformed Model

Premultiplication of (6.46) by the instruments \mathbf{Z}' yields the transformed model

$$\mathbf{Z}'\mathbf{y} = \mathbf{Z}'\mathbf{X}\beta + \mathbf{Z}'\mathbf{u}. \quad (6.47)$$

This transformed model is often used as motivation for the IV estimator when $r = K$, since ignoring $\mathbf{Z}'\mathbf{u}$ since $N^{-1}\mathbf{Z}'\mathbf{u} \rightarrow \mathbf{0}$ and solving yields $\widehat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$.

Here instead we consider the overidentified case. Conditional on \mathbf{Z} the error $\mathbf{Z}'\mathbf{u}$ has mean zero and variance $\sigma^2\mathbf{Z}'\mathbf{Z}$ given the assumptions after (6.46). The efficient GLS estimator of $\boldsymbol{\beta}$ in model (6.46) is then

$$\widehat{\boldsymbol{\beta}} = [\mathbf{X}'\mathbf{Z}(\sigma^2\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\sigma^2\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}, \quad (6.48)$$

which equals the 2SLS estimator in (6.44) since the multipliers σ^2 cancel out. More generally, note that if the transformed model (6.47) is instead estimated by WLS with weighting matrix \mathbf{W}_N then the more general estimator (6.37) is obtained.

Theil's Interpretation

Theil (1953) proposed estimation by OLS regression of the original model (6.46), except that the regressors \mathbf{X} are replaced by a prediction $\widehat{\mathbf{X}}$ that is asymptotically uncorrelated with the error term.

Suppose that in the **reduced form model** the regressors \mathbf{X} are a linear combination of the instruments plus some error, so that

$$\mathbf{X} = \mathbf{Z}\Pi + \mathbf{v}, \quad (6.49)$$

where Π is a $K \times r$ matrix. Multivariate OLS regression of \mathbf{X} on \mathbf{Z} yields estimator $\widehat{\Pi} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ and OLS predictions $\widehat{\mathbf{X}} = \mathbf{Z}\widehat{\Pi}$ or

$$\widehat{\mathbf{X}} = \mathbf{P}_{\mathbf{Z}}\mathbf{X},$$

where $\mathbf{P}_{\mathbf{Z}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$. OLS regression of \mathbf{y} on $\widehat{\mathbf{X}}$ rather than \mathbf{y} on \mathbf{X} yields estimator

$$\widehat{\boldsymbol{\beta}}_{\text{Theil}} = (\widehat{\mathbf{X}}'\widehat{\mathbf{X}})^{-1}\widehat{\mathbf{X}}'\mathbf{y}. \quad (6.50)$$

Theil's interpretation permits computation by two OLS regressions, with the first-stage OLS giving $\widehat{\mathbf{X}}$ and the second-stage OLS giving $\widehat{\boldsymbol{\beta}}$, leading to the term **two-stage least-squares estimator**.

To establish consistency of this estimator reexpress the linear model (6.46) as

$$\mathbf{y} = \widehat{\mathbf{X}}\boldsymbol{\beta} + (\mathbf{X} - \widehat{\mathbf{X}})\boldsymbol{\beta} + \mathbf{u}.$$

The second-stage OLS regression of \mathbf{y} on $\widehat{\mathbf{X}}$ yields a consistent estimator of $\boldsymbol{\beta}$ if the regressor $\widehat{\mathbf{X}}$ is asymptotically uncorrelated with the composite error term $(\mathbf{X} - \widehat{\mathbf{X}})\boldsymbol{\beta} + \mathbf{u}$. If $\widehat{\mathbf{X}}$ were any proxy variable there is no reason for this to hold; however, here $\widehat{\mathbf{X}}$ is uncorrelated with $(\mathbf{X} - \widehat{\mathbf{X}})$ as an OLS prediction is orthogonal to the OLS residual. Thus $\text{plim } N^{-1}\widehat{\mathbf{X}}'(\mathbf{X} - \widehat{\mathbf{X}})\boldsymbol{\beta} = \mathbf{0}$. Also,

$$N^{-1}\widehat{\mathbf{X}}'\mathbf{u} = N^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{u} = N^{-1}\mathbf{X}'\mathbf{Z}(N^{-1}\mathbf{Z}'\mathbf{Z})^{-1}N^{-1}\mathbf{Z}'\mathbf{u}.$$

Then $\widehat{\mathbf{X}}$ is asymptotically uncorrelated with \mathbf{u} provided \mathbf{Z} is a valid instrument so that $\text{plim } N^{-1}\mathbf{Z}'\mathbf{u} = \mathbf{0}$. This consistency result for $\widehat{\boldsymbol{\beta}}_{\text{Theil}}$ depends heavily on the linearity of the model and does not generalize to nonlinear models.

Theil's estimator in (6.50) equals the 2SLS estimator defined earlier in (6.44). We have

$$\begin{aligned}\hat{\beta}_{\text{Theil}} &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{P}_Z'\mathbf{P}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_Z\mathbf{y} \\ &= (\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_Z\mathbf{y},\end{aligned}$$

the 2SLS estimator, using $\mathbf{P}_Z'\mathbf{P}_Z = \mathbf{P}_Z$ in the final equality.

Care is needed in implementing 2SLS using Theil's method. The second-stage OLS will give the wrong standard errors, even if errors are homoskedastic, as it will estimate σ^2 using the second-stage OLS regression residuals $(\mathbf{y} - \hat{\mathbf{X}}\hat{\beta})$ rather than the actual residuals $(\mathbf{y} - \mathbf{X}\hat{\beta})$. In practice one may also make adjustment for heteroskedastic errors. It is much easier to use a program that offers 2SLS as an option and directly computes (6.44) and the associated variance matrix given in Table 6.2.

The 2SLS interpretation does not always carry over to nonlinear models, as detailed in Section 6.5.4. The GMM interpretation does, and for this reason it is emphasized here more than Theil's original derivation of linear 2SLS.

Theil actually considered a model where only some of the regressors \mathbf{X} are endogenous and the remaining are exogenous. The preceding analysis still applies, provided all the exogenous components of \mathbf{X} are included in the instruments \mathbf{Z} . Then the first-stage OLS regression of the exogenous regressors on the instruments fits perfectly and the predictions of the exogenous regressors equal their actual values. So in practice at the first-stage just the endogenous variables are regressed on the instruments, and the second-stage regression is of \mathbf{y} on the exogenous regressors and the first-stage predictions of the endogenous regressors.

Basmann's Interpretation

Basmann (1957) proposed using as instruments the OLS reduced form predictions $\hat{\mathbf{X}} = \mathbf{P}_Z\mathbf{X}$ for the simple IV estimator in the just-identified case, since there are then exactly as many instruments $\hat{\mathbf{X}}$ as regressors \mathbf{X} . This yields

$$\hat{\beta}_{\text{Basmann}} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y}. \quad (6.51)$$

This is consistent since $\text{plim } N^{-1}\hat{\mathbf{X}}'\mathbf{u} = \mathbf{0}$, as already shown for Theil's estimator.

The estimator (6.51) actually equals the 2SLS estimator defined in (6.44), since $\hat{\mathbf{X}}' = \mathbf{X}'\mathbf{P}_Z$.

This IV approach will lead to correct standard errors and can be extended to non-linear settings.

6.4.4. Alternatives to Standard IV Estimators

The IV-based optimal GMM and 2SLS estimators presented in Section 6.4.2 are the standard estimators used when regressors are endogenous. Chernozhukov and Hansen (2005) present an IV estimator for quantile regression.

Here we briefly discuss leading alternative estimators that have received renewed interest given the poor finite-sample properties of 2SLS with weak instruments detailed in Section 4.9. We focus on single-equation linear models. At this stage there is no method that is relatively efficient yet has small bias in small samples.

Limited-Information Maximum Likelihood

The **limited-information maximum likelihood (LIML) estimator** is obtained by joint ML estimation of the single equation (6.46) plus the reduced form for the endogenous regressors in the right-hand side of (6.46) assuming homoskedastic normal errors. For details see Greene (2003, p. 402) or Davidson and MacKinnon (1993, pp. 644–651). More generally the k class of estimators (see, for example, Greene, 2003, p. 403) includes LIML, 2SLS, and OLS.

The LIML estimator due to Anderson and Rubin (1949) predates the 2SLS estimator. Unlike 2SLS, the LIML estimator is invariant to the normalization used in a simultaneous equations system. Moreover, LIML and 2SLS are asymptotically equivalent given homoskedastic errors. Yet LIML is rarely used as it is more difficult to implement and harder to explain than 2SLS. Bekker (1994) presents small-sample results for LIML and a generalization of LIML. See also Hahn and Hausman (2002).

Split-Sample IV

Begin with Basmann's interpretation of 2SLS as an IV estimator given in (6.51). Substituting for \mathbf{y} from (6.46) yields

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\widehat{\mathbf{X}}'\mathbf{X})^{-1}\widehat{\mathbf{X}}'\mathbf{u}.$$

By assumption $\text{plim } N^{-1}\mathbf{Z}'\mathbf{u} = \mathbf{0}$ so $\text{plim } N^{-1}\widehat{\mathbf{X}}'\mathbf{u} = \mathbf{0}$ and $\widehat{\boldsymbol{\beta}}$ is consistent. However, correlation between \mathbf{X} and \mathbf{u} , the reason for IV estimation, means that $\widehat{\mathbf{X}} = \mathbf{P}_Z\mathbf{X}$ is correlated with \mathbf{u} . Thus $E[\widehat{\mathbf{X}}'\mathbf{u}] \neq \mathbf{0}$, which leads to bias in the IV estimator. This bias arises from using $\widehat{\mathbf{X}} = \mathbf{Z}\widehat{\boldsymbol{\Pi}}$ rather than $\widehat{\mathbf{X}} = \mathbf{Z}\boldsymbol{\Pi}$ as the instrument.

An alternative is to instead use as instrument predictions $\widetilde{\mathbf{X}}$, which have the property that $E[\widetilde{\mathbf{X}}'\mathbf{u}] = \mathbf{0}$ in addition to $\text{plim } N^{-1}\widetilde{\mathbf{X}}'\mathbf{u} = \mathbf{0}$, and use estimator

$$\widetilde{\boldsymbol{\beta}} = (\widetilde{\mathbf{X}}'\mathbf{X})^{-1}\widetilde{\mathbf{X}}'\mathbf{y}.$$

Since $E[\widetilde{\mathbf{X}}'\mathbf{u}] = \mathbf{0}$ does not imply $E[(\widetilde{\mathbf{X}}'\mathbf{X})^{-1}\widetilde{\mathbf{X}}'\mathbf{u}] = \mathbf{0}$, this estimator will still be biased, but the bias may be reduced.

Angrist and Krueger (1995) proposed obtaining such instruments by splitting the sample into two subsamples $(\mathbf{y}_1, \mathbf{X}_1, \mathbf{Z}_1)$ and $(\mathbf{y}_2, \mathbf{X}_2, \mathbf{Z}_2)$. The first sample is used to obtain estimate $\widehat{\boldsymbol{\Pi}}_1$ from regression of \mathbf{X}_1 on \mathbf{Z}_1 . The second sample is used to obtain the IV estimator where the instrument $\widetilde{\mathbf{X}}_2 = \mathbf{Z}_2\widehat{\boldsymbol{\Pi}}_1$ uses $\widehat{\boldsymbol{\Pi}}_1$ obtained from the separate first sample. Angrist and Krueger (1995) define the **unbiased split-sample IV estimator** as

$$\widetilde{\boldsymbol{\beta}}_{USSIV} = (\widetilde{\mathbf{X}}_2'\mathbf{X}_2)^{-1}\widetilde{\mathbf{X}}_2'\mathbf{y}_2.$$

The **split-sample IV estimator** $\tilde{\beta}_{SSIV} = (\tilde{\mathbf{X}}_2' \tilde{\mathbf{X}}_2)^{-1} \tilde{\mathbf{X}}_2' \mathbf{y}_2$ is a variant based on Theil's interpretation of 2SLS. These estimators have finite-sample bias toward zero, unlike 2SLS, which is biased toward OLS. However, considerable efficiency loss occurs because only half the sample is used at the final stage.

Jackknife IV

A more efficient variant of this estimator implements a similar procedure but generates instruments observation by observation.

Let the subscript $(-i)$ denote the leave-one-out operation that drops the i th observation. Then for the i th observation we obtain estimate $\widehat{\Pi}_i$ from regression of $\mathbf{X}_{(-i)}$ on $\mathbf{Z}_{(-i)}$ and use as instrument $\tilde{\mathbf{x}}'_i = \mathbf{z}'_i \widehat{\Pi}_i$. Repeating N times gives an instrument vector denoted $\tilde{\mathbf{X}}_{(-i)}$ with i th row $\tilde{\mathbf{x}}'_i$. This leads to the **jackknife IV estimator**

$$\tilde{\beta}_{JIV} = (\tilde{\mathbf{X}}'_{(-i)} \mathbf{X})^{-1} \tilde{\mathbf{X}}'_{(-i)} \mathbf{y}_2.$$

This estimator was originally proposed by Phillips and Hale (1977). Angrist, Imbens and Krueger (1999) and Blomquist and Dahlberg (1999) called it a jackknife estimator since the jackknife (see Section 11.5.5) is a leave-one-out method for bias reduction. The computational burden of obtaining the N jackknife predicted values $\tilde{\mathbf{x}}_i$ is modest by use of the recursive formula given in Section 11.5.5. The Monte Carlo evidence given in the two recent papers is mixed, however, indicating a potential for bias reduction but also an increase in the variance. So the jackknife version may not be better than the conventional version in terms of mean-square error. The earlier paper by Phillips and Hale (1977) presents analytical results that the finite-sample bias of the JIV estimator is smaller than that of 2SLS only for appreciably overidentified models with $r > 2(K + 1)$. See also Hahn, Hausman and Kuersteiner (2001).

Independently Weighted 2SLS

A related method to split-sample IV is the independently weighted GMM estimator of Altonji and Segal (1996) given in Section 6.3.5. Splitting the sample into G groups and specializing to linear IV yields the **independently weighted IV estimator**

$$\widehat{\beta}_{IWIV} = \frac{1}{G} \sum_{g=1}^G \left[\mathbf{X}'_g \mathbf{Z}_g \widehat{\mathbf{S}}_{(-g)}^{-1} \mathbf{Z}'_g \mathbf{X}_g \right]^{-1} \mathbf{X}'_g \mathbf{Z}_g \widehat{\mathbf{S}}_{(-g)}^{-1} \mathbf{Z}'_g \mathbf{y}_g,$$

where $\widehat{\mathbf{S}}_{(-g)}$ is computed using $\widehat{\mathbf{S}}$ defined in (6.40) except that observations from the g th group are excluded. In a panel application Ziliak (1997) found that the independently weighted IV estimator performed much better than the unbiased split-sample IV estimator.

6.5. Nonlinear Instrumental Variables

Nonlinear IV methods, notably nonlinear 2SLS proposed by Amemiya (1974), permit consistent estimates of nonlinear regression models in situations where the NLS

estimator is inconsistent because to regressors are correlated with the error term. We present these methods as a straightforward extension of the GMM approach for linear models.

Unlike the linear case the estimators have no explicit formula, but the asymptotic distribution can be obtained as a special case of the Section 6.3 results. This section presents single-equation results, with systems results given in Section 6.10.4. A fundamentally important result is that a natural extension of Theil's 2SLS method for linear models to nonlinear models can lead to inconsistent parameter estimates (see Section 6.5.4). Instead, the GMM approach should be used.

An alternative nonlinearity can arise when the model for the dependent variable is a linear model, but the reduced form for the endogenous regressor(s) is a nonlinear model owing to special features of the dependent variable. For example, the endogenous regressor may be a count or a binary outcome. In that case the linear methods of the previous section still apply. One approach is to ignore the special nature of the endogenous regressor and just do regular linear 2SLS or optimal GMM. Alternatively, obtain fitted values for the endogenous regressor by appropriate nonlinear regression, such as Poisson regression on all the instruments if the endogenous regressor is a count, and then do regular linear IV using this fitted value as the instrument for the count, following Basmann's approach. Both estimators are consistent, though they have different asymptotic distributions. The first simpler approach is the usual procedure.

6.5.1. Nonlinear GMM with Instruments

Consider the quite general nonlinear regression model where the error term may be additive or nonadditive (see Section 6.2.2). Thus

$$u_i = r(y_i, \mathbf{x}_i, \boldsymbol{\beta}), \quad (6.52)$$

where the nonlinear model with additive error is the special case

$$u_i = y_i - g(\mathbf{x}_i, \boldsymbol{\beta}), \quad (6.53)$$

where $g(\cdot)$ is a specified function. The estimators given in Section 6.2.2 are inconsistent if $E[u_i | \mathbf{x}_i] \neq 0$.

Assume the existence of r instruments \mathbf{z} , where $r \geq K$, that satisfy

$$E[u_i | \mathbf{z}_i] = 0. \quad (6.54)$$

This is the same conditional moment condition as in the linear case, except that $u_i = r(y_i, \mathbf{x}_i, \boldsymbol{\beta})$ rather than $u_i = y_i - \mathbf{x}'_i \boldsymbol{\beta}$.

Nonlinear GMM Estimator

By the law of iterated expectations, (6.54) leads to

$$E[\mathbf{z}_i u_i] = \mathbf{0}. \quad (6.55)$$

The GMM estimator minimizes the quadratic form in the corresponding sample moment condition.

In matrix notation let \mathbf{u} denote the $N \times 1$ error vector with i th entry u_i given in (6.52) and let \mathbf{Z} to be an $N \times r$ matrix of instruments with i th row \mathbf{z}'_i . Then $\sum_i \mathbf{z}_i u_i = \mathbf{Z}'\mathbf{u}$ and the **GMM estimator in the nonlinear IV model** $\hat{\beta}_{\text{GMM}}$ minimizes

$$\mathcal{Q}_N(\beta) = \left(\frac{1}{N} \mathbf{u}' \mathbf{Z} \right) \mathbf{W}_N \left(\frac{1}{N} \mathbf{Z}' \mathbf{u} \right), \quad (6.56)$$

where \mathbf{W}_N is an $r \times r$ weighting matrix. Unlike linear GMM, the first-order conditions do not lead to a closed-form solution for $\hat{\beta}_{\text{GMM}}$.

Distribution of Nonlinear GMM Estimator

The GMM estimator is consistent for β given (6.54) and asymptotically normally distributed with estimated asymptotic variance

$$\hat{\nabla} [\hat{\beta}_{\text{GMM}}] = N [\hat{\mathbf{D}}' \mathbf{Z} \mathbf{W}_N \mathbf{Z}' \hat{\mathbf{D}}]^{-1} [\hat{\mathbf{D}}' \mathbf{Z} \mathbf{W}_N \hat{\mathbf{S}} \mathbf{W}_N \mathbf{Z}' \hat{\mathbf{D}}] [\hat{\mathbf{D}}' \mathbf{Z} \mathbf{W}_N \mathbf{Z}' \hat{\mathbf{D}}]^{-1} \quad (6.57)$$

using the results from Section 6.3.3 with $\mathbf{h}(\cdot) = \mathbf{z}\mathbf{u}$, where $\hat{\mathbf{S}}$ is given in the following and $\hat{\mathbf{D}}$ is an $N \times K$ matrix of derivatives of the error term

$$\hat{\mathbf{D}} = \left. \frac{\partial \mathbf{u}}{\partial \beta'} \right|_{\hat{\beta}_{\text{GMM}}}. \quad (6.58)$$

With nonadditive errors, $\hat{\mathbf{D}}$ has i th row $\partial r(y_i, \mathbf{x}_i, \beta)/\partial \beta'|_{\hat{\beta}}$. With additive errors, $\hat{\mathbf{D}}$ has i th row $\partial g(\mathbf{x}_i, \beta)/\partial \beta'|_{\hat{\beta}}$, ignoring the minus sign that cancels out in (6.57).

For independent heteroskedastic errors,

$$\hat{\mathbf{S}} = N^{-1} \sum_i \hat{u}_i^2 \mathbf{z}_i \mathbf{z}'_i, \quad (6.59)$$

similar to the linear case except now $\hat{u}_i = r(y_i, \mathbf{x}, \hat{\beta})$ or $\hat{u}_i = y_i - g(\mathbf{x}, \hat{\beta})$.

The asymptotic variance of the GMM estimator in the nonlinear model is therefore the same as that in the linear case given in (6.39), with the change that the regressor matrix \mathbf{X} is replaced by the derivative $\partial \mathbf{u}/\partial \beta'|_{\hat{\beta}}$. This is exactly the same change as observed in Section 5.8 in going from linear to nonlinear least squares. By analogy with linear IV, the **rank condition for identification** is that $\text{plim } N^{-1} \mathbf{Z}' \partial \mathbf{u}/\partial \beta'|_{\beta_0}$ is of rank K and the weaker **order condition** is that $r \geq K$.

6.5.2. Different Nonlinear GMM Estimators.

Two leading specializations of the GMM estimator, which differ in the choice of weighting matrix, are optimal GMM that sets $\mathbf{W}_N = \hat{\mathbf{S}}^{-1}$ and nonlinear two-stage least squares (NL2SLS) that sets $\mathbf{W}_N = (\mathbf{Z}' \mathbf{Z})^{-1}$. Table 6.3 summarizes these estimators and their associated variance matrices, assuming independent heteroskedastic errors, and gives results for general \mathbf{W}_N and results for nonlinear IV in the just-identified model.

Table 6.3. GMM Estimators in Nonlinear IV Model and Their Asymptotic Variance^a

Estimator	Definition and Asymptotic Variance
GMM (general \mathbf{W}_N)	$Q_{\text{GMM}}(\beta) = \mathbf{u}' \mathbf{Z} \mathbf{W}_N \mathbf{Z}' \mathbf{u}$ $\widehat{\mathbf{V}}[\widehat{\beta}] = N[\widehat{\mathbf{D}}' \mathbf{Z} \mathbf{W}_N \mathbf{Z}' \widehat{\mathbf{D}}]^{-1} [\widehat{\mathbf{D}}' \mathbf{Z} \mathbf{W}_N \widehat{\mathbf{S}} \mathbf{W}_N \mathbf{Z}' \widehat{\mathbf{D}}] [\widehat{\mathbf{D}}' \mathbf{Z} \mathbf{W}_N \mathbf{Z}' \widehat{\mathbf{D}}]^{-1}$
Optimal GMM ($\mathbf{W}_N = \widehat{\mathbf{S}}^{-1}$)	$Q_{\text{OGMM}}(\beta) = \mathbf{u}' \mathbf{Z} \widehat{\mathbf{S}}^{-1} \mathbf{Z}' \mathbf{u}$ $\widehat{\mathbf{V}}[\widehat{\beta}] = N[\widehat{\mathbf{D}}' \mathbf{Z} \widehat{\mathbf{S}}^{-1} \mathbf{Z}' \widehat{\mathbf{D}}]^{-1}$
NL2SLS ($\mathbf{W}_N = [N^{-1} \mathbf{Z}' \mathbf{Z}]^{-1}$)	$Q_{\text{NL2SLS}}(\beta) = \mathbf{u}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{u}$ $\widehat{\mathbf{V}}[\widehat{\beta}] = N[\widehat{\mathbf{D}}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \widehat{\mathbf{D}}]^{-1} [\widehat{\mathbf{D}}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \widehat{\mathbf{S}} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \widehat{\mathbf{D}}]$ $\times [\widehat{\mathbf{D}}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \widehat{\mathbf{D}}]^{-1}$ $\widehat{\mathbf{V}}[\widehat{\beta}] = s^2 [\widehat{\mathbf{D}}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \widehat{\mathbf{D}}]^{-1}$ if homoskedastic errors
NLIV (just-identified)	$\widehat{\beta}_{\text{NLIV}}$ solves $\mathbf{Z}' \mathbf{u} = \mathbf{0}$ $\widehat{\mathbf{V}}[\widehat{\beta}] = N(\mathbf{Z}' \widehat{\mathbf{D}})^{-1} \widehat{\mathbf{S}} (\widehat{\mathbf{D}}' \mathbf{Z})^{-1}$

^a Equations are for a nonlinear regression model with error \mathbf{u} defined in (6.53) or (6.52) and instruments \mathbf{Z} . $\widehat{\mathbf{D}}$ is the derivative of the error vector with respect to β' evaluated at $\widehat{\beta}$ and simplifies for models with additive error to the derivative of the conditional mean function with respect to β' evaluated at $\widehat{\beta}$. $\widehat{\mathbf{S}}$ is defined in (6.59). All variance matrix estimates assume errors that are independent across observations and heteroskedastic, aside from the simplification for homoskedastic errors given for the NL2SLS estimator.

Nonlinear Instrumental Variables

In the just-identified case one can directly use the sample moment conditions corresponding to (6.55). This yields the **method of moments estimator in the nonlinear IV model** $\widehat{\beta}_{\text{NLIV}}$ that solves

$$\frac{1}{N} \sum_{i=1}^N \mathbf{z}_i u_i = \mathbf{0}, \quad (6.60)$$

or equivalently $\mathbf{Z}' \mathbf{u} = \mathbf{0}$ with asymptotic variance matrix given in Table 6.3.

Nonlinear estimators are often computed using iterative methods that obtain an optimum to an objective function rather than solve nonlinear systems of estimating equations. For the just-identified case $\widehat{\beta}_{\text{NLIV}}$ can be computed as a GMM estimator minimizing (6.56) with any choice of weighting matrix, most simply $\mathbf{W}_N = \mathbf{I}$, leading to the same estimate.

Optimal Nonlinear GMM

For overidentified models the optimal GMM estimator uses weighting matrix $\mathbf{W}_N = \widehat{\mathbf{S}}^{-1}$. The **optimal GMM estimator in the nonlinear IV model** $\widehat{\beta}_{\text{OGMM}}$ therefore minimizes

$$Q_N(\beta) = \left(\frac{1}{N} \mathbf{u}' \mathbf{Z} \right) \widehat{\mathbf{S}}^{-1} \left(\frac{1}{N} \mathbf{Z}' \mathbf{u} \right). \quad (6.61)$$

The estimated asymptotic variance matrix given in Table 6.3 is of relatively simple form as (6.57) simplifies when $\mathbf{W}_N = \widehat{\mathbf{S}}^{-1}$.

As in the linear case the optimal GMM estimator is a two-step estimator when errors are heteroskedastic. In computing the estimated variance one can use $\widehat{\mathbf{S}}$ as presented in Table 6.3, but it is more common to instead use an estimator $\widetilde{\mathbf{S}}$, say, that is also computed using (6.59) but evaluates the residual at the optimal GMM estimator rather than the first-step estimate used to form $\widehat{\mathbf{S}}$ in (6.61).

Nonlinear 2SLS

A special case of the GMM estimator with instruments sets $\mathbf{W}_N = (\mathbf{N}^{-1}\mathbf{Z}'\mathbf{Z})^{-1}$ in (6.56). This gives the **nonlinear two-stage least-squares** estimator $\widehat{\beta}_{\text{NL2SLS}}$ that minimizes

$$Q_N(\beta) = \frac{1}{N} \mathbf{u}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{u}. \quad (6.62)$$

This estimator has the attraction of being the optimal GMM estimator if errors are homoskedastic, as then $\widehat{\mathbf{S}} = s^2 \mathbf{Z}' \mathbf{Z} / N$, where s^2 is a consistent estimate of the constant $V[u|z]$ so $\widehat{\mathbf{S}}^{-1}$ is a multiple of $(\mathbf{Z}' \mathbf{Z})^{-1}$.

With homoskedastic error this estimator has the simpler estimated asymptotic variance given in Table 6.3, a result often given in textbooks. However, in microeconomics applications it is common to permit heteroskedastic errors and use the more complicated robust estimate also given in Table 6.3.

The NL2SLS estimator, proposed by Amemiya (1974), was an important precursor to GMM. The estimator can be motivated along similar lines to the first motivation for linear 2SLS given in Section 6.4.3. Thus premultiply the model error \mathbf{u} by the instruments \mathbf{Z}' to obtain $\mathbf{Z}' \mathbf{u}$, where $E[\mathbf{Z}' \mathbf{u}] = \mathbf{0}$ since $E[\mathbf{u}|Z] = \mathbf{0}$. Then do nonlinear GLS regression. Assuming homoskedastic errors this minimizes

$$Q_N(\beta) = \mathbf{u}' \mathbf{Z} [\sigma^2 \mathbf{Z}' \mathbf{Z}]^{-1} \mathbf{Z}' \mathbf{u},$$

as $V[\mathbf{u}|Z] = \sigma^2 \mathbf{I}$ implies $V[\mathbf{Z}' \mathbf{u}|Z] = \sigma^2 \mathbf{Z}' \mathbf{Z}$. This objective function is just a scalar multiple of (6.62).

The Theil two-stage interpretation of linear 2SLS does not always carry over to nonlinear models (see Section 6.5.4). Moreover, NL2SLS is clearly a one-step estimator. Amemiya chose the name NL2SLS because, as in the linear case, it permits consistent estimation using instrumental variables. The name should not be taken literally, and clearer terms are **nonlinear IV** or **nonlinear generalized IV estimation**.

Instrument Choice in Nonlinear Models

The preceding estimators presume the existence of instruments such that $E[u|z] = \mathbf{0}$ and that estimation is best if based on the unconditional moment condition $E[z u] = \mathbf{0}$.

Consider the nonlinear model with additive error so that $u = y - g(x, \beta)$. To be relevant the instrument must be correlated with the regressors x ; yet to be valid it cannot be a direct causal variable for y . From the variance matrix given in (6.57) it is actually correlation of z with $\partial g / \partial \beta$ rather than just x that matters, to ensure that $\widehat{\mathbf{D}}' \mathbf{Z}$ should be large. Weak instruments concerns are just as relevant here as in the linear case studied in Section 4.9.

Given likely heteroskedasticity the optimal moment condition on which to base estimation, given $E[u|\mathbf{z}] = \mathbf{0}$, is not $E[\mathbf{zu}] = \mathbf{0}$. From Section 6.3.7, however, the optimal moment condition requires additional moment assumptions that are difficult to make, so it is standard to use $E[\mathbf{zu}] = \mathbf{0}$ as has been done here.

An alternative way to control for heteroskedasticity is to base GMM estimation on an error term defined to be close to homoskedastic. For example, with count data rather than use $u = y - \exp(\mathbf{x}'\beta)$, work with the standardized error $u^* = u/\sqrt{\exp(\mathbf{x}'\beta)}$ (see Section 6.2.2). Note, however, that $E[u^*|\mathbf{z}] = 0$ and $E[u|\mathbf{z}] = 0$ are different assumptions.

Often just one component of \mathbf{x} is correlated with u . Then, as in the linear case, the exogenous components can be used as instruments for themselves and the challenge is to find an additional instrument that is uncorrelated with u . There are some nonlinear applications that arise from formal economic models as in Section 6.2.7, in which case the many subcomponents of the information set are available as instruments.

6.5.3. Poisson IV Example

The Poisson regression model with exogenous regressors specifies $E[y|\mathbf{x}] = \exp(\mathbf{x}'\beta)$. This can be viewed as a model with additive error $u = y - \exp(\mathbf{x}'\beta)$. If regressors are endogenous then $E[u|\mathbf{x}] \neq 0$ and the Poisson MLE will then be inconsistent. Consistent estimation assumes the existence of instruments \mathbf{z} that satisfy $E[u|\mathbf{z}] = 0$ or, equivalently,

$$E[y - \exp(\mathbf{x}'\beta)|\mathbf{z}] = 0.$$

The preceding results can be directly applied. The objective function is

$$Q_N(\beta) = \left[N^{-1} \sum_i \mathbf{z}_i u_i \right]' \mathbf{W}_N \left[N^{-1} \sum_i \mathbf{z}_i u_i \right],$$

where $u_i = y_i - \exp(\mathbf{x}'_i\beta)$. The first-order conditions are then

$$\left[\sum_i \exp(\mathbf{x}'_i\beta) \mathbf{x}_i \mathbf{z}'_i \right] \mathbf{W}_N \left[\sum_i \mathbf{z}_i (y_i - \exp(\mathbf{x}'_i\beta)) \right] = \mathbf{0}.$$

The asymptotic distribution is given in Table 6.3, with $\widehat{\mathbf{D}}' \mathbf{Z} = \sum_i e^{\mathbf{x}'_i\widehat{\beta}} \mathbf{x}_i \mathbf{z}'_i$ since $\partial g/\partial \beta = \exp(\mathbf{x}'\beta)\mathbf{x}$ and $\widehat{\mathbf{S}}$ defined in (6.39) with $\widehat{u}_i = y_i - \exp(\mathbf{x}'_i\widehat{\beta})$. The optimal GMM and NL2SLS estimators differ in whether the weighting matrix is $\widehat{\mathbf{S}}^{-1}$ or $(N^{-1}\mathbf{Z}'\mathbf{Z})^{-1}$, where $\mathbf{Z}'\mathbf{Z} = \sum_i \mathbf{z}_i \mathbf{z}'_i$.

An alternative consistent estimator follows the Basmann approach. First, estimate by OLS the reduced form $\mathbf{x}_i = \boldsymbol{\Pi} \mathbf{z}_i + \mathbf{v}_i$ giving K predictions $\widehat{\mathbf{x}}_i = \widehat{\boldsymbol{\Pi}} \mathbf{z}_i$. Second, estimate by nonlinear IV as in (6.60) with instruments $\widehat{\mathbf{x}}_i$ rather than \mathbf{z}_i . Given the OLS formula for $\widehat{\boldsymbol{\Pi}}$ this estimator solves

$$\left[\sum_i \mathbf{x}_i \mathbf{z}'_i \right] \left[\sum_i \mathbf{z}_i \mathbf{z}'_i \right]^{-1} \left[\sum_i (y_i - \exp(\mathbf{x}'_i\beta)) \mathbf{z}_i \right] = \mathbf{0}.$$

This estimator differs from the NL2SLS estimator because the first term in the left-hand side differs. Potential problems with instead generalizing Theil's method for linear models are detailed in the next section.

Similar issues arise in nonlinear models other than Poisson regression, such as models for binary data.

6.5.4. Two-Stage Estimation in Nonlinear Models

The usual interpretation of linear 2SLS can fail in nonlinear models. Thus suppose y has mean $g(\mathbf{x}, \beta)$ and there are instruments \mathbf{z} for the regressors \mathbf{x} . Then OLS regression of \mathbf{x} on instruments \mathbf{z} to get fitted values $\hat{\mathbf{x}}$ followed by NLS regression of y on $g(\hat{\mathbf{x}}, \beta)$ can lead to inconsistent parameter estimates of β , as we now demonstrate. Instead, one needs to use the NL2SLS estimator presented in the previous section.

Consider the following simple model, based on one presented in Amemiya (1984), that is nonlinear in variables though still linear in parameters. Let

$$\begin{aligned} y &= \beta x^2 + u, \\ x &= \pi z + v, \end{aligned} \tag{6.63}$$

where the zero-mean errors u and v are correlated. The regressor x^2 is endogenous, since x is a function of v and by assumption u and v are correlated. As a result the OLS estimator of β is inconsistent. If z is generated independently of the other random variables in the model it is a valid instrument as it is clearly then independent of u but correlated with x .

The IV estimator is $\hat{\beta}_{\text{IV}} = (\sum_i z_i x_i^2)^{-1} \sum_i z_i y_i$. This can be implemented by a regular IV regression of y on x^2 with instrument z . Some algebra shows that, as expected, $\hat{\beta}_{\text{IV}}$ equals the nonlinear IV estimator defined in (6.60).

Suppose instead we perform the following two-stage least-squares estimation. First, regress x on z to get $\hat{x} = \hat{\pi}z$ and then regress y on \hat{x}^2 . Then $\hat{\beta}_{\text{2SLS}} = (\sum_i \hat{x}_i^2 \hat{x}_i^2)^{-1} \sum_i \hat{x}_i^2 y_i$, where \hat{x}_i^2 is the square of the prediction \hat{x}_i obtained from OLS regression of x on z . This yields an inconsistent estimate. Adapting the proof for the linear case in Section 6.4.3 we have

$$\begin{aligned} y_i &= \beta x_i^2 + u_i \\ &= \beta \hat{x}_i^2 + w_i, \end{aligned}$$

where $w_i = \beta(x_i^2 - \hat{x}_i^2) + u_i$. An OLS regression of y_i on \hat{x}_i^2 is inconsistent for β because the regressor \hat{x}_i^2 is asymptotically correlated with the composite error term w_i . Formally, $(x_i^2 - \hat{x}_i^2) = (\pi z_i + v_i)^2 - (\hat{\pi}z_i)^2 = \pi^2 z_i^2 + 2\pi z_i v_i + v_i^2 - \hat{\pi}^2 z_i^2$ implies, using $\text{plim } \hat{\pi} = \pi$ and some algebra, that $\text{plim } N^{-1} \sum_i \hat{x}_i^2 (x_i^2 - \hat{x}_i^2) = \text{plim } N^{-1} \sum_i \pi^2 z_i^2 v_i^2 \neq 0$ even if z_i and v_i are independent. Hence $\text{plim } N^{-1} \sum_i \hat{x}_i^2 w_i \neq \text{plim } N^{-1} \sum_i \hat{x}_i^2 \beta(x_i - \hat{x}_i)^2 = 0$.

A variation that is consistent, however, is to regress x^2 rather than x on z at the first stage and use the prediction $\hat{x}^2 \neq (\hat{x})^2$ at the second stage. It can be shown that this equals $\hat{\beta}_{\text{IV}}$. The instrument for x^2 needs to be the fitted value for x^2 rather than the square of the fitted value for x .

This example generalizes to other nonlinear models where the nonlinearity is in regressors only, so that

$$y = \mathbf{g}(\mathbf{x})' \beta + u,$$

Table 6.4. Nonlinear Two-Stage Least-Squares Example^a

Variable	Estimator		
	OLS	NL2SLS	Two-Stage
x^2	1.189 (0.025)	0.960 (0.046)	1.642 (0.172)
R^2	0.88	0.85	0.80

^a The dgp given in the text has true coefficient equal to one. The sample size is $N = 200$.

where $\mathbf{g}(\mathbf{x})$ is a nonlinear function of \mathbf{x} . Common examples are use of powers and natural logarithm. Suppose $E[u|\mathbf{z}] = 0$. Inconsistent estimates are obtained by regressing \mathbf{x} on \mathbf{z} to get predictions $\hat{\mathbf{x}}$, and then regressing y on $\mathbf{g}(\hat{\mathbf{x}})$. Consistent estimates can be obtained by instead regressing $\mathbf{g}(\mathbf{x})$ on \mathbf{z} to get predictions $\hat{\mathbf{g}}(\mathbf{x})$, and then regressing y on $\hat{\mathbf{g}}(\mathbf{x})$ at the second stage. We use $\hat{\mathbf{g}}(\mathbf{x})$ rather than $\mathbf{g}(\hat{\mathbf{x}})$ as instrument for $\mathbf{g}(\mathbf{x})$. Even then the second-stage regression gives invalid standard errors as OLS output will use residuals $\hat{u} = y - \hat{\mathbf{g}}(\mathbf{x})'\hat{\beta}$ rather than $\hat{u} = y - \mathbf{g}(\mathbf{x})'\hat{\beta}$. It is best to directly use a GMM or NL2SLS command.

More generally models may be nonlinear in both variables and parameters. Consider a single-index model with additive error, so that

$$y = g(\mathbf{x}'\beta) + u.$$

Inconsistent estimates may be obtained by OLS of \mathbf{x} on \mathbf{z} to get predictions $\hat{\mathbf{x}}$, and then NLS regression of y on $g(\hat{\mathbf{x}}'\beta)$. Either GMM or NL2SLS needs to be used. Essentially, for consistency we want $\hat{g}(\mathbf{x}'\beta)$, not $g(\hat{\mathbf{x}}'\beta)$.

NL2SLS Example

We consider NL2SLS estimation in a model with a simple nonlinearity resulting from the square of an endogenous variable appearing as a regressor, as in the previous section.

The dgp is (6.63), so $y = \beta x^2 + u$ and $x = \pi z + v$, where $\beta = 1$, and $\pi = 1$, and $z = 1$ for all observations and (u, v) are joint normal with means 0, variances 1, and correlation 0.8. A sample of size 200 is drawn. Results are shown in Table 6.4.

The nonlinearity here is quite mild with the square of x rather than x appearing as regressor. Interest lies in estimating its coefficient β . The OLS estimator is inconsistent, whereas NL2SLS is consistent. The two-stage method where first an OLS regression of x on z is used to form \hat{x} and then an OLS regression of y on $(\hat{x})^2$ is performed that yields an estimate that is more than two standard errors from the true value of $\beta = 1$. The simulation also indicates a loss in goodness of fit and precision with larger standard errors and lower R^2 , similar to linear IV.

6.6. Sequential Two-Step m-Estimation

Sequential two-step estimation procedures are estimation procedures where the estimate of a parameter of ultimate interest is based on initial estimation of an unknown parameter. An example is feasible GLS when the error has conditional variance $\exp(\mathbf{z}'\gamma)$. Given an estimate $\tilde{\gamma}$ of γ , the FGLS estimator $\hat{\beta}$ solves $\sum_{i=1}^N (y_i - \mathbf{x}_i'\hat{\beta})^2 / \exp(\mathbf{z}_i'\tilde{\gamma})$. A second example is the Heckman two-step estimator given in Section 16.10.2.

These estimators are attractive as they can provide a relatively simple way to obtain consistent parameter estimates. However, for valid statistical inference it may be necessary to adjust the asymptotic variance of the second-step estimator to allow for the first-step estimation. We present results for the special case where the estimating equations for both the first- and second-step estimators set a sample average to zero, which is the case for m-estimators, method of moments, and estimating equations estimators.

Partition the parameter vector θ into θ_1 and θ_2 , with ultimate interest in θ_2 . The model is estimated sequentially by first obtaining $\hat{\theta}_1$ that solves $\sum_{i=1}^N \mathbf{h}_{1i}(\hat{\theta}_1) = \mathbf{0}$ and then, given $\hat{\theta}_1$, obtaining $\hat{\theta}_2$ that solves $N^{-1} \sum_{i=1}^N \mathbf{h}_{2i}(\hat{\theta}_1, \hat{\theta}_2) = \mathbf{0}$. In general the distribution of $\hat{\theta}_2$ given estimation of $\hat{\theta}_1$ differs from, and is more complicated than, the distribution of $\hat{\theta}_2$ if θ_1 is known. Statistical inference is invalid if it fails to take into account this complication, except in some special cases given at the end of this section.

The following derivation is given in Newey (1984), with similar results obtained by Murphy and Topel (1985) and Pagan (1986). The two-step estimator can be rewritten as a one-step estimator where (θ_1, θ_2) jointly solve the equations

$$\begin{aligned} N^{-1} \sum_{i=1}^N \mathbf{h}_1(\mathbf{w}_i, \hat{\theta}_1) &= \mathbf{0}, \\ N^{-1} \sum_{i=1}^N \mathbf{h}_2(\mathbf{w}_i, \hat{\theta}_1, \hat{\theta}_2) &= \mathbf{0}. \end{aligned} \quad (6.64)$$

Defining $\theta = (\theta'_1 \quad \theta'_2)'$ and $\mathbf{h}_i = (\mathbf{h}'_{1i} \quad \mathbf{h}'_{2i})'$, we can write the equations as

$$N^{-1} \sum_{i=1}^N \mathbf{h}(\mathbf{w}_i, \hat{\theta}) = \mathbf{0}.$$

In this setup it is assumed that $\dim(\mathbf{h}_1) = \dim(\theta_1)$ and $\dim(\mathbf{h}_2) = \dim(\theta_2)$, so that the number of estimating equations equals the number of parameters. Then (6.64) is an estimating equations estimator or MM estimator.

Consistency requires that $\text{plim } N^{-1} \sum_i \mathbf{h}(\mathbf{w}_i, \theta_0) = \mathbf{0}$, where $\theta_0 = [\theta'_{10} \quad \theta'_{20}]$. This condition should be satisfied if $\hat{\theta}_1$ is consistent for θ_{10} in the first step, and if second-step estimation of $\hat{\theta}_2$ with θ_{10} known (rather than estimated by $\hat{\theta}_1$) would lead to a consistent estimate of θ_{20} . Within a method of moments framework we require $E[\mathbf{h}_1(\theta_1)] = \mathbf{0}$ and $E[\mathbf{h}_2(\theta_1, \theta_2)] = \mathbf{0}$. We assume that consistency is established.

For the asymptotic distribution we apply the general result that

$$\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{G}_0^{-1}\mathbf{S}_0(\mathbf{G}_0^{-1})'],$$

6.6. SEQUENTIAL TWO-STEP M-ESTIMATION

where \mathbf{G}_0 and \mathbf{S}_0 are defined in Proposition 6.1. Partition \mathbf{G}_0 and \mathbf{S}_0 in a similar way to the partitioning of $\boldsymbol{\theta}$ and \mathbf{h}_i . Then

$$\mathbf{G}_0 = \lim \frac{1}{N} \sum_{i=1}^N \mathbf{E} \begin{bmatrix} \partial \mathbf{h}_{1i} / \partial \boldsymbol{\theta}'_1 & \mathbf{0} \\ \partial \mathbf{h}_{2i} / \partial \boldsymbol{\theta}'_1 & \partial \mathbf{h}_{2i} / \partial \boldsymbol{\theta}'_2 \end{bmatrix} = \begin{bmatrix} \mathbf{G}_{11} & \mathbf{0} \\ \mathbf{G}_{21} & \mathbf{G}_{22} \end{bmatrix},$$

using $\partial \mathbf{h}_{1i}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}'_2 = \mathbf{0}$ since $\mathbf{h}_{1i}(\boldsymbol{\theta})$ is not a function of $\boldsymbol{\theta}_2$ from (6.64). Since \mathbf{G}_0 , \mathbf{G}_{11} , and \mathbf{G}_{22} are square matrices

$$\mathbf{G}_0^{-1} = \begin{bmatrix} \mathbf{G}_{11}^{-1} & \mathbf{0} \\ -\mathbf{G}_{22}^{-1} \mathbf{G}_{21} \mathbf{G}_{11}^{-1} & \mathbf{G}_{22}^{-1} \end{bmatrix}.$$

Clearly,

$$\mathbf{S}_0 = \lim \frac{1}{N} \sum_{i=1}^N \mathbf{E} \begin{bmatrix} \mathbf{h}_{1i} \mathbf{h}_{1i}' & \mathbf{h}_{1i} \mathbf{h}_{2i}' \\ \mathbf{h}_{2i} \mathbf{h}_{1i}' & \mathbf{h}_{2i} \mathbf{h}_{2i}' \end{bmatrix} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix}.$$

The asymptotic variance of $\widehat{\boldsymbol{\theta}}_2$ is the (2, 2) submatrix of the variance matrix of $\widehat{\boldsymbol{\theta}}$. After some algebra, we get

$$\mathbf{V}[\widehat{\boldsymbol{\theta}}_2] = \mathbf{G}_{22}^{-1} \left\{ \begin{array}{l} \mathbf{S}_{22} + \mathbf{G}_{21} [\mathbf{G}_{11}^{-1} \mathbf{S}_{11} \mathbf{G}_{11}^{-1}] \mathbf{G}_{21}' \\ -\mathbf{G}_{21} \mathbf{G}_{11}^{-1} \mathbf{S}_{12} - \mathbf{S}_{21} \mathbf{G}_{11}^{-1} \mathbf{G}_{21}' \end{array} \right\} \mathbf{G}_{22}^{-1}. \quad (6.65)$$

The usual computer output yields standard errors that are incorrect and understate the true standard errors, since $\mathbf{V}[\widehat{\boldsymbol{\theta}}_2]$ is then assumed to be $\mathbf{G}_{22}^{-1} \mathbf{S}_{22} \mathbf{G}_{22}^{-1}$, which can be shown to be smaller than the true variance given in (6.65).

There is no need to account for additional variability in the second-step caused by estimation in the first step in the special case that $\mathbf{E}[\partial \mathbf{h}_{2i}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}_1] = \mathbf{0}$, as then $\mathbf{G}_{21} = \mathbf{0}$ and $\mathbf{V}[\widehat{\boldsymbol{\theta}}_2]$ in (6.65) reduces to $\mathbf{G}_{22}^{-1} \mathbf{S}_{22} \mathbf{G}_{22}^{-1}$.

A well-known example of $\mathbf{G}_{21} = \mathbf{0}$ is FGLS. Then for heteroskedastic errors

$$\mathbf{h}_{2i}(\boldsymbol{\theta}) = \frac{\mathbf{x}_{2i}(y_i - \mathbf{x}'_i \boldsymbol{\theta}_2)}{\sigma(\mathbf{x}_i, \boldsymbol{\theta}_1)},$$

where $\mathbf{V}[y_i | \mathbf{x}_i] = \sigma^2(\mathbf{x}_i, \boldsymbol{\theta}_1)$, and

$$\mathbf{E}[\partial \mathbf{h}_{2i}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}_1] = \mathbf{E} \left[-\mathbf{x}_{2i} \frac{(y_i - \mathbf{x}'_i \boldsymbol{\theta}_2)}{\sigma(\mathbf{x}_i, \boldsymbol{\theta}_1)^2} \frac{\partial \sigma(\mathbf{x}_i, \boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_1} \right],$$

which equals zero since $\mathbf{E}[y_i | \mathbf{x}_i] = \mathbf{x}'_i \boldsymbol{\theta}_2$. Furthermore, for FGLS consistency of $\widehat{\boldsymbol{\theta}}_2$ does not require that $\widehat{\boldsymbol{\theta}}_1$ be consistent since $\mathbf{E}[\mathbf{h}_{2i}(\boldsymbol{\theta})] = \mathbf{0}$ just requires that $\mathbf{E}[y_i | \mathbf{x}_i] = \mathbf{x}'_i \boldsymbol{\theta}_2$, which does not depend on $\boldsymbol{\theta}_1$.

A second example of $\mathbf{G}_{21} = \mathbf{0}$ is ML estimation with a block diagonal matrix so that $\mathbf{E}[\partial^2 \mathcal{L}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}'_2] = \mathbf{0}$. This is the case for example for regression under normality, where $\boldsymbol{\theta}_1$ are the variance parameters and $\boldsymbol{\theta}_2$ are the regression parameters.

In other examples, however, $\mathbf{G}_{21} \neq \mathbf{0}$ and the more cumbersome expression (6.65) needs to be used. This is done automatically by computer packages for some standard two-step estimators, most notably Heckman's two-step estimator of the sample selection model given in Section 16.5.4. Otherwise, $\mathbf{V}[\widehat{\boldsymbol{\theta}}_2]$ needs to be computed manually. Many of the components come from earlier estimation. In particular, $\mathbf{G}_{11}^{-1} \mathbf{S}_{11} \mathbf{G}_{11}^{-1}$ is

the robust variance matrix of $\widehat{\theta}_1$ and $\mathbf{G}_{22}^{-1}\mathbf{S}_{22}\mathbf{G}_{22}^{-1}$ is the robust variance matrix estimate of $\widehat{\theta}_2$ that incorrectly ignores the estimation error in $\widehat{\theta}_1$. For data independent over i the subcomponents of the \mathbf{S}_0 submatrix are consistently estimated by $\widehat{\mathbf{S}}_{jk} = N^{-1} \sum_i \widehat{\mathbf{h}}_{ji} \widehat{\mathbf{h}}_{ki}'$, $j, k = 1, 2$. This leaves computation of $\widehat{\mathbf{G}}_{21} = N^{-1} \sum_i \partial \mathbf{h}_{2i} / \partial \boldsymbol{\theta}_1|_{\widehat{\theta}}$ as the main challenge.

A recommended simpler approach is to obtain bootstrap standard errors (see Section 16.2.5), or directly jointly estimate $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in the combined model (6.64), assuming access to a GMM routine.

These simpler approaches can also be applied to sequential estimators that are GMM estimators rather than m-estimators. Then combining the two estimators will lead to a set of conditions more complicated than (6.64) and we no longer get (6.65). However, one can still bootstrap or estimate jointly rather than sequentially.

6.7. Minimum Distance Estimation

Minimum distance estimation provides a way to estimate structural parameters $\boldsymbol{\theta}$ that are a specified function of reduced form parameters $\boldsymbol{\pi}$, given a consistent estimate $\widehat{\boldsymbol{\pi}}$ of $\boldsymbol{\pi}$.

A standard reference is Ferguson (1958). Rothenberg (1973) applied this method to linear simultaneous equations models, though the alternative methods given in Section 6.9.6 are the standard methods used. Minimum distance estimation is most often used in panel data analysis. In the initial work by Chamberlain (1982, 1984) (see Section 22.2.7) he lets $\widehat{\boldsymbol{\pi}}$ be OLS estimates from linear regression of the current-period dependent variable on regressors in all periods. Subsequent applications to covariance structures (see Section 22.5.4) let $\widehat{\boldsymbol{\pi}}$ be estimated variances and autocovariances of the panel data. See also the indirect inference method (Section 12.6).

Suppose that the relationship between q structural parameters and $r > q$ reduced form parameters is that $\boldsymbol{\pi}_0 = \mathbf{g}(\boldsymbol{\theta}_0)$. Further suppose that we have a consistent estimate $\widehat{\boldsymbol{\pi}}$ of the reduced form parameters. An obvious estimator is $\widehat{\boldsymbol{\theta}}$ such that $\widehat{\boldsymbol{\pi}} = \mathbf{g}(\widehat{\boldsymbol{\theta}})$, but this is infeasible since $q < r$. Instead, the **minimum distance (MD) estimator** $\widehat{\boldsymbol{\theta}}_{\text{MD}}$ minimizes with respect to $\boldsymbol{\theta}$ the objective function

$$Q_N(\boldsymbol{\theta}) = (\widehat{\boldsymbol{\pi}} - \mathbf{g}(\boldsymbol{\theta}))' \mathbf{W}_N (\widehat{\boldsymbol{\pi}} - \mathbf{g}(\boldsymbol{\theta})), \quad (6.66)$$

where \mathbf{W}_N is an $r \times r$ weighting matrix.

If $\widehat{\boldsymbol{\pi}} \xrightarrow{P} \boldsymbol{\pi}_0$ and $\mathbf{W}_N \xrightarrow{P} \mathbf{W}_0$, where \mathbf{W}_0 is finite positive semidefinite then $Q_N(\widehat{\boldsymbol{\theta}}) \xrightarrow{P} Q_0(\boldsymbol{\theta}) = (\boldsymbol{\pi}_0 - \mathbf{g}(\boldsymbol{\theta}))' \mathbf{W}_0 (\boldsymbol{\pi}_0 - \mathbf{g}(\boldsymbol{\theta}))$. It follows that $\boldsymbol{\theta}_0$ is locally identified if $\text{Rank}[\mathbf{W}_0 \times \partial \mathbf{g}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}'] = q$, while consistency essentially requires that $\boldsymbol{\pi}_0 = \mathbf{g}(\boldsymbol{\theta}_0)$.

For the MD estimator $\sqrt{N}(\widehat{\boldsymbol{\theta}}_{\text{MD}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{V}[\widehat{\boldsymbol{\theta}}_{\text{MD}}]]$, where

$$\mathbf{V}[\widehat{\boldsymbol{\theta}}_{\text{MD}}] = (\mathbf{G}_0' \mathbf{W}_0 \mathbf{G}_0)^{-1} (\mathbf{G}_0' \mathbf{W}_0 \mathbf{V}[\widehat{\boldsymbol{\pi}}] \mathbf{W}_0 \mathbf{G}_0) (\mathbf{G}_0' \mathbf{W}_0 \mathbf{G}_0)^{-1}, \quad (6.67)$$

$\mathbf{G}_0 = \partial \mathbf{g}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}'|_{\boldsymbol{\theta}_0}$, and it is assumed that the reduced form parameters $\widehat{\boldsymbol{\pi}}$ have limit distribution $\sqrt{N}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{V}[\widehat{\boldsymbol{\pi}}]]$. More efficient reduced form estimators lead to more efficient MD estimators, since smaller $\mathbf{V}[\widehat{\boldsymbol{\pi}}]$ leads to smaller $\mathbf{V}[\widehat{\boldsymbol{\theta}}_{\text{MD}}]$ in (6.67).

To obtain the result (6.67), begin with the following rescaling of the first-order conditions for the MD estimator:

$$G_N(\widehat{\boldsymbol{\theta}})' \mathbf{W}_N \sqrt{N}(\widehat{\boldsymbol{\pi}} - \mathbf{g}(\widehat{\boldsymbol{\theta}})) = \mathbf{0}, \quad (6.68)$$

where $G_N(\boldsymbol{\theta}) = \partial \mathbf{g}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}'$. An exact first-order Taylor series expansion about $\boldsymbol{\theta}_0$ yields

$$\sqrt{N} \mathbf{h}(\widehat{\boldsymbol{\pi}} - \mathbf{g}(\widehat{\boldsymbol{\theta}})) = \sqrt{N}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0) - G_N(\boldsymbol{\theta}^+) \sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \quad (6.69)$$

where $\boldsymbol{\theta}^+$ lies between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$ and we have used $\mathbf{g}(\boldsymbol{\theta}_0) = \boldsymbol{\pi}_0$. Substituting (6.69) back into (6.68) and solving for $\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ yields

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = [G_N(\widehat{\boldsymbol{\theta}})' \mathbf{W}_N G_N(\boldsymbol{\theta}^+)]^{-1} G_N(\widehat{\boldsymbol{\theta}})' \mathbf{W}_N \sqrt{N}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0), \quad (6.70)$$

which leads directly to (6.67).

For given reduced form estimator $\widehat{\boldsymbol{\pi}}$, the most efficient MD estimator uses weighting matrix $\mathbf{W}_N = \widehat{\mathbf{V}}[\widehat{\boldsymbol{\pi}}]^{-1}$ in (6.66). This estimator is called the **optimal MD (OMD) estimator**, and sometimes the **minimum chi-square estimator** following Ferguson (1958).

A common alternative special case is the **equally weighted minimum distance (EWMD) estimator**, which sets $\mathbf{W}_N = \mathbf{I}$. This is less efficient than the OMD estimator, but it does not have the finite-sample bias problems analogous to those discussed in Section 6.3.5 that arise when the optimal weighting matrix is used. The EWMD estimator can be simply obtained by NLS regression of $\widehat{\boldsymbol{\pi}}_j$ on $g_j(\widehat{\boldsymbol{\theta}})$, $j = 1, \dots, r$, since minimizing $(\widehat{\boldsymbol{\pi}} - \mathbf{g}(\widehat{\boldsymbol{\theta}}))'(\widehat{\boldsymbol{\pi}} - \mathbf{g}(\widehat{\boldsymbol{\theta}}))$ yields the same first-order conditions as those in (6.68) with $\mathbf{W}_N = \mathbf{I}$.

The maximized value of the objective function for the OMD is chi-squared distributed. Specifically,

$$(\widehat{\boldsymbol{\pi}} - \mathbf{g}(\widehat{\boldsymbol{\theta}}_{\text{OMD}}))' \widehat{\mathbf{V}}[\widehat{\boldsymbol{\pi}}]^{-1} (\widehat{\boldsymbol{\pi}} - \mathbf{g}(\widehat{\boldsymbol{\theta}}_{\text{OMD}})) \quad (6.71)$$

is asymptotically distributed as $\chi^2(r - q)$ under $H_0 : \mathbf{g}(\boldsymbol{\theta}_0) = \boldsymbol{\pi}_0$. This provides a model specification test analogous to the OIR test of Section 6.3.8.

The MD estimator is qualitatively similar to the GMM estimator. The GMM framework is the standard one employed. MD estimation is most often used in panel studies of covariance structures, since then $\widehat{\boldsymbol{\pi}}$ comprises easily estimated sample moments (variances and covariances) that can then be used to obtain $\widehat{\boldsymbol{\theta}}$.

6.8. Empirical Likelihood

The MM and GMM approaches do not require complete specification of the conditional density. Instead, estimation is based on moment conditions of the form $E[\mathbf{h}(y, \mathbf{x}, \boldsymbol{\theta})] = \mathbf{0}$. The empirical likelihood approach, due to Owen (1988), is an alternative estimation procedure based on the same moment condition.

An attraction of the empirical likelihood estimator is that, although it is asymptotically equivalent to the GMM estimator, it has different finite-sample properties, and in some examples it outperforms the GMM estimator.

6.8.1. Empirical Likelihood Estimation of Population Mean

We begin with empirical likelihood in the case of a scalar iid random variable y with density $f(y)$ and sample likelihood function $\prod_i f(y_i)$. The complication considered here is that the density $f(y)$ is not specified, so the usual ML approach is not possible.

A completely nonparametric approach seeks to estimate the density $f(y)$ evaluated at each of the sample values of y . Let $\pi_i = f(y_i)$ denote the probability that the i th observation on y takes the realized value y_i . Then the goal is to maximize the so-called empirical likelihood function $\prod_i \pi_i$, or equivalently to maximize the empirical log-likelihood function $N^{-1} \sum_i \ln \pi_i$, which is a multinomial model with no structure placed on π_i . This log-likelihood is unbounded, unless a constraint is placed on the range of values taken by π_i . The normalization used is that $\sum_i \pi_i = 1$. This yields the standard estimate of the cumulative distribution function in the fully nonparametric case, as we now demonstrate.

The empirical likelihood estimator maximizes with respect to $\boldsymbol{\pi}$ and η the Lagrangian

$$\mathcal{L}_{\text{EL}}(\boldsymbol{\pi}, \eta) = \frac{1}{N} \sum_{i=1}^N \ln \pi_i - \eta \left(\sum_{i=1}^N \pi_i - 1 \right), \quad (6.72)$$

where $\boldsymbol{\pi} = [\pi_1 \dots \pi_N]'$ and η is a Lagrange multiplier. Although the data y_i do not explicitly appear in (6.72) they appear implicitly as $\pi_i = f(y_i)$. Setting the derivatives with respect to π_i ($i = 1, \dots, N$), and η to zero and solving yields $\hat{\pi}_i = 1/N$ and $\eta = 1$. Thus the estimated density function $\hat{f}(y)$ has mass $1/N$ at each of the realized values y_i , $i = 1, \dots, N$. The resulting distribution function is $\hat{F}(y) = N^{-1} \sum_{i=1}^N \mathbf{1}(y \leq y_i)$, where $\mathbf{1}(A) = 1$ if event A occurs and 0 otherwise. $\hat{F}(y)$ is just the usual empirical distribution function.

Now introduce parameters. As a simple example, suppose we introduce the moment restriction that $E[y - \mu] = 0$, where μ is the unknown population mean. In the empirical likelihood context this population moment is replaced by a sample moment, where the sample moment weights sample values by the probabilities π_i . Thus we introduce the constraint that $\sum_i \pi_i(y_i - \mu) = 0$. The Lagrangian for the maximum empirical likelihood estimator is

$$\mathcal{L}_{\text{EL}}(\boldsymbol{\pi}, \eta, \lambda, \mu) = \frac{1}{N} \sum_{i=1}^N \ln \pi_i - \eta \left(\sum_{i=1}^N \pi_i - 1 \right) - \lambda \sum_{i=1}^N \pi_i(y_i - \mu), \quad (6.73)$$

where η and λ are Lagrange multipliers.

Begin by differentiating the Lagrangian with respect to π_i ($i = 1, \dots, N$), η , and λ but not μ . Setting these derivatives to zero yields equations that are functions of μ . Solving leads to the solution $\pi_i = \pi_i(\mu)$ and hence an empirical likelihood $N^{-1} \sum_i \ln \pi_i(\mu)$ that is then maximized with respect to μ . This solution method leads to nonlinear equations that need to be solved numerically.

For this particular problem an easier way to solve for μ is to note that the maximized value of $\mathcal{L}(\boldsymbol{\pi}, \eta, \lambda, \mu)$ must be less than or equal to $N^{-1} \sum_i \ln N^{-1}$, since this is the maximized value without the last constraint. However, $\mathcal{L}(\boldsymbol{\pi}, \eta, \lambda, \mu)$ equals

$N^{-1} \sum_i \ln N^{-1}$ if $\pi_i = 1/N$ and $\hat{\mu} = N^{-1} \sum_i y_i = \bar{y}$. So the maximum empirical likelihood estimator of the population mean is the sample mean.

6.8.2. Empirical Likelihood Estimation of Regression Parameters

Now consider regression data that are iid over i . The only structure placed on the model are r moment conditions

$$E[\mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta})] = \mathbf{0}, \quad (6.74)$$

where $\mathbf{h}(\cdot)$ and \mathbf{w}_i are defined in Section 6.3.1. For example, $\mathbf{h}(\mathbf{w}, \boldsymbol{\theta}) = \mathbf{x}(y - \mathbf{x}'\boldsymbol{\theta})$ for OLS estimation and $\mathbf{h}(y, \mathbf{x}, \boldsymbol{\theta}) = (\partial g/\partial \boldsymbol{\theta})(y - g(\mathbf{x}, \boldsymbol{\theta}))$ for NLS estimation.

The empirical likelihood approach maximizes the empirical likelihood function $N^{-1} \sum_i \ln \pi_i$ subject to the constraint $\sum_i \pi_i = 1$ (see (6.72)) and the additional sample constraint based on the population moment condition (6.74) that

$$\sum_{i=1}^N \pi_i \mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}) = \mathbf{0}. \quad (6.75)$$

Thus we maximize with respect to $\boldsymbol{\pi}$, η , $\boldsymbol{\lambda}$, and $\boldsymbol{\theta}$

$$\mathcal{L}_{EL}(\boldsymbol{\pi}, \eta, \boldsymbol{\lambda}, \boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^N \ln \pi_i - \eta \left(\sum_{i=1}^N \pi_i - 1 \right) - \boldsymbol{\lambda}' \sum_{i=1}^N \pi_i \mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}), \quad (6.76)$$

where the Lagrangian multipliers are a scalar η and column vector $\boldsymbol{\lambda}$ of the same dimension as $\mathbf{h}(\cdot)$.

First, concentrate out the N parameters π_1, \dots, π_N . Differentiating $\mathcal{L}(\boldsymbol{\pi}, \eta, \boldsymbol{\lambda}, \boldsymbol{\theta})$ with respect to π_i yields $1/(N\pi_i) - \eta - \boldsymbol{\lambda}' \mathbf{h}_i = 0$. Then we obtain $\eta = 1$ by multiplying by π_i and summing over i and using $\sum_i \pi_i \mathbf{h}_i = \mathbf{0}$. It follows that

$$\pi_i(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \frac{1}{N(1 + \boldsymbol{\lambda}' \mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}))}. \quad (6.77)$$

The problem is now reduced to a maximization problem with respect to $(r + q)$ variables $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$, the Lagrangian multipliers associated with the r moment conditions (6.74), and the q parameters $\boldsymbol{\theta}$.

Solution at this stage requires numerical methods, even for just-identified models. One can maximize with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ the function $N^{-1} \sum_i \ln[1/N(1 + \boldsymbol{\lambda}' \mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}))]$.

Alternatively, first concentrate out $\boldsymbol{\lambda}$. Differentiating $\mathcal{L}(\boldsymbol{\pi}(\boldsymbol{\theta}, \boldsymbol{\lambda}), \eta, \boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$ yields $\sum_i \pi_i \mathbf{h}_i = \mathbf{0}$. Define $\boldsymbol{\lambda}(\boldsymbol{\theta})$ to be the implicit solution to the system of $\dim(\boldsymbol{\lambda})$ equations

$$\sum_{i=1}^N \frac{1}{N(1 + \boldsymbol{\lambda}' \mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}))} \mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}) = \mathbf{0}.$$

In implementation numerical methods are needed to obtain $\boldsymbol{\lambda}(\boldsymbol{\theta})$. Then (6.77) becomes

$$\pi_i(\boldsymbol{\theta}) = \frac{1}{N(1 + \boldsymbol{\lambda}(\boldsymbol{\theta})' \mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}))}. \quad (6.78)$$

By substituting (6.78) into the empirical likelihood function $N^{-1} \sum_i \ln \pi_i$, the empirical log-likelihood function evaluated at θ becomes

$$\mathcal{L}_{\text{EL}}(\theta) = -N^{-1} \sum_{i=1}^N \ln[N(1 + \lambda(\theta)' \mathbf{h}(\mathbf{w}_i, \theta))].$$

The **maximum empirical likelihood (MEL) estimator** $\hat{\theta}_{\text{MEL}}$ maximizes this function with respect to θ .

Qin and Lawless (1994) show that

$$\sqrt{N}(\hat{\theta}_{\text{MEL}} - \theta_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{A}(\theta_0)^{-1} \mathbf{B}(\theta_0) \mathbf{A}(\theta_0)^{-1}],$$

where $\mathbf{A}(\theta_0) = \text{plimE}[\partial \mathbf{h}(\theta)/\partial \theta'|_{\theta_0}]$ and $\mathbf{B}(\theta_0) = \text{plimE}[\mathbf{h}(\theta)\mathbf{h}(\theta)'|_{\theta_0}]$. This is the same limit distribution as the method of moments (see (6.13)). In finite samples $\hat{\theta}_{\text{MEL}}$ differs from $\hat{\theta}_{\text{GMM}}$, however, and inference is based on sample estimates

$$\begin{aligned}\hat{\mathbf{A}} &= \sum_{i=1}^N \hat{\pi}_i \left. \frac{\partial \mathbf{h}_i'}{\partial \theta} \right|_{\hat{\theta}}, \\ \hat{\mathbf{B}} &= \sum_{i=1}^N \hat{\pi}_i \mathbf{h}_i(\hat{\theta}) \mathbf{h}_i(\hat{\theta})'\end{aligned}$$

that weight by the estimated probabilities $\hat{\pi}_i$ rather than the proportions $1/N$.

Imbens (2002) provides a recent survey of empirical likelihood that contrasts empirical likelihood with GMM. Variations include replacing $N^{-1} \sum_i \ln \pi_i$ in (6.26) by $N^{-1} \sum_i \pi_i \ln \pi_i$. Empirical likelihood is computationally more burdensome; see Imbens (2002) for a discussion. The advantage is that the asymptotic theory provides a better finite-sample approximation to the distribution of the empirical likelihood estimator than it does to that for the GMM estimator. This is pursued further in Section 11.6.4.

6.9. Linear Systems of Equations

The preceding estimation theory covers single-equation estimation methods used in the majority of applied studies. We now consider joint estimation of several equations. Equations linear in parameters with an additive error are presented in this section, with extensions to nonlinear systems given in the subsequent section.

The main advantage of joint estimation is the gain in efficiency that results from incorporation of correlation in unobservables across equations for a given individual. Additionally, joint estimation may be necessary if there are restrictions on parameters across equations. With exogenous regressors systems estimation is a minor extension of single-equation OLS and GLS estimation, whereas with endogenous regressors it is single-equation IV methods that are adapted.

One leading example is systems of equations such as those for observed demand of several commodities at a point in time for many individuals. For seemingly unrelated regression all regressors are exogenous whereas for simultaneous equations models some regressors are endogenous.

A second leading example is panel data, where a single equation is observed at several points in time for many individuals, and each time period is treated as a separate equation. By viewing a panel data model as an example of a system it is possible to improve efficiency, obtain panel robust standard errors, and derive instruments when some regressors are endogenous.

Many econometrics texts provide lengthy presentations of linear systems. The treatment here is very brief. It is mainly directed toward generalization to nonlinear systems (see Section 6.10) and application to panel data (see Chapters 21–23).

6.9.1. Linear Systems of Equations

The single-equation linear model is given by $y_i = \mathbf{x}'_i \beta + u_i$, where y_i and u_i are scalars and \mathbf{x}_i and β are column vectors. The **multiple-equation linear model**, or **multivariate linear model**, with G dependent variables is given by

$$y_i = \mathbf{X}_i \beta + \mathbf{u}_i, \quad i = 1, \dots, N, \quad (6.79)$$

where \mathbf{y}_i and \mathbf{u}_i are $G \times 1$ vectors, \mathbf{X}_i is a $G \times K$ matrix, and β is a $K \times 1$ column vector.

Throughout this section we make the cross-section assumption that the error vector \mathbf{u}_i is independent over i , so $E[\mathbf{u}_i \mathbf{u}'_j] = \mathbf{0}$ for $i \neq j$. However, components of \mathbf{u}_i for given i may be correlated and have variances and covariances that vary over i , leading to conditional error variance matrix for the i th individual

$$\Omega_i = E[\mathbf{u}_i \mathbf{u}'_i | \mathbf{X}_i]. \quad (6.80)$$

There are various ways that a multiple-equation model may arise. At one extreme the seemingly unrelated equations model combines G equations, such as demands for different consumer goods, where parameters vary across equations and regressors may or may not vary across equations. At the other extreme the linear panel data combines G periods of data for the same equation, with parameters that are constant across periods and regressors that may or may not vary across periods. These two cases are presented in detail in Sections 6.9.3 and 6.9.4.

Stacking (6.79) over N individuals gives

$$\begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_N \end{bmatrix} \beta + \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_N \end{bmatrix}, \quad (6.81)$$

or

$$\mathbf{y} = \mathbf{X} \beta + \mathbf{u}, \quad (6.82)$$

where \mathbf{y} and \mathbf{u} are $NG \times 1$ vectors and \mathbf{X} is a $NG \times K$ matrix.

The results given in the following can be obtained by treating the stacked model (6.82) in the same way as in the single-equation case. Thus the OLS estimator is $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and in the just-identified case with instrument matrix \mathbf{Z} the IV estimator is $\hat{\beta} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$. The only real change is that the usual cross-section assumption of a diagonal error variance matrix is replaced by assumption of a block-diagonal error

matrix. This block-diagonality needs to be accommodated in computing the estimated variance matrix of a systems estimator and in forming feasible GLS estimators and efficient GMM estimators.

6.9.2. Systems OLS and FGLS Estimation

An OLS estimation of the system (6.82) yields the **systems OLS estimator** $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Using (6.81) it follows immediately that

$$\hat{\beta}_{\text{SOLS}} = \left[\sum_{i=1}^N \mathbf{X}_i' \mathbf{X}_i \right]^{-1} \sum_{i=1}^N \mathbf{X}_i' \mathbf{y}_i. \quad (6.83)$$

The estimator is asymptotically normal and, assuming the data are independent over i , the usual robust sandwich result applies and

$$\hat{\mathbb{V}}[\hat{\beta}_{\text{SOLS}}] = \left[\sum_{i=1}^N \mathbf{X}_i' \mathbf{X}_i \right]^{-1} \sum_{i=1}^N \mathbf{X}_i' \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i' \mathbf{X}_i \left[\sum_{i=1}^N \mathbf{X}_i' \mathbf{X}_i \right]^{-1}, \quad (6.84)$$

where $\hat{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{X}_i \hat{\beta}$. This variance matrix estimate permits conditional variances and covariances of the errors to differ across individuals.

Given correlation of the components of the error vector for a given individual, more efficient estimation is possible by GLS or FGLS. If observations are independent over i , the **systems GLS estimator** is systems OLS applied to the transformed system

$$\Omega_i^{-1/2} \mathbf{y}_i = \Omega_i^{-1/2} \mathbf{X}_i \beta + \Omega_i^{-1/2} \mathbf{u}_i, \quad (6.85)$$

where Ω_i is the error variance matrix defined in (6.80). The transformed error $\Omega_i^{-1/2} \mathbf{u}_i$ has mean zero and variance

$$\begin{aligned} \mathbb{E} \left[\left(\Omega_i^{-1/2} \mathbf{u}_i \right)' \left(\Omega_i^{-1/2} \mathbf{u}_i \right) | \mathbf{X}_i \right] &= \Omega_i^{-1/2} \mathbb{E} [\mathbf{u}_i' \mathbf{u}_i | \mathbf{X}_i] \Omega_i^{-1/2} \\ &= \Omega_i^{-1/2} \Omega_i \Omega_i^{-1/2} \\ &= \mathbf{I}_G. \end{aligned}$$

So the transformed system has errors that are homoskedastic and uncorrelated over G equations and OLS is efficient.

To implement this estimator, a model for Ω_i needs to be specified, say $\Omega_i = \Omega_i(\gamma)$. Then perform systems OLS estimation in the transformed system where Ω_i is replaced by $\Omega_i(\hat{\gamma})$, where $\hat{\gamma}$ is a consistent estimate of γ . This yields the **systems feasible GLS (SFGLS) estimator**

$$\hat{\beta}_{\text{SFGLS}} = \left[\sum_{i=1}^N \mathbf{X}_i' \hat{\Omega}_i^{-1} \mathbf{X}_i \right]^{-1} \sum_{i=1}^N \mathbf{X}_i' \hat{\Omega}_i^{-1} \mathbf{y}_i. \quad (6.86)$$

This estimator is asymptotically normal and to guard against possible misspecification of $\Omega_i(\gamma)$ we can use the robust sandwich estimate of the variance matrix

$$\widehat{\mathbf{V}}[\widehat{\beta}_{\text{SFGLS}}] = \left[\sum_{i=1}^N \mathbf{X}'_i \widehat{\Omega}_i^{-1} \mathbf{X}_i \right]^{-1} \sum_{i=1}^N \mathbf{X}'_i \widehat{\Omega}_i^{-1} \widehat{\mathbf{u}}_i \widehat{\mathbf{u}}'_i \widehat{\Omega}_i^{-1} \mathbf{X}_i \left[\sum_{i=1}^N \mathbf{X}'_i \widehat{\Omega}_i^{-1} \mathbf{X}_i \right]^{-1}, \quad (6.87)$$

where $\widehat{\Omega}_i = \Omega_i(\widehat{\gamma})$.

The most common specification used for Ω_i is to assume that it does not vary over i . Then $\Omega_i = \Omega$ is a $G \times G$ matrix that can be consistently estimated for finite G and $N \rightarrow \infty$ by

$$\widehat{\Omega} = \frac{1}{N} \sum_{i=1}^N \widehat{\mathbf{u}}_i \widehat{\mathbf{u}}'_i, \quad (6.88)$$

where $\widehat{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{X}_i \widehat{\beta}_{\text{SOLS}}$. Then the SFGLS estimator is (6.86) with $\widehat{\Omega}$ instead of $\widehat{\Omega}_i$, and after some algebra the SFGLS estimator can also be written as

$$\widehat{\beta}_{\text{SFGLS}} = \left[\mathbf{X}' (\widehat{\Omega}^{-1} \otimes \mathbf{I}_N) \mathbf{X} \right]^{-1} \mathbf{X}' (\widehat{\Omega}^{-1} \otimes \mathbf{I}_N) \mathbf{y}', \quad (6.89)$$

where \otimes denotes the Kronecker product. The assumption that $\Omega_i = \Omega$ rules out, for example, heteroskedasticity over i . This is a strong assumption, and in many applications it is best to use robust standard errors calculated using (6.87), which gives correct standard errors even if Ω_i does vary over i .

6.9.3. Seemingly Unrelated Regressions

The **seemingly unrelated regressions (SUR) model** specifies the g th of G equations for the i th of N individuals to be given by

$$y_{ig} = \mathbf{x}'_{ig} \beta_g + u_{ig}, \quad g = 1, \dots, G, \quad i = 1, \dots, N, \quad (6.90)$$

where \mathbf{x}_{ig} are regressors that are assumed to be exogenous and β_g are $K_g \times 1$ parameter vectors. For example, for demand data on G goods for N individuals, y_{ig} may be the i th individual's expenditure on good g or budget share for good g . In all that follows G is assumed fixed and reasonably small while $N \rightarrow \infty$. Note that we use the subscript order y_{ig} as results then transfer easily to panel data with variable y_{it} (see Section 6.9.4). Other authors use the reverse order y_{gi} .

The SUR model was proposed by Zellner (1962). The term seemingly unrelated regressions is deceptive, as clearly the equations are related if the errors u_{ig} in different equations are correlated. For the SUR model the relationship between y_{ig} and y_{ih} is indirect; it comes through correlation in the errors across different equations.

Estimation combines observations over both equations and individuals. For microeconometrics applications, where independence over i is assumed, it is most convenient to first stack all equations for a given individual. Stacking all G equations for the i th

individual we get

$$\begin{bmatrix} y_{i1} \\ \vdots \\ y_{iG} \end{bmatrix} = \begin{bmatrix} \mathbf{x}'_{i1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{x}'_{iG} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_G \end{bmatrix} + \begin{bmatrix} u_{i1} \\ \vdots \\ u_{iG} \end{bmatrix}, \quad (6.91)$$

which is of the form $\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{u}_i$ in (6.79), where \mathbf{y}_i and \mathbf{u}_i are $G \times 1$ vectors with g th entries y_{ig} and u_{ig} , \mathbf{X}_i is a $G \times K$ matrix with g th row $[\mathbf{0} \cdots \mathbf{x}'_{ig} \cdots \mathbf{0}]$, and $\beta = [\beta'_1 \dots \beta'_G]'$ is a $K \times 1$ vector where $K = K_1 + \dots + K_G$. Some authors instead first stack all individuals for a given equation, leading to different algebraic expressions for the same estimators.

Given the definitions of \mathbf{X}_i and \mathbf{y}_i it is easy to show that $\hat{\beta}_{\text{SOLS}}$ in (6.83) is

$$\begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_G \end{bmatrix} = \begin{bmatrix} \left[\sum_{i=1}^N \mathbf{x}_{i1} \mathbf{x}'_{i1} \right]^{-1} \sum_{i=1}^N \mathbf{x}_{i1} y_{i1} \\ \vdots \\ \left[\sum_{i=1}^N \mathbf{x}_{iG} \mathbf{x}'_{iG} \right]^{-1} \sum_{i=1}^N \mathbf{x}_{iG} y_{iG} \end{bmatrix},$$

so that systems OLS is the same as separate **equation-by-equation OLS**. As might be expected a priori, if the only link across equations is the error and the errors are treated as being uncorrelated then joint estimation reduces to single-equation estimation.

A better estimator is the feasible GLS estimator defined in (6.86) using $\hat{\Omega}$ in (6.88) and statistical inference based on the asymptotic variance given in (6.87). This estimator is generally more efficient than systems OLS, though it can be shown to collapse to OLS if the errors are uncorrelated across equations or if exactly the same regressors appear in each equation.

Seemingly unrelated regression models may impose **cross-equation parameter restrictions**. For example, a symmetry restriction may imply that the coefficient of the second regressor in the first equation equals the coefficient of the first regressor in the second equation. If such restrictions are equality restrictions one can easily estimate the model by appropriate redefinition of \mathbf{X}_i and β given in (6.79). For example, if there are two equations and the restriction is that $\beta_2 = -\beta_1$ then define $\mathbf{X}_i = [\mathbf{x}_{i1} \ -\mathbf{x}_{i2}]'$ and $\beta = \beta_1$. Alternatively, one can estimate using systems extensions of single-equation OLS and GLS with linear restrictions on the parameters.

Also, in systems of equations it is possible that the variance matrix of the error vector \mathbf{u}_i is singular, as a result of **adding-up constraints**. For example, suppose y_{ig} is the i th budget share, and the model is $y_{ig} = \alpha_g + \mathbf{z}'_i \beta_g + u_{ig}$, where the same regressors appear in each equation. Then $\sum_g y_{ig} = 1$ since budget shares sum to one, which requires $\sum_g \alpha_g = 1$, $\sum_g \beta_g = \mathbf{0}$, and $\sum_g u_{ig} = 0$. The last restriction means Ω_i is singular and hence noninvertible. One can eliminate one equation, say the last, and estimate the model by systems estimation applied to the remaining $G - 1$ equations. Then the parameter estimates for the G th equation can be obtained using the adding-up constraint. For example, $\hat{\alpha}_G = 1 - (\hat{\alpha}_1 + \dots + \hat{\alpha}_{G-1})$. It is also possible to impose equality restrictions on the parameters in this setup. A literature exists on methods that ensure that estimates obtained are invariant to the equation deleted; see, for example, Berndt and Savin (1975).

6.9.4. Panel Data

Another leading application of systems GLS methods is to panel data, where a scalar dependent variable is observed in each of T time periods for N individuals. Panel data can be viewed as a system of equations, either T equations for N individuals or N equations for T time periods. In microeconomics we assume a short panel, with T small and $N \rightarrow \infty$ so it is natural to set it up as a scalar dependent variable y_{it} , where the g th equation in the preceding discussion is now interpreted as the t th time period and $G = T$.

A simple panel data model is

$$y_{it} = \mathbf{x}'_{it}\beta + u_{it}, \quad t = 1, \dots, T, \quad i = 1, \dots, N, \quad (6.92)$$

a specialization of (6.90) with β now constant. Then in (6.79) the regressor matrix becomes $\mathbf{X}_i = [\mathbf{x}_{i1} \cdots \mathbf{x}_{iT}]'$. After some algebra the systems OLS estimator defined in (6.83) can be reexpressed as

$$\hat{\beta}_{\text{POLS}} = \left[\sum_{i=1}^N \sum_{t=1}^T \mathbf{x}_{it} \mathbf{x}'_{it} \right]^{-1} \sum_{i=1}^N \sum_{t=1}^T \mathbf{x}_{it} y_{it}. \quad (6.93)$$

This estimator is called the **pooled OLS estimator** as it pools or combines the cross-section and time-series aspects of the data.

The pooled estimator is obtained simply by OLS estimation of y_{it} on \mathbf{x}_{it} . However, if u_{it} are correlated over t for given i , the default OLS standard errors that assume independence of the error over both i and t are invalid and can be greatly downward biased. Instead, statistical inference should be based on the robust form of the covariance matrix given in (6.84). This is detailed in Section 21.2.3. In practice models more complicated than (6.92) that include individual specific effects are estimated (see Section 21.2).

6.9.5. Systems IV Estimation

Estimation of a single linear equation with endogenous regressors was presented in Section 6.4. Now we extend this to the multivariate linear model (6.79) when $E[\mathbf{u}_i | \mathbf{X}_i] \neq \mathbf{0}$. Brundy and Jorgenson (1971) considered IV estimation applied to the system of equations to produce estimates that are both consistent and efficient.

We assume the existence of a $G \times r$ matrix of instruments \mathbf{Z}_i that satisfy $E[\mathbf{u}_i | \mathbf{Z}_i] = \mathbf{0}$ and hence

$$E[\mathbf{Z}'_i(\mathbf{y}_i - \mathbf{X}_i\beta)] = \mathbf{0}. \quad (6.94)$$

These instruments can be used to obtain consistent parameter estimates using single-equation IV methods, but joint equation estimation can improve efficiency. The **systems GMM estimator** minimizes

$$Q_N(\beta) = \left[\sum_{i=1}^N \mathbf{Z}'_i(\mathbf{y}_i - \mathbf{X}_i\beta) \right]' \mathbf{W}_N \left[\sum_{i=1}^N \mathbf{Z}'_i(\mathbf{y}_i - \mathbf{X}_i\beta) \right], \quad (6.95)$$

where \mathbf{W}_N is an $r \times r$ weighting matrix. Performing some algebra yields

$$\widehat{\boldsymbol{\beta}}_{\text{SGMM}} = [\mathbf{X}' \mathbf{Z} \mathbf{W}_N \mathbf{Z}' \mathbf{X}]^{-1} [\mathbf{X}' \mathbf{Z} \mathbf{W}_N \mathbf{Z}' \mathbf{y}], \quad (6.96)$$

where \mathbf{X} is an $NG \times K$ matrix obtained by stacking $\mathbf{X}_1, \dots, \mathbf{X}_N$ (see (6.81)) and \mathbf{Z} is an $NG \times r$ matrix obtained by similarly stacking $\mathbf{Z}_1, \dots, \mathbf{Z}_N$. The systems GMM estimator has exactly the same form as (6.37), and the asymptotic variance matrix is that given in (6.39). It follows that a robust estimate of the variance matrix is

$$\widehat{V}[\widehat{\boldsymbol{\beta}}_{\text{SGMM}}] = N [\mathbf{X}' \mathbf{Z} \mathbf{W}_N \mathbf{Z}' \mathbf{X}]^{-1} [\mathbf{X}' \mathbf{Z} \mathbf{W}_N \widehat{\mathbf{S}} \mathbf{W}_N \mathbf{Z}' \mathbf{X}] [\mathbf{X}' \mathbf{Z} \mathbf{W}_N \mathbf{Z}' \mathbf{X}]^{-1}, \quad (6.97)$$

where, in the systems case and assuming independence over i ,

$$\widehat{\mathbf{S}} = \frac{1}{N} \sum_{i=1}^N \mathbf{Z}_i' \widehat{\mathbf{u}}_i \widehat{\mathbf{u}}_i' \mathbf{Z}_i. \quad (6.98)$$

Several choices of weighting matrix receive particular attention.

First, the **optimal systems GMM estimator** is (6.96) with $\mathbf{W}_N = \widehat{\mathbf{S}}^{-1}$, where $\widehat{\mathbf{S}}$ is defined in (6.98). The variance matrix then simplifies to

$$\widehat{V}[\widehat{\boldsymbol{\beta}}_{\text{OSGMM}}] = N [\mathbf{X}' \mathbf{Z} \widehat{\mathbf{S}}^{-1} \mathbf{Z}' \mathbf{X}]^{-1}.$$

This estimator is the most efficient GMM estimator based on moment conditions (6.94). The efficiency gain arises from two factors: (1) systems estimation, which permits errors in different equations to be correlated, so that $V[\mathbf{u}_i | \mathbf{Z}_i]$ is not restricted to being block diagonal, and (2) an allowance for quite general heteroskedasticity and correlation, so that Ω_i can vary over i .

Second, the **systems 2SLS estimator** arises when $\mathbf{W}_N = (N^{-1} \mathbf{Z}' \mathbf{Z})^{-1}$. Consider the SUR model defined in (6.91), with some of the regressors \mathbf{x}_{ig} now endogenous. Then systems 2SLS reduces to equation-by-equation 2SLS, with instruments \mathbf{z}_g for the g th equation, if we define the instrument matrix to be

$$\mathbf{Z}_i = \begin{bmatrix} \mathbf{z}'_{i1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{z}'_{iG} \end{bmatrix}. \quad (6.99)$$

In many applications $\mathbf{z}_1 = \mathbf{z}_2 = \dots = \mathbf{z}_g$ so that a common set of instruments is used in all equations, but we need not restrict analysis to this case. For the panel data model (6.92) systems 2SLS reduces to pooled 2SLS if we define $\mathbf{Z}_i = [\mathbf{z}_{i1} \cdots \mathbf{z}_{iT}]'$.

Third, suppose that $V[\mathbf{u}_i | \mathbf{Z}_i]$ does not vary over i , so that $V[\mathbf{u}_i | \mathbf{Z}_i] = \Omega$. This is a systems analogue of the single-equation assumption of homoskedasticity. Then as with (6.88) a consistent estimate of Ω is $\widehat{\Omega} = N^{-1} \sum_i \widehat{\mathbf{u}}_i \widehat{\mathbf{u}}_i'$, where $\widehat{\mathbf{u}}_i$ are residuals based on a consistent IV estimator such as systems 2SLS. Then the optimal GMM estimator is (6.96) with $\mathbf{W}_N = \mathbf{I}_N \otimes \widehat{\Omega}$. This estimator should be contrasted with the three-stage least-squares estimator presented at the end of the next section.

6.9.6. Linear Simultaneous Equations Systems

The linear simultaneous equations model, introduced in Section 2.4, is a very important model that is often presented in considerable length in introductory graduate-level econometrics courses. In this section we provide a very brief self-contained summary. The discussion of identification overlaps with that in Chapter 2. Due to the presence of endogenous variables OLS and SUR estimators are inconsistent. Consistent estimation methods are placed in the context of GMM estimation, even though the standard methods were developed well before GMM.

The linear **simultaneous equations model** specifies the g th of G equations for the i th of N individuals to be given by

$$y_{ig} = \mathbf{z}'_{ig} \boldsymbol{\gamma}_g + \mathbf{Y}'_{ig} \boldsymbol{\beta}_g + u_{ig}, \quad g = 1, \dots, G, \quad (6.100)$$

where the order of subscripts is that of Section 6.9 rather than Section 2.4, \mathbf{z}_g is a vector of exogenous regressors that are assumed to be uncorrelated with the error term u_g and \mathbf{Y}_g is a vector that contains a subset of the dependent variables $y_1, \dots, y_{g-1}, y_{g+1}, \dots, y_G$ of the other $G - 1$ equations. \mathbf{Y}_g is endogenous as it is correlated with model errors. The model for the i th individual can equivalently be written as

$$\mathbf{y}'_i \mathbf{B} + \mathbf{z}'_i \boldsymbol{\Gamma} = \mathbf{u}_i, \quad (6.101)$$

where $\mathbf{y}_i = [y_{i1} \dots y_{iG}]'$ is a $G \times 1$ vector of endogenous variables, \mathbf{z}_i is an $r \times 1$ vector of exogenous variables that is the union of $\mathbf{z}_{i1}, \dots, \mathbf{z}_{iG}$, $\mathbf{u}_i = [u_{i1} \dots u_{iG}]'$ is a $G \times 1$ error vector, \mathbf{B} is a $G \times G$ parameter matrix with diagonal entries unity, $\boldsymbol{\Gamma}$ is an $r \times G$ parameter matrix, and some of the entries in \mathbf{B} and $\boldsymbol{\Gamma}$ are constrained to be unity. It is assumed that \mathbf{u}_i is iid over i with mean $\mathbf{0}$ and variance matrix $\boldsymbol{\Sigma}$.

The model (6.101) is called the **structural form** with different restrictions on \mathbf{B} and $\boldsymbol{\Gamma}$ corresponding to different structures. Solving for the endogenous variables as a function of the exogenous variables yields the **reduced form**

$$\begin{aligned} \mathbf{y}_i &= -\mathbf{z}'_i \boldsymbol{\Gamma} \mathbf{B}^{-1} + \mathbf{u}_i \mathbf{B}^{-1} \\ &= \mathbf{z}'_i \boldsymbol{\Pi} + \mathbf{v}_i, \end{aligned} \quad (6.102)$$

where $\boldsymbol{\Pi} = -\boldsymbol{\Gamma} \mathbf{B}^{-1}$ is the $r \times G$ matrix of reduced form parameters and $\mathbf{v}_i = \mathbf{u}_i \mathbf{B}^{-1}$ is the reduced form error vector with variance $\boldsymbol{\Omega} = (\mathbf{B}^{-1})' \boldsymbol{\Sigma} \mathbf{B}^{-1}$.

The reduced form can be consistently estimated by OLS, yielding estimates of $\boldsymbol{\Pi} = -\boldsymbol{\Gamma} \mathbf{B}^{-1}$ and $\boldsymbol{\Omega} = (\mathbf{B}^{-1})' \boldsymbol{\Sigma} \mathbf{B}^{-1}$. The problem of identification, see Section 2.5, is one of whether these lead to unique estimates of the structural form parameters \mathbf{B} , $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$. This requires some parameter restrictions since without restrictions \mathbf{B} , $\boldsymbol{\Gamma}$, and $\boldsymbol{\Sigma}$ contain G^2 more parameters than $\boldsymbol{\Pi}$ and $\boldsymbol{\Omega}$. A necessary condition for **identification** of parameters in the g th equation is the **order condition** that the number of exogenous variables excluded from the g th equation must be at least equal to the number of endogenous variables included. This is the same as the order condition given in Section 6.4.1. For example, if \mathbf{Y}_{ig} in (6.100) has one component, so there is one endogenous variable in the equation, then at least one of the components of \mathbf{x}_i must not be included. This will ensure that there are as many instruments as regressors.

A sufficient condition for identification is the stronger **rank condition**. This is given in many books such as Greene's (2003) and for brevity is not given here. Other restrictions, such as covariance restrictions, may also lead to identification.

Given identification, the structural model parameters can be consistently estimated by separate estimation of each equation by **two-stage least squares** defined in (6.44). The same set of instruments \mathbf{z}_i is used for each equation. In the g th equation the sub-component \mathbf{z}_{ig} is used as instrument for itself and the remainder of \mathbf{z}_i is used as instrument for \mathbf{Y}_{ig} .

More efficient systems estimates are obtained using the **three-stage least-squares (3SLS) estimator** of Zellner and Theil (1962), which assumes errors are homoskedastic but are correlated across equations. First, estimate the reduced form coefficients $\boldsymbol{\Pi}$ in (6.102) by OLS regression of \mathbf{y} on \mathbf{z} . Second, obtain the 2SLS estimates by OLS regression of (6.100), where \mathbf{Y}_g is replaced by the reduced form predictions $\widehat{\mathbf{Y}}_g = \mathbf{z}'\widehat{\boldsymbol{\Pi}}_G$. This is OLS regression of y_g on $\widehat{\mathbf{Y}}_g$ and \mathbf{z}_g , or equivalently of y_g on $\widehat{\mathbf{x}}_g$, where $\widehat{\mathbf{x}}_g$ are the predictions of \mathbf{Y}_g and \mathbf{z}_g from OLS regression on \mathbf{z} . Third, obtain the 3SLS estimates by systems OLS regression of y_g on $\widehat{\mathbf{x}}_g$, $g = 1, \dots, G$. Then from (6.89)

$$\widehat{\boldsymbol{\theta}}_{3SLS} = \left[\widehat{\mathbf{X}}' \left(\widehat{\boldsymbol{\Omega}}^{-1} \otimes \mathbf{I}_N \right) \widehat{\mathbf{X}} \right]^{-1} \widehat{\mathbf{X}}' \left(\widehat{\boldsymbol{\Omega}}^{-1} \otimes \mathbf{I}_N \right) \mathbf{y},$$

where $\widehat{\mathbf{X}}$ is obtained by first forming a block-diagonal matrix $\widehat{\mathbf{X}}_i$ with diagonal blocks $\widehat{\mathbf{x}}_{i1}, \dots, \widehat{\mathbf{x}}_{iG}$ and then stacking $\widehat{\mathbf{X}}_1, \dots, \widehat{\mathbf{X}}_N$, and $\widehat{\boldsymbol{\Omega}} = N^{-1} \sum_i \widehat{\mathbf{u}}_i \widehat{\mathbf{u}}_i'$ with $\widehat{\mathbf{u}}_i$ the residual vectors calculated using the 2SLS estimates.

This estimator coincides with the systems GMM estimator with $\mathbf{W}_N = \mathbf{I}_N \otimes \widehat{\boldsymbol{\Omega}}$ in the case that the systems GMM estimator uses the same instruments in every equation. Otherwise, 3SLS and systems GMM differ, though both yield consistent estimates if $E[\mathbf{u}_i | \mathbf{z}_i] = \mathbf{0}$.

6.9.7. Linear Systems ML Estimation

The systems estimators for the linear model are essentially LS or IV estimators with inference based on robust standard errors. Now additionally assume normally distributed iid errors, so that $\mathbf{u}_i \sim \mathcal{N}[\mathbf{0}, \boldsymbol{\Omega}]$.

For systems with exogenous regressors the resulting MLE is asymptotically equivalent to the GLS estimator. These estimators do use different estimators of $\boldsymbol{\Omega}$ and hence $\boldsymbol{\beta}$, however, so that there are small-sample differences between the MLE and the GLS estimator. For example, see Chapter 21 for the random effects panel data model.

For the linear SEM (6.101), the **limited information maximum likelihood** estimator, a single-equation ML estimator, is asymptotically equivalent to 2SLS. The **full information maximum likelihood** estimator, the systems MLE, is asymptotically equivalent to 3SLS. See, for example, Schmidt (1976) and Greene (2003).

6.10. Nonlinear Sets of Equations

We now consider systems of equations that are nonlinear in parameters. For example, demand equation systems obtained from a specified direct or indirect utility may be

nonlinear in parameters. More generally, if a nonlinear model is appropriate for a dependent variable studied in isolation, for example a logit or Poisson model, then any joint model for two or more such variables will necessarily be nonlinear.

We begin with a discussion of fully parametric joint modeling, before focusing on partially parametric modeling. As in the linear case we present models with exogenous regressors before considering the complication of endogenous regressors.

6.10.1. Nonlinear Systems ML Estimation

Maximum likelihood estimation for a single dependent variable was presented in Section 5.6. These results can be immediately applied to joint models of several dependent variables, with the very minor change that the single dependent variable conditional density $f(y_i|\mathbf{x}_i, \boldsymbol{\theta})$ becomes $f(\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\theta})$, where \mathbf{y}_i denotes the vector of dependent variables, \mathbf{X}_i denotes all the regressors, and $\boldsymbol{\theta}$ denotes all the parameters.

For example, if $y_1 \sim \mathcal{N}[\exp(\mathbf{x}'_1\boldsymbol{\beta}_1), \sigma_1^2]$ and $y_2 \sim \mathcal{N}[\exp(\mathbf{x}'_2\boldsymbol{\beta}_2), \sigma_2^2]$ then a suitable joint model may be to assume that (y_1, y_2) are bivariate normal with means $\exp(\mathbf{x}'_1\boldsymbol{\beta}_1)$ and $\exp(\mathbf{x}'_2\boldsymbol{\beta}_2)$, variances σ_1^2 and σ_2^2 , and correlation ρ .

For data that are not normally distributed there can be challenges in specifying and selecting a sufficiently flexible joint distribution. For example, for univariate counts a standard starting model is the negative binomial (see Chapter 20). However, in extending this to a bivariate or multivariate model for counts there are several alternative bivariate negative binomial models to choose from. These might differ, for example, as to whether the univariate conditional distribution or the univariate marginal distribution is negative binomial. In contrast the multivariate normal distribution has conditional and marginal distributions that are both normal. All of these multivariate negative binomial distributions place some restrictions on the range of correlation such as restricting to positive correlation, whereas for the multivariate normal there is no such restriction.

Fortunately, modern computational advances permit richer models to be specified. For example, a reasonably flexible model for correlated bivariate counts is to assume that, conditional on unobservables ε_1 and ε_2 , y_1 is Poisson with mean $\exp(\mathbf{x}'_1\boldsymbol{\beta}_1 + \varepsilon_1)$ and y_2 is Poisson with mean $\exp(\mathbf{x}'_2\boldsymbol{\beta}_2 + \varepsilon_2)$. An estimable bivariate distribution can be obtained by assuming that the unobservables ε_1 and ε_2 are bivariate normal and integrating them out. There is no closed-form solution for this bivariate distribution, but the parameters can nonetheless be estimated using the method of maximum simulated likelihood presented in Section 12.4.

A number of examples of nonlinear joint models are given throughout Part 4 of the book. The simplest joint models can be inflexible, so consistency can rely on distributional assumptions that are too restrictive. However, there is generally no theoretical impediment to specifying more flexible models that can be estimated using computationally intensive methods.

In particular, two leading methods for generating rich multivariate parametric models are presented in detail in Section 19.3. These methods are given in the context of duration data models, but they have much wider applicability. First, one can introduce correlated **unobserved heterogeneity**, as in the bivariate count example just given.

Second, one can use **copulas**, which provide a way to generate a joint distribution given specified univariate marginals.

For ML estimation a simpler though less efficient quasi-ML approach is to specify separate parametric models for y_1 and y_2 and obtain ML estimates assuming independence of y_1 and y_2 but then do statistical inference permitting y_1 and y_2 to be correlated. This has been presented in Section 5.7.5. In the remainder of this section we consider such partially parametric approaches.

The challenges became greater if there is endogeneity, so that a dependent variable in one equation appears as a regressor in another equation. Few models for nonlinear simultaneous equations exist, aside from nonlinear regression models with additive errors that are normally distributed.

6.10.2. Nonlinear Systems of Equations

For linear regression the movement from single equation to multiple equations is clear as the starting point is the linear model $y = \mathbf{x}'\beta + u$ and estimation is by least squares. Efficient systems estimation is then by systems GLS estimation. For nonlinear models there can be much more variety in the starting point and estimation method.

We define the **multivariate nonlinear model** with G dependent variables to be

$$\mathbf{r}(\mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\beta}) = \mathbf{u}_i, \quad (6.103)$$

where \mathbf{y}_i and \mathbf{u}_i are $G \times 1$ vectors, $\mathbf{r}(\mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\beta})$ is a $G \times 1$ vector function, \mathbf{X}_i is a $G \times L$ matrix, and $\boldsymbol{\beta}$ is a $K \times 1$ column vector. Throughout this section we make the cross-section assumption that the error vector \mathbf{u}_i is independent over i , but components of \mathbf{u}_i for given i may be correlated with variances and covariances that vary over i .

One example of (6.103) is a **nonlinear seemingly unrelated regression model**. Then the g th of G equations for the i th of N individuals is given by

$$r_g(y_{ig}, \mathbf{x}_{ig}, \boldsymbol{\beta}_g) = u_{ig}, \quad g = 1, \dots, G. \quad (6.104)$$

For example, $u_{ig} = y_{ig} - \exp(\mathbf{x}'_{ig}\boldsymbol{\beta}_g)$. Then \mathbf{u}_i and $\mathbf{r}(\cdot)$ in (6.103) are $G \times 1$ vectors with g th entries u_{ig} and $r_g(\cdot)$, \mathbf{X}_i is the same block-diagonal matrix as that defined in (6.91), and $\boldsymbol{\beta}$ is obtained by stacking $\boldsymbol{\beta}_1$ to $\boldsymbol{\beta}_G$.

A second example is a **nonlinear panel data model**. Then for individual i in period t

$$r(y_{it}, \mathbf{x}_{it}, \boldsymbol{\beta}) = u_{it}, \quad t = 1, \dots, T. \quad (6.105)$$

Then \mathbf{u}_i and $\mathbf{r}(\cdot)$ in (6.103) are $T \times 1$ vectors, so $G = T$, with t th entries u_{it} and $r(y_{it}, \mathbf{x}_{it}, \boldsymbol{\beta})$. The panel model differs from the SUR model by having the same function $r(\cdot)$ and parameters $\boldsymbol{\beta}$ in each period.

6.10.3. Nonlinear Systems Estimation

When the regressors \mathbf{X}_i in the model (6.103) are exogenous

$$\mathbb{E}[\mathbf{u}_i | \mathbf{X}_i] = \mathbf{0}, \quad (6.106)$$

where \mathbf{u}_i is the error term defined in (6.103). We assume that the error term is independent over i , and the variance matrix is

$$\Omega_i = \mathbb{E}[\mathbf{u}_i \mathbf{u}'_i | \mathbf{X}_i]. \quad (6.107)$$

Additive Errors

Systems estimation is a straightforward adaptation of systems OLS and FGLS estimation of the linear models when the nonlinear model is additive in the error term, so that (6.103) specializes to

$$\mathbf{u}_i = \mathbf{y}_i - \mathbf{g}(\mathbf{X}_i, \beta). \quad (6.108)$$

Then the **systems NLS estimator** minimizes the sum of squared residuals $\sum_i \mathbf{u}'_i \mathbf{u}_i$, whereas the **systems FGNLS estimator** minimizes

$$Q_N(\beta) = \sum_i \mathbf{u}'_i \widehat{\Omega}_i^{-1} \mathbf{u}_i, \quad (6.109)$$

where we specify a model $\Omega_i(\gamma)$ for Ω_i and $\widehat{\Omega}_i = \Omega_i(\widehat{\gamma})$. To guard against possible misspecification of Ω_i one can use robust standard errors that essentially require only that \mathbf{u}_i is independent and satisfies (6.106). Then the estimated variance of the systems FGNLS estimator is the same as that for the linear systems FGLS estimator in (6.87), with \mathbf{X}_i replaced by $\partial \mathbf{g}(\mathbf{y}_i, \beta) / \partial \beta' |_{\widehat{\beta}}$ and now $\widehat{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{g}(\mathbf{X}_i, \widehat{\beta})$. The estimated variance of the simpler systems NLS estimator is obtained by additionally replacing $\widehat{\Omega}_i$ by \mathbf{I}_G .

The main challenge can be specifying a useful model for Ω_i . As an example, suppose we wish to jointly model two count data variables. In Chapter 20 we show that a standard model for counts, a little more general than the Poisson model, specifies the conditional mean to be $\exp(\mathbf{x}'\beta)$ and the conditional variance to be a multiple of $\exp(\mathbf{x}'\beta)$. Then a joint model might specify $\mathbf{u} = [u_1 \ u_2]'$, where $u_1 = y_1 - \exp(\mathbf{x}'_1 \beta_1)$ and $u_2 = y_2 - \exp(\mathbf{x}'_2 \beta_2)$. The variance matrix Ω_i then has diagonal entries $\alpha_1 \exp(\mathbf{x}'_1 \beta_1)$ and $\alpha_2 \exp(\mathbf{x}'_2 \beta_2)$, and one possible parameterization for the covariance is $\alpha_3 [\exp(\mathbf{x}'_1 \beta_1) \exp(\mathbf{x}'_2 \beta_2)]^{1/2}$. The estimate $\widehat{\Omega}_i$ then requires estimates of $\beta_1, \beta_2, \alpha_1, \alpha_2$, and α_3 that may be obtained from first-step single-equation estimation.

Nonadditive Errors

With nonadditive errors least-squares regression is no longer appropriate, as shown in the single-equation case in Section 6.2.2. Wooldridge (2002) presents consistent method of moments estimation.

The conditional moment restriction (6.106) leads to many possible unconditional moment conditions that can be used for estimation. The obvious starting point is to base estimation on the moment conditions $\mathbb{E}[\mathbf{X}'_i \mathbf{u}_i] = \mathbf{0}$. However, other moment conditions may be used. We more generally consider estimation based on K moment conditions

$$\mathbb{E}[\mathbf{R}(\mathbf{X}_i, \beta)' \mathbf{u}_i] = \mathbf{0}, \quad (6.110)$$

where $\mathbf{R}(\mathbf{X}_i, \beta)$ is a $K \times G$ matrix of functions of \mathbf{X}_i and β . The specification of $\mathbf{R}(\mathbf{X}_i, \beta)$ and possible dependence on β are discussed in the following.

By construction there are as many moment conditions as parameters. The **systems method of moments estimator** $\hat{\beta}_{\text{SMM}}$ solves the corresponding sample moment conditions

$$\frac{1}{N} \sum_{i=1}^N \mathbf{R}(\mathbf{X}_i, \beta)' \mathbf{r}(\mathbf{y}_i, \mathbf{X}_i, \hat{\beta}_{\text{SMM}}) = \mathbf{0}, \quad (6.111)$$

where in practice $\mathbf{R}(\mathbf{X}_i, \beta)$ is evaluated at a first-step estimate $\tilde{\beta}$. This estimator is asymptotically normal with variance matrix

$$\widehat{\mathbf{V}}[\hat{\beta}_{\text{SMM}}] = \left[\sum_{i=1}^N \widehat{\mathbf{D}}_i' \widehat{\mathbf{R}}_i \right]^{-1} \sum_{i=1}^N \widehat{\mathbf{R}}_i' \widehat{\mathbf{u}}_i \widehat{\mathbf{u}}_i' \widehat{\mathbf{R}}_i \left[\sum_{i=1}^N \widehat{\mathbf{R}}_i' \widehat{\mathbf{D}}_i \right]^{-1}, \quad (6.112)$$

where $\widehat{\mathbf{D}}_i = \partial \mathbf{r}_i / \partial \beta'|_{\tilde{\beta}}$, $\widehat{\mathbf{R}}_i = \mathbf{R}(\mathbf{X}_i, \tilde{\beta})$, and $\widehat{\mathbf{u}}_i = \mathbf{r}(\mathbf{y}_i, \mathbf{X}_i, \tilde{\beta}_{\text{SMM}})$.

The main issue is specification of $\mathbf{R}(\mathbf{X}, \beta)$ in (6.110). From Section 6.3.7, the most efficient estimator based on (6.106) specifies

$$\mathbf{R}^*(\mathbf{X}_i, \beta) = E \left[\frac{\partial \mathbf{r}(\mathbf{y}_i, \mathbf{X}_i, \beta)'}{\partial \beta} | \mathbf{X}_i \right] \Omega_i^{-1}. \quad (6.113)$$

In general the first expectation on the right-hand side requires strong distributional assumptions, making optimal estimation difficult.

Simplification does occur, however, if the nonlinear model is one with additive error defined in (6.108). Then $\mathbf{R}^*(\mathbf{X}_i, \beta) = \partial \mathbf{g}(\mathbf{X}_i, \beta)' / \partial \beta \times \Omega_i^{-1}$, and the estimating equations (6.110) become

$$N^{-1} \sum_{i=1}^N \frac{\partial \mathbf{g}(\mathbf{X}_i, \beta)'}{\partial \beta} \Omega_i^{-1} (\mathbf{y}_i - \mathbf{X}'_i \hat{\beta}_{\text{SMM}}) = \mathbf{0}.$$

This estimator is asymptotically equivalent to the systems FGNLS estimator that minimizes (6.109).

6.10.4. Nonlinear Systems IV Estimation

When the regressors \mathbf{X}_i in the model (6.103) are endogenous, so that $E[\mathbf{u}_i | \mathbf{X}_i] \neq \mathbf{0}$, we assume the existence of a $G \times r$ matrix of instruments \mathbf{Z}_i such that

$$E[\mathbf{u}_i | \mathbf{Z}_i] = \mathbf{0}, \quad (6.114)$$

where \mathbf{u}_i is the error term defined in (6.103). We assume that the error term is independent over i , and the variance matrix is $\Omega_i = E[\mathbf{u}_i \mathbf{u}_i' | \mathbf{Z}_i]$. For the nonlinear SUR model \mathbf{Z}_i is as defined in (6.99).

The approach is similar to that used in the preceding section for the systems MM estimator, with the additional complication that now there may be a surplus of instruments leading to a need for GMM estimation rather than just MM estimation. Conditional moment restriction (6.106) leads to many possible unconditional moment conditions that can be used for estimation. Here we follow many others in basing estimation

on the moment conditions $E[\mathbf{Z}'_i \mathbf{u}_i] = \mathbf{0}$. Then a **systems GMM estimator** minimizes

$$Q_N(\beta) = \left[\sum_{i=1}^N \mathbf{Z}'_i \mathbf{r}(\mathbf{y}_i, \mathbf{X}_i, \beta) \right]' \mathbf{W}_N \left[\sum_{i=1}^N \mathbf{Z}'_i \mathbf{r}(\mathbf{y}_i, \mathbf{X}_i, \beta) \right]. \quad (6.115)$$

This estimator is asymptotically normal with estimated variance

$$\widehat{\mathbb{V}}[\widehat{\beta}_{\text{SGMM}}] = N [\widehat{\mathbf{D}}' \mathbf{Z} \mathbf{W}_N \mathbf{Z}' \widehat{\mathbf{D}}]^{-1} [\widehat{\mathbf{D}}' \mathbf{Z} \mathbf{W}_N \widehat{\mathbf{S}} \mathbf{W}_N \mathbf{Z}' \widehat{\mathbf{D}}] [\widehat{\mathbf{D}}' \mathbf{Z} \mathbf{W}_N \mathbf{Z}' \widehat{\mathbf{D}}]^{-1}, \quad (6.116)$$

where $\widehat{\mathbf{D}}' \mathbf{Z} = \sum_i \partial \mathbf{r}'_i / \partial \beta|_{\widehat{\beta}} \mathbf{Z}_i$ and $\widehat{\mathbf{S}} = N^{-1} \sum_i \mathbf{Z}_i \widehat{\mathbf{u}}_i \widehat{\mathbf{u}}'_i \mathbf{Z}'_i$ and we assume \mathbf{u}_i is independent over i with variance matrix $V[\mathbf{u}_i | \mathbf{X}_i] = \Omega_i$.

The choice $\mathbf{W}_N = [N^{-1} \sum_i \mathbf{Z}_i \mathbf{Z}'_i]^{-1}$ corresponds to NL2SLS in the case that $\mathbf{r}(\mathbf{y}_i, \mathbf{X}_i, \beta)$ is obtained from a nonlinear SUR model. The choice $\mathbf{W}_N = [N^{-1} \sum_i \mathbf{Z}_i \widehat{\Omega} \mathbf{Z}'_i]^{-1}$, where $\widehat{\Omega} = N^{-1} \sum_i \widehat{\mathbf{u}}_i \widehat{\mathbf{u}}'_i$, is called **nonlinear 3SLS** (NL3SLS) and is the most efficient estimator based on the moment condition $E[\mathbf{Z}'_i \mathbf{u}_i] = \mathbf{0}$ in the special case that $\Omega_i = \Omega$. The choice $\mathbf{W}_N = \widehat{\mathbf{S}}^{-1}$ gives the most efficient estimator under the more general assumption that Ω_i may vary with i . As usual, however, moment conditions other than $E[\mathbf{Z}'_i \mathbf{u}_i] = \mathbf{0}$ may lead to more efficient estimators.

6.10.5. Nonlinear Simultaneous Equations Systems

The **nonlinear simultaneous equations model** specifies that the g th of G equations for the i th of N individuals is given by

$$u_{ig} = r_g(\mathbf{y}_i, \mathbf{x}_{ig}, \beta_g), \quad g = 1, \dots, G. \quad (6.117)$$

This is the nonlinear SUR model with regressors that now include dependent variables from other equations. Unlike the linear SEM, there are few practically useful results to help ensure that a nonlinear SEM is identified.

Given identification, consistent estimates can be obtained using the GMM estimators presented in the previous section. Alternatively, we can assume that $\mathbf{u}_i \sim \mathcal{N}[\mathbf{0}, \Omega]$ and obtain the **nonlinear full-information maximum likelihood estimator**. In a departure from the linear SEM, the nonlinear full-information MLE in general has an asymptotic distribution that differs from NL3SLS, and consistency of the nonlinear full-information MLE requires that the errors are actually normally distributed. For details see Amemiya (1985).

Handling endogeneity in nonlinear models can be complicated. Section 16.8 considers simultaneity in Tobit models, where analysis is simpler when the model is linear in the latent variables. Section 20.6.2 considers a more highly nonlinear example, endogenous regressors in count data models.

6.11. Practical Considerations

Ideally GMM could be implemented using an econometrics package, requiring little more difficulty and knowledge than that needed, say, for nonlinear least-squares estimation with heteroskedastic errors. However, not all leading econometrics packages

provide a broad GMM module. Depending on the specific application, GMM estimation may require a switch to a more suitable package or use of a matrix programming language along with familiarity with the algebra of GMM.

A common application of GMM is IV estimation. Most econometrics packages include linear IV but not all include nonlinear IV estimators. The default standard errors may assume homoskedastic errors rather than being heteroskedastic-robust. As already emphasized in Chapter 4, it can be difficult to obtain instruments that are uncorrelated with the error yet reasonably correlated with the regressor or, in the nonlinear case, the appropriate derivative of the error with respect to parameters.

Econometrics packages usually include linear systems but not nonlinear systems. Again, default standard errors may not be robust to heteroskedasticity.

6.12. Bibliographic Notes

Textbook treatments of GMM include chapters by Davidson and MacKinnon (1993, 2004), Hamilton (1994), and Greene (2003). The more recent books by Hayashi (2000) and Wooldridge (2002) place considerable emphasis on GMM estimation. Bera and Bilias (2002) provide a synthesis and history of many of the estimators presented in Chapters 5 and 6.

- 6.3** The original reference for GMM is Hansen (1982). A good explanation of optimal moments for GMM is given in the appendix of Arellano (2003). The October 2002 issue of *Journal of Business and Economic Statistics* is devoted to GMM estimation.
- 6.4** The classic treatment of linear IV estimation by Sargan (1958) is a key precursor to GMM.
- 6.5** The nonlinear 2SLS estimator introduced by Amemiya (1974) generalizes easily to the GMM estimator.
- 6.6** Standard references for sequential two-step estimation are Newey (1984), Murphy and Topel (1985), and Pagan (1986).
- 6.7** A standard reference for minimum distance estimation is Chamberlain (1982).
- 6.8** A good overview of empirical likelihood is provided by Mittelhammer, Judge, and Miller (2000) and key references are Owen (1988, 2001) and Qin and Lawless (1994). Imbens (2002) provides a review and application of this relatively new method.
- 6.9** Texts such as Greene's (2003) provide a more detailed coverage of systems estimation than that provided here, especially for linear seemingly unrelated regressions and linear simultaneous equations models.
- 6.10** Amemiya (1985) presents nonlinear simultaneous equations in detail.

Exercises

- 6–1** For the gamma regression model of Exercise 5.2, $E[y|\mathbf{x}] = \exp(\mathbf{x}'\beta)$ and $V[y|\mathbf{x}] = (\exp(\mathbf{x}'\beta))^2/2$.
- (a) Show that these conditions imply that $E[\mathbf{x}\{(y - \mathbf{x}'\beta)^2 - (\exp(\mathbf{x}'\beta))^2/2\}] = \mathbf{0}$.
 - (b) Use the moment condition in part (a) to form a method of moments estimator $\widehat{\beta}_{MM}$.
 - (c) Give the asymptotic distribution of $\widehat{\beta}_{MM}$ using result (6.13).
 - (d) Suppose we use the moment condition $E[\mathbf{x}(y - \exp(\mathbf{x}'\beta))]$ in addition to that in part (a). Give the objective function for a GMM estimator of β .

6–2 Consider the linear regression model for data independent over i with $y_i = \mathbf{x}'_i\beta + u_i$. Suppose $E[u_i|\mathbf{x}_i] \neq 0$ but there are available instruments \mathbf{z}_i with $E[u_i|\mathbf{z}_i] = 0$ and $V[u_i|\mathbf{z}_i] = \sigma_i^2$, where $\dim(\mathbf{z}) > \dim(\mathbf{x})$. We consider the GMM estimator $\hat{\beta}$ that minimizes

$$Q_N(\beta) = [N^{-1} \sum_i \mathbf{z}_i(y_i - \mathbf{x}'_i\beta)]' \mathbf{W}_N [N^{-1} \sum_i \mathbf{z}_i(y_i - \mathbf{x}'_i\beta)].$$

- (a) Derive the limit distribution of $\sqrt{N}(\hat{\beta} - \beta_0)$ using the general GMM result (6.11).
- (b) State how to obtain a consistent estimate of the asymptotic variance of $\hat{\beta}$.
- (c) If errors are homoskedastic what choice of \mathbf{W}_N would you use? Explain your answer.
- (d) If errors are heteroskedastic what choice of \mathbf{W}_N would you use? Explain your answer.

6–3 Consider the Laplace intercept-only example at the end of Section 6.3.6, so $y = \mu + u$. Then GMM estimation is based on $E[\mathbf{h}(\mu)] = \mathbf{0}$, where $\mathbf{h}(\mu) = [(y - \mu), (y - \mu)^3]'$.

- (a) Using knowledge of the central moments of y given in Section 6.3.6, show that $\mathbf{G}_0 = E[\partial \mathbf{h}/\partial \mu] = [-1, -6]'$ and that $\mathbf{S}_0 = E[\mathbf{h}\mathbf{h}']$ has diagonal entries 2 and 720 and off-diagonal entries 24.
- (b) Hence show that $\mathbf{G}_0' \mathbf{S}_0^{-1} \mathbf{G}_0 = 252/432$.
- (c) Hence show that $\hat{\mu}_{\text{OGMM}}$ has asymptotic variance $1.7143/N$.
- (d) Show that the GMM estimator of μ with $\mathbf{W} = \mathbf{I}_2$ has asymptotic variance $19.14/N$.

6–4 This question uses the probit model but requires little knowledge of the model. Let y denote a binary variable that takes value 0 or 1 according to whether or not an event occurs, let \mathbf{x} denote a regressor vector, and assume independent observations.

- (a) Suppose $E[y|\mathbf{x}] = \Phi(\mathbf{x}'\beta)$, where $\Phi(\cdot)$ is the standard normal cdf. Show that $E[(y - \Phi(\mathbf{x}'\beta))\mathbf{x}] = \mathbf{0}$. Hence give the estimating equations for a method of moments estimator for β .
- (b) Will this estimator yield the same estimates as the probit MLE? [For just this part you need to read Section 14.3.]
- (c) Give a GMM objective function corresponding to the estimator in part (a). That is, give an objective function that yields the same first-order conditions, up to a full-rank matrix transformation, as those obtained in part (a).
- (d) Now suppose that because of endogeneity in some of the components $E[y|\mathbf{x}] \neq \Phi(\mathbf{x}'\beta)$. Assume there exists a vector \mathbf{z} , $\dim(\mathbf{z}) > \dim(\mathbf{x})$, such that $E[y - \Phi(\mathbf{x}'\beta)|\mathbf{z}] = 0$. Give the objective function for a consistent estimator of β . The estimator need not be fully efficient.
- (e) For your estimator in part (d) give the asymptotic distribution of the estimator. State clearly any assumptions made on the dgp to obtain this result.
- (f) Give the weighting matrix, and a way to calculate it, for the optimal GMM estimator in part (d).
- (g) Give a real-world example of part (d). That is, give a meaningful example of a probit model with endogenous regressor(s) and valid instrument(s). State the dependent variable, the endogenous regressor(s), and the instrument(s) used to permit consistent estimation. [This part is surprisingly difficult.]

- 6–5** Suppose we impose the constraint that $E[\mathbf{w}_i] = \mathbf{g}(\theta)$, where $\dim[\mathbf{w}] > \dim[\theta]$.
- Obtain the objective function for the GMM estimator.
 - Obtain the objective function for the minimum distance estimator (see Section 6.7) with $\pi = E[\mathbf{w}_i]$ and $\hat{\pi} = \bar{\mathbf{w}}$.
 - Show that MD and GMM are equivalent in this example.
- 6–6** The MD estimator (see Section 6.7) uses the restriction $\pi - \mathbf{g}(\theta) = \mathbf{0}$. Suppose more generally that the restriction is $\mathbf{h}(\theta, \pi) = \mathbf{0}$ and we estimate using the **generalized MD estimator** that minimizes $Q_N(\theta) = \mathbf{h}(\theta, \hat{\pi})' \mathbf{W}_N \mathbf{h}(\theta, \hat{\pi})$. Adapt (6.68)–(6.70) to show that (6.67) holds with $\mathbf{G}_0 = \partial \mathbf{h}(\theta, \pi)/\partial \theta|_{\theta_0, \pi_0}$ and $V[\hat{\pi}]$ replaced by $\mathbf{H}_0' V[\hat{\pi}] \mathbf{H}_0$, where $\mathbf{H}_0 = \partial \mathbf{h}(\theta, \pi)/\partial \pi|_{\theta_0, \pi_0}$.
- 6–7** For data generated from the dgp given in Section 6.6.4 with $N = 1,000$, obtain NL2SLS estimates and compare these to the two-stage estimates.

Hypothesis Tests

7.1. Introduction

In this chapter we consider tests of hypotheses, possibly nonlinear in the parameters, using estimators appropriate for nonlinear models.

The distribution of test statistics can be obtained using the same statistical theory as that used for estimators, since test statistics like estimators are *statistics*, that is, functions of the sample. Given appropriate linearization of estimators and hypotheses, the results closely resemble to those for testing linear restrictions in the linear regression model. The results rely on asymptotic theory, however, and exact t - and F -distributed test statistics for the linear model under normality are replaced by test statistics that are asymptotically standard normal distributed (z -tests) or chi-square distributed.

There are two main practical concerns in hypothesis testing. First, tests may have the wrong size, so that in testing at a nominal significance level of, say, 5%, the actual probability of rejection of the null hypothesis may be much more or less than 5%. Such a wrong size is almost certain to arise in moderate size samples as the underlying asymptotic distribution theory is only an approximation. One remedy is the bootstrap method, introduced in this chapter but sufficiently important and broad to be treated separately in Chapter 11. Second, tests may have low power, so that there is low probability of rejecting the null hypothesis when it should be rejected. This potential weakness of tests is often neglected. Size and power are given more prominence here than in most textbook treatments of testing.

The Wald test, the most widely used testing procedure, is defined in Section 7.2. Section 7.3 additionally presents the likelihood ratio test and score or Lagrange multiplier tests, applicable when estimation is by ML. The various tests are illustrated in Section 7.4. Section 7.5 extends these tests to estimators other than ML, including robust forms of tests. Sections 7.6, 7.7, and 7.8 present, respectively, test power, Monte Carlo simulation methods, and the bootstrap.

Methods for determining model specification and selection, rather than hypothesis tests per se, are given separate treatment in Chapter 8.

7.2. Wald Test

The Wald test, due to Wald (1943), is the preeminent hypothesis test in microeconomics. It requires estimation of the unrestricted model, that is, the model without imposition of the restrictions of the null hypothesis. The Wald test is widely used because modern software usually permits estimation of the unrestricted model even if it is more complicated than the restricted model, and modern software increasingly provides robust variance matrix estimates that permit Wald tests under relatively weak distributional assumptions. The usual statistics for tests of statistical significance of regressors reported by computer packages are examples of Wald test statistics.

This section presents the Wald test of nonlinear hypotheses in considerable detail, presenting both theory and examples. The closely related delta method, used to form confidence intervals or regions for nonlinear functions of parameters, is also presented. A weakness of the Wald test – its lack of invariance to algebraically equivalent parameterizations of the null hypothesis – is detailed at the end of the section.

7.2.1. Linear Hypotheses in Linear Models

We first review standard linear model results, as the Wald test is a generalization of the usual test for linear restrictions in the linear regression model.

The null and alternative hypotheses for a two-sided test of linear restrictions on the regression parameters in the linear regression model $\mathbf{y} = \mathbf{X}'\boldsymbol{\beta} + \mathbf{u}$ are

$$\begin{aligned} H_0 &: \mathbf{R}\boldsymbol{\beta}_0 - \mathbf{r} = \mathbf{0}, \\ H_a &: \mathbf{R}\boldsymbol{\beta}_0 - \mathbf{r} \neq \mathbf{0}, \end{aligned} \tag{7.1}$$

where in the notation used here there are h restrictions, \mathbf{R} is an $h \times K$ matrix of constants of full rank h , $\boldsymbol{\beta}$ is the $K \times 1$ parameter vector, \mathbf{r} is an $h \times 1$ vector of constants, and $h \leq K$.

For example, a joint test that $\beta_1 = 1$ and $\beta_2 - \beta_3 = 2$ when $K = 4$ can be expressed as (7.1) with

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 \end{bmatrix}, \quad \mathbf{r} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

The Wald test of $\mathbf{R}\boldsymbol{\beta}_0 - \mathbf{r} = \mathbf{0}$ is a test of closeness to zero of the sample analogue $\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{r}$, where $\widehat{\boldsymbol{\beta}}$ is the unrestricted OLS estimator. Under the strong assumption that $\mathbf{u} \sim \mathcal{N}[\mathbf{0}, \sigma_0^2 \mathbf{I}]$, the estimator $\widehat{\boldsymbol{\beta}} \sim \mathcal{N}[\boldsymbol{\beta}_0, \sigma_0^2 (\mathbf{X}'\mathbf{X})^{-1}]$ and so

$$\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{r} \sim \mathcal{N}[\mathbf{0}, \sigma_0^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}'],$$

under H_0 , where $\mathbf{R}\boldsymbol{\beta}_0 - \mathbf{r} = \mathbf{0}$ has led to simplification to a mean of $\mathbf{0}$. Taking the quadratic form leads to the test statistic

$$W_1 = (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{r})' [\sigma_0^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{r}),$$

which is exactly $\chi^2(h)$ distributed under H_0 . In practice the test statistic W_1 cannot be calculated, however, as σ_0^2 is not known.

In large samples replacing σ_0^2 by its estimate s^2 does not affect the limit distribution of W_1 , since this is equivalent to premultiplication of W_1 by σ_0^2/s^2 and $\text{plim}(\sigma_0^2/s^2) = 1$ (see the Transformation Theorem A.12). Thus

$$W_2 = (\mathbf{R}\hat{\beta} - r)' [s^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{R}\hat{\beta} - r) \quad (7.2)$$

converges to the $\chi^2(h)$ distribution under H_0 .

The test statistic W_2 is chi-square distributed only asymptotically. In this linear example with normal errors an alternative exact small-sample result can be obtained. A standard result derived in many introductory texts is that

$$W_3 = W_2/h$$

is exactly $F(h, N - K)$ distributed under H_0 , if $s^2 = (N - K)^{-1} \sum_i \hat{u}_i^2$, where \hat{u}_i is the OLS residual. This is the familiar **F -test statistic**, which is often reexpressed in terms of sums of squared residuals.

Exact results such as that for W_3 are not possible in nonlinear models, and even in linear models they require very strong assumptions. Instead, the nonlinear analogue of W_2 is employed, with distributional results that are asymptotic only.

7.2.2. Nonlinear Hypotheses

We consider hypothesis tests of h restrictions, possibly **nonlinear in parameters**, on the $q \times 1$ parameter vector θ , where $h \leq q$. For linear regression $\theta = \beta$ and $q = K$.

The **null** and **alternative hypotheses** for a two-sided test are

$$\begin{aligned} H_0 : \mathbf{h}(\theta_0) &= \mathbf{0}, \\ H_a : \mathbf{h}(\theta_0) &\neq \mathbf{0}, \end{aligned} \quad (7.3)$$

where $\mathbf{h}(\cdot)$ is a $h \times 1$ vector function of θ . Note that $\mathbf{h}(\theta)$ in this chapter is used to denote the restrictions of the null hypothesis. This should not be confused with the use of $\mathbf{h}(\mathbf{w}, \theta)$ in the previous chapter to denote the moment conditions used to form an MM or GMM estimator.

Familiar linear examples include tests of statistical significance of a single coefficient, $h(\theta) = \theta_j = 0$, and tests of subsets of coefficients, $\mathbf{h}(\theta) = \theta_2 = \mathbf{0}$. A nonlinear example of a single restriction is $h(\theta) = \theta_1/\theta_2 - 1 = 0$. These examples are studied in later sections.

It is assumed that $h(\theta)$ is such that the $h \times q$ matrix

$$\mathbf{R}(\theta) = \frac{\partial \mathbf{h}(\theta)}{\partial \theta'} \quad (7.4)$$

is of full rank h when evaluated at $\theta = \theta_0$. This assumption is equivalent to linear independence of restrictions in the linear model, in which case $\mathbf{R}(\theta) = \mathbf{R}$ does not depend on θ and has rank h . It is also assumed that the parameters are not at the **boundary of the parameter space** under the null hypothesis. This rules out, for example, testing $H_0 : \theta_1 = 0$ if the model requires $\theta_1 \geq 0$.

7.2.3. Wald Test Statistic

The intuition behind the Wald test is very simple. The obvious test of whether $\mathbf{h}(\theta_0) = \mathbf{0}$ is to obtain estimate $\widehat{\theta}$ without imposing the restrictions and see whether $\mathbf{h}(\widehat{\theta}) \simeq \mathbf{0}$. If $\mathbf{h}(\widehat{\theta}) \xrightarrow{a} \mathcal{N}[\mathbf{0}, V[\mathbf{h}(\widehat{\theta})]]$ under H_0 then the test statistic

$$W = \mathbf{h}(\widehat{\theta})' [V[\mathbf{h}(\widehat{\theta})]]^{-1} \mathbf{h}(\widehat{\theta}) \xrightarrow{a} \chi^2(h).$$

The only complication is finding $V[\mathbf{h}(\widehat{\theta})]$, which will depend on the restrictions $\mathbf{h}(\cdot)$ and the estimator $\widehat{\theta}$.

By a first-order Taylor series expansion (see section 7.2.4) under the null hypothesis, $\mathbf{h}(\widehat{\theta})$ has the same limit distribution as $\mathbf{R}(\theta_0)(\widehat{\theta} - \theta_0)$, where $\mathbf{R}(\theta)$ is defined in (7.4). Then $\mathbf{h}(\widehat{\theta})$ is asymptotically normal under H_0 with mean zero and variance matrix $\mathbf{R}(\theta_0)V[\widehat{\theta}]\mathbf{R}(\theta_0)'$. A consistent estimate is $\widehat{\mathbf{R}}N^{-1}\widehat{\mathbf{C}}\widehat{\mathbf{R}}'$, where $\widehat{\mathbf{R}} = \mathbf{R}(\widehat{\theta})$ and it is assumed that the estimator $\widehat{\theta}$ is root- N consistent with

$$\sqrt{N}(\widehat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{C}_0], \quad (7.5)$$

and $\widehat{\mathbf{C}}$ is any consistent estimate of \mathbf{C}_0 .

Common Versions of the Wald Test

The preceding discussion leads to the **Wald test statistic**

$$W = N\widehat{\mathbf{h}}'[\widehat{\mathbf{R}}\widehat{\mathbf{C}}\widehat{\mathbf{R}}']^{-1}\widehat{\mathbf{h}}, \quad (7.6)$$

where $\widehat{\mathbf{h}} = \mathbf{h}(\widehat{\theta})$ and $\widehat{\mathbf{R}} = \partial\mathbf{h}(\theta)/\partial\theta'|_{\widehat{\theta}}$. An equivalent expression is $W = \widehat{\mathbf{h}}'[\widehat{\mathbf{R}}\widehat{V}[\widehat{\theta}]\widehat{\mathbf{R}}']^{-1}\widehat{\mathbf{h}}$, where $\widehat{V}[\widehat{\theta}] = N^{-1}\widehat{\mathbf{C}}$ is the estimated asymptotic variance of $\widehat{\theta}$.

The test statistic W is asymptotically $\chi^2(h)$ distributed under H_0 . So H_0 is rejected against H_a at significance level α if $W > \chi^2_\alpha(h)$ and is not rejected otherwise. Equivalently, H_0 is rejected at level α if the **p-value**, which equals $\Pr[\chi^2(h) > W]$, is less than α .

One can also implement the Wald test statistic as an F -test. The **Wald asymptotic F-statistic**

$$F = W/h \quad (7.7)$$

is asymptotically $F(h, N - q)$ distributed. This yields the same *p*-value as W in (7.6) as $N \rightarrow \infty$ though in finite samples the *p*-values will differ. For nonlinear models it is most common to report W , though F is also used in the hope that it might provide a better approximation in small samples.

For a test of just one restriction, the square root of the Wald chi-square test is a standard normal test statistic. This result is useful as it permits testing a one-sided hypothesis. Specifically, for scalar $h(\theta)$ the **Wald z-test statistic** is

$$W_z = \frac{\widehat{h}}{\sqrt{\widehat{\mathbf{r}}N^{-1}\widehat{\mathbf{C}}\widehat{\mathbf{r}}}}, \quad (7.8)$$

where $\widehat{h} = h(\widehat{\theta})$ and $\widehat{\mathbf{r}} = \partial h(\theta)/\partial\theta'|_{\widehat{\theta}}$ is a $1 \times k$ vector. Result (7.6) implies that W_z is asymptotically standard normal distributed under H_0 . Equivalently, W_z is

asymptotically t distributed with $(N - q)$ degrees of freedom, since the t goes to the normal as $N \rightarrow \infty$. So W_z can also be a **Wald t -test statistic**.

Discussion

The Wald test statistic (7.6) for the nonlinear case has the same form as the linear model statistic W_2 given in (7.2). The estimated deviation from the null hypothesis is $\hat{\mathbf{h}}(\hat{\boldsymbol{\theta}})$ rather than $(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})$. The matrix \mathbf{R} is replaced by the estimated derivative matrix $\hat{\mathbf{R}}$, and the assumption that \mathbf{R} is of full rank is replaced by the assumption that \mathbf{R}_0 is of full rank. Finally, the estimated asymptotic variance of the estimator is $N^{-1}\hat{\mathbf{C}}$ rather than $s^2(\mathbf{X}'\mathbf{X})^{-1}$.

There is a range of possible consistent estimates of \mathbf{C}_0 (see Section 5.5.2), leading in practice to different computed values of W or F or W_z that are asymptotically equivalent. In particular, \mathbf{C}_0 is often of the sandwich form $\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}$, consistently estimated by a robust estimate $\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}$. An advantage of the Wald test is that it is easy to robustify to ensure valid statistical inference under relatively weak distributional assumptions, such as potentially heteroskedastic errors.

Rejection of H_0 is more likely the larger is W or F or, for two-sided tests, W_z . This happens the further $\hat{\mathbf{h}}(\hat{\boldsymbol{\theta}})$ is from the null hypothesis value $\mathbf{0}$; the more efficient the estimator $\hat{\boldsymbol{\theta}}$, since then $\hat{\mathbf{C}}$ is small; and the larger the sample size since then N^{-1} is small. The last result is a consequence of testing at unchanged significance level α as sample size increases. In principle one could decrease α as the sample size is increased. Such penalties for fully parametric models are presented in Section 8.5.1.

7.2.4. Derivation of the Wald Statistic

By an exact first-order Taylor series expansion around $\boldsymbol{\theta}_0$

$$\mathbf{h}(\hat{\boldsymbol{\theta}}) = \mathbf{h}(\boldsymbol{\theta}_0) + \left. \frac{\partial \mathbf{h}}{\partial \boldsymbol{\theta}'} \right|_{\boldsymbol{\theta}^+} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

for some $\boldsymbol{\theta}^+$ between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. It follows that

$$\sqrt{N}(\mathbf{h}(\hat{\boldsymbol{\theta}}) - \mathbf{h}(\boldsymbol{\theta}_0)) = \mathbf{R}(\boldsymbol{\theta}^+) \sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

where $\mathbf{R}(\boldsymbol{\theta})$ is defined in (7.4), which implies that

$$\sqrt{N}(\mathbf{h}(\hat{\boldsymbol{\theta}}) - \mathbf{h}(\boldsymbol{\theta}_0)) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{R}_0 \mathbf{C}_0 \mathbf{R}_0'] \quad (7.9)$$

by direct application of the limit normal product rule (Theorem A.7) as $\mathbf{R}(\boldsymbol{\theta}^+) \xrightarrow{P} \mathbf{R}_0 = \mathbf{R}(\boldsymbol{\theta}_0)$ and using the limit distribution for $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ given in (7.5).

Under the null hypothesis (7.9) simplifies since $\mathbf{h}(\boldsymbol{\theta}_0) = \mathbf{0}$, and hence

$$\sqrt{N}\mathbf{h}(\hat{\boldsymbol{\theta}}) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{R}_0 \mathbf{C}_0 \mathbf{R}_0'] \quad (7.10)$$

under H_0 . One could in theory use this multivariate normal distribution to define a rejection region, but it is much simpler to transform to a chi-square distribution. Recall that $\mathbf{z} \sim \mathcal{N}[\mathbf{0}, \Omega]$ with Ω of full rank implies $\mathbf{z}'\Omega^{-1}\mathbf{z} \sim \chi^2(\dim(\Omega))$. Then (7.10)

implies that

$$N\mathbf{h}(\hat{\boldsymbol{\theta}})'[\mathbf{R}_0 \mathbf{C}_0 \mathbf{R}_0']^{-1}\mathbf{h}(\hat{\boldsymbol{\theta}}) \xrightarrow{d} \chi^2(h),$$

under H_0 , where the matrix inverse in this expression exists by the assumptions that \mathbf{R}_0 and \mathbf{C}_0 are of full rank. The Wald statistic defined in (7.6) is obtained upon replacing \mathbf{R}_0 and \mathbf{C}_0 by consistent estimates.

7.2.5. Wald Test Examples

The most common tests are tests of one or more exclusion restrictions. We also provide an example of test of a nonlinear hypothesis.

Tests of Exclusion Restrictions

Consider the exclusion restrictions that the last h components of $\boldsymbol{\theta}$ are equal to zero. Then $\mathbf{h}(\boldsymbol{\theta}) = \boldsymbol{\theta}_2 = \mathbf{0}$ where we partition $\boldsymbol{\theta} = (\boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2)'$. It follows that

$$\mathbf{R}(\boldsymbol{\theta}) = \frac{\partial \mathbf{h}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} = \begin{bmatrix} \frac{\partial \boldsymbol{\theta}_2}{\partial \boldsymbol{\theta}'_1} & \frac{\partial \boldsymbol{\theta}_2}{\partial \boldsymbol{\theta}'_2} \end{bmatrix} = [\mathbf{0} \quad \mathbf{I}_h],$$

where $\mathbf{0}$ is a $(q - h) \times q$ matrix of zeros and \mathbf{I}_h is an $h \times h$ identity matrix, so

$$\mathbf{R}(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\mathbf{R}(\boldsymbol{\theta})' = [\mathbf{0} \quad \mathbf{I}_h] \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_h \end{bmatrix} = \mathbf{C}_{22}.$$

The Wald test statistic for **exclusion restrictions** is therefore

$$W = \boldsymbol{\theta}_2'[N^{-1}\widehat{\mathbf{C}}_{22}]^{-1}\boldsymbol{\theta}_2, \quad (7.11)$$

where $N^{-1}\widehat{\mathbf{C}}_{22} = \widehat{\mathbf{V}}[\boldsymbol{\theta}_2]$, and is asymptotically distributed as $\chi^2(h)$ under H_0 .

This test statistic is a generalization of the test of subsets of regressors in the linear regression model. In that case small-sample results are available if errors are normally distributed and the related F-test is instead used.

Tests of Statistical Significance

Tests of significance of a single coefficient are tests of whether or not θ_j , the j th component of $\boldsymbol{\theta}$, differs from zero. Then $h(\boldsymbol{\theta}) = \theta_j$ and $\mathbf{r}(\boldsymbol{\theta}) = \partial h / \partial \boldsymbol{\theta}'$ is a vector of zeros except for a j th entry of 1, so (7.8) simplifies to

$$W_z = \frac{\widehat{\theta}_j}{\text{se}[\widehat{\theta}_j]}, \quad (7.12)$$

where $\text{se}[\widehat{\theta}_j] = \sqrt{N^{-1}\widehat{c}_{jj}}$ is the standard error of $\widehat{\theta}_j$ and \widehat{c}_{jj} is the j th diagonal entry in $\widehat{\mathbf{C}}$.

The test statistic W_z in (7.12) is often called a “**t-statistic**”, owing to results for the linear regression model under normality, but strictly speaking it is an asymptotic “**z-statistic**.”

For a **two-sided test** of $H_0 : \theta_{j0} = 0$ against $H_a : \theta_{j0} \neq 0$, H_0 is rejected at significance level α if $|W_z| > z_{\alpha/2}$ and is not rejected otherwise. This yields exactly the same results as the Wald chi-square test, since $W_z^2 = W$, where W is defined in (7.6), and $z_{\alpha/2}^2 = \chi^2(1)$.

Often there is prior information about the sign of θ_j . Then one should use a **one-sided hypothesis test**. For example, suppose it is felt based on economic reasoning or past studies that $\theta_j > 0$. It makes a difference whether $\theta_j > 0$ is specified to be the null or the alternative hypothesis. For one-sided tests it is customary to specify the claim made as the alternative hypothesis, as it can be shown that then stronger evidence is required to support the claim. Here $H_0 : \theta_{j0} \leq 0$ is rejected against $H_a : \theta_{j0} > 0$ at significance level α if $W_z > z_\alpha$. Similarly, for a claim that $\theta_j < 0$, test $H_0 : \theta_{j0} \geq 0$ against $H_a : \theta_{j0} < 0$ and reject H_0 at significance level α if $W_z < -z_\alpha$.

Computer output usually gives the p -value for a two-sided test, but in many cases it is more appropriate to use a one-sided test. If $\hat{\theta}_j$ has the “correct” sign then the p -value for the one-sided test is half that reported for a two-sided test.

Tests of Nonlinear Restriction

Consider a test of the single nonlinear restriction

$$H_0 : h(\boldsymbol{\theta}) = \theta_1/\theta_2 - 1 = 0.$$

Then $\mathbf{R}(\boldsymbol{\theta})$ is a $1 \times q$ vector with first element $\partial h / \partial \theta_1 = 1/\theta_2$, second element $\partial h / \partial \theta_2 = -\theta_1/\theta_2^2$, and remaining elements zero. By letting \widehat{c}_{jk} denote the jk th element of $\widehat{\mathbf{C}}$, (7.6) becomes

$$\mathbf{W} = N \left(\frac{\widehat{\theta}_1}{\widehat{\theta}_2} - 1 \right)^2 \left(\begin{bmatrix} \frac{1}{\widehat{\theta}_2} & -\frac{\widehat{\theta}_1}{\widehat{\theta}_2^2} \mathbf{0} \end{bmatrix} \begin{bmatrix} \widehat{c}_{11} & \widehat{c}_{12} & \cdots \\ \widehat{c}_{21} & \widehat{c}_{22} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} 1/\widehat{\theta}_2 \\ -\widehat{\theta}_1/\widehat{\theta}_2^2 \\ \mathbf{0} \end{bmatrix} \right)^{-1},$$

where $\mathbf{0}$ is a $(q-2) \times q$ matrix of zeros, yielding

$$\mathbf{W} = N[\widehat{\theta}_2(\widehat{\theta}_1 - \widehat{\theta}_2)]^2 (\widehat{\theta}_2^2 \widehat{c}_{11} - 2\widehat{\theta}_1 \widehat{\theta}_2 \widehat{c}_{12} + \widehat{\theta}_1^2 \widehat{c}_{22})^{-1}, \quad (7.13)$$

which is asymptotically $\chi^2(1)$ distributed under H_0 . Equivalently, $\sqrt{\mathbf{W}}$ is asymptotically standard normal distributed.

7.2.6. Tests in Misspecified Models

Most treatments of hypothesis testing, including that given in Chapters 7 and 8 of this book, assume that the null hypothesis model is correctly specified, aside from relatively minor misspecification that does not affect estimator consistency but requires robustification of standard errors.

In practice this is a considerable oversimplification. For example, in testing for heteroskedastic errors it is assumed that this is the only respect in which the regression is deficient. However, if the conditional mean is misspecified then the true size of the test will differ from the nominal size, even asymptotically. Moreover, asymptotic

equivalence of tests, such as that for the Wald, likelihood ratio, and Lagrange multiplier tests, will no longer hold. The better specified the model, however, the more useful are the tests.

Also, note that tests often have some power against hypotheses other than the explicitly stated alternative hypothesis. For example, suppose the null hypothesis model is $y = \beta_1 + \beta_2 x + u$, where u is homoskedastic. A test of whether to also include z as a regressor will also have some power against the alternative that the model is nonlinear in x , for example $y = \beta_1 + \beta_2 x + \beta_3 x^2 + u$, if x and z are correlated. Similarly, a test against heteroskedastic errors will also have some power against nonlinearity in x . Rejection of the null hypothesis does not mean that the alternative hypothesis model is the only possible model.

7.2.7. Joint Versus Separate Tests

In applied work one often wants to know which coefficients out of a set of coefficients are “significant.” When there are several hypotheses under test, one can either do a **joint test** or simultaneous test of all hypotheses of interest or perform **separate tests** of the hypotheses.

A leading example in linear regression concerns the use of separate t -tests for testing the null hypotheses $H_{10} : \beta_1 = 0$ and $H_{20} : \beta_2 = 0$ versus using an F -test of the joint hypothesis $H_0 : \beta_1 = \beta_2 = 0$, where throughout the alternative is that at least one of the parameters does not equal zero. The F -test is an explicit joint test, with rejection of H_0 if the estimated point $(\hat{\beta}_1, \hat{\beta}_2)$ falls outside an elliptical probability contour. Alternatively, the two separate t -tests can be conducted. This procedure is an implicit joint test, called an **induced test** (Savin, 1984). The separate tests reject H_0 if either H_{10} or H_{20} is rejected, which occurs if $(\hat{\beta}_1, \hat{\beta}_2)$ falls outside a rectangle whose boundaries are the critical values of the two test statistics. Even if the same significance level is used to test H_0 , so that the ellipse and rectangles have the same area, the rejection regions for the joint and separate tests differ and there is a potential for a conflict between them. For example, $(\hat{\beta}_1, \hat{\beta}_2)$ may lie within the ellipse but outside the rectangle.

Let e_1 and e_2 denote the event of type I error (see Section 7.5.1) in the two separate tests, and let $e_I = e_1 \cup e_2$ denote the event of a type I error in the induced joint test. Then $\Pr[e_I] = \Pr[e_1] + \Pr[e_2] - \Pr[e_1 \cap e_2]$, which implies that

$$\alpha_I \leq \alpha_1 + \alpha_2, \tag{7.14}$$

where α_I , α_1 , and α_2 denote the sizes of, respectively, the induced joint test, the first separate test, and the second separate test. In the special case where the separate tests are statistically independent, $\Pr[e_1 \cap e_2] = \Pr[e_1] \Pr[e_2] = \alpha_1 \alpha_2$ and hence $\alpha_I = \alpha_1 + \alpha_2 - \alpha_1 \alpha_2$. For a typically low value of α_1 and α_2 , such as .05 or .01, $\alpha_1 \alpha_2$ is very small and the upper bound (7.14) is a good indicator of the size of the test.

A substantial literature on induced tests examines the problem of choosing critical values for the separate tests such that the induced test has a known size. We do not pursue this issue at length but mention the **Bonferroni t -test** as an example. The critical values of this test have been tabulated; see Savin (1984).

Statistically independent tests arise in linear regression with orthogonal regressors and in likelihood-based testing (see Section 7.3) if relevant parts of the information matrix are diagonal. Then the induced joint test statistic is based on the two statistically independent separate test statistics, whereas the explicit joint null test statistic is the sum of the two separate test statistics. The joint null may be rejected because either one component or both components of the null are rejected. The use of separate tests will reveal which situation applies.

In the more general case of correlated regressors or a nondiagonal information matrix, the explicit joint test suffers from the disadvantage that the rejection of the null does not indicate the source of the rejection. If the induced joint test is used then setting the size of the test requires some variant of the Bonferroni test or approximation using the upper bound in (7.14). Similar issues also arise when separate tests are applied sequentially, with each stage conditioned on the outcome of the previous stage. Section 18.7.1 presents an example with discussion of a joint test of two hypotheses where the two components of the test are correlated.

7.2.8. Delta Method for Confidence Intervals

The method used to derive the Wald test statistic is called the **delta method**, as Taylor series approximation of $\mathbf{h}(\hat{\theta})$ entails taking the derivative of $\mathbf{h}(\theta)$. This method can also be used to obtain the distribution of a nonlinear combination of parameters and hence form confidence intervals or regions.

One example is estimating the ratio θ_1/θ_2 by $\hat{\theta}_1/\hat{\theta}_2$. A second example is prediction of the conditional mean $g(\mathbf{x}'\beta)$, say, using $g(\mathbf{x}'\hat{\beta})$. A third example is the estimated elasticity with respect to change in one component of \mathbf{x} .

Confidence Intervals

Consider inference on the parameter vector $\gamma = \mathbf{h}(\theta)$ that is estimated by

$$\hat{\gamma} = \mathbf{h}(\hat{\theta}), \quad (7.15)$$

where the limit distribution of $\sqrt{N}(\hat{\theta} - \theta_0)$ is that given in (7.5). Then direct application of (7.9) yields $\sqrt{N}(\hat{\gamma} - \gamma_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{R}(\theta_0)\mathbf{C}_0\mathbf{R}(\theta_0)']$, where $\mathbf{R}(\theta)$ is defined in (7.4). Equivalently, we say that $\hat{\gamma}$ is asymptotically normally distributed with estimated asymptotic variance matrix

$$\widehat{\mathbf{V}}[\hat{\gamma}] = \widehat{\mathbf{R}}N^{-1}\widehat{\mathbf{C}}\widehat{\mathbf{R}}', \quad (7.16)$$

a result that can be used to form confidence intervals or regions.

In particular, a $100(1 - \alpha)\%$ **confidence interval for the scalar parameter** γ is

$$\gamma \in \hat{\gamma} \pm z_{\alpha/2}se[\hat{\gamma}], \quad (7.17)$$

where

$$se[\hat{\gamma}] = \sqrt{\widehat{\mathbf{R}}N^{-1}\widehat{\mathbf{C}}\widehat{\mathbf{R}}'}, \quad (7.18)$$

where $\widehat{\mathbf{r}} = \mathbf{r}(\hat{\theta})$ and $\mathbf{r}(\theta) = \partial\gamma/\partial\theta' = \partial h(\theta)/\partial\theta'$.

Confidence Interval Examples

As an example, suppose that $E[y|\mathbf{x}] = \exp(\mathbf{x}'\boldsymbol{\beta})$ and we wish to obtain a confidence interval for the predicted conditional mean when $\mathbf{x} = \mathbf{x}_p$. Then $h(\boldsymbol{\beta}) = \exp(\mathbf{x}'_p\boldsymbol{\beta})$, so $\partial h/\partial\boldsymbol{\beta}' = \exp(\mathbf{x}'_p\boldsymbol{\beta})\mathbf{x}_p$ and (7.18) yields

$$se[\exp(\mathbf{x}'_p\hat{\boldsymbol{\beta}})] = \exp(\mathbf{x}'_p\hat{\boldsymbol{\beta}})\sqrt{\mathbf{x}'_p N^{-1}\hat{\mathbf{C}}\mathbf{x}_p},$$

where $\hat{\mathbf{C}}$ is a consistent estimate of the variance matrix in the limit distribution of $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$.

As a second example, suppose we wish to obtain a confidence interval for e^β rather than for β , a scalar coefficient. Then $h(\boldsymbol{\beta}) = e^\beta$, so $\partial h/\partial\beta = e^\beta$ and (7.18) yields $se[e^{\hat{\beta}}] = e^{\hat{\beta}}se[\hat{\beta}]$. This yields a 95% confidence interval for e^β of $e^{\hat{\beta}} \pm 1.96e^{\hat{\beta}}se[\hat{\beta}]$.

The delta method is not always the best method to obtain a confidence interval, because it restricts the confidence interval to being symmetric about $\hat{\gamma}$. Moreover, in the preceding example the confidence interval can include negative values even though $e^\beta > 0$. An alternative confidence interval is obtained by exponentiation of the terms in the confidence interval for β . Then

$$\begin{aligned} & \Pr[\hat{\beta} - 1.96se[\hat{\beta}] < \beta < \hat{\beta} + 1.96se[\hat{\beta}]] = 0.95 \\ & \Rightarrow \Pr[\exp(\hat{\beta} - 1.96se[\hat{\beta}]) < e^\beta < \exp(\hat{\beta} + 1.96se[\hat{\beta}])] = 0.95. \end{aligned}$$

This confidence interval has the advantage of being asymmetric and including only positive values. This transformation is often used for confidence intervals for slope parameters in binary outcome models and in duration models. The approach can be generalized to other transformations $\gamma = h(\theta)$, provided $h(\cdot)$ is monotonic.

7.2.9. Lack of Invariance of the Wald Test

The Wald test statistic is easily obtained, provided estimates of the unrestricted model can be obtained, and is no less powerful than other possible test procedures, as discussed in later sections. For these reasons it is the most commonly used test procedure.

However, the Wald test has a fundamental problem: It is not invariant to algebraically equivalent parameterizations of the null hypothesis. For example, consider the example of Section 7.2.5. Then $H_0 : \theta_1/\theta_2 - 1 = 0$ can equivalently be expressed as $H_0 : \theta_1 - \theta_2 = 0$, leading to Wald chi-square test statistic

$$W^* = N(\hat{\theta}_1 - \hat{\theta}_2)^2 (\hat{c}_{11} - 2\hat{c}_{12} + \hat{c}_{22})^{-1}, \quad (7.19)$$

which differs from W in (7.13). The statistics W and W^* can differ substantially in finite samples, even though asymptotically they are equivalent. The small-sample difference can be quite substantial, as demonstrated in a Monte Carlo exercise by Gregory and Veall (1985), who considered a very similar example. For tests with nominal size 0.05, one variant of the Wald test had actual size between 0.04 and 0.06 across all simulations, so asymptotic theory provided a good small-sample approximation, whereas an alternative asymptotically equivalent variant of the Wald test had actual size that in some simulations exceeded 0.20.

Phillips and Park (1988) explained the differences by showing that, although different representations of the null hypothesis restrictions have the same chi-square distribution using conventional asymptotic methods, they have different asymptotic distributions using a more refined asymptotic theory based on Edgeworth expansions (see Section 11.4.3). Furthermore, in particular settings such as the previous example, the Edgeworth expansions can be used to indicate parameterizations of H_0 and regions of the parameter space where the usual asymptotic theory is likely to provide a poor small-sample approximation.

The lesson is that care is needed when nonlinear restrictions are being tested. As a robustness check one can perform several Wald tests using different algebraically equivalent representations of the null hypothesis restrictions. If these lead to substantially different conclusions there may be a problem. One solution is to perform a bootstrap version of the Wald test. This can provide better small-sample performance and eliminate much of the difference between Wald tests that use different representations of H_0 , because from Section 11.4.4 the bootstrap essentially implements an Edgeworth expansion. A second solution is to use other testing methods, given in the next section, that are invariant to different representations of H_0 .

7.3. Likelihood-Based Tests

In this section we consider hypothesis testing when the likelihood function is known, that is, the distribution is fully specified. There are then three classical statistical techniques for testing hypotheses – the Wald test, the likelihood ratio (LR) test, and the Lagrange multiplier (LM) test. A fourth test, the **C(α) test**, due to Neyman (1959), is less commonly used and is not presented here; see Davidson and MacKinnon (1993). All four tests are asymptotically equivalent, so one chooses among them based on ease of computation and on finite-sample performance. We also do not cover the **smooth test** of Neyman (1937), which Bera and Ghosh (2002) argue is optimal and is as fundamental as the other tests.

These results assume correct specification of the likelihood function. Extension to tests based on quasi-ML estimators, as well as on m-estimators and efficient GMM estimators, is given in Section 7.5.

7.3.1. Wald, Likelihood Ratio, and Lagrange Multiplier (Score) Tests

Let $L(\theta)$ denote the likelihood function, the joint conditional density of \mathbf{y} given \mathbf{X} and parameters θ . We wish to test the null hypothesis given in (7.3) that $\mathbf{h}(\theta_0) = \mathbf{0}$.

Tests other than the Wald test require estimation that imposes the restrictions of the null hypothesis. Define the estimators

$$\begin{aligned}\widehat{\theta}_u &\text{ (unrestricted MLE),} \\ \widetilde{\theta}_r &\text{ (restricted MLE).}\end{aligned}\tag{7.20}$$

The **unrestricted MLE** $\widehat{\theta}_u$ maximizes $\ln L(\theta)$; it was more simply denoted $\widehat{\theta}$ in earlier discussion of the Wald test. The **restricted MLE** $\widetilde{\theta}_r$ maximizes the Lagrangian

$\ln L(\boldsymbol{\theta}) - \boldsymbol{\lambda}' \mathbf{h}(\boldsymbol{\theta})$, where $\boldsymbol{\lambda}$ is an $h \times 1$ vector of Lagrangian multipliers. In the simple case of exclusion restrictions $\mathbf{h}(\boldsymbol{\theta}) = \boldsymbol{\theta}_2 = \mathbf{0}$, where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$, the restricted MLE is $\tilde{\boldsymbol{\theta}}_r = (\tilde{\boldsymbol{\theta}}_{1r}', \mathbf{0}')$, where $\tilde{\boldsymbol{\theta}}_{1r}'$ is obtained simply as the maximum with respect to $\boldsymbol{\theta}_1$ of the restricted likelihood $\ln L(\boldsymbol{\theta}_1, \mathbf{0})$ and $\mathbf{0}$ is a $(q-h) \times 1$ vector of zeros.

We motivate and define the three test statistics here, with derivation deferred to Section 7.3.3. All three test statistics converge in distribution to $\chi^2(h)$ under H_0 . So H_0 is rejected at significance level α if the computed test statistic exceeds $\chi_\alpha^2(h)$. Equivalently, reject H_0 at level α if $p \leq \alpha$, where $p = \Pr[\chi^2(h) > t]$ is the p -value and t is the computed value of the test statistic.

Likelihood Ratio Test

The motivation for the LR test statistic is that if H_0 is true, the unconstrained and constrained maxima of the log-likelihood function should be the same. This suggests using a function of the difference between $\ln L(\hat{\boldsymbol{\theta}}_u)$ and $\ln L(\tilde{\boldsymbol{\theta}}_r)$.

Implementation requires obtaining the limit distribution of this difference. It can be shown that twice the difference is asymptotically chi-square distributed under H_0 . This leads immediately to the **likelihood ratio test** statistic

$$\text{LR} = -2 [\ln L(\tilde{\boldsymbol{\theta}}_r) - \ln L(\hat{\boldsymbol{\theta}}_u)]. \quad (7.21)$$

Wald Test

The motivation for the Wald test is that if H_0 is true, the unrestricted MLE $\hat{\boldsymbol{\theta}}_u$ should satisfy the restrictions of H_0 , so $h(\hat{\boldsymbol{\theta}}_u)$ should be close to zero.

Implementation requires obtaining the asymptotic distribution of $\mathbf{h}(\hat{\boldsymbol{\theta}}_u)$. The general form of the Wald test is given in (7.6). Specialization occurs for the MLE because by the IM equality $\mathbf{V}[\hat{\boldsymbol{\theta}}_u] = -N^{-1}\mathbf{A}_0^{-1}$, where

$$\mathbf{A}_0 = \text{plim } N^{-1} \frac{\partial^2 \ln L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \Big|_{\boldsymbol{\theta}_0}. \quad (7.22)$$

This leads to the **Wald test** statistic

$$W = -N \hat{\mathbf{h}}' [\hat{\mathbf{R}} \hat{\mathbf{A}}^{-1} \hat{\mathbf{R}}']^{-1} \hat{\mathbf{h}}, \quad (7.23)$$

where $\hat{\mathbf{h}} = \mathbf{h}(\hat{\boldsymbol{\theta}}_u)$, $\hat{\mathbf{R}} = \mathbf{R}(\hat{\boldsymbol{\theta}}_u)$, $\mathbf{R}(\boldsymbol{\theta}) = \partial \mathbf{h}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}'$, and $\hat{\mathbf{A}}$ is a consistent estimate of \mathbf{A}_0 . The minus sign appears since \mathbf{A}_0 is negative definite.

Lagrange Multiplier Test or Score Test

One motivation for the LM test statistic is that the gradient $\partial \ln L / \partial \boldsymbol{\theta}|_{\hat{\boldsymbol{\theta}}_u} = \mathbf{0}$ at the maximum of the likelihood function. If H_0 is true, then this maximum should also occur at the restricted MLE (i.e., $\partial \ln L / \partial \boldsymbol{\theta}|_{\tilde{\boldsymbol{\theta}}_r} \simeq \mathbf{0}$) because imposing the constraint will have little impact on the estimated value of $\boldsymbol{\theta}$. Using this motivation LM is called the **score test** because $\partial \ln L / \partial \boldsymbol{\theta}$ is the score vector.

An alternative motivation is to measure the closeness to zero of the **Lagrange multipliers** of the constrained optimization problem for the restricted MLE. Maximizing

$\ln L(\theta) - \lambda' \mathbf{h}(\theta)$ with respect to θ implies that

$$\frac{\partial \ln L}{\partial \theta} \Big|_{\tilde{\theta}_r} = \frac{\partial \mathbf{h}(\theta)'}{\partial \theta} \Big|_{\tilde{\theta}_r} \times \tilde{\lambda}_r. \quad (7.24)$$

It follows that tests based on the estimated Lagrange multipliers $\tilde{\lambda}_r$ are equivalent to tests based on the score $\partial \ln L / \partial \theta|_{\tilde{\theta}_r}$, since $\partial \mathbf{h} / \partial \theta'$ is assumed to be of full rank.

Implementation requires obtaining the asymptotic distribution of $\partial \ln L / \partial \theta|_{\tilde{\theta}_r}$. This leads to the **Lagrange multiplier test** or **score test** statistic

$$LM = -N^{-1} \frac{\partial \ln L}{\partial \theta'} \Big|_{\tilde{\theta}_r} \tilde{\mathbf{A}}^{-1} \frac{\partial \ln L}{\partial \theta} \Big|_{\tilde{\theta}_r}, \quad (7.25)$$

where $\tilde{\mathbf{A}}$ is a consistent estimate of \mathbf{A}_0 in (7.22) evaluated at $\tilde{\theta}_r$ rather than $\hat{\theta}_u$.

The **LM test**, due to Aitchison and Silvey (1958) and Silvey (1959), is equivalent to the **score test**, due to Rao (1947). The test statistic LM is usually derived by obtaining an analytical expression for the score rather than the Lagrange multipliers. Econometricians usually call the test an LM test, even though a clearer terminology is to call it a score test.

Discussion

Good intuition is provided by the expository graphical treatment of the three tests by Buse (1982) that views all three tests as measuring the change in the log-likelihood. Here we provide a verbal summary.

Consider scalar parameter and a Wald test of whether $\theta_0 - \theta^* = 0$. Then a given departure of $\hat{\theta}_u$ from θ^* will translate into a larger change in $\ln L$, the more curved is the log-likelihood function. A natural measure of curvature is the second derivative $H(\theta) = \partial^2 \ln L / \partial \theta^2$. This suggests $W = -(\hat{\theta}_u - \theta^*)^2 H(\hat{\theta}_u)$. The statistic W in (7.23) can be viewed as a generalization to vector θ and more general restrictions $\mathbf{h}(\theta_0)$ with $N\hat{\mathbf{A}}$ measuring the curvature.

For the score test Buse shows that a given value of $\partial \ln L / \partial \theta|_{\tilde{\theta}_r}$ translates into a larger change in $\ln L$, the less curved is the log-likelihood function. This leads to use of $(N\tilde{\mathbf{A}})^{-1}$ in (7.25). And the statistic LR directly compares the log-likelihoods.

An Illustration

To illustrate the three tests consider an iid example with $y_i \sim \mathcal{N}[\mu_0, 1]$ and test of $H_0 : \mu_0 = \mu^*$. Then $\hat{\mu}_u = \bar{y}$ and $\tilde{\mu}_r = \mu^*$.

For the LR test, $\ln L(\mu) = -\frac{N}{2} \ln 2\pi - \frac{1}{2} \sum_i (y_i - \mu)^2$ and some algebra yields

$$LR = 2[\ln L(\bar{y}) - \ln L(\mu^*)] = N(\bar{y} - \mu^*)^2.$$

The Wald test is based on whether $\bar{y} - \mu^* \simeq 0$. Here it is easy to show that $\bar{y} - \mu^* \sim \mathcal{N}[0, 1/N]$ under H_0 , leading to the quadratic form

$$W = (\bar{y} - \mu^*)[1/N]^{-1}(\bar{y} - \mu^*).$$

This simplifies to $N(\bar{y} - \mu^*)^2$ and so here $W = LR$.

The LM test is based on closeness to zero of $\partial \ln L(\mu)/\partial \mu|_{\mu^*} = \sum_i (y_i - \mu)|_{\mu^*} = N(\bar{y} - \mu^*)$. This is just a rescaling of $(\bar{y} - \mu^*)$ so $\text{LM} = W$. More formally, $\tilde{A}(\mu^*) = -1$ since $\partial^2 \ln L(\mu)/\partial \mu^2 = -N$ and (7.25) yields

$$\text{LM} = N^{-1}(N(\bar{y} - \mu^*))[1]^{-1}(N(\bar{y} - \mu^*)).$$

This also simplifies to $N(\bar{y} - \mu^*)^2$ and verifies that $\text{LM} = W = LR$.

Despite their quite different motivations, the three test statistics are equivalent here. This exact equivalence is special to this example with constant curvature owing to a log-likelihood quadratic in μ . More generally the three test statistics differ in finite samples but are equivalent asymptotically (see Section 7.3.4).

7.3.2. Poisson Regression Example

Consider testing exclusion restrictions in the Poisson regression model introduced in Section 5.2. This example is mainly pedagogical as in practice one should perform statistical inference for count data under weaker distributional assumptions than those of the Poisson model (see Chapter 20).

If y given \mathbf{x} is Poisson distributed with conditional mean $\exp(\mathbf{x}'\beta)$ then the log-likelihood function is

$$\ln L(\beta) = \sum_{i=1}^N \{-\exp(\mathbf{x}'_i\beta) + y_i \mathbf{x}'_i\beta - \ln y_i!\}. \quad (7.26)$$

For h exclusion restrictions the null hypothesis is $H_0 : \mathbf{h}(\beta) = \beta_2 = \mathbf{0}$, where $\beta = (\beta'_1, \beta'_2)'$.

The unrestricted MLE $\widehat{\beta}$ maximizes (7.26) with respect to β and has first-order conditions $\sum_i (y_i - \exp(\mathbf{x}'_i\beta))\mathbf{x}_i = \mathbf{0}$. The limit variance matrix is $-\mathbf{A}^{-1}$, where

$$\mathbf{A} = -\text{plim } N^{-1} \sum_i \exp(\mathbf{x}'_i\beta) \mathbf{x}_i \mathbf{x}'_i.$$

The restricted MLE is $\widetilde{\beta} = (\widetilde{\beta}'_1, \mathbf{0}')'$, where $\widetilde{\beta}_1$ maximizes (7.26) with respect to β_1 , with $\mathbf{x}'_i\beta$ replaced by $\mathbf{x}'_{1i}\beta_1$ since $\beta_2 = \mathbf{0}$. Thus $\widetilde{\beta}_1$ solves the first-order conditions $\sum_i (y_i - \exp(\mathbf{x}'_{1i}\beta_1))\mathbf{x}_{1i} = \mathbf{0}$.

The LR test statistic (7.21) is easily calculated from the fitted log-likelihoods of the restricted and unrestricted models.

The Wald test statistic for exclusion restrictions from Section 7.2.5 is $W = -N\widehat{\beta}_2'\widehat{\mathbf{A}}^{22}\widehat{\beta}_2$, where $\widehat{\mathbf{A}}^{22}$ is the (2,2) block of $\widehat{\mathbf{A}}^{-1}$ and $\widehat{\mathbf{A}} = -N^{-1} \sum_i \exp(\mathbf{x}'_i\widehat{\beta}) \mathbf{x}_i \mathbf{x}'_i$.

The LM test is based on $\partial \ln L(\beta)/\partial \beta = \sum_i \mathbf{x}_i(y_i - \exp(\mathbf{x}'_i\beta))$. At the restricted MLE this equals $\sum_i \mathbf{x}_i \widetilde{u}_i$, where $\widetilde{u}_i = y_i - \exp(\mathbf{x}'_{1i}\beta_1)$ is the residual from estimation of the restricted model. The LM test statistic (7.25) is

$$\text{LM} = \left[\sum_{i=1}^N \mathbf{x}_i \widetilde{u}_i \right]' \left[\sum_{i=1}^N \exp(\mathbf{x}'_{1i}\widetilde{\beta}_1) \mathbf{x}_i \mathbf{x}'_i \right]^{-1} \left[\sum_{i=1}^N \mathbf{x}_i \widetilde{u}_i \right]. \quad (7.27)$$

Some further simplification is possible since $\sum_i \mathbf{x}_{1i} \widetilde{u}_i = \mathbf{0}$ from the first-order conditions for the restricted MLE given earlier. The LM test here is based on the correlation between the omitted regressors and the residual, a result that is extended to other examples in Section 7.3.5.

In general it can be difficult to obtain an algebraic expression for the LM test. For standard applications of the LM test this has been done and is incorporated into computer packages. Computation by auxiliary regression may also be possible (see Section 3.5).

7.3.3. Derivation of Tests

The distribution of the Wald test was formally derived in Section 7.2.4. Proofs for the likelihood ratio and Lagrange multiplier tests are more complicated and we merely sketch them here.

Likelihood Ratio Test

For simplicity consider the special case where the null hypothesis is $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}$, so that there is no estimation error in $\tilde{\boldsymbol{\theta}}_r = \bar{\boldsymbol{\theta}}$. Taking a second-order Taylor series expansion of $\ln L(\bar{\boldsymbol{\theta}})$ about $\ln L(\hat{\boldsymbol{\theta}}_u)$ yields

$$\ln L(\bar{\boldsymbol{\theta}}) = \ln L(\hat{\boldsymbol{\theta}}_u) + \frac{\partial \ln L}{\partial \boldsymbol{\theta}'} \Big|_{\hat{\boldsymbol{\theta}}_u} (\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_u) + \frac{1}{2} (\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_u)' \frac{\partial^2 \ln L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \Big|_{\hat{\boldsymbol{\theta}}_u} (\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_u) + R,$$

where R is a remainder term. Since $\partial \ln L / \partial \boldsymbol{\theta}|_{\hat{\boldsymbol{\theta}}_u} = \mathbf{0}$ by the first-order conditions, this implies upon rearrangement that

$$-2 [\ln L(\bar{\boldsymbol{\theta}}) - \ln L(\hat{\boldsymbol{\theta}}_u)] = -(\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_u)' \frac{\partial^2 \ln L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \Big|_{\hat{\boldsymbol{\theta}}_u} (\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_u) + R. \quad (7.28)$$

The right-hand side of (7.28) is $\chi^2(h)$ under $H_0 : \boldsymbol{\theta} = \bar{\boldsymbol{\theta}}$ since by standard results $\sqrt{N}(\hat{\boldsymbol{\theta}}_u - \bar{\boldsymbol{\theta}}) \xrightarrow{d} \mathcal{N}[\mathbf{0}, -[\text{plim } N^{-1} \partial^2 \ln L / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}']^{-1}]$. For derivation of the limit distribution of LR in the general case see, for example, Amemiya (1985, p. 143).

A reason for preferring LR is that by the Neyman–Pearson (1933) lemma the uniformly most powerful test for testing a simple null hypothesis versus simple alternative hypothesis is a function of the likelihood ratio $L(\tilde{\boldsymbol{\theta}}_r)/L(\hat{\boldsymbol{\theta}}_u)$, though not necessarily the specific function $-2 \ln(L(\tilde{\boldsymbol{\theta}}_r)/L(\hat{\boldsymbol{\theta}}_u))$ that equals LR given in (7.21) and gives the test statistic its name.

LM or Score Test

By a first-order Taylor series expansion

$$\frac{1}{\sqrt{N}} \frac{\partial \ln L}{\partial \boldsymbol{\theta}} \Big|_{\tilde{\boldsymbol{\theta}}_r} = \frac{1}{\sqrt{N}} \frac{\partial \ln L}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}_0} + \frac{1}{N} \frac{\partial^2 \ln L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \sqrt{N} (\tilde{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_0),$$

and both terms in the right-hand side contribute to the limit distribution. Then the $\chi^2(h)$ distribution of LM defined in (7.25) follows since it can be shown that

$$\mathbf{R}_0 \mathbf{A}_0^{-1} \frac{1}{\sqrt{N}} \frac{\partial \ln L}{\partial \boldsymbol{\theta}} \Big|_{\tilde{\boldsymbol{\theta}}_r} \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{R}_0 \mathbf{A}_0^{-1} \mathbf{B}_0 \mathbf{A}_0^{-1} \mathbf{R}_0'], \quad (7.29)$$

where details are provided in Wooldridge (2002, p. 365), for example, and \mathbf{R}_0 and \mathbf{A}_0 are defined in (7.4) and (7.22) and

$$\mathbf{B}_0 = \text{plim } N^{-1} \left. \frac{\partial \ln L}{\partial \boldsymbol{\theta}} \frac{\partial \ln L}{\partial \boldsymbol{\theta}'} \right|_{\boldsymbol{\theta}_0}. \quad (7.30)$$

Result (7.29) leads to a chi-square statistic that is much more complicated than (7.25), but simplification to (7.25) then occurs by the information matrix equality.

7.3.4. Which Test?

Choice of test procedure is usually made based on existence of robust versions, finite-sample performance, and ease of computation.

Asymptotic Equivalence

All three test statistics are asymptotically distributed as $\chi^2(h)$ under H_0 . Furthermore, all three can be shown to be noncentral $\chi^2(h; \lambda)$ distributed with the same noncentrality parameter under local alternatives. Details are provided for the Wald test in Section 7.6.3. So the tests all have the same asymptotic power against local alternatives.

The finite-sample distributions of the three statistics differ. In the linear regression model with normality, a variant of the Wald test statistic for h linear restrictions on $\boldsymbol{\theta}$ exactly equals the $F(h, N - K)$ statistic (see Section 7.2.1) whereas no analytical results exist for the LR and LM statistics. More generally, in nonlinear models exact small-sample results do not exist.

In some cases an ordering of the values taken by the three test statistics can be obtained. In particular for tests of linear restrictions in the linear regression model under normality, Berndt and Savin (1977) showed that $\text{Wald} \geq \text{LR} \geq \text{LM}$. This result is of little theoretical consequence, as the test least likely to reject under the null will have the smallest actual size but also the smallest power. However, it is of practical consequence for the linear model, as it means when testing at fixed nominal size α that the Wald test will always reject H_0 more often than the LR, which in turn will reject more often than the LM test. The Wald test would be preferred by a researcher determined to reject H_0 . This result is restricted to linear models.

Invariance to Reparameterization

The Wald test is not invariant to algebraically equivalent parameterizations of the null hypothesis (see Section 7.2.9) whereas the LR test is invariant. Some but not all versions of the LM test are invariant. The LM test is generally invariant if the expected Hessian (see Section 5.5.2) is used to estimate \mathbf{A}_0 and not invariant if the Hessian is used. The test LM^* defined later in (7.34) is invariant. The lack of invariance for the Wald test is a major weakness.

Robust Versions

In some cases with misspecified density the quasi-MLE (see Section 5.7) remains consistent. The Wald test is then easily robustified (see Section 7.2). The LM test can be robustified with more difficulty; see (7.38) in Section 7.5.1 for a general result for m-estimators and Section 8.4 for some robust LM test examples. The LR test is no longer chi-square distributed, except in a special case given later in (7.39). Instead, the LR test is a mixture of chi-squares (see Section 8.5.3).

Convenience

Convenience in computation is also a consideration. LR requires estimation of the model twice, once with and once without the restrictions of the null hypothesis. If done by a package, it is easily implemented as one need only read off the printed log-likelihood routinely printed out, subtract, and multiply by 2. Wald requires estimation only under H_a and is best to use when the unrestricted model is easy to estimate. For example, this is the case for restrictions on the parameters of the conditional mean in nonlinear models such as NLS, probit, Tobit, and logit. The LM statistic requires estimation only under H_0 and is best to use when the restricted model is easy to estimate. Examples are tests for autocorrelation and heteroskedasticity, where it is easiest to estimate the null hypothesis model that does not have these complications.

The Wald test is often used for tests of statistical significance whereas the LM test is often used for tests of correct model specification.

7.3.5. Interpretation and Computation of the LM test

Lagrange multiplier tests have the additional advantages of simple interpretation in some leading examples and computation by auxiliary regression.

In this section attention is restricted to the usual cross-section data case of a scalar dependent variable independent over i , so that $\partial \ln L(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} = \sum_i s_i(\boldsymbol{\theta})$, where

$$s_i(\boldsymbol{\theta}) = \frac{\partial \ln f(y_i | \mathbf{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \quad (7.31)$$

is the contribution of the i th observation to the score vector of the unrestricted model. From (7.25) the LM test is a test of the closeness to zero of $\sum_i s_i(\tilde{\boldsymbol{\theta}}_r)$.

Simple Interpretation of the LM Test

Suppose that the density is such that $\mathbf{s}(\boldsymbol{\theta})$ factorizes as

$$\mathbf{s}(\boldsymbol{\theta}) = \mathbf{g}(\mathbf{x}, \boldsymbol{\theta}) r(y, \mathbf{x}, \boldsymbol{\theta}) \quad (7.32)$$

for some $q \times 1$ vector function $\mathbf{g}(\cdot)$ and scalar function $r(y, \mathbf{x}, \boldsymbol{\theta})$, the latter of which may be interpreted as a generalized residual because y appears in $r(\cdot)$ but not $\mathbf{g}(\cdot)$. For example, for Poisson regression $\partial \ln f / \partial \boldsymbol{\theta} = \mathbf{x}(y - \exp(\mathbf{x}'\boldsymbol{\beta}))$.

Given (7.32) and independence over i , $\partial \ln L / \partial \theta |_{\tilde{\theta}_r} = \sum_i \tilde{g}_i \tilde{r}_i$, where $\tilde{g}_i = g(\mathbf{x}_i, \tilde{\theta}_r)$ and $\tilde{r}_i = r(y_i, \mathbf{x}_i, \tilde{\theta}_r)$. The LM test can therefore be simply interpreted as a score test of the correlation between \tilde{g}_i and the residual \tilde{r}_i . This interpretation was given in Section 7.3.2 for the LM test with Poisson regression, where $\tilde{g}_i = \mathbf{x}_i$ and $\tilde{r}_i = y_i - \exp(\mathbf{x}'_i \tilde{\beta}_1)$.

The partition (7.32) will arise whenever $f(y)$ is based on a one-parameter density. In particular, many common likelihood models are based on one-parameter LEF densities, with parameter μ then modeled as a function of \mathbf{x} and β . In the LEF case $r(y, \mathbf{x}, \theta) = (y - E[y|\mathbf{x}])$ (see Section 5.7.3), so the generalized residual $r(\cdot)$ in (7.32) is then the usual residual.

More generally a partition similar to (7.32) will also arise when $f(y)$ is based on a two-parameter density, the information matrix is block diagonal in the two parameters, and the two parameters in turn depend on regressors and parameter vectors β and α that are distinct. Then LM tests on β are tests of correlation of $\tilde{g}_{\beta i}$ and $\tilde{r}_{\beta i}$, where $s(\beta) = g_\beta(\mathbf{x}, \theta) r_\beta(y, \mathbf{x}, \theta)$, with similar interpretation for LM tests on α .

A leading example is linear regression under normality with two parameters μ and σ^2 modeled as $\mu = \mathbf{x}'\beta$ and $\sigma^2 = \alpha$ or $\sigma^2 = \sigma^2(\mathbf{z}, \alpha)$. For exclusion restrictions in linear regression under normality, $s_i(\beta) = \mathbf{x}_i(y_i - \mathbf{x}'_i \beta)$ and the LM test is one of correlation between regressors \mathbf{x}_i and the restricted model residual $\tilde{u}_i = y_i - \mathbf{x}'_i \tilde{\beta}_1$. For tests of heteroskedasticity with $\sigma_i^2 = \exp(\alpha_1 + \mathbf{z}'_i \alpha_2)$, $s_i(\alpha) = \frac{1}{2} \mathbf{z}_i((y_i - \mathbf{x}'_i \beta)^2 / \sigma_i^2 - 1)$, and the LM test is one of correlation between \mathbf{z}_i and the squared residual $\tilde{u}_i^2 = (y_i - \mathbf{x}'_i \tilde{\beta})^2$, since σ_i^2 is constant under the null hypothesis that $\alpha_2 = \mathbf{0}$.

Outer Product of the Gradient Versions of the LM Test

Now return to the general $s_i(\theta)$ defined in (7.31). We show in the following that an asymptotically equivalent version of the LM test statistic (7.25) can be obtained by running the **auxiliary regression** or artificial regression

$$1 = \tilde{s}_i' \gamma + v_i, \quad (7.33)$$

where $\tilde{s}_i = s_i(\tilde{\theta}_r)$, and computing

$$\text{LM}^* = N R_u^2, \quad (7.34)$$

where R_u^2 is the uncentered R^2 defined after (7.36). LM^* is asymptotically $\chi^2(h)$ under H_0 . Equivalently, LM^* equals ESS_u , the uncentered explained sum of squares (the sum of squares of the fitted values), or equals $N - \text{RSS}$, where RSS is the residual sum of squares, from regression (7.33).

This result can be easy to implement as in many applications it can be quite simple to analytically obtain $s_i(\theta)$, generate data for the q components $\tilde{s}_{1i}, \dots, \tilde{s}_{qi}$, and regress 1 on $\tilde{s}_{1i}, \dots, \tilde{s}_{qi}$. Note that here $f(y_i | \mathbf{x}_i, \theta)$ in (7.31) is the density of the unrestricted model.

For the exclusion restrictions in the Poisson model example in Section 7.3.2, $s_i(\beta) = (y_i - \exp(\mathbf{x}'_i \beta)) \mathbf{x}_i$ and $\mathbf{x}'_i \tilde{\beta}_r = \mathbf{x}'_{1i} \tilde{\beta}_{1r}$. It follows that LM^* can be computed

as NR_u^2 from regressing 1 on $(y_i - \exp(\mathbf{x}'_{1i} \tilde{\beta}_{1r}))\mathbf{x}_i$, where \mathbf{x}_i contains both \mathbf{x}_{1i} and \mathbf{x}_{2i} , and $\tilde{\beta}_{1r}$ is obtained from Poisson regression of y_i on \mathbf{x}_{1i} alone.

Equations (7.33) and (7.34) require only independence over i . Other auxiliary regressions are possible if further structure is assumed. In particular, specialize to cases where $\mathbf{s}(\boldsymbol{\theta})$ factorizes as in (7.32), and define $r(y, \mathbf{x}, \boldsymbol{\theta})$ so that $V[r(y, \mathbf{x}, \boldsymbol{\theta})] = 1$. Then an alternative asymptotically equivalent version of the LM test is NR_u^2 from regression of \tilde{r}_i on $\tilde{\mathbf{g}}_i$. This includes LM tests for linear regression under normality, such as the Breusch–Pagan LM test for heteroskedasticity.

These alternative versions of the LM test are called **outer-product-of-the-gradient** versions of the LM test, as they replace $-\mathbf{A}_0$ in (7.22) by an outer-product-of-the-gradient (OPG) estimate or BHHH estimate of \mathbf{B}_0 . Although they are easily computed, OPG variants of LM tests can have poor small-sample properties with large size distortions. This has discouraged use of the OPG form of the LM test. These small-sample problems can be greatly reduced by bootstrapping (see Section 11.6.3). Davidson and MacKinnon (1984) propose double-length auxiliary regressions that also perform better in finite samples.

Derivation of the OPG Version

To derive LM^* , first note that in (7.25), $\partial \ln L(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}|_{\tilde{\boldsymbol{\theta}}_r} = \sum \tilde{\mathbf{s}}_i$. Second, by the information matrix equality $\mathbf{A}_0 = -\mathbf{B}_0$ and, from Section 5.5.2, \mathbf{B}_0 can be consistently estimated under H_0 by the OPG estimate or BHHH estimate $N^{-1} \sum \tilde{\mathbf{s}}_i \tilde{\mathbf{s}}_i'$. Combining, these results gives an asymptotically equivalent version of the LM test statistic (7.25):

$$\text{LM}^* = \left(\sum_{i=1}^N \tilde{\mathbf{s}}_i \right) \left[\sum_{i=1}^N \tilde{\mathbf{s}}_i \tilde{\mathbf{s}}_i' \right]^{-1} \left(\sum_{i=1}^N \tilde{\mathbf{s}}_i \right). \quad (7.35)$$

This statistic can be computed from an auxiliary regression of 1 on $\tilde{\mathbf{s}}_i$ as follows. Define \mathbf{S} to be the $N \times q$ matrix with i th row $\tilde{\mathbf{s}}_i'$, and define \mathbf{I} to be the $N \times 1$ vector of ones. Then

$$\text{LM}^* = \mathbf{I}' \mathbf{S} [\mathbf{S}' \mathbf{S}]^{-1} \mathbf{S}' \mathbf{I} = \text{ESS}_u = NR_u^2. \quad (7.36)$$

In general for regression of \mathbf{y} on \mathbf{X} the **uncentered explained sums of squares** (ESS_u) is $\mathbf{y}' \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y}$, which is exactly of the form (7.36), whereas the **uncentered R^2** is $R_u^2 = \mathbf{y}' \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y} / \mathbf{y}' \mathbf{y}$, which here is (7.36) divided by $\mathbf{I}' \mathbf{I} = N$. The term *uncentered* is used because in R_u^2 division is by the sum of squared deviations of \mathbf{y} around zero rather than around the sample mean.

7.4. Example: Likelihood-Based Hypothesis Tests

The various test procedures – Wald, LR, and LM – are illustrated using generated data from the dgp $\mathbf{y} | \mathbf{x}$ Poisson distributed with mean $\exp(\beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4)$, where $\beta_1 = 0$ and $\beta_2 = \beta_3 = \beta_4 = 0.1$ and the three regressors are iid draws from $\mathcal{N}[0, 1]$.

HYPOTHESIS TESTS

Table 7.1. Test Statistics for Poisson Regression Example^a

Null Hypothesis	Test Statistic				ln L	Result at level 0.05
	Wald	LR	LM	LM*		
$H_{10} : \beta_3 = 0$	5.904 (0.015)	5.754 (0.016)	5.916 (0.015)	6.218 (0.013)	-241.648	Reject
$H_{20} : \beta_3 = 0, \beta_4 = 0$	8.570 (0.014)	8.302 (0.016)	8.575 (0.014)	9.186 (0.010)	-242.922	Reject
$H_{30} : \beta_3 - \beta_4 = 0$	0.293 (0.588)	0.293 (0.589)	0.293 (0.588)	0.315 (0.575)	-238.918	Do not reject
$H_{40} : \beta_3/\beta_4 - 1 = 0$	0.158 (0.691)	0.293 (0.589)	0.293 (0.588)	0.315 (0.575)	-238.918	Do not reject

^a The dgp for y is the Poisson distribution with parameter $\exp(0.0 + 0.1x_2 + 0.1x_3 + 0.1x_4)$ and sample size $N = 200$. Test statistics are given with associated p -values in parentheses. Tests of the second hypothesis are $\chi^2(2)$ and the other tests are $\chi^2(1)$ distributed. Log-likelihoods for restricted ML estimation are also given; the log-likelihood in the unrestricted model is -238.772.

Poisson regression of y on an intercept, x_2 , x_3 , and x_4 for a generated sample of size 200 yielded unrestricted MLE

$$\widehat{E}[y|x] = \exp(-0.165 - 0.028x_2 + 0.163x_3 + 0.103x_4),$$

where associated t -statistics are given in parentheses and the unrestricted log-likelihood is -238.772.

The analysis tests four different hypotheses, detailed in the first column of Table 7.1. The estimator is nonlinear, whereas the hypotheses are examples of, respectively, single exclusion restriction, multiple exclusion restriction, linear restrictions, and nonlinear restrictions. The remainder of the table gives four asymptotically equivalent test statistics of these hypotheses and their associated p -values. For this sample all tests reject the first two hypotheses and do not reject the remaining two, at significance level 0.05.

The Wald test statistic is computed using (7.23). This requires estimation of the unrestricted model, given previously, to obtain the variance matrix estimate of the unrestricted MLE. Wald tests of different hypotheses then require computation of different \mathbf{h} and \mathbf{R} and simplify in some cases. The Wald chi-square test of the single exclusion restriction is just the square of the usual t -test, with $2.43^2 \simeq 5.90$. The Wald test statistic of the joint exclusion restrictions is detailed in Section 7.2.5. Here x_3 is statistically significant and x_4 is statistically insignificant, whereas jointly x_3 and x_4 are statistically significant at level 0.05. The Wald test for the third hypothesis is given in (7.19) and leads to nonrejection. The third and fourth hypotheses are equivalent, since $\beta_3/\beta_4 - 1 = 0$ implies $\beta_3 = \beta_4$, but the Wald test statistic for the fourth hypothesis, given in (7.13), differs from (7.19). The statistic (7.13) was calculated using matrix operations, as most packages will at best calculate Wald tests of linear hypotheses.

The LR test statistic is especially easy to compute, using (7.21), given estimation of the restricted model. For the first three hypotheses the restricted model is

estimated by Poisson regression of y on, respectively, regressors $(1, x_2, x_4)$, $(1, x_2)$, and $(1, x_2, x_3 + x_4)$, where the third regression uses $\beta_3 x_3 + \beta_4 x_4 = \beta_3(x_3 + x_4)$ if $\beta_3 = \beta_4$. As an example of the LR test, for the second hypothesis $\text{LR} = -2[-238.772 - (-242.922)] = 8.30$. The fourth restricted model in theory requires ML estimation subject to nonlinear constraints on the parameters, which few packages do. However, constrained ML estimation is invariant to the way the restrictions are expressed, so here the same estimates are obtained as for the third restricted model, leading to the same LR test statistic.

The LM test statistic is computed using (7.25), which for the Poisson model specializes to (7.27). This statistic is computed using matrix commands, with different restrictions leading to the different restricted MLE estimates $\tilde{\beta}$. As for the LR test, the LM test is invariant to transformations, so the LM tests of the third and fourth hypotheses are equivalent.

An asymptotically equivalent version of the LM test statistic is the statistic LM^* given in (7.35). This can be computed as the explained sum of squares from the auxiliary regression (7.33). For the Poisson model $s_{ji} = \partial \ln f(y_i)/\partial \beta_j = (y_i - \exp(\mathbf{x}_i'\beta))x_{ji}$, with evaluation at the appropriate restricted MLE for the hypothesis under consideration. The statistic LM^* is simpler to compute than LM, though like LM it requires restricted ML estimates.

In this example with generated data the various test statistics are very similar. This is not always the case. In particular, the test statistic LM^* can have poorer finite-sample size properties than LM, even if the dgp is known. Also, in applications with real data the dgp is unlikely to be perfectly specified, leading to divergence of the various test statistics even in infinitely large samples.

7.5. Tests in Non-ML Settings

The Wald test is the standard test to use in non-ML settings. From Section 7.2 it is a general testing procedure that can always be implemented, using an appropriate sandwich estimator of the variance matrix of the parameter estimates. The only limitation is that in some applications unrestricted estimation may be much more difficult to perform than restricted estimation.

The LM or score test, based on departures from zero of the gradient vector of the unrestricted model evaluated at the restricted estimates, can also be generalized to non-ML estimators. The form of the LM test, however, is usually considerably more complicated than in the ML case. Moreover, the simplest forms of the LM test statistic based on auxiliary regressions are usually not robust to distributional misspecification.

The LR test is based on the difference between the maximized values of the objective function with and without restrictions imposed. This usually does not generalize to objective functions other than the likelihood function, as this difference is usually not chi-square distributed.

For completeness we provide a condensed presentation of extension of the ML tests to m-estimators and to efficient GMM estimators. As already noted, in most applications use of the simpler Wald test is sufficient.

7.5.1. Tests Based on m-Estimators

Tests for m-estimators are straightforward extensions of those for ML estimators, except that it is no longer possible to use the information matrix equality to simplify the test statistics and the LR test generalizes in only very special cases. The resulting test statistics are asymptotically $\chi^2(h)$ distributed under $H_0 : \mathbf{h}(\boldsymbol{\theta}) = \mathbf{0}$ and have the same noncentral chi-square distribution under local alternatives.

Consider m-estimators that maximize $Q_N(\boldsymbol{\theta}) = N^{-1} \sum_i q_i(\boldsymbol{\theta})$ with first-order conditions $N^{-1} \sum_i \mathbf{s}_i(\boldsymbol{\theta}) = \mathbf{0}$. Define the $q \times q$ matrices $\mathbf{A}(\boldsymbol{\theta}) = N^{-1} \sum_i \partial \mathbf{s}_i(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}'$ and $\mathbf{B}(\boldsymbol{\theta}) = N^{-1} \sum_i \mathbf{s}_i(\boldsymbol{\theta}) \mathbf{s}_i(\boldsymbol{\theta})'$ and the $h \times q$ matrix $\mathbf{R}(\boldsymbol{\theta}) = \partial \ln \mathbf{h}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}'$. Let $\hat{\boldsymbol{\theta}}_u$ and $\tilde{\boldsymbol{\theta}}_r$ denote unrestricted and restricted estimators, respectively, and let $\hat{\mathbf{A}} = \mathbf{A}(\hat{\boldsymbol{\theta}}_u)$ and $\tilde{\mathbf{A}} = \mathbf{A}(\tilde{\boldsymbol{\theta}}_r)$ with similar notation for \mathbf{B} and \mathbf{R} . Finally, let $\hat{\mathbf{h}} = \mathbf{h}(\hat{\boldsymbol{\theta}}_u)$ and $\tilde{\mathbf{s}}_i = \mathbf{s}_i(\tilde{\boldsymbol{\theta}}_r)$.

The Wald test statistic is based on closeness of $\hat{\mathbf{h}}$ to zero. Here

$$W = \hat{\mathbf{h}}' \left[\hat{\mathbf{R}} N^{-1} \hat{\mathbf{A}}^{-1} \hat{\mathbf{B}} \hat{\mathbf{A}}^{-1} \hat{\mathbf{R}}' \right]^{-1} \hat{\mathbf{h}}, \quad (7.37)$$

since from Section 5.5.1 the robust variance matrix estimate for $\hat{\boldsymbol{\theta}}_u$ is $N^{-1} \hat{\mathbf{A}}^{-1} \hat{\mathbf{B}} \hat{\mathbf{A}}^{-1}$. Packages with the option of robust standard errors use this more general form to compute Wald tests of statistical significance.

Let $\mathbf{g}(\boldsymbol{\theta}) = \partial \ln Q_N(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ denote the gradient vector, and let $\tilde{\mathbf{g}} = \mathbf{g}(\tilde{\boldsymbol{\theta}}_r) = \sum_i \tilde{\mathbf{s}}_i$. The LM test statistic is based on the closeness of $\tilde{\mathbf{g}}$ to $\mathbf{0}$ and is given by

$$LM = N \tilde{\mathbf{g}}' \left[\tilde{\mathbf{A}}^{-1} \tilde{\mathbf{R}}' \left(\tilde{\mathbf{R}} \tilde{\mathbf{A}}^{-1} \tilde{\mathbf{B}} \tilde{\mathbf{A}}'^{-1} \tilde{\mathbf{R}} \right)^{-1} \tilde{\mathbf{R}} \tilde{\mathbf{A}}^{-1} \right]^{-1} \tilde{\mathbf{g}}, \quad (7.38)$$

a result obtained by forming a chi-square test statistic based on (7.29), where $N \tilde{\mathbf{g}}$ replaces $|\partial \ln L / \partial \boldsymbol{\theta}|_{\tilde{\boldsymbol{\theta}}_r}$. This test is clearly not as simple to implement as a robust Wald test. Some examples of computation of the robust form of LM tests are given in Section 8.4. The standard implementations of LM tests in computer packages are often not robust versions of the LM test.

The LR test does not generalize easily. It does generalize to m-estimators if $\mathbf{B}_0 = -\alpha \mathbf{A}_0$ for some scalar α , a weaker version of the IM equality. In such special cases the quasi-likelihood ratio (QLR) test statistic is

$$QLR = -2N [Q_N(\tilde{\boldsymbol{\theta}}_r) - Q_N(\hat{\boldsymbol{\theta}}_u)] / \hat{\alpha}_u, \quad (7.39)$$

where $\hat{\alpha}_u$ is a consistent estimate of α obtained from unrestricted estimation (see Wooldridge, 2002, p. 370). The condition $\mathbf{B}_0 = -\alpha \mathbf{A}_0$ holds for generalized linear models (see Section 5.7.4). Then the statistic QLR is equivalent to the difference of deviances for the restricted and unrestricted models, a generalization of the F -test based on the difference between restricted and unrestricted sum of squared residuals for OLS and NLS estimation with homoskedastic errors. For general quasi-ML estimation, with $\mathbf{B}_0 \neq -\alpha \mathbf{A}_0$, the LR test statistic can be distributed as a weighted sum of chi-squares (see Section 8.5.3).

7.5.2. Tests Based on Efficient GMM Estimators

For GMM the various test statistics are simplest for efficient GMM, meaning GMM estimation using the optimal weighting matrix. This poses no great practical restriction as the optimal weighting matrix can always be estimated, as detailed in Section 6.3.5.

Consider GMM estimation based on the moment condition $E[\mathbf{m}_i(\boldsymbol{\theta})] = \mathbf{0}$. (Note the change in notation from Chapter 6: $\mathbf{h}(\boldsymbol{\theta})$ is being used in the current chapter to denote the restrictions under H_0 .) Using the notation introduced in Section 6.3.5, the efficient unrestricted GMM estimator $\hat{\boldsymbol{\theta}}_u$ minimizes $Q_N(\boldsymbol{\theta}) = \mathbf{g}_N(\boldsymbol{\theta})' \mathbf{S}_N^{-1} \mathbf{g}_N(\boldsymbol{\theta})$, where $\mathbf{g}_N(\boldsymbol{\theta}) = N^{-1} \sum_i \mathbf{m}_i(\boldsymbol{\theta})$ and \mathbf{S}_N is consistent for $\mathbf{S}_0 = V[\mathbf{g}_N(\boldsymbol{\theta})]$. The restricted GMM estimator $\tilde{\boldsymbol{\theta}}_r$ is assumed to minimize $Q_N(\boldsymbol{\theta})$ with the same weighting matrix \mathbf{S}_N^{-1} , subject to the restriction $\mathbf{h}(\boldsymbol{\theta}) = \mathbf{0}$.

The three following test statistics, summarized by Newey and West (1987a) are asymptotically $\chi^2(h)$ distributed under $H_0 : \mathbf{h}(\boldsymbol{\theta}) = \mathbf{0}$ and have the same noncentral chi-square distribution under local alternatives.

The Wald test statistic as usual is based on closeness of $\hat{\mathbf{h}}$ to zero. This yields

$$W = \hat{\mathbf{h}}' \left[\hat{\mathbf{R}} N^{-1} (\hat{\mathbf{G}}' \mathbf{S}^{-1} \hat{\mathbf{G}})^{-1} \hat{\mathbf{R}} \right]^{-1} \hat{\mathbf{h}}, \quad (7.40)$$

since the variance of the efficient GMM estimator is $N^{-1} (\hat{\mathbf{G}}' \mathbf{S}^{-1} \hat{\mathbf{G}})^{-1}$ from Section 6.3.5, where $\mathbf{G}_N(\boldsymbol{\theta}) = \partial \mathbf{g}_N(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}'$ and the carat denotes evaluation at $\hat{\boldsymbol{\theta}}_u$.

The first-order conditions of efficient GMM are $\hat{\mathbf{G}}' \mathbf{S}^{-1} \tilde{\mathbf{g}} = \mathbf{0}$. The LM statistic tests whether this gradient vector is close to zero when instead evaluated at $\tilde{\boldsymbol{\theta}}_r$, leading to

$$LM = N \tilde{\mathbf{g}}' \mathbf{S}^{-1} \tilde{\mathbf{G}} (\tilde{\mathbf{G}}' \mathbf{S}^{-1} \tilde{\mathbf{G}})^{-1} \tilde{\mathbf{G}}' \mathbf{S}^{-1} \tilde{\mathbf{g}}, \quad (7.41)$$

where the tilda denotes evaluation at $\tilde{\boldsymbol{\theta}}_r$ and we use the Section 6.3.3 assumption that $\sqrt{N} \mathbf{g}_N(\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}[0, \mathbf{S}_0]$, so $\sqrt{N} \mathbf{G}' \mathbf{S}^{-1} \mathbf{g} \xrightarrow{d} \mathcal{N}[\mathbf{0}, \text{plim } N^{-1} \mathbf{G}' \mathbf{S}^{-1} \mathbf{G}]$.

For the efficient GMM estimator the difference in maximized values of the objective function can also be compared, leading to the difference test statistic

$$D = N [Q_N(\tilde{\boldsymbol{\theta}}_r) - Q_N(\hat{\boldsymbol{\theta}}_u)]. \quad (7.42)$$

Like W and LM , the statistic D is asymptotically $\chi^2(h)$ distributed under $H_0 : \mathbf{h}(\boldsymbol{\theta}) = \mathbf{0}$.

Even in the likelihood case, this last statistic differs from the LR statistic because it uses a different objective function. The MLE minimizes $Q_N(\boldsymbol{\theta}) = -N^{-1} \sum_i \ln f(y_i | \boldsymbol{\theta})$. From Section 6.3.7, the asymptotically equivalent efficient GMM estimator instead minimizes the quadratic form $Q_N(\boldsymbol{\theta}) = N^{-1} (\sum_i s_i(\boldsymbol{\theta}))' (\sum_i s_i(\boldsymbol{\theta}))$, where $s_i(\boldsymbol{\theta}) = \partial \ln f(y_i | \boldsymbol{\theta}) / \partial \boldsymbol{\theta}$. The statistic D can be used in general, provided the GMM estimator used is the efficient GMM estimator, whereas the LR test can only be generalized for some special cases of m-estimators mentioned after (7.39).

For MM estimators, that is, in the just-identified GMM model, $D = LM = N Q_N(\tilde{\boldsymbol{\theta}}_r)$, so the LM and difference tests are equivalent. For D this simplification occurs because $\mathbf{g}_N(\hat{\boldsymbol{\theta}}_u) = \mathbf{0}$ and so $Q_N(\hat{\boldsymbol{\theta}}_u) = 0$. For LM simplification occurs in (7.41) as then $\tilde{\mathbf{G}}_N$ is invertible.

7.6. Power and Size of Tests

The remaining sections of this chapter study two limitations in using the usual computer output to test hypotheses.

First, a test can have little ability to discriminate between the null and alternative hypotheses. Then the test has low power, meaning there is a low probability of rejecting the null hypothesis when it is false. Standard computer output does not calculate test power, but it can be evaluated using asymptotic methods (see this section) or finite-sample Monte Carlo methods (see Section 7.7). If a major contribution of an empirical paper is the rejection or nonrejection of a particular hypothesis, there is no reason for the paper not to additionally present the power of the test against some meaningful alternative hypothesis.

Second, the true size of the test may differ substantially from the nominal size of the test obtained from asymptotic theory. The rule of thumb that sample size $N > 30$ is sufficient for asymptotic theory to provide a good approximation for inference on a single variable does not extend to models with regressors. Poor approximation is most likely in the tails of the approximating distribution, but the tails are used to obtain critical values of tests at common significance levels such as 5%. In practice the critical value for a test statistic obtained from large-sample approximation is often smaller than the correct critical value based on the unknown true distribution. Small-sample refinements are attempts to get closer to the exact critical value. For linear regression under normality exact critical values can be obtained, using the t rather than z and the F rather than χ^2 distribution, but similar results are not exact for nonlinear regression. Instead, small-sample refinements may be obtained through Monte Carlo methods (see Section 7.7) or by use of the bootstrap (see Section 7.8 and Chapter 11).

With modern computers it is relatively easy to correct the size and investigate the power of tests used in an applied study. We present this neglected topic in some detail.

7.6.1. Test Size and Power

Hypothesis tests lead to either rejection or nonrejection of the null hypothesis. Correct decisions are made if H_0 is rejected when H_0 is false or if H_0 is not rejected when H_0 is true.

There are also two possible incorrect decisions: (1) rejecting H_0 when H_0 is true, called a **type I error**, and (2) nonrejection of H_0 when H_0 is false, called a **type II error**. Ideally the probabilities of both errors will be low, but in practice decreasing the probability of one type of error comes at the expense of increasing the probability of the other. The classical hypothesis testing solution is to fix the probability of a type I error at a particular level, usually 0.05, while leaving the probability of a type II error unspecified.

Define the **size of a test or significance level**

$$\begin{aligned}\alpha &= \Pr[\text{type I error}] \\ &= \Pr[\text{reject } H_0 | H_0 \text{ true}],\end{aligned}\tag{7.43}$$

with common choices of α being 0.01, 0.05, or 0.10. A hypothesis is rejected if the test statistic falls into a rejection region defined so that the test significance level equals the specified value of α . A closely related equivalent method computes the ***p-value*** of a test, the marginal significance level at which the null hypothesis is just rejected, and rejects H_0 if the *p*-value is less than the specified value of α . Both methods require only knowledge of the distribution of the test statistic under the null hypothesis, presented in Section 7.2 for the Wald test statistic.

Consideration should also be given to the probability of a type II error. The **power of a test** is defined to be

$$\begin{aligned}\text{Power} &= \Pr[\text{reject } H_0 | H_a \text{ true}] \\ &= 1 - \Pr[\text{accept } H_0 | H_a \text{ true}] \\ &= 1 - \Pr[\text{Type II error}].\end{aligned}\tag{7.44}$$

Ideally, test power is close to one since then the probability of a type II error is close to zero. Determining the power requires knowledge of the distribution of the test statistic under H_a .

Analysis of test power is typically ignored in empirical work, except that test procedures are usually chosen to be ones that are known theoretically to have power that, for given level α , is high relative to other alternative test statistics. Ideally, the **uniformly most powerful (UMP)** test is used. This is the test that has the greatest power, for given level α , for all alternative hypotheses. UMP tests do exist when testing a simple null hypothesis against a simple alternative hypothesis. Then the Neyman–Pearson lemma gives the result that the UMP test is a function of the likelihood ratio. For more general testing situations involving composite hypotheses there is usually no UMP test, and further restrictions are placed such as UMP one-sided tests. In practice, power considerations are left to theoretical econometricians who use theory and simulations applied to various testing procedures to suggest which testing procedures are the most powerful.

It is nonetheless possible to determine test power in any given application. In the following we detail how to compute the asymptotic power of the Wald test, which equals that of the LR and LM tests in the fully parametric case.

7.6.2. Local Alternative Hypotheses

Since power is the probability of rejecting H_0 when H_a is true, the computation of power requires obtaining the distribution of the test statistic under the alternative hypothesis. For a Wald chi-square test at significance level α the power equals $\Pr[W > \chi^2_\alpha(h) | H_a]$. Calculation of this probability requires specification of a particular alternative hypothesis, because $H_a : \mathbf{h}(\boldsymbol{\theta}) \neq \mathbf{0}$ is very broad.

The obvious choice is the **fixed alternative** $\mathbf{h}(\boldsymbol{\theta}) = \boldsymbol{\delta}$, where $\boldsymbol{\delta}$ is an $h \times 1$ finite vector of nonzero constants. The quantity $\boldsymbol{\delta}$ is sometimes referred to as the hypothesis error, and larger hypothesis errors lead to greater power. For a fixed alternative the Wald test statistic asymptotically has power one as it rejects the null hypothesis all the time. To see this note that if $\mathbf{h}(\boldsymbol{\theta}) = \boldsymbol{\delta}$ then the Wald test statistic becomes

infinite, since

$$\begin{aligned} W &= \widehat{\mathbf{h}}(\widehat{\mathbf{R}}N^{-1}\widehat{\mathbf{C}}\widehat{\mathbf{R}}')^{-1}\widehat{\mathbf{h}} \\ &\xrightarrow{P} \delta' (\mathbf{R}_0 N^{-1} \mathbf{C}_0 \mathbf{R}_0')^{-1} \delta, \end{aligned}$$

using $\widehat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}_0$, so $\widehat{\mathbf{h}} = \mathbf{h}(\widehat{\boldsymbol{\theta}}_u) \xrightarrow{P} \mathbf{h}(\boldsymbol{\theta}) = \boldsymbol{\delta}$, and $\widehat{\mathbf{C}} \xrightarrow{P} \mathbf{C}_0$. It follows that $W \xrightarrow{P} \infty$ since all the terms except N are finite and nonzero. This infinite value leads to H_0 being always rejected, as it should be, and hence having perfect power of one.

The Wald test statistic is therefore a **consistent test statistic**, that is, one whose power goes to one as $N \rightarrow \infty$. Many test statistics are consistent, just as many estimators are consistent. More stringent criteria are needed to discriminate among the test statistics, just as relative efficiency is used to choose among estimators.

For estimators that are root- N consistent, we consider a **sequence of local alternatives**

$$H_a : \mathbf{h}(\boldsymbol{\theta}) = \boldsymbol{\delta}/\sqrt{N}, \quad (7.45)$$

where $\boldsymbol{\delta}$ is a vector of fixed constants with $\boldsymbol{\delta} \neq \mathbf{0}$. This sequence of alternative hypotheses, called **Pitman drift**, gets closer to the null hypothesis value of zero as the sample size gets larger, at the same rate \sqrt{N} as used to scale up $\widehat{\boldsymbol{\theta}}$ to get a nondegenerate distribution for the consistent estimator. The alternative hypothesis value of $\mathbf{h}(\boldsymbol{\theta})$ therefore moves toward zero at a rate that negates any improved efficiency with increased sample size. For a much more detailed account of local alternatives and related literatures see McManus (1991).

7.6.3. Asymptotic Power of the Wald Test

Under the sequence of local alternatives (7.45) the Wald test statistic has a nondegenerate distribution, the **noncentral chi-square distribution**. This permits determination of the power of the Wald test.

Specifically, as is shown in Section 7.7.4, under H_a the Wald statistic W defined in (7.6) is asymptotically $\chi^2(h; \lambda)$ distributed, where $\chi^2(h; \lambda)$ denotes the **noncentral chi-square distribution with noncentrality parameter**

$$\lambda = \frac{1}{2} \boldsymbol{\delta}' (\mathbf{R}_0 \mathbf{C}_0 \mathbf{R}_0')^{-1} \boldsymbol{\delta}, \quad (7.46)$$

and \mathbf{R}_0 and \mathbf{C}_0 are defined in (7.4) and (7.5). The **power of the Wald test**, the probability of rejecting H_0 given the local alternative H_a is true, is therefore

$$\text{Power} = \Pr[W > \chi_{\alpha}^2(h) | W \sim \chi_{\alpha}^2(h; \lambda)]. \quad (7.47)$$

Figure 7.1 plots power against λ for tests of a scalar hypothesis ($h = 1$) at the commonly used sizes or significance levels of 10%, 5%, and 1%. For λ close to zero the power equals the size, and for large λ the power goes to one.

These features hold also for $h > 1$. In particular power is monotonically increasing in the noncentrality parameter λ defined in (7.46). Several general results follow.

First, power is increasing in the distance between the null and alternative hypotheses, as then $\boldsymbol{\delta}$ and hence λ increase.

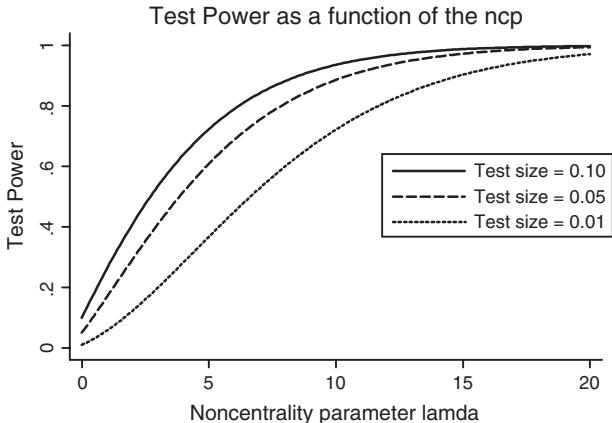


Figure 7.1: Power of Wald chi-square test with one degree of freedom for three different test sizes as the noncentrality parameter ranges from 0 to 20.

Second, for given alternative δ power increases with efficiency of the estimator $\widehat{\theta}$, as then \mathbf{C}_0 is smaller and hence λ is larger.

Third, as the size of the test increases power increases and the probability of a type II error decreases.

Fourth, if several different test statistics are all $\chi^2(h)$ under the null hypothesis and noncentral- $\chi^2(h)$ under the alternative, the preferred test statistic is that with the highest noncentrality parameter λ since then power is the highest. Furthermore, two tests that have the same noncentrality parameter are asymptotically equivalent under local alternatives.

Finally, in actual applications one can calculate the power as a function of δ . Specifically, for a specified alternative δ , an estimated noncentrality parameter $\widehat{\lambda}$ can be computed using (7.46) using parameter estimate $\widehat{\theta}$ with associated estimates $\widehat{\mathbf{R}}$ and $\widehat{\mathbf{C}}$. Such power calculations are illustrated in Section 7.6.5.

7.6.4. Derivation of Asymptotic Power

To obtain the distribution of \mathbf{W} under H_a , begin with the Taylor series expansion result (7.9). This simplifies to

$$\sqrt{N}\mathbf{h}(\widehat{\theta}) \xrightarrow{d} \mathcal{N}[\delta, \mathbf{R}_0 \mathbf{C}_0 \mathbf{R}'_0], \quad (7.48)$$

under H_a , since then $\sqrt{N}\mathbf{h}(\theta) = \delta$. Thus a quadratic form centered at δ would be chi-square distributed under H_a .

The Wald test statistic \mathbf{W} defined in (7.6) instead forms a quadratic form centered at $\mathbf{0}$ and is no longer chi-squared distributed under H_a . In general if $\mathbf{z} \sim \mathcal{N}[\boldsymbol{\mu}, \boldsymbol{\Omega}]$, where $\text{rank}(\boldsymbol{\Omega}) = h$, then $\mathbf{z}'\boldsymbol{\Omega}^{-1}\mathbf{z} \sim \chi^2(h; \lambda)$, where $\chi^2(h; \lambda)$ denotes the noncentral chi-square distribution with noncentrality parameter $\lambda = \frac{1}{2}\boldsymbol{\mu}'\boldsymbol{\Omega}^{-1}\boldsymbol{\mu}$. Applying this result to (7.48) yields

$$N\mathbf{h}(\widehat{\theta})'(\mathbf{R}_0 \mathbf{C}_0 \mathbf{R}'_0)^{-1}\mathbf{h}(\widehat{\theta}) \xrightarrow{d} \chi^2(h; \lambda), \quad (7.49)$$

under H_a , where λ is defined in (7.49).

7.6.5. Calculation of Asymptotic Power

To shed light on how power changes with δ , consider tests of coefficient significance in the scalar case. Then the noncentrality parameter defined in (7.46) is

$$\lambda = \frac{\delta^2}{2c} \simeq \frac{\left(\delta/\sqrt{N}\right)^2}{2(\text{se}[\hat{\theta}])^2}, \quad (7.50)$$

where the approximation arises because of estimation of c , the limit variance of $\sqrt{N}(\hat{\theta} - \theta)$, by $N(\text{se}[\hat{\theta}])^2$, where $\text{se}[\hat{\theta}]$ is the standard error of $\hat{\theta}$.

Consider a Wald chi-square test of $H_0 : \theta = 0$ against the alternative hypothesis that θ is within a standard errors of zero, that is, against

$$H_a : \theta = a \times \text{se}[\hat{\theta}],$$

where $\text{se}[\hat{\theta}]$ is treated here as a constant. Then δ/\sqrt{N} in (7.45) equals $a \times \text{se}[\hat{\theta}]$, so that (7.50) simplifies to $\lambda = a^2/2$. Thus the Wald test is asymptotically $\chi^2_0(1; \lambda)$ under H_a where $\lambda = a^2/2$.

From Figure 7.1 it is clear for the common case of significance level tests at 5% that if $a = 2$ the power is well below 0.5, if $a = 4$ the power is around 0.5, and if $a = 6$ the power is still below 0.9. A borderline test of statistical significance can therefore have low power against alternatives that are many standard errors from zero. Intuitively, if $\hat{\theta} = 2\text{se}[\hat{\theta}]$ then a test of $\theta = 0$ against $\theta = 4\text{se}[\hat{\theta}]$ has power of approximately 0.5, because a 95% confidence interval for θ is approximately $(0, 4\text{se}[\hat{\theta}])$, implying that values of $\theta = 0$ or $\theta = 4\text{se}[\hat{\theta}]$ are just as likely.

As a more concrete example, suppose θ measures the percentage increase in wage resulting from a training program, and that a study finds $\hat{\theta} = 6$ with $\text{se}[\hat{\theta}] = 4$. Then the Wald test at 5% significance level leads to nonrejection of H_0 , since $W = (6/4)^2 = 2.25 < \chi^2_{.05}(1) = 3.96$. The conclusion of such a study will often state that the training program is not statistically significant. One should not interpret this as meaning that there is a high probability that the training program has no effect, however, as this test has low power. For example, the preceding analysis indicates that a test of $H_0 : \theta = 0$ against $H_a : \theta = 16$, a relatively large training effect, has power of only 0.5, since $4 \times \text{se}[\hat{\theta}] = 16$. Reasons for low power include small sample size, large model error variance, and small spread in the regressors.

In simple cases, solving the inverse problem of estimating the minimum sample size needed to achieve a given desired level of power is possible. This is especially popular in medical studies.

Andrews (1989) gives a more formal treatment of using the noncentrality parameter to determine regions of the parameter space against which a test in an empirical setting is likely to have low power. He provides many applied examples where it is easy to determine that tests have low power against meaningful alternatives.

7.7. Monte Carlo Studies

Our discussion of statistical inference has so far relied on asymptotic results. For small samples analytical results are rarely available, aside from tests of linear restrictions in

the linear regression model under normality. Small-sample results can nonetheless be obtained by performing a Monte Carlo study.

7.7.1. Overview

An example of a **Monte Carlo study** of the small-sample properties of a test statistic is the following. Set the sample size N to 40, say, and randomly generate 10,000 samples of size 40 under the H_0 model. For each replication (sample) form the test statistic of interest and test H_0 , rejecting H_0 if the test statistic falls in the rejection region, usually determined by asymptotic results.

The **true size** or **actual size** of the test statistic is simply the fraction of replications for which the test statistic falls in the rejection region. Ideally, this is close to the **nominal size**, which is the chosen significance level of the test. For example, if testing at 5% the nominal test size is 0.05 and the true size is hopefully close to 0.05.

Determining test power in small samples requires additional simulation, with samples generated under one or more particular specification of the possible models that lie in the composite alternative hypothesis H_a . The **power** is calculated as the fraction of replications for that the null hypothesis is rejected, using either the same test as used in determining the true size, or a **size-corrected version** of the test that uses a rejection region such that the nominal size equals the true size.

Monte Carlo studies are simple to implement, but there are many subtleties involved in designing a good Monte Carlo study. For an excellent discussion see Davidson and MacKinnon (1993).

7.7.2. Monte Carlo Details

As an example of a Monte Carlo study we consider statistical inference on the slope coefficient in a probit model. The following analysis does not rely on knowledge of the probit model.

The data-generating process is a probit model, with binary regressor y equal to one with probability

$$\Pr[y = 1|\mathbf{x}] = \Phi(\beta_1 + \beta_2 x),$$

where $\Phi(\cdot)$ is the standard normal cdf, $x \sim \mathcal{N}[0, 1]$, and $(\beta_1, \beta_2) = (0, 1)$.

The data (y, x) are easily generated for this dgp. The regressor x is first obtained as a random draw from the standard normal distribution. Then, from Section 14.4.2 the dependent variable y is set equal to 1 if $x + u > 0$ and is set to 0 otherwise, where u is a random draw from the standard normal. For this dgp $y = 1$ roughly half the time and $y = 0$ the other half.

In each simulation N new observations of both x and y are drawn, and the MLE from probit regression of y on x is obtained. An alternative is to use the same N draws of the regressor x in each simulation and only redraw y . The former setup corresponds to simple random sampling and the latter corresponds to analysis conditional on x or “fixed in repeated trials”; see Section 4.4.7.

Monte Carlo studies often consider a range of sample sizes. Here we simply set $N = 40$. Programs can be checked by also setting a very large value of N ,

say $N = 10,000$, as then Monte Carlo results should be very close to asymptotic results.

Numerous simulations are needed to determine actual test size, because this depends on behavior in the tails of the distribution rather than the center. If S simulations are run for a test of true size α , then the proportion of times the null hypothesis is correctly rejected is an outcome from S binomial trials with mean α and variance $\alpha(1 - \alpha)/S$. So 95% of Monte Carlos will estimate the test size to be in the interval $\alpha \pm 1.96\sqrt{\alpha(1 - \alpha)/S}$. A mere 100 simulations is not enough since, for example, this interval is (0.007, 0.093) when $\alpha = 0.05$. For 10,000 simulations the 95% interval is much more precise, equalling (0.008, 0.012), (0.046, 0.054), (0.094, 0.106), and (0.192, 0.208) for α equal to, respectively, 0.01, 0.05, 0.10, and 0.20. Here $S = 10,000$ simulations are used.

A problem that can arise in Monte Carlo simulations is that for some simulation samples the model may not be estimable. For example, consider linear regression on just an intercept and an indicator variable. If the indicator variable happens to always take the same value, say 0, in a simulation sample then its coefficient cannot be separately identified from that for the intercept. A similar problem arises in the probit and other binary outcome models, if all ys are 0 or all ys are 1 in a simulation sample. The standard procedure, which can be criticized, is to drop such simulation samples, and to write computer code that permits the simulation loop to continue when such a problem arises. In this example the problem did not arise with $N = 40$, but it did for $N = 30$.

7.7.3. Small-Sample Bias

Before moving to testing we look at the small-sample properties of the MLE $\hat{\beta}_2$ and its estimated standard error $\text{se}[\hat{\beta}_2]$.

Across the 10,000 simulations $\hat{\beta}_2$ had mean 1.201 and standard deviation 0.452, whereas $\text{se}[\hat{\beta}_2]$ had mean 0.359. The MLE is therefore biased upward in small samples, as the average of $\hat{\beta}_2$ is considerably greater than $\beta_2 = 1$. The standard errors are biased downward in small samples since the average of $\text{se}[\hat{\beta}_2]$ is considerably smaller than the standard deviation of $\hat{\beta}_2$.

7.7.4. Test Size

We consider a two-sided test of $H_0 : \beta_2 = 1$ against $H_a : \beta_2 \neq 1$, using the Wald test

$$z = W_z = \frac{\hat{\beta}_2 - 1}{\text{se}[\hat{\beta}_2]},$$

where $\text{se}[\hat{\beta}_2]$ is the standard error of the MLE estimated using the variance matrix given in Section 14.3.2, which is minus the inverse of the expected Hessian. Given the dgp, asymptotically z is standard normal distributed and z^2 is chi-squared distributed. The goal is to find how well this approximates the small-sample distribution.

Figure 7.2 gives the density for the $S = 10,000$ computed values of z , where the density is plotted using the kernel density estimate of Chapter 9 rather than a histogram. This is superimposed on the standard normal density. Clearly the asymptotic result is not exact, especially in the upper tail where the difference is clearly large enough to

Table 7.2. Wald Test Size and Power for Probit Regression Example^a

Nominal Size (α)	Actual Size	Actual Power	Asymptotic Power
0.01	0.005	0.007	0.272
0.05	0.029	0.226	0.504
0.10	0.081	0.608	0.628
0.20	0.192	0.858	0.755

^a The dgp for y is the Probit with $\Pr[y = 1] = \Phi(0 + \beta_2 x)$ and sample size $N = 40$. The test is a two-sided Wald test of whether or not the slope coefficient equals 1. Actual size is calculated from $S = 10,000$ simulations with $\beta_2 = 1$ and power is calculated from 10,000 simulations with $\beta_2 = 2$.

lead to size distortions when testing at, say, 5%. Also, across the simulations z has mean $0.114 \neq 0$ and standard deviation $0.956 \neq 1$.

The first two columns of Table 7.2 give the nominal size and the actual size of the Wald test for nominal sizes $\alpha = 0.01, 0.05, 0.10$, and 0.20 . The actual size is the proportion of the 10,000 simulations in which $|z| > z_{\alpha/2}$, or equivalently that $z^2 > \chi^2_{\alpha}(1)$. Clearly the actual size of the test is much less than the nominal size for $\alpha \leq 0.10$. An ad hoc small-sample correction is to instead assume that z is t distributed with 38 degrees of freedom, and reject if $|z| > t_{\alpha/2}(38)$. However, this leads to even smaller actual size, since $t_{\alpha/2}(38) > z_{\alpha/2}$.

The Monte Carlo simulations can also be used to obtain size-corrected critical values. Thus the lower and upper 2.5 percentiles of the 10,000 simulated values of z are -1.905 and 2.003 . It follows that an asymmetric rejection region with actual size 0.05 is $z < -1.905$ and $z > 2.003$, a larger rejection region than $|z_2| > 1.960$.

7.7.5. Test Power

We consider power of the Wald test under $H_a : \beta_2 = 2$. We would expect the power to be reasonable because this value of β_2 lies two to three standard errors away from the

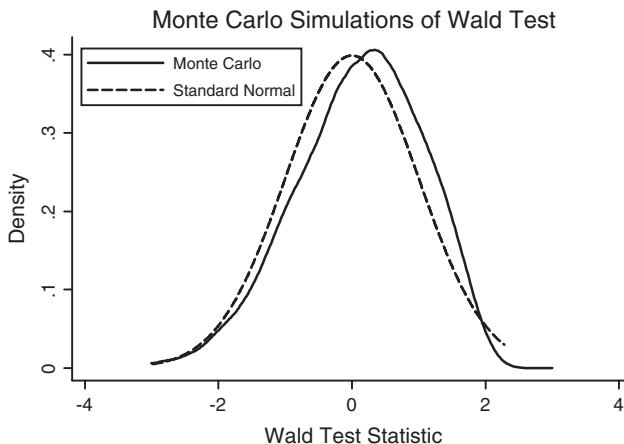


Figure 7.2: Density of Wald test statistic that slope coefficient equals one computed by Monte Carlo simulation with standard normal density also plotted for comparison. Data are generated from a probit regression model.

null hypothesis value of $\beta_2 = 1$, given that $\text{se}[\hat{\beta}_2]$ has average value 0.359. The actual and nominal power of the Wald test are given in the last two columns of Table 7.2.

The actual power is obtained in the same way as actual size, being the proportion of the 10,000 simulations in which $|z| > z_{\alpha/2}$. The only change is that, in generating y in the simulation, $\beta_2 = 2$ rather than 1. The actual power is very low for $\alpha = 0.01$ and 0.05, cases where the actual size is much less than the nominal size.

The nominal power of the Wald test is determined using the asymptotic non-central $\chi^2(1, \lambda)$ distribution under H_a , where from (7.50) $\lambda = \frac{1}{2}(\delta/\sqrt{N})^2/\text{se}[\hat{\beta}_2]^2 = \frac{1}{2} \times 1^2/0.359^2 \simeq 3.88$, since the local alternative is that $H_a : \beta_2 - 1 = \delta/\sqrt{N}$, so $\delta/\sqrt{N} = 1$ for $\beta_2 = 2$. The asymptotic result is not exact, but it does provide a useful estimate of the power for $\alpha = 0.10$ and 0.20, cases where the true size closely matches the nominal size.

7.7.6. Monte Carlo in Practice

The preceding discussion has emphasized use of the Monte Carlo analysis to calculate test power and size. A Monte Carlo analysis can also be very useful for determining small-sample bias in an estimator and, by setting N large, for determining that an estimator is actually consistent. Such Monte Carlo routines are very simple to run using current computer packages.

A Monte Carlo analysis can be applied to real data if the conditional distribution of y given \mathbf{x} is fully parametrized. For example, consider a probit model estimated with real data. In each simulation the regressors are set at their sample values, if the sampling framework is one of fixed regressors in repeated samples, while a new set of values for the binary dependent variable y needs to be generated. This will depend on what values of the parameters β are used. Let $\hat{\beta}_1, \dots, \hat{\beta}_K$ denote the probit estimates from the original sample and consider a Wald test of $H_0 : \beta_j = 0$. To calculate test size, generate S simulation samples by setting $\beta_k = \hat{\beta}_k$ for $j \neq k$ and setting $\beta_j = 0$, and then calculate the proportion of simulations in which $H_0 : \beta_j = 0$ is rejected. To estimate the power of the Wald test against a specific alternative $H_a : \beta_j = 1$, say, generate y with $\beta_k = \hat{\beta}_k$ for $j \neq k$ and $\beta_j = 1$ in generating y , and calculate the proportion of simulations in which $H_0 : \beta_j = 0$ is rejected.

In practice much microeconometric analysis is based on estimators that are not based on fully parametric models. Then additional distributional assumptions are needed to perform a Monte Carlo analysis.

Alternatively, power can be calculated using asymptotic methods rather than finite-sample methods. Additionally the bootstrap, presented next, can be used to obtain size using a more refined asymptotic theory.

7.8. Bootstrap Example

The bootstrap is a variant of Monte Carlo simulation that has the attraction of being implementable with fewer parametric assumptions and with little additional program

code beyond that required to estimate the model in the first place. Essential ingredients for the bootstrap to be valid are that the estimator actually has a limit distribution and that the bootstrap resamples quantities that are iid.

The bootstrap has two general uses. First, it can be used as an alternative way to compute statistics without asymptotic refinement. This is particularly useful for computing standard errors when analytical formulas are complex. Second, it can be used to implement a refinement of the usual asymptotic theory that may provide a better finite-sample approximation to the distribution of test statistics.

We illustrate the bootstrap to implement a Wald test, ahead of a complete treatment in Chapter 11.

7.8.1. Inference Using Standard Asymptotics

Consider again a probit example with binary regressor y equal to one with probability $p = \Phi(\gamma + \beta x)$, where $\Phi(\cdot)$ is the standard normal cdf. Interest lies in testing $H_0 : \beta = 1$ against $H_a : \beta \neq 1$ at significance level 0.05. The analysis here does not require knowledge of the probit model.

One sample of size $N = 30$ is generated. Probit ML estimation yields $\hat{\beta} = 0.817$ and $s_{\hat{\beta}} = 0.294$, where the standard error is based on $-\hat{A}^{-1}$, so the test statistic $z = (1 - 0.817)/0.294 = -0.623$.

Using standard asymptotic theory we obtain 5% critical values of -1.96 and 1.96 , since $z_{.025} = 1.96$, and H_0 is not rejected.

7.8.2. Bootstrap without Asymptotic Refinement

The departure point of the bootstrap method is to resample from an approximation to the population; see Section 11.2.1. The paired bootstrap does so by resampling from the original sample.

Thus form B pseudo-samples of size N by drawing with replacement from the original data $\{(y_i, x_i), i = 1, \dots, N\}$. For example, the first pseudo-sample of size 30 may have (y_1, x_1) once, (y_2, x_2) not at all, (y_3, x_3) twice, and so on. This yields B estimates $\hat{\beta}_1^*, \dots, \hat{\beta}_B^*$ of the parameter of interest β , that can be used to estimate features of the distribution of the original estimate $\hat{\beta}$.

For example, suppose the computer program used to estimate a probit model reports $\hat{\beta}$ but not the standard error $s_{\hat{\beta}}$. The bootstrap solves this problem since we can use the estimated standard deviation $s_{\hat{\beta}, \text{boot}}$ of $\hat{\beta}_1^*, \dots, \hat{\beta}_B^*$ from the B bootstrap pseudo-samples. Given this standard error estimate it is possible to perform a Wald hypothesis test on β .

For the probit Wald test example, the resulting bootstrap estimate of the standard error of $\hat{\beta}$ is 0.376, leading to $z = (1 - 0.817)/0.376 = -0.487$. Since -0.487 lies in $(-1.96, 1.96)$ we do not reject H_0 at 5%.

This use of the bootstrap to test hypotheses does not lead to size improvements in small samples. However, it can lead to great time savings in many applications if it is difficult to otherwise obtain the standard errors for an estimator.

7.8.3. Bootstrap with Asymptotic Refinement

Some bootstraps can lead to a better asymptotic approximation to the distribution of z . This is likely to lead to finite-sample critical values that are better in the sense that the actual size is likely to be closer to the nominal size of 0.05. Details are provided in Chapter 11. Here we illustrate the method.

Again form B pseudo-samples of size N by drawing with replacement from the original data. Estimate the probit model in each pseudo-sample and for the b th pseudo-sample compute $z_b^* = (\hat{\beta}_b^* - \hat{\beta})/s_{\hat{\beta}_b^*}$, where $\hat{\beta}$ is the original estimate. The bootstrap distribution for the original test statistic z is then the empirical distribution of z_1^*, \dots, z_B^* rather than the standard normal. The lower and upper 2.5 percentiles of this empirical distribution give the bootstrap critical values.

For the example here with $B = 1,000$ the lower and upper 2.5 percentiles of the empirical bootstrap distribution of z were found to be -2.62 and 1.83 . The bootstrap critical values for testing at 5% are then -2.62 and 1.83 , rather than the usual ± 1.96 . Since the initial sample test statistic $z = -0.623$ lies in $(-2.62, 1.83)$ we do not reject $H_0 : \beta = 1$. A bootstrap p -value can also be computed.

Unlike the bootstrap in the previous section, an asymptotic improvement occurs here because the studentized test statistic z is asymptotically pivotal (see Section 11.2.3) whereas the estimator $\hat{\beta}$ is not.

7.9. Practical Considerations

Microeconomics research places emphasis on statistical inference based on minimal distributional assumptions, using robust estimates of the variance matrix of an estimator. There is no sense in robust inference, however, if failure of distributional assumptions leads to the more serious complication of estimator inconsistency as can happen for some though not all ML estimators.

Many packages provide a “robust” standard errors option in estimator commands. In microeconomics packages robust often means heteroskedastic consistent and does not guard against other complications such as clustering, see Section 24.5, that can also lead to invalid statistical inference.

Robust inference is usually implemented using a Wald test. The Wald test has the weakness of invariance to reparametrization of nonlinear hypotheses, though this may be diminished by performing an appropriate bootstrap. Standard auxiliary regressions for the LM test and implementations of LM tests on computer packages are usually not robustified, though in some cases relatively simple robustification of the LM test is possible (see Section 8.4).

The power of tests can be weak. Ideally, power against some meaningful alternative would be reported. Failing this, as Section 7.6 indicates, one should be careful about overstating the conclusions from a hypothesis test unless parameters are very precisely estimated.

The finite sample size of tests derived from asymptotic theory is also an issue. The bootstrap method, detailed in Chapter 11, has the potential to yield hypothesis tests and confidence intervals with much better finite-sample properties.

Statistical inference can be quite fragile, so these issues are of importance to the practitioner. Consider a two-tailed Wald test of statistical significance when $\hat{\theta} = 1.96$, and assume the test statistic is indeed standard normal distributed. If $s_{\hat{\theta}} = 1.0$ then $t = 1.96$ and the p -value is 0.050. However, the true p -value is a much higher 0.117 if the standard error was underestimated by 20% (so correct $t = 1.57$), and a much lower 0.014 if the standard error was overestimated by 20% (so $t = 2.35$).

7.10. Bibliographic Notes

The econometrics texts by Gouriéroux and Monfort (1989) and Davidson and MacKinnon (1993) give quite lengthy treatment of hypothesis testing. The presentation here considers only equality restrictions. For tests of inequality restrictions see Gouriéroux, Holly, and Monfort (1982) for the linear case and Wolak (1991) for the nonlinear case. For hypothesis testing when the parameters are at the boundary of the parameter space under the null hypothesis the tests can break down; see Andrews (2001).

- 7.3 A useful graphical treatment of the three classical test procedures is given by Buse (1982).
- 7.5 Newey and West (1987a) present extension of the classical tests to GMM estimation.
- 7.6 Davidson and MacKinnon (1993) give considerable discussion of power and explain the distinction between explicit and implicit null and alternative hypotheses.
- 7.7 For Monte Carlo studies see Davidson and MacKinnon (1993) and Hendry (1984).
- 7.8 The bootstrap method due to Efron (1979) is detailed in Chapter 11.

Exercises

- 7–1** Suppose a sample yields estimates $\hat{\theta}_1 = 5$, $\hat{\theta}_2 = 3$ with asymptotic variance estimates 4 and 2 and the correlation coefficient between $\hat{\theta}_1$ and $\hat{\theta}_2$ equals 0.5. Assume asymptotic normality of the parameter estimates.
- (a) Test $H_0 : \theta_1 e^{\theta_2} = 100$ against $H_a : \theta_1 \neq 100$ at level 0.05.
 - (b) Obtain a 95% confidence interval for $\gamma = \theta_1 e^{\theta_2}$.
- 7–2** Consider NLS regression for the model $y = \exp(\alpha + \beta x) + \varepsilon$, where α , β , and x are scalars and $\varepsilon \sim \mathcal{N}[0, 1]$. Note that for simplicity $\sigma_\varepsilon^2 = 1$ and need not be estimated. We want to test $H_0 : \beta = 0$ against $H_a : \beta \neq 0$.
- (a) Give the first-order conditions for the unrestricted MLE of α and β .
 - (b) Give the asymptotic variance matrix for the unrestricted MLE of α and β .
 - (c) Give the explicit solution for the restricted MLE of α and β .
 - (d) Give the auxiliary regression to compute the OPG form of the LM test.
 - (e) Give the complete expression for the original form of the LM test. Note that it involves derivatives of the unrestricted log-likelihood evaluated at the restricted MLE of α and β . [This is more difficult than parts (a)–(d).]

- 7–3** Suppose we wish to choose between two nested parametric models. The relationship between the densities of the two models is that $g(y|x, \beta, \alpha = 0) = f(y|x, \beta)$, where for simplicity both β and α are scalars. If g is the correct density then the MLE of β based on density f is inconsistent. A test of model f against model g is a test of $H_0 : \alpha = 0$ against $H_a : \alpha \neq 0$. Suppose ML estimation yields the following results: (1) model f : $\hat{\beta} = 5.0$, $se[\hat{\beta}] = 0.5$, and $\ln L = -106$; (2) model g : $\hat{\beta} = 3.0$, $se[\hat{\beta}] = 1.0$, $\hat{\alpha} = 2.5$, $se[\hat{\alpha}] = 1.0$, and $\ln L = -103$. Not all of the

HYPOTHESIS TESTS

following tests are possible given the preceding information. If there is enough information, perform the tests and state your conclusions. If there is not enough information, then state this.

- (a) Perform a Wald test of H_0 at level 0.05.
- (b) Perform a Lagrange multiplier test of H_0 at level 0.05.
- (c) Perform a likelihood ratio test of H_0 at level 0.05.
- (d) Perform a Hausman test of H_0 at level 0.05.

7–4 Consider test of $H_0 : \mu = 0$ against $H_a : \mu \neq 0$ at nominal size 0.05 when the dgp is $y \sim \mathcal{N}[\mu, 100]$, so the standard deviation is 10, and the sample size is $N = 10$. The test statistic is the usual t -test statistic $t = \hat{\mu}/\sqrt{s/10}$, where $s^2 = (1/9) \sum_i (y_i - \bar{y})^2$. Perform 1,000 simulations to answer the following.

- (a) Obtain the actual size of the t -test if the correct finite-sample critical values $\pm t_{025}(8) = \pm 2.306$ are used. Is there size distortion?
- (b) Obtain the actual size of the t -test if the asymptotic approximation critical values $\pm z_{025} = \pm 1.960$ are used. Is there size distortion?
- (c) Obtain the power of the t -test against the alternative $H_a : \mu = 1$, when the critical values $\pm t_{025}(8) = \pm 2.306$ are used. Is the test powerful against this particular alternative?

7–5 Use the health expenditure data of Section 16.6. The model is a probit regression of DMED, an indicator variable for positive health expenditures, against the 17 regressors listed in the second paragraph of Section 16.6. You should obtain the estimates given in the first column of Table 16.1. Consider joint test of the statistical significance of the self-rated health indicators HLTHG, HLTHF, and HLTHP at level 0.05.

- (a) Perform a Wald test.
- (b) Perform a likelihood ratio test.
- (c) Perform an auxiliary regression to implement an LM test. [This will require some additional coding.]

CHAPTER 8

Specification Tests and Model Selection

8.1. Introduction

Two important practical aspects of microeconometric modeling are determining whether a model is correctly specified and selecting from alternative models. For these purposes it is often possible to use the hypothesis testing methods presented in the previous chapter, especially when models are nested. In this chapter we present several other methods.

First, m-tests such as conditional moment tests are tests of whether moment conditions imposed by a model are satisfied. The approach is similar in spirit to GMM, except that the moment conditions are not imposed in estimation and are instead used for testing. Such tests are conceptually very different from the hypothesis tests of Chapter 7, as there is no explicit statement of an alternative hypothesis model.

Second, Hausman tests are tests of the difference between two estimators that are both consistent if the model is correctly specified but diverge if the model is incorrectly specified.

Third, tests of nonnested models require special methods because the usual hypothesis testing approach can only be applied when one model is nested within another.

Finally, it can be useful to compute and report statistics of model adequacy that are not test statistics. For example, an analogue of R^2 may be used to measure the goodness of fit of a nonlinear model.

Ideally, these methods are used in a cycle of model specification, estimating, testing, and evaluation. This cycle can move from a general model toward a specific model, or from a specific model to a more general one that is felt to capture the most important features of the data.

Section 8.2 presents m-tests, including conditional moment tests, the information matrix test, and chi-square goodness of fit tests. The Hausman test is presented in Section 8.3. Tests for several common misspecifications are discussed in Section 8.4. Discrimination between nonnested models is the focus of Section 8.5. Commonly used convenient implementations of the tests of Sections 8.2–8.5 can rely on strong distributions and/or perform poorly in finite samples. These concerns have discouraged use

of some of these tests, but such concerns are outdated because in many cases the bootstrap methods presented in Chapter 11 can correct for these weaknesses. Section 8.6 considers the consequences of testing a model on subsequent inference. Model diagnostics are presented in the stand-alone Section 8.7.

8.2. m-Tests

m-Tests, such as conditional moment tests, are a general specification testing procedure that encompasses many common specification tests. The tests are easily implemented using auxiliary regressions when estimation is by ML, a situation where tests of model assumptions are especially desirable. Implementation is usually more difficult when estimators are instead based on minimal distributional assumptions.

We first introduce the test statistic and computational methods, followed by leading examples and an illustration of the tests.

8.2.1. m-Test Statistic

Suppose a model implies the **population moment condition**

$$H_0 : E[\mathbf{m}_i(\mathbf{w}_i, \boldsymbol{\theta})] = \mathbf{0}, \quad (8.1)$$

where \mathbf{w} is a vector of observables, usually the dependent variable y and regressors \mathbf{x} and sometimes additional variables \mathbf{z} , $\boldsymbol{\theta}$ is a $q \times 1$ vector of parameters, and $\mathbf{m}_i(\cdot)$ is an $h \times 1$ vector. A simple example is that $E[(y - \mathbf{x}'\boldsymbol{\beta})\mathbf{z}] = \mathbf{0}$ if \mathbf{z} can be omitted in the linear model $y = \mathbf{x}'\boldsymbol{\beta} + u$. Especially for fully parametric models there are many candidates for $\mathbf{m}_i(\cdot)$.

An **m-test** is a test of the closeness to zero of the corresponding **sample moment**

$$\widehat{\mathbf{m}}_N(\widehat{\boldsymbol{\theta}}) = N^{-1} \sum_{i=1}^N \mathbf{m}_i(\mathbf{w}_i, \widehat{\boldsymbol{\theta}}). \quad (8.2)$$

This approach is similar to that for the Wald test, where $\mathbf{h}(\boldsymbol{\theta}) = \mathbf{0}$ is tested by testing the closeness to zero of $\mathbf{h}(\widehat{\boldsymbol{\theta}})$.

A test statistic is obtained by a method similar to that detailed in Section 7.2.4 for the Wald test. In Section 8.2.3 it is shown that if (8.1) holds then

$$\sqrt{N}\widehat{\mathbf{m}}_N(\widehat{\boldsymbol{\theta}}) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{V}_{\mathbf{m}}], \quad (8.3)$$

where $\mathbf{V}_{\mathbf{m}}$ defined later in (8.10) is more complicated than in the case of the Wald test because $\mathbf{m}_i(\mathbf{w}_i, \widehat{\boldsymbol{\theta}})$ has two sources of stochastic variation as both \mathbf{w}_i and $\widehat{\boldsymbol{\theta}}$ are random.

A chi-square test statistic can then be obtained by taking the corresponding quadratic form. Thus the **m-test statistic** for (8.1) is

$$M = N\widehat{\mathbf{m}}_N(\widehat{\boldsymbol{\theta}})' \widehat{\mathbf{V}}_{\mathbf{m}}^{-1} \widehat{\mathbf{m}}_N(\widehat{\boldsymbol{\theta}}), \quad (8.4)$$

which is asymptotically $\chi^2(\text{rank}[\mathbf{V}_{\mathbf{m}}])$ distributed if the moment conditions (8.1) are correct. An m-test rejects the moment conditions (8.1) at significance level α if $M > \chi^2_\alpha(h)$ and does not reject otherwise.

A complication is that \mathbf{V}_m may not be of full rank h . For example, this is the case if the estimator $\hat{\theta}$ itself sets a linear combination of components of $\hat{\mathbf{m}}_N(\hat{\theta})$ to $\mathbf{0}$. In some cases, such as the OIR test, $\hat{\mathbf{V}}_m$ is still of full rank and M can be computed but the chi-square test statistic has only $\text{rank}[\mathbf{V}_m]$ degrees of freedom. In other cases $\hat{\mathbf{V}}_m$ itself is not of full rank. Then it is simplest to drop $(h - \text{rank}[\mathbf{V}_m])$ of the moment conditions and perform an m-test using just this subset of the moment conditions. Alternatively, the full set of moment conditions can be used, but $\hat{\mathbf{V}}_m^{-1}$ in (8.4) is replaced by $\hat{\mathbf{V}}_m^-$, the generalized inverse of $\hat{\mathbf{V}}_m$. The Moore–Penrose generalized inverse \mathbf{V}^- of a matrix \mathbf{V} satisfies $\mathbf{V}\mathbf{V}^-\mathbf{V} = \mathbf{V}$, $\mathbf{V}^-\mathbf{V}\mathbf{V}^- = \mathbf{V}^-$, $(\mathbf{V}\mathbf{V}^-)' = \mathbf{V}\mathbf{V}^-$, and $(\mathbf{V}^-\mathbf{V})' = \mathbf{V}^-\mathbf{V}$. When \mathbf{V}_m is less than full rank then strictly speaking (8.3) no longer holds, since the multivariate normal requires full rank \mathbf{V}_m , but (8.4) still holds given these adjustments.

The m-test approach is conceptually very simple. The moment restriction (8.1) is rejected if a quadratic form in the sample estimate (8.2) is far enough from zero. The challenges are in calculating M since $\hat{\mathbf{V}}_m$ can be quite complex (see Section 8.2.2), selecting moments $\mathbf{m}(\cdot)$ to test (see Sections 8.2.3–8.2.6 for leading examples), and interpreting reasons for rejection of (8.1) (see Section 8.2.8).

8.2.2. Computation of the m-Statistic

There are several ways to compute the m-statistic.

First, one can always *directly compute* $\hat{\mathbf{V}}_m$, and hence M , using the consistent estimates of the components of \mathbf{V}_m given in Section 8.2.3. Most practitioners shy away from this approach as it entails matrix computations.

Second, the *bootstrap* can always be used (see Section 11.6.3), since the bootstrap can provide an estimate of \mathbf{V}_m that controls for all sources of variation in $\hat{\mathbf{m}}_N(\hat{\theta}) = N^{-1} \sum_i \mathbf{m}_i(\mathbf{w}_i, \hat{\theta})$.

Third, in some cases *auxiliary regressions* similar to those for the LM test given in Section 7.3.5 can be run to compute asymptotically equivalent versions of M that do not require computation of $\hat{\mathbf{V}}_m$. These auxiliary regressions may in turn be bootstrapped to obtain an asymptotic refinement (see Section 11.6.3). We present several leading auxiliary regressions.

Auxiliary Regressions Using the ML Estimator

Model specification tests are especially desirable when inference is done within the likelihood framework, as in general any misspecification of the density can lead to inconsistency of the MLE. Fortunately, an m-test is easily implemented when estimation is by maximum likelihood.

Specifically, when $\hat{\theta}$ is the MLE, generalizing the LM test result of Section 7.3.5 (see Section 8.2.3) yields an asymptotically equivalent version of the m-test is obtained from the **auxiliary regression**

$$1 = \hat{\mathbf{m}}_i' \delta + \hat{\mathbf{s}}_i' \gamma + u_i, \quad (8.5)$$

where $\hat{\mathbf{m}}_i = \mathbf{m}_i(y_i, \mathbf{x}_i, \hat{\boldsymbol{\theta}}_{\text{ML}})$, $\hat{s}_i = \partial \ln f(y_i | \mathbf{x}_i, \boldsymbol{\theta}) / \partial \boldsymbol{\theta} |_{\hat{\boldsymbol{\theta}}_{\text{ML}}}$ is the contribution of the i th observation to the score and $f(y_i | \mathbf{x}_i, \boldsymbol{\theta})$ is the conditional density function, by calculating

$$M^* = N R_u^2, \quad (8.6)$$

where R_u^2 is the uncentered R^2 defined at the end of Section 7.3.5. Equivalently, M^* equals ESS_u , the uncentered explained sum of squares (the sum of squares of the fitted values) from regression (8.5), or M^* equals $N - \text{RSS}$, where RSS is the residual sum of squares from regression (8.5). M^* is asymptotically $\chi^2(h)$ under H_0 .

The test statistic M^* is called the **outer product of the gradient** form of the m-test, and it is a generalization of the auxiliary regression for the LM test (see Section 7.3.5). Although the OPG form can be easily computed, it has poor small-sample properties with large size distortions. Similar to the LM test, however, these small-sample problems can be greatly reduced by using bootstrap methods (see Section 11.6.3).

The test statistic M^* may also be appropriate in some non-ML settings. The auxiliary regression is applicable whenever $E[\partial \mathbf{m} / \partial \boldsymbol{\theta}'] = -E[\mathbf{m} \mathbf{s}']$ (see Section 8.2.3). By the generalized IM equality (see Section 5.6.3), this condition holds for the MLE when expectation is with respect to the specified density $f(\cdot)$. It can also hold under weaker distributional assumptions in some cases.

Auxiliary Regressions When $E[\partial \mathbf{m} / \partial \boldsymbol{\theta}'] = \mathbf{0}$

In some applications $\mathbf{m}_i(\mathbf{w}_i, \boldsymbol{\theta})$ satisfies

$$E\left[\partial \mathbf{m}_i(\mathbf{w}_i, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}' \Big|_{\boldsymbol{\theta}_0}\right] = \mathbf{0}, \quad (8.7)$$

in addition to (8.1).

Then it can be shown that the asymptotic distribution of $\sqrt{N} \hat{\mathbf{m}}_N(\hat{\boldsymbol{\theta}})$ is the same as that of $\sqrt{N} \mathbf{m}_N(\boldsymbol{\theta}_0)$, so $\mathbf{V}_{\mathbf{m}} = \text{plim } N^{-1} \sum_i \mathbf{m}_{i0} \mathbf{m}_{i0}'$, which can be consistently estimated by $\hat{\mathbf{V}}_{\mathbf{m}} = N^{-1} \sum_i \hat{\mathbf{m}}_i \hat{\mathbf{m}}_i'$. The test statistic can be computed in a similar manner to (8.5), except the **auxiliary regression** is more simply

$$1 = \hat{\mathbf{m}}_i' \boldsymbol{\delta} + u_i, \quad (8.8)$$

with test statistic M^{**} equal to N times the uncentered R^2 .

This auxiliary regression is valid for any root- N consistent estimator $\hat{\boldsymbol{\theta}}$, not just the MLE, provided (8.7) holds. The condition (8.7) is met in a few examples; see Section 8.2.9 for an example.

Even if (8.7) does not hold the simpler regression (8.8) might still be run as a guide, as it places a lower bound on the correct value of M , the m-test statistic. If this simpler regression leads to rejection then (8.1) is certainly rejected.

Other Auxiliary Regressions

Alternative auxiliary regressions to (8.5) and (8.8) are possible if $\mathbf{m}(y, \mathbf{x}, \boldsymbol{\theta})$ and $\mathbf{s}(y, \mathbf{x}, \boldsymbol{\theta})$ can be appropriately factorized.

First, if $\mathbf{s}(y, \mathbf{x}, \boldsymbol{\theta}) = \mathbf{g}(\mathbf{x}, \boldsymbol{\theta})r(y, \mathbf{x}, \boldsymbol{\theta})$ and $\mathbf{m}(y, \mathbf{x}, \boldsymbol{\theta}) = \mathbf{h}(\mathbf{x}, \boldsymbol{\theta})r(y, \mathbf{x}, \boldsymbol{\theta})$ for some common scalar function $r(\cdot)$ with $V[r(y, \mathbf{x}, \boldsymbol{\theta})] = 1$ and estimation is by ML, then an asymptotically equivalent regression to (8.5) is NR_u^2 from regression of \hat{r}_i on $\hat{\mathbf{g}}_i$ and $\hat{\mathbf{h}}_i$.

Second, if $\mathbf{m}(y, \mathbf{x}, \boldsymbol{\theta}) = \mathbf{h}(\mathbf{x}, \boldsymbol{\theta})v(y, \mathbf{x}, \boldsymbol{\theta})$ for some scalar function $v(\cdot)$ with $V[v(y, \mathbf{x}, \boldsymbol{\theta})] = 1$ and $E[\partial\mathbf{m}/\partial\boldsymbol{\theta}'] = \mathbf{0}$, then an asymptotically equivalent regression to (8.8) is NR_u^2 from regression of \hat{v}_i on $\hat{\mathbf{h}}_i$. For further details see Wooldridge (1991).

Additional auxiliary regressions exist in special settings. Examples are given in Section 8.4, and White (1994) gives a quite general treatment.

8.2.3. Derivations for the m-Test Statistic

To avoid the need to compute \mathbf{V}_m , the variance matrix in (8.3), m-tests are usually implemented using auxiliary regressions or bootstrap methods. For completeness this section derives the actual expression for \mathbf{V}_m and provides justification for the auxiliary regressions (8.5) and (8.8).

The key is obtaining the distribution of $\widehat{\mathbf{m}}_N(\widehat{\boldsymbol{\theta}})$ defined in (8.2). This is complicated because $\mathbf{m}_N(\widehat{\boldsymbol{\theta}})$ is stochastic for two reasons: the random variables \mathbf{w}_i and evaluation at the estimator $\widehat{\boldsymbol{\theta}}$.

Assume that $\widehat{\boldsymbol{\theta}}$ is an m-estimator or estimating equations estimator that solves

$$\frac{1}{N} \sum_{i=1}^N \mathbf{s}_i(\mathbf{w}_i, \widehat{\boldsymbol{\theta}}) = \mathbf{0}, \quad (8.9)$$

for some function $\mathbf{s}(\cdot)$, here not necessarily $\partial \ln f(y|\mathbf{x}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}$, and make the usual cross-section assumption that data are independent over i . Then we shall show that $\sqrt{N}\widehat{\mathbf{m}}_N(\widehat{\boldsymbol{\theta}}) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{V}_m]$, as in (8.3), where

$$\mathbf{V}_m = \mathbf{H}_0 \mathbf{J}_0 \mathbf{H}'_0, \quad (8.10)$$

the $h \times (h + q)$ matrix

$$\mathbf{H}_0 = [\mathbf{I}_h \ - \mathbf{C}_0 \mathbf{A}_0^{-1}], \quad (8.11)$$

where $\mathbf{C}_0 = \text{plim } N^{-1} \sum_i \partial \mathbf{m}_{i0} / \partial \boldsymbol{\theta}'$ and $\mathbf{A}_0 = \text{plim } N^{-1} \sum_i \partial \mathbf{s}_{i0} / \partial \boldsymbol{\theta}'$, and the $(h + q) \times (h + q)$ matrix

$$\mathbf{J}_0 = \text{plim } N^{-1} \begin{bmatrix} \sum_{i=1}^N \mathbf{m}_{i0} \mathbf{m}'_{i0} & \sum_{i=1}^N \mathbf{m}_{i0} \mathbf{s}'_{i0} \\ \sum_{i=1}^N \mathbf{s}_{i0} \mathbf{m}'_{i0} & \sum_{i=1}^N \mathbf{s}_{i0} \mathbf{s}'_{i0} \end{bmatrix}, \quad (8.12)$$

where $\mathbf{m}_{i0} = \mathbf{m}_i(\mathbf{w}_i, \boldsymbol{\theta}_0)$ and $\mathbf{s}_{i0} = \mathbf{s}_i(\mathbf{w}_i, \boldsymbol{\theta}_0)$.

To derive (8.10), take a first-order Taylor series expansion around $\boldsymbol{\theta}_0$ to obtain

$$\sqrt{N}\widehat{\mathbf{m}}_N(\widehat{\boldsymbol{\theta}}) = \sqrt{N}\mathbf{m}_N(\boldsymbol{\theta}_0) + \frac{\partial \mathbf{m}_N(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'} \sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(1). \quad (8.13)$$

For $\widehat{\boldsymbol{\theta}}$ defined in (8.9) this implies that

$$\sqrt{N}\widehat{\mathbf{m}}_N(\widehat{\boldsymbol{\theta}}) = \frac{1}{\sqrt{N}} \sum_{i=1}^N \mathbf{m}_i(\boldsymbol{\theta}_0) - \mathbf{C}_0 \mathbf{A}_0^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^N \mathbf{s}_{i0} + o_p(1), \quad (8.14)$$

where we use $\mathbf{m}_N = N^{-1} \sum_i \mathbf{m}_i$, $\partial \mathbf{m}_N / \partial \boldsymbol{\theta}' = N^{-1} \sum_i \partial \mathbf{m}_i / \partial \boldsymbol{\theta}' \xrightarrow{P} \mathbf{C}_0$, and $\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ has the same limit distribution as $\mathbf{A}_0^{-1} N^{-1/2} \sum_i \mathbf{s}_{i0}$ by applying the usual first-order Taylor series expansion to (8.9). Equation (8.14) can be written as

$$\sqrt{N} \widehat{\mathbf{m}}_N(\widehat{\boldsymbol{\theta}}) = [\mathbf{I}_h - \mathbf{C}_0 \mathbf{A}_0^{-1}] \begin{bmatrix} \frac{1}{\sqrt{N}} \sum_{i=1}^N \mathbf{m}_{i0} \\ \frac{1}{\sqrt{N}} \sum_{i=1}^N \mathbf{s}_{i0} \end{bmatrix} + o_p(1). \quad (8.15)$$

Equation (8.10) follows by application of the limit normal product rule (Theorem A.17) as the second term in the product in (8.15) has limit normal distribution under H_0 with mean $\mathbf{0}$ and variance \mathbf{J}_0 .

To compute \mathbf{M} in (8.4), a consistent estimate $\widehat{\mathbf{V}}_{\mathbf{m}}$ for $\mathbf{V}_{\mathbf{m}}$ can be obtained by replacing each component of $\mathbf{V}_{\mathbf{m}}$ by a consistent estimate. For example, \mathbf{C}_0 can be consistently estimated by $\widehat{\mathbf{C}} = N^{-1} \sum_i \partial \mathbf{m}_i / \partial \boldsymbol{\theta}'|_{\widehat{\boldsymbol{\theta}}}$, and so on. Although this can always be done, using auxiliary regressions is easier when they are available.

First, consider the auxiliary regression (8.5) when $\widehat{\boldsymbol{\theta}}$ is the MLE. By the generalized IM equality (see Section 5.6.3) $E[\partial \mathbf{m}_{i0} / \partial \boldsymbol{\theta}'] = -E[\mathbf{m}_{i0} \mathbf{s}'_{i0}]$, where for the MLE we specialize to $\mathbf{s}_i = \partial \ln f(y_i, \mathbf{x}_i, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}'$. Considerable simplification occurs since then $\mathbf{C}_0 = -\text{plim} N^{-1} \sum_i \mathbf{m}_{i0} \mathbf{s}'_{i0}$ and $\mathbf{A}_0 = -\text{plim} N^{-1} \sum_i \mathbf{s}_{i0} \mathbf{s}'_{i0}$, which also appear in the \mathbf{J}_0 matrix. This leads to the OPG form of the test. For further details see Newey (1985) or Pagan and Vella (1989).

Second, for the auxiliary regression (8.8), note that if $E[\partial \mathbf{m}_{i0} / \partial \boldsymbol{\theta}'] = \mathbf{0}$ then $\mathbf{C}_0 = \mathbf{0}$, so $\mathbf{H}_0 = [\mathbf{I}_h \ \mathbf{0}]$ and hence $\mathbf{H}_0 \mathbf{J}_0 \mathbf{H}'_0 = \text{plim} N^{-1} \sum_i \mathbf{m}_{i0} \mathbf{m}'_{i0}$.

8.2.4. Conditional Moment Tests

Conditional moment tests, due to Newey (1985) and Tauchen (1985), are m-tests of unconditional moment restrictions that are obtained from an underlying conditional moment restriction.

As an example, consider the linear regression model $y = \mathbf{x}'\beta + u$. A standard assumption for consistency of the OLS estimator is that the error has conditional mean zero, or equivalently the conditional moment restriction

$$E[y - \mathbf{x}'\beta | \mathbf{x}] = \mathbf{0}. \quad (8.16)$$

In Chapter 6 we considered using some of the implied unconditional moment restrictions as the basis of MM or GMM estimation. In particular (8.16) implies that $E[\mathbf{x}(y - \mathbf{x}'\beta)] = \mathbf{0}$. Solving the corresponding sample moment condition $\sum_i \mathbf{x}_i(y_i - \mathbf{x}'_i \beta) = \mathbf{0}$ leads to the OLS estimator for β . However, (8.16) implies many other moment conditions that are not used in estimation. Consider the unconditional moment restriction

$$E[\mathbf{g}(\mathbf{x})(y - \mathbf{x}'\beta)] = \mathbf{0},$$

where the vector $\mathbf{g}(\mathbf{x})$ should differ from \mathbf{x} , already used in OLS estimation. For example, $\mathbf{g}(\mathbf{x})$ may contain the squares and cross-products of the components of the regressor vector \mathbf{x} . This suggests a test based on whether or not the corresponding sample moment $\widehat{\mathbf{m}}_N(\widehat{\beta}) = N^{-1} \sum_i \mathbf{g}(\mathbf{x}_i)(y_i - \mathbf{x}'_i \widehat{\beta})$ is close to zero.

More generally, consider the conditional moment restriction

$$E[r(y, \mathbf{x}, \boldsymbol{\theta})|\mathbf{x}] = \mathbf{0}, \quad (8.17)$$

for some scalar function $r(\cdot)$. The **conditional (CM) moment test** is an m-test based on the implied unconditional moment restrictions

$$E[\mathbf{g}(\mathbf{x})r(y, \mathbf{x}, \boldsymbol{\theta})] = \mathbf{0}, \quad (8.18)$$

where $\mathbf{g}(\mathbf{x})$ and/or $r(y, \mathbf{x}, \boldsymbol{\theta})$ are chosen so that these restrictions are not already used in estimation.

Likelihood-based models lead to many potential restrictions. For less than fully parametric models examples of $r(y, \mathbf{x}, \boldsymbol{\theta})$ include $y - \mu(\mathbf{x}, \boldsymbol{\theta})$, where $\mu(\cdot)$ is the specified conditional mean function, and $(y - \mu(\mathbf{x}, \boldsymbol{\theta}))^2 - \sigma^2(\mathbf{x}, \boldsymbol{\theta})$, where $\sigma^2(\mathbf{x}, \boldsymbol{\theta})$ is a specified conditional variance function.

8.2.5. White's Information Matrix Test

For ML estimation the information matrix equality implies moment restrictions that may be used in an m-test, as they are usually not imposed in obtaining the MLE.

Specifically, from Section 5.6.3 the IM equality implies

$$E[\text{Vech}[\mathbf{D}_i(y_i, \mathbf{x}_i, \boldsymbol{\theta}_0)]] = \mathbf{0}, \quad (8.19)$$

where the $q \times q$ matrix \mathbf{D}_i is given by

$$\mathbf{D}_i(y_i, \mathbf{x}_i, \boldsymbol{\theta}_0) = \frac{\partial^2 \ln f_i}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} + \frac{\partial \ln f_i}{\partial \boldsymbol{\theta}} \frac{\partial \ln f_i}{\partial \boldsymbol{\theta}'}, \quad (8.20)$$

and the expectation is taken with respect to the assumed conditional density $f_i = f(y_i | \mathbf{x}_i, \boldsymbol{\theta})$. Here Vech is the vector-half operator that stacks the columns of the matrix \mathbf{D}_i in the same way as the Vec operator, except that only the $q(q+1)/2$ unique elements of the symmetric matrix \mathbf{D}_i are stacked.

White (1982) proposed the **information matrix test** of whether the corresponding sample moment

$$\widehat{\mathbf{d}}_N(\widehat{\boldsymbol{\theta}}) = N^{-1} \sum_{i=1}^N \text{Vech}[\mathbf{D}_i(y_i, \mathbf{x}_i, \widehat{\boldsymbol{\theta}}_{\text{ML}})] \quad (8.21)$$

is close to zero. Using (8.4) the IM test statistic is

$$\text{IM} = N \widehat{\mathbf{d}}_N(\widehat{\boldsymbol{\theta}})' \widehat{\mathbf{V}}^{-1} \widehat{\mathbf{d}}_N(\widehat{\boldsymbol{\theta}}), \quad (8.22)$$

where the expression for $\widehat{\mathbf{V}}$ given in White (1982) is quite complicated. A much easier way to implement the test, due to Lancaster (1984) and Chesher (1984), is to use the auxiliary regression (8.5), which is applicable since the MLE is used in (8.21).

The IM test can also be applied to a subset of the restrictions in (8.19). This should be done if q is large as then the number of restrictions $q(q+1)/2$ being tested is very large.

Large values of the IM test statistic lead to rejection of the restrictions of the IM equality and the conclusion that the density is incorrectly specified. In general

this means that the ML estimator is inconsistent. In some special cases, detailed in Section 5.7, the MLE may still be consistent though standard errors need then to be based on the sandwich form of the variance matrix.

8.2.6. Chi-Square Goodness-of-Fit Test

A useful specification test for fully parametric models is to compare predicted probabilities with sample relative frequencies. The model is a poor one if these differ considerably.

Begin with discrete iid random variable y that can take one of J possible values with probabilities p_1, p_2, \dots, p_J , $\sum_{j=1}^J p_j = 1$. The correct specification of the probabilities can be tested by testing the equality of theoretical frequencies Np_j to the observed frequencies $N\bar{p}_j$, where \bar{p}_j is the fraction of the sample that takes the j th possible value. The **Pearson chi-square goodness-of-fit test (PCGF) statistic** is

$$\text{PCGF} = \sum_{j=1}^J \frac{(N\bar{p}_j - Np_j)^2}{Np_j}. \quad (8.23)$$

This statistic is asymptotically $\chi^2(J - 1)$ distributed under the null hypothesis that the probabilities p_1, p_2, \dots, p_J are correct. The test can be extended to permit the probabilities to be predicted from regression (see Exercise 8.2). Consider a multinomial model for discrete y with probabilities $p_{ij} = p_{ij}(\mathbf{x}_i, \boldsymbol{\theta})$. Then p_j in (8.23) is replaced by $\hat{p}_j = N^{-1} \sum_i F_j(\mathbf{x}_i, \hat{\boldsymbol{\theta}})$ and if $\hat{\boldsymbol{\theta}}$ is the multinomial MLE we again get a chi-square distribution, but with reduced number of degrees of freedom ($J - \dim(\boldsymbol{\theta}) - 1$) resulting from the estimation of $\boldsymbol{\theta}$ (see Andrews, 1988a).

For regression models other than multinomial models, the statistic PCGF in (8.23) can be computed by grouping y into cells, but the statistic PCGF is then no longer chi-square distributed. Instead, a closely related m-test statistic is used. To derive this statistic, break the range of y into J mutually exclusive cells, where the J cells span all possible values of y . Let $d_{ij}(y_i)$ be an indicator variable equal to one if $y_i \in$ cell j and equal to zero otherwise. Let $p_{ij}(\mathbf{x}_i, \boldsymbol{\theta}) = \int_{y_i \in \text{cell } j} f(y_i | \mathbf{x}_i, \boldsymbol{\theta}) dy_i$ be the predicted probability that observation i falls in cell j , where $f(y|\mathbf{x}, \boldsymbol{\theta})$ is the conditional density of y and to begin with we assume the parameter vector $\boldsymbol{\theta}$ is known. If the conditional density is correctly specified, then

$$E[d_{ij}(y_i) - p_{ij}(\mathbf{x}_i, \boldsymbol{\theta})] = 0, \quad j = 1, \dots, J. \quad (8.24)$$

Stacking all J moments in obvious vector notation, we have

$$E[\mathbf{d}_i(y_i) - \mathbf{p}_i(\mathbf{x}_i, \boldsymbol{\theta})] = \mathbf{0}, \quad (8.25)$$

where \mathbf{d}_i and \mathbf{p}_i are $J \times 1$ vectors with j th entries d_{ij} and p_{ij} . This suggests an m-test of the closeness to zero of the corresponding sample moment

$$\widehat{\mathbf{d}}_N(\hat{\boldsymbol{\theta}}) = N^{-1} \sum_{i=1}^N (\mathbf{d}_i(y_i) - \mathbf{p}_i(\mathbf{x}_i, \hat{\boldsymbol{\theta}})), \quad (8.26)$$

which is the difference between the vector of sample relative frequencies $N^{-1} \sum_i \mathbf{d}_i$ and the vector of predicted frequencies $N^{-1} \sum_i \widehat{\mathbf{p}}_i$. Using (8.5) we obtain the

chi-square goodness-of-fit (CGF) test statistic of Andrews (1988a, 1988b):

$$\text{CGF} = N \widehat{\mathbf{d}}\mathbf{p}_N(\widehat{\boldsymbol{\theta}})' \widehat{\mathbf{V}}^{-1} \widehat{\mathbf{d}}\mathbf{p}_N(\widehat{\boldsymbol{\theta}}), \quad (8.27)$$

where the expression for $\widehat{\mathbf{V}}$ is quite complicated. The CGF test statistic is easily computed using the auxiliary regression (8.5), with $\widehat{\mathbf{m}}_i = \mathbf{d}_i - \widehat{\mathbf{p}}_i$. This auxiliary regression is appropriate here because a fully parametric model is being tested and so $\widehat{\boldsymbol{\theta}}$ will be the MLE.

One of the categories needs to be dropped because of the restriction that probabilities sum to one, yielding a test statistic that is asymptotically $\chi^2(J - 1)$ under the null hypothesis that $f(y|\mathbf{x}, \boldsymbol{\theta})$ is correctly specified. Further categories may need to be dropped in some special cases, such as the multinomial example already discussed after (8.23). In addition to reporting the calculated test statistic it can be informative to report the components of $N^{-1} \sum_i \mathbf{d}_i$ and $N^{-1} \sum_i \widehat{\mathbf{p}}_i$.

The relevant asymptotic theory is provided by Andrews (1988a), with a simpler presentation and several applications given in Andrews (1988b). For simplicity we presented cells determined by the range of y , but the partitioning can be on both y and \mathbf{x} . Cells should be chosen so that no cell has only a few observations. For further details and a history of this test see these articles.

For continuous random variable y in the iid case a more general test than the SCGF test is the Kolmogorov test; this uses the entire distribution of y , not just cells formed from y . Andrews (1997) presents a regression version of the Kolmogorov test, but it is much more difficult to implement than the CGF test.

8.2.7. Test of Overidentifying Restrictions

Tests of overidentifying assumptions (see Section 6.3.8) are examples of m-tests.

In the notation of Chapter 6, the GMM estimator is based on the assumption that $E[\mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}_0)] = \mathbf{0}$. If the model is overidentified, then only q of these moment restrictions are used in estimation, leading to $(r - q)$ linearly dependent orthogonality conditions, where $r = \dim[\mathbf{h}(\cdot)]$, that can be used to form an m-test. Then we use \mathbf{M} in (8.4), where $\widehat{\mathbf{m}}_N = N^{-1} \sum_i \mathbf{h}(\mathbf{w}_i, \widehat{\boldsymbol{\theta}})$. As shown in Section 6.3.9, if $\widehat{\boldsymbol{\theta}}$ is the optimal GMM estimator then $\widehat{\mathbf{m}}_N(\widehat{\boldsymbol{\theta}})' \widehat{\mathbf{S}}_N^{-1} \widehat{\mathbf{m}}_N(\widehat{\boldsymbol{\theta}})$, where $\widehat{\mathbf{S}}_N = N^{-1} \sum_{i=1}^N \widehat{\mathbf{h}}_i \widehat{\mathbf{h}}_i'$, is asymptotically $\chi^2(r - q)$ distributed. A more intuitive linear IV example is given in Section 8.4.4.

8.2.8. Power and Consistency of Conditional Moment Tests

Because there is no explicit alternative hypothesis, m-tests differ from the tests of Chapter 7.

Several authors have given examples where the IM test can be shown to be equivalent to a conventional LM test of null against alternative hypotheses. Chesher (1984) interpreted the IM test as a test for random parameter heterogeneity. For the linear model under normality, A. Hall (1987) showed that subcomponents of the IM test correspond to LM tests of heteroskedasticity, symmetry, and kurtosis. Cameron and

Trivedi (1998) give some additional examples and reference to results for the linear exponential family.

More generally, m-tests can be interpreted in a conditional moment framework as follows. Begin with an added variable test in a linear regression model. Suppose we want to test whether $\beta_2 = \mathbf{0}$ in the model $y = \mathbf{x}'_1\beta_1 + \mathbf{x}'_2\beta_2 + u$. This is a test of $H_0 : E[y - \mathbf{x}'_1\beta_1 | \mathbf{x}] = 0$ against $H_a : E[y - \mathbf{x}'_1\beta_1 | \mathbf{x}] = \mathbf{x}'_2\beta_2$. The most powerful test of $H_0 : \beta_2 = \mathbf{0}$ in regression of $y - \mathbf{x}'_1\beta_1$ on \mathbf{x}_2 is based on the efficient GLS estimator

$$\widehat{\beta}_2 = \left[\sum_{i=1}^N \frac{\mathbf{x}_{2i}\mathbf{x}'_{2i}}{\sigma_i^2} \right]^{-1} \sum_{i=1}^N \frac{\mathbf{x}_{2i}(y_i - \mathbf{x}'_{1i}\beta_1)}{\sigma_i^2},$$

where $\sigma_i^2 = V[y_i | \mathbf{x}_i]$ under H_0 and independence over i is assumed. This test is equivalent to a test based on the second sum alone, which is an m-test of

$$E\left[\frac{\mathbf{x}_{2i}(y_i - \mathbf{x}'_{1i}\beta_1)}{\sigma_i^2}\right] = \mathbf{0}. \quad (8.28)$$

Reversing the process, we can interpret an m-test based on (8.28) as a CM test of $H_0 : E[y - \mathbf{x}'_1\beta_1 | \mathbf{x}] = 0$ against $H_a : E[y - \mathbf{x}'_1\beta_1 | \mathbf{x}] = \mathbf{x}'_2\beta_2$. Also, an m-test based on $E[\mathbf{x}_2(y - \mathbf{x}'_1\beta_1)] = \mathbf{0}$ can be interpreted as a CM test of $H_0 : E[y - \mathbf{x}'_1\beta_1 | \mathbf{x}] = 0$ against $H_a : E[y - \mathbf{x}'_1\beta_1 | \mathbf{x}] = \sigma_{y|\mathbf{x}}^2 \mathbf{x}'_2\beta_2$, where $\sigma_{y|\mathbf{x}}^2 = V[y | \mathbf{x}]$ under H_0 .

More generally, suppose we start with the conditional moment restriction

$$E[r(y_i, \mathbf{x}_i, \boldsymbol{\theta}) | \mathbf{x}_i] = 0, \quad (8.29)$$

for some scalar function $r(\cdot)$. Then an m-test based on the unconditional moment restriction

$$E[\mathbf{g}(\mathbf{x}_i)r(y_i, \mathbf{x}_i, \boldsymbol{\theta})] = \mathbf{0} \quad (8.30)$$

can be interpreted as a CM test with null and alternative hypotheses

$$\begin{aligned} H_0 &: E[r(y_i, \mathbf{x}_i, \boldsymbol{\theta}) | \mathbf{x}_i] = 0, \\ H_a &: E[r(y_i, \mathbf{x}_i, \boldsymbol{\theta}) | \mathbf{x}_i] = \sigma_i^2 \mathbf{g}(\mathbf{x}_i)' \boldsymbol{\gamma}, \end{aligned} \quad (8.31)$$

where $\sigma_i^2 = V[r(y_i, \mathbf{x}_i, \boldsymbol{\theta}) | \mathbf{x}_i]$ under H_0 .

This approach gives a guide to the directions in which a CM test has power. Although (8.30) suggests power is in the general direction of $\mathbf{g}(\mathbf{x})$, from (8.31) a more precise statement is that it is instead the direction of $\mathbf{g}(\mathbf{x})$ multiplied by the variance of $r(y, \mathbf{x}, \boldsymbol{\theta})$. The distinction is important because many cross-section applications this variance is not constant across observations. For further details and references see Cameron and Trivedi (1998), who call this a regression-based CM test. The approach generalizes to vector $\mathbf{r}(\cdot)$, though with more cumbersome algebra.

An m-test is a test of a finite number of moment conditions. It is therefore possible to construct a dgp for which the underlying conditional moment condition, such as that in (8.29), is false yet the moment conditions are satisfied. Then the CM test is inconsistent as it fails to reject with probability one as $N \rightarrow \infty$. Bierens (1990) proposed a way to specify $\mathbf{g}(\mathbf{x})$ in (8.30) that ensures a **consistent conditional moment test**, for tests of functional form in the nonlinear regression model where $r(y, \mathbf{x}, \boldsymbol{\theta}) = y - f(\mathbf{x}, \boldsymbol{\theta})$.

Ensuring the consistency of the test does not, however, ensure that it will have high power against particular alternatives.

8.2.9. m-Tests Example

To illustrate various m-tests we consider the Poisson regression model introduced in Section 5.2, with Poisson density $f(y) = e^{-\mu} \mu^y / y!$ and $\mu = \exp(\mathbf{x}'\boldsymbol{\beta})$.

We wish to test

$$H_0 : E[\mathbf{m}(y, \mathbf{x}, \boldsymbol{\beta})] = \mathbf{0},$$

for various choices of $\mathbf{m}(\cdot)$. This test will be conducted under the assumption that the dgp is indeed the specified Poisson density.

Auxiliary Regressions

Since estimation is by ML we can use the m-test statistic M^* computed as N times the uncentered R^2 from auxiliary regression (8.5), where

$$1 = \widehat{\mathbf{m}}(y_i, \mathbf{x}_i, \widehat{\boldsymbol{\beta}})' \boldsymbol{\delta} + (y_i - \exp(\mathbf{x}'\widehat{\boldsymbol{\beta}})) \mathbf{x}'_i \boldsymbol{\gamma} + u_i, \quad (8.32)$$

since $\widehat{\mathbf{s}} = |\partial \ln f(y)/\partial \boldsymbol{\beta}|_{\widehat{\boldsymbol{\beta}}} = (y - \exp(\mathbf{x}'\widehat{\boldsymbol{\beta}}))\mathbf{x}$ and $\widehat{\boldsymbol{\beta}}$ is the MLE. Under H_0 the test is $\chi^2(\dim(\mathbf{m}))$ distributed.

An alternative is the M^{**} statistic from auxiliary regression

$$1 = \widehat{\mathbf{m}}(y, \mathbf{x}, \mathbf{z}, \widehat{\boldsymbol{\beta}})' \boldsymbol{\delta} + u. \quad (8.33)$$

This test is asymptotically equivalent to LM^* if $\mathbf{m}(\cdot)$ is such that $E[\partial \mathbf{m}/\partial \boldsymbol{\beta}] = \mathbf{0}$, but otherwise it is not chi-squared distributed.

Moments Tested

Correct specification of the conditional mean function, that is, $E[y - \exp(\mathbf{x}'\boldsymbol{\beta})|\mathbf{x}] = 0$, can be tested by an m-test of

$$E[(y - \exp(\mathbf{x}'\boldsymbol{\beta}))\mathbf{z}] = \mathbf{0},$$

where \mathbf{z} may be a function of \mathbf{x} . For the Poisson and other LEF models, \mathbf{z} cannot equal \mathbf{x} because the first-order conditions for $\widehat{\boldsymbol{\beta}}_{ML}$ impose the restriction that $\sum_i (y_i - \exp(\mathbf{x}'_i \widehat{\boldsymbol{\beta}})) \mathbf{x}_i = \mathbf{0}$, leading to $M = 0$ if $\mathbf{z} = \mathbf{x}$. Instead, \mathbf{z} could include squares and cross-products of the regressors.

Correct specification of the variance may also be tested, as the Poisson distribution implies conditional mean-variance equality. Since $V[y|\mathbf{x}] - E[y|\mathbf{x}] = 0$, with $E[y|\mathbf{x}] = \exp(\mathbf{x}'\boldsymbol{\beta})$, this suggests an m-test of

$$E[(y - \exp(\mathbf{x}'\boldsymbol{\beta}))^2 - \exp(\mathbf{x}'\boldsymbol{\beta})\mathbf{x}] = \mathbf{0}.$$

A variation instead tests

$$E[(y - \exp(\mathbf{x}'\boldsymbol{\beta}))^2 - y]\mathbf{x} = \mathbf{0},$$

as $E[y|\mathbf{x}] = \exp(\mathbf{x}'\boldsymbol{\beta})$. Then $\mathbf{m}(\boldsymbol{\beta}) = \{(y - \exp(\mathbf{x}'\boldsymbol{\beta}))^2 - y\}\mathbf{x}$ has the property that $E[\partial\mathbf{m}/\partial\boldsymbol{\beta}] = \mathbf{0}$, so (8.7) holds and the alternative regression (8.33) yields an asymptotically equivalent test to the regression (8.32).

A standard specification test for parametric models is the IM test. For the Poisson density, \mathbf{D} defined in (8.19) becomes $\mathbf{D}(y, \mathbf{x}, \boldsymbol{\beta}) = \{(y - \exp(\mathbf{x}'\boldsymbol{\beta}))^2 - y\}\mathbf{x}\mathbf{x}'$, and we test

$$E[\{(y - \exp(\mathbf{x}'\boldsymbol{\beta}))^2 - y\}\text{Vech}[\mathbf{x}\mathbf{x}']] = \mathbf{0}.$$

Clearly for the Poisson example the IM test is a test of the first and second moment conditions implied by the Poisson model, a result that holds more generally for LEF models. The test statistic M^{**} is asymptotically equivalent to M^* since here $E[\partial\mathbf{m}/\partial\boldsymbol{\beta}] = \mathbf{0}$.

The Poisson assumption can also be tested using a chi-square goodness-of-fit test. For example, since few counts exceed three in the subsequent simulation example, form four cells corresponding to $y = 0, 1, 2$, and 3 or more, where in implementing the test the cell with $y = 3$ or more are dropped because probabilities sum to one. So for $j = 0, \dots, 2$ compute indicator $d_{ij} = 1$ if $y_i = j$ and $d_{ij} = 0$ otherwise and compute predicted probability $\hat{p}_{ij} = e^{-\hat{\mu}_i} \hat{\mu}_i^j / j!$, where $\hat{\mu}_i = \exp(\mathbf{x}'_i \hat{\boldsymbol{\beta}})$. Then test

$$E[(\mathbf{d} - \mathbf{p})] = \mathbf{0},$$

where $\mathbf{d}_i = [d_{i0}, d_{i1}, d_{i2}]$ and $\mathbf{p}_i = [p_{i0}, p_{i1}, p_{i2}]$ by the auxiliary regression (8.33) where $\hat{\mathbf{m}}_i = \mathbf{d}_i - \hat{\mathbf{p}}_i$.

Simulation Results

Data were generated from a Poisson model with mean $E[y|\mathbf{x}] = \exp(\beta_1 + \beta_2 x_2)$, where $x_2 \sim \mathcal{N}[0, 1]$ and $(\beta_1, \beta_2) = (0, 1)$. Poisson ML regression of y on \mathbf{x} for a sample of size 200 yielded

$$\widehat{E}[y|x] = \exp(-0.165 + \frac{1.124x_2}{(0.089)}),$$

where associated standard errors are in parentheses.

The results of the various M-tests are given in Table 8.1.

Table 8.1. Specification m-Tests for Poisson Regression Example^a

Test Type	H_0 where $\mu = \exp(\mathbf{x}'\boldsymbol{\beta})$	M^*	dof	p-value	M^{**}
1. Correct mean	$E[(y - \mu)x_2^2] = 0$	3.27	1	0.07	0.44
2. Variance = mean	$E[\{(y - \mu)^2 - \mu\}\mathbf{x}] = \mathbf{0}$	2.43	2	0.30	1.89
3. Variance = mean	$E[\{(y - \mu)^2 - y\}\mathbf{x}] = \mathbf{0}$	2.43	2	0.30	2.41
4. Information Matrix	$E[\{(y - \mu)^2 - y\}\text{Vech}[\mathbf{x}\mathbf{x}']] = \mathbf{0}$	2.95	3	0.40	2.73
5. Chi-square GOF	$E[\mathbf{d} - \mathbf{p}] = \mathbf{0}$	2.50	3	0.48	0.75

^a The dgp for y is the Poisson distribution with mean parameter $\exp(0 + x_2)$ and sample size $N = 200$. The m-test statistic M^* is chi-squared with degrees of freedom given in the dof column and p-value given in the p-value column. The alternative test statistic M^{**} is valid for tests 3 and 4 only.

8.3. HAUSMAN TEST

As an example of computation of M^* using (8.32) consider the IM test. Since $\mathbf{x} = [1, x_2]'$ and $\text{Vech}[\mathbf{xx}'] = [1, x_2, x_2^2]',$ the auxiliary regression is of 1 on $\{(y - \hat{\mu})^2 - y\},$ $\{(y - \hat{\mu})^2 - y\}x_2,$ $\{(y - \hat{\mu})^2 - y\}x_2^2,$ $(y - \hat{\mu}),$ and $(y - \hat{\mu})x_2$ and yields uncentered $R^2 = 0.01473$ and $N = 200,$ leading to $M^* = 2.95.$ The same value of M^* is obtained directly from the uncentered explained sum of squares of 2.95, and indirectly as N minus 197.05, the residual sum of squares from this regression. The test statistic is $\chi^2(3)$ distributed with $p = 0.40,$ so the null hypothesis is not rejected at significance level 0.05.

For the chi-square goodness-of-fit test the actual frequencies are, respectively, 0.435, 0.255, and 0.110; and the corresponding predicted frequencies are 0.429, 0.241, and 0.124. This yields PCGF = 0.47 using (8.23), but this statistic is not chi-squared as it does not control for error in estimating $\hat{\beta}.$ The auxiliary regression for the correct statistic CGF in (8.27) leads to $M^* = 2.50,$ which is chi-square distributed.

In this simulation all five moment conditions are not rejected at level 0.05 since the p -value for M^* exceeds 0.05. This is as expected, as the data in this simulation example are generated from the specified density so that tests at level 0.05 should reject only 5% of the time. The alternative statistic M^{**} is valid only for tests 3 and 4 since only then does $E[\partial\mathbf{m}/\partial\beta] = \mathbf{0};$ otherwise, it only provides a lower bound for $M.$

8.3. Hausman Test

Tests based on comparisons between two different estimators are called Hausman tests, after Hausman (1978), or Wu–Hausman tests or even Durbin–Wu–Hausman tests after Wu (1973) and Durbin (1954) who proposed similar tests.

8.3.1. Hausman Test

Consider a test for endogeneity of a regressor in a single equation. Two alternative estimators are the OLS and 2SLS estimators, where the 2SLS estimator uses instruments to control for possible endogeneity of the regressor. If there is endogeneity then OLS is inconsistent, so the two estimators will have different probability limit. If there is no endogeneity both estimators are consistent, so the two estimators have the same probability limit. This suggests testing for endogeneity by testing for difference between the OLS and 2SLS estimators, see Section 8.4.3 for further discussion.

More generally, consider two estimators $\hat{\theta}$ and $\tilde{\theta}.$ We consider the testing situation where

$$\begin{aligned} H_0 &: \text{plim}(\hat{\theta} - \tilde{\theta}) = \mathbf{0}, \\ H_a &: \text{plim}(\hat{\theta} - \tilde{\theta}) \neq \mathbf{0}. \end{aligned} \tag{8.34}$$

Assume the difference between the two root- N consistent estimators is also root- N consistent under H_0 with mean $\mathbf{0}$ and a limit normal distribution, so that

$$\sqrt{N}(\hat{\theta} - \tilde{\theta}) \xrightarrow{d} \mathcal{N}[\mathbf{0}, V_H],$$

where V_H denotes the variance matrix in the limiting distribution. Then the Hausman test statistic

$$H = (\hat{\theta} - \tilde{\theta})'(N^{-1}\hat{V}_H)^{-1}(\hat{\theta} - \tilde{\theta}) \quad (8.35)$$

is asymptotically $\chi^2(q)$ distributed under H_0 . We reject H_0 at level α if $H > \chi_{\alpha}^2(q)$.

In some applications, such as tests of endogeneity, $V[\hat{\theta} - \tilde{\theta}]$ is of less than full rank. Then the generalized inverse is used in (8.35) and the chi-square test has degrees of freedom equal to the rank of $V[\hat{\theta} - \tilde{\theta}]$.

The Hausman test can be applied to just a subset of the parameters. For example, interest may lie solely in the coefficient of the possibly endogenous regressor and whether it changes in moving from OLS to 2SLS. Then just one component of θ is used and the test statistic is $\chi^2(1)$ distributed. As in other settings, this test on a subset of parameters can lead to a conclusion different from that of a test on all parameters.

8.3.2. Computation of the Hausman Test

Computing the Hausman test is easy in principle but difficult in practice owing to the need to obtain a consistent estimate of V_H , the limit variance matrix of $\sqrt{N}(\hat{\theta} - \tilde{\theta})$. In general

$$N^{-1}V_H = V[\hat{\theta} - \tilde{\theta}] = V[\hat{\theta}] + V[\tilde{\theta}] - 2\text{Cov}[\hat{\theta}, \tilde{\theta}]. \quad (8.36)$$

The first two quantities are readily computed from the usual output, but the third is not.

Computation for Fully Efficient Estimator under the Null Hypothesis

Although the essential null and alternative hypotheses of the Hausman test are as in (8.34), in applications there is usually a specific null hypothesis model and alternative hypothesis in mind. For example, in comparing OLS and 2SLS estimators the null hypothesis model has all regressors exogenous whereas the alternative hypothesis model permits some regressors to be endogenous.

If $\hat{\theta}$ is the efficient estimator in the null hypothesis model, then $\text{Cov}[\hat{\theta}, \tilde{\theta}] = V[\hat{\theta}]$. For proof see Exercise 8.3. This implies $V[\hat{\theta} - \tilde{\theta}] = V[\tilde{\theta}] - V[\hat{\theta}]$, so

$$H = (\hat{\theta} - \tilde{\theta})' (\hat{V}[\tilde{\theta}] - \hat{V}[\hat{\theta}])^{-1} (\hat{\theta} - \tilde{\theta}). \quad (8.37)$$

This statistic has the considerable advantage of requiring only the estimated asymptotic variance matrices of the parameter estimates $\hat{\theta}$ and $\tilde{\theta}$. It is helpful to use a program that permits saving parameter and variance matrix estimates and computation using matrix commands.

For example, this simplification can be applied to endogeneity tests in a linear regression model if the errors are assumed to be homoskedastic. Then $\hat{\theta}$ is the OLS estimator that is fully efficient under the null hypothesis of no endogeneity, and $\tilde{\theta}$ is the 2SLS estimator. Care is needed, however, to ensure the consistent estimates of the variance matrices are such that $\hat{V}[\tilde{\theta}] - \hat{V}[\hat{\theta}]$ is positive definite (see Ruud, 1984). In

8.3. HAUSMAN TEST

in the OLS–2SLS comparison the variance matrix estimators $\widehat{V}[\tilde{\theta}]$ and $\widehat{V}[\widehat{\theta}]$ should use the same estimate of the error variance σ^2 .

Version (8.37) of the Hausman test is especially easy to calculate by hand if θ is a scalar, or if only one component of the parameter vector is tested. Then

$$H = (\widehat{\theta} - \tilde{\theta})^2 / (\tilde{s}^2 - \widehat{s}^2)$$

is $\chi^2(1)$ distributed, where \widehat{s} and \tilde{s} are the reported standard errors of $\widehat{\theta}$ and $\tilde{\theta}$.

Auxiliary Regressions

In some leading cases the Hausman test can be more simply computed as a standard test for the significance of a subset of regressors in an augmented OLS regression, derived under the assumption that $\widehat{\theta}$ is fully efficient. Examples are given in Section 8.4.3 and in Section 21.4.3.

Robust Hausman Tests

The simpler version (8.37) of the Hausman test, and standard auxiliary regressions, requires the strong distributional assumption that $\widehat{\theta}$ is fully efficient. This is counter to the approach of performing robust inference under relatively weak distributional assumptions.

Direct estimation of $\text{Cov}[\widehat{\theta}, \tilde{\theta}]$ and hence V_H is in principle possible. Suppose $\widehat{\theta}$ and $\tilde{\theta}$ are m-estimators that solve $\sum_i \mathbf{h}_{1i}(\widehat{\theta}) = \mathbf{0}$ and $\sum_i \mathbf{h}_{2i}(\tilde{\theta}) = \mathbf{0}$. Define $\widehat{\delta} = [\widehat{\theta}, \tilde{\theta}]$. Then $V[\widehat{\delta}] = \mathbf{G}_0^{-1} \mathbf{S}_0 (\mathbf{G}_0^{-1})'$, where \mathbf{G}_0 and \mathbf{S}_0 are defined in Section 6.6, with the simplification that here $\mathbf{G}_{12} = \mathbf{0}$. The desired $V[\widehat{\theta} - \tilde{\theta}] = \mathbf{R} V[\widehat{\delta}] \mathbf{R}'$, where $\mathbf{R} = [\mathbf{I}_q, -\mathbf{I}_q]$. Implementation can require additional coding that may be application specific.

A simpler approach is to bootstrap (see Section 11.6.3), though care is needed in some applications to ensure use of the correct degrees of freedom in the chi-square test.

Another possible approach for less than fully efficient $\widehat{\theta}$ is to use an auxiliary regression that is appropriate in the efficient case but to perform the subsets of regressors test using robust standard errors. This robust test is simple to implement and will have power in testing the misspecification of interest, though it may not necessarily be equivalent to the Hausman test that uses the more general form of H given in (8.35). An example is given in Section 21.4.3.

Finally, bounds can be calculated that do not require computation of $\text{Cov}[\widehat{\theta}, \tilde{\theta}]$. For scalar random variables, $\text{Cov}[x, y] \leq s_x s_y$. For the scalar case this suggests an upper bound for H of $N(\widehat{\theta} - \tilde{\theta})^2 / (\tilde{s}^2 + \widehat{s}^2 - 2\tilde{s}\widehat{s})$, where $\tilde{s}^2 = \widehat{V}[\widehat{\theta}]$ and $\widehat{s}^2 = \widehat{V}[\tilde{\theta}]$. A lower bound for H is $N(\widehat{\theta} - \tilde{\theta})^2 / (\tilde{s}^2 + \widehat{s}^2)$, under the assumption that $\widehat{\theta}$ and $\tilde{\theta}$ are positively correlated. In practice, however, these bounds are quite wide.

8.3.3. Power of the Hausman Test

The Hausman test is a quite general procedure that does not explicitly state an alternative hypothesis and therefore need not have high power against particular alternatives.

For example, consider tests of exclusion restrictions in fully parametric models. Denote the null hypothesis $H_0 : \theta_2 = \mathbf{0}$, where θ is partitioned as $(\theta'_1, \theta'_2)'$. An obvious specification test is a Hausman test of the difference $\widehat{\theta}_1 - \tilde{\theta}_1$, where $(\widehat{\theta}_1, \widehat{\theta}_2)$ is the unrestricted MLE and $(\tilde{\theta}_1, \mathbf{0})$ is the restricted MLE of θ . Holly (1982) showed that this Hausman test coincides with a classical test (Wald, LR, or LM) of $H_0 : \mathcal{I}_{11}^{-1} \mathcal{I}_{12} \theta_2 = \mathbf{0}$, where $\mathcal{I}_{ij} = E[\partial^2 \mathcal{L}(\theta_1, \theta_2) / \partial \theta_i \partial \theta_j]$, rather than of $H_0 : \theta_2 = \mathbf{0}$. The two tests coincide if \mathcal{I}_{12} is of full column rank and $\dim(\theta_1) \geq \dim(\theta_2)$, as then $\mathcal{I}_{11}^{-1} \mathcal{I}_{12} \theta_2 = \mathbf{0}$ iff $\theta_2 = \mathbf{0}$. Otherwise, they can differ. Clearly, the Hausman test will have no power against H_0 if the information matrix is block diagonal as then $\mathcal{I}_{12} = \mathbf{0}$. Holly (1987) extended analysis to nonlinear hypotheses.

8.4. Tests for Some Common Misspecifications

In this section we present tests for some common model misspecifications. Attention is focused on test statistics that can be computed using auxiliary regressions, using minimal assumptions to permit inference robust to heteroskedastic errors.

8.4.1. Tests for Omitted Variables

Omitted variables usually lead to inconsistent parameter estimates, except for special cases such as an omitted regressor in the linear model that is uncorrelated with the other regressors. It is therefore important to test for potential omitted variables.

The Wald test is most often used as it is usually no more difficult to estimate the model with omitted variables included than to estimate the restricted model with omitted variables excluded. Furthermore, this test can use robust sandwich standard errors, though this really only makes sense if the estimator retains consistency in situations where robust sandwich errors are necessary.

If attention is restricted to ML estimation an alternative is to estimate models with and without the potentially irrelevant regressors and perform an LR test.

Robust forms of the LM test can be easily computed in some settings. For example, consider a test of $H_0 : \beta_2 = \mathbf{0}$ in the Poisson model with mean $\exp(\mathbf{x}'_1 \beta_1 + \mathbf{x}'_2 \beta_2)$. The LM test statistic is based on the score statistic $\sum_i \mathbf{x}_i \tilde{u}_i$, where $\tilde{u}_i = y_i - \exp(\mathbf{x}'_{1i} \tilde{\beta}_1)$ (see Section 7.3.2). Now a heteroskedastic robust estimate for the variance of $N^{-1/2} \sum_i \mathbf{x}_i u_i$, where $u_i = y_i - E[y_i | \mathbf{x}_i]$, is $N^{-1} \sum_i u_i^2 \mathbf{x}_i \mathbf{x}'_i$, and it can be shown that

$$LM^+ = \left[\sum_{i=1}^n \mathbf{x}_i \tilde{u}_i \right]' \left[\sum_{i=1}^n \tilde{u}_i^2 \mathbf{x}_i \mathbf{x}'_i \right]^{-1} \left[\sum_{i=1}^n \mathbf{x}_i \tilde{u}_i \right]$$

is a robust LM test statistic that does not require the Poisson restriction that $V[u_i | \mathbf{x}_i] = \exp(\mathbf{x}'_{1i} \beta_1)$ under H_0 . This can be computed as N times the uncentered R^2 from regression of 1 on $\mathbf{x}_{1i} \tilde{u}_i$ and $\mathbf{x}_{2i} \tilde{u}_i$. Such robust LM tests are possible more generally for assumed models in the linear exponential family, as the score statistic in such models is again a weighted average of a residual \tilde{u}_i (see Wooldridge, 1991). This class includes OLS, and adaptations are also possible when estimation is by 2SLS or by NLS; see Wooldridge (2002).

8.4.2. Tests for Heteroskedasticity

Parameter estimates in linear or nonlinear regression models of the conditional mean estimated by LS or IV methods retain their consistency in the presence of heteroskedasticity. The only correction needed is to the standard errors of these estimates. This does not require modeling heteroskedasticity, as heteroskedastic-robust standard errors can be computed under minimal distributional assumptions using the result of White (1980). So there is little need to test for heteroskedasticity, unless estimator efficiency is of great concern. Nonetheless, we summarize some results on tests for heteroskedasticity.

We begin with LS estimation of the linear regression model $y = \mathbf{x}'\beta + u$. Suppose heteroskedasticity is modeled by $V[u|\mathbf{x}] = g(\alpha_1 + \mathbf{z}'\alpha_2)$, where \mathbf{z} is usually a subset of \mathbf{x} and $g(\cdot)$ is often the exponential function. The literature focuses on tests of $H_0 : \alpha_2 = \mathbf{0}$ using the LM approach because, unlike Wald and LR tests, these require only OLS estimation of β . The standard LM test of Breusch and Pagan (1979) depends heavily on the assumption of normally distributed errors, as it uses the restriction that $E[u^4|\mathbf{x}^4] = 3\sigma^4$ under H_0 . Koenker (1981) proposed a more robust version of the LM test, NR^2 from regression of \hat{u}_i^2 on 1 and \mathbf{z}_i , where \hat{u}_i is the OLS residual. This test requires the weaker assumption that $E[u^4|\mathbf{x}]$ is constant. Like the Breusch–Pagan test it is invariant to choice of the function $g(\cdot)$. The White (1980a) test for heteroskedasticity is equivalent to this LM test, with $\mathbf{z} = \text{Vech}[\mathbf{x}\mathbf{x}']$. The test can be further generalized to let $E[u^4|\mathbf{x}]$ vary with \mathbf{x} , though constancy may be a reasonable assumption for the test since H_0 already specifies that $E[u^2|\mathbf{x}]$ is constant.

Qualitatively similar results carry over to nonlinear models of the conditional mean that assume a particular form of heteroskedasticity that may be tested for misspecification. For example, the Poisson regression model sets $V[y|\mathbf{x}] = \exp(\mathbf{x}'\beta)$. More generally, for models in the linear exponential family, the quasi-MLE is consistent despite misspecified heteroskedasticity and qualitatively similar results to those here apply. Then valid inference is possible even if the model for heteroskedasticity is misspecified, provided the robust standard errors presented in Section 5.7.4 are used. If one still wishes to test for correct specification of heteroskedasticity then robust LM tests are possible (see Wooldridge, 1991).

Heteroskedasticity can lead to the more serious consequence of inconsistency of parameter estimates in some nonlinear models. A leading example is the Tobit model (see Chapter 16), a linear regression model with normal homoskedastic errors that becomes nonlinear as the result of censoring or truncation. Then testing for heteroskedasticity becomes more important. A model for $V[u|\mathbf{x}]$ can be specified and Wald, LR, or LM tests can be performed or m-tests for heteroskedasticity can be used (see Pagan and Vella, 1989).

8.4.3. Hausman Tests for Endogeneity

Instrumental variables estimators should only be used where there is a need for them, since LS estimators are more efficient if all regressors are exogenous and from Section 4.9 this loss of efficiency can be substantial. It can therefore be useful to test

whether IV methods are needed. A **test for endogeneity of regressors** compares IV estimates with LS estimates. If regressors are endogenous then in the limit these estimates will differ, whereas if regressors are exogenous the two estimators will not differ. Thus large differences between LS and IV estimates can be interpreted as evidence of endogeneity.

This example provides the original motivation for the Hausman test. Consider the linear regression model

$$y = \mathbf{x}'_1 \beta_1 + \mathbf{x}'_2 \beta_2 + u, \quad (8.38)$$

where \mathbf{x}_1 is potentially endogenous and \mathbf{x}_2 is exogenous. Let $\hat{\beta}$ be the OLS estimator and $\tilde{\beta}$ be the 2SLS estimator in (8.38). Assuming homoskedastic errors so that OLS is efficient under the null hypothesis of no endogeneity, a Hausman test of endogeneity of \mathbf{x}_1 can be calculated using the test statistic H defined in (8.37). Because $V[\hat{\beta}] - V[\tilde{\beta}]$ can be shown to be not of full rank, however, a generalized inverse is needed and the degrees of freedom are $\dim(\beta_1)$ rather than $\dim(\beta)$.

Hausman (1978) showed that the test can more simply be implemented by test of $\gamma = \mathbf{0}$ in the augmented OLS regression

$$y = \mathbf{x}'_1 \beta_1 + \mathbf{x}'_2 \beta_2 + \widehat{\mathbf{x}}'_1 \gamma + u,$$

where $\widehat{\mathbf{x}}_1$ is the predicted value of the endogenous regressors \mathbf{x}_1 from reduced form multivariate regression of \mathbf{x}_1 on the instruments \mathbf{z} . Equivalently, we can test $\gamma = \mathbf{0}$ in the augmented OLS regression

$$y = \mathbf{x}'_1 \beta_1 + \mathbf{x}'_2 \beta_2 + \widehat{\mathbf{v}}'_1 \gamma + u,$$

where $\widehat{\mathbf{v}}_1$ is the residual from the reduced form multivariate regression of \mathbf{x}_1 on the instruments \mathbf{z} . Intuition for these tests is that if u in (8.38) is uncorrelated with \mathbf{x}_1 and \mathbf{x}_2 , then $\gamma = \mathbf{0}$. If instead u is correlated with \mathbf{x}_1 , then this will be picked up by significance of additional transformations of \mathbf{x}_1 such as $\widehat{\mathbf{x}}_1$ and $\widehat{\mathbf{v}}_1$.

For cross-section data it is customary to presume heteroskedastic errors. Then the OLS estimator $\hat{\beta}$ is inefficient in (8.38) and the simpler version (8.37) of the Hausman test cannot be used. However, the preceding augmented OLS regressions can still be used, provided $\gamma = \mathbf{0}$ is tested using the heteroskedastic-consistent estimate of the variance matrix. This should actually be equivalent to the Hausman test, as from Davidson and MacKinnon (1993, p. 239) $\widehat{\gamma}_{OLS}$ in these augmented regressions equals $\mathbf{A}_N(\hat{\beta} - \tilde{\beta})$, where \mathbf{A}_N is a full-rank matrix with finite probability limit.

Additional Hausman tests for endogeneity are possible. Suppose $y = \mathbf{x}'_1 \beta_1 + \mathbf{x}'_2 \beta_2 + \mathbf{x}'_3 \beta_3 + u$, where \mathbf{x}_1 is potentially endogenous \mathbf{x}_2 is assumed to be endogenous, and \mathbf{x}_3 is assumed to be exogenous. Then endogeneity of \mathbf{x}_1 can be tested by comparing the 2SLS estimator with just \mathbf{x}_2 instrumented to the 2SLS estimator with both \mathbf{x}_1 and \mathbf{x}_2 instrumented. The Hausman test can also be generalized to nonlinear regression models, with OLS replaced by NLS and 2SLS replaced by NL2SLS. Davidson and MacKinnon (1993) present augmented regressions that can be used to compute the relevant Hausman test, assuming homoskedastic errors. Mroz (1987) provides a good application of endogeneity tests including examples of computation of $V[\hat{\theta} - \tilde{\theta}]$ when $\hat{\theta}$ is not efficient.

8.4.4. OIR Tests for Exogeneity

If an IV estimator is used then the instruments must be exogenous for the IV estimator to be consistent. For just-identified models it is not possible to test for instrument exogeneity. Instead, a priori arguments need to be used to justify instrument validity. Some examples are given in Section 4.8.2. For overidentified models, however, a test for exogeneity of instruments is possible.

We begin with linear regression. Then $y = \mathbf{x}'\beta + u$ and instruments \mathbf{z} are valid if $E[u|\mathbf{z}] = \mathbf{0}$ or if $E[\mathbf{z}u] = \mathbf{0}$. An obvious test of $H_0 : E[\mathbf{z}u] = \mathbf{0}$ is based on departures of $N^{-1} \sum_i \mathbf{z}_i \hat{u}_i$ from zero. In the just-identified case the IV estimator solves $N^{-1} \sum_i \mathbf{z}_i \hat{u}_i = 0$ so this test is not useful. In the overidentified case the overidentifying restrictions test presented in Section 6.3.8 is

$$\text{OIR} = \hat{\mathbf{u}}' \mathbf{Z} \hat{\mathbf{S}}^{-1} \mathbf{Z}' \hat{\mathbf{u}}, \quad (8.39)$$

where $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\beta}$, $\hat{\beta}$ is the optimal GMM estimator that minimizes $\mathbf{u}' \mathbf{Z} \hat{\mathbf{S}}^{-1} \mathbf{Z}' \mathbf{u}$, and $\hat{\mathbf{S}}$ is consistent for $\text{plim } N^{-1} \sum_i u_i^2 \mathbf{z}_i \mathbf{z}_i'$. The OIR test of Hansen (1982) is an extension of a test proposed by Sargan (1958) for linear IV, and the test statistic (8.39) is often called a **Sargan test**. If OIR is large then the moment conditions are rejected and the IV estimator is inconsistent. Rejection of H_0 is usually interpreted as evidence that the instruments \mathbf{z} are endogenous, but it could also be evidence of model misspecification so that in fact $y \neq \mathbf{x}'\beta + u$. In either case rejection indicates problems for the IV estimator.

As formally derived in Section 6.3.9, OIR is distributed as $\chi^2(r - K)$ under H_0 , where $(r - K)$ is the number of overidentifying restrictions. To gain some intuition for this result it is useful to specialize to homoskedastic errors. Then $\hat{\mathbf{S}} = \hat{\sigma}^2 \mathbf{Z}' \mathbf{Z}$, where $\hat{\sigma}^2 = \hat{\mathbf{u}}' \hat{\mathbf{u}} / (N - K)$, so

$$\text{OIR} = \frac{\hat{\mathbf{u}}' \mathbf{P}_{\mathbf{Z}} \hat{\mathbf{u}}}{\hat{\mathbf{u}}' \hat{\mathbf{u}} / (N - K)},$$

where $\mathbf{P}_{\mathbf{Z}} = \mathbf{Z}(\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}'$. Thus OIR is a ratio of quadratic forms in $\hat{\mathbf{u}}$. Under H_0 the numerator has probability limit $\sigma^2(r - K)$ and the denominator has $\text{plim } \hat{\sigma}^2 = \sigma^2$, so the ratio is centered on $r - K$, but this is the mean of a $\chi^2(r - K)$ random variable.

The test statistic in (8.39) extends immediately to nonlinear regression, by simply defining $u_i = y - g(\mathbf{x}, \beta)$ or $u = r(y, \mathbf{x}, \beta)$ as in Section 6.5, and to linear systems and panel estimators by appropriate definition of \mathbf{u} (see Sections 6.9 and 6.10).

For linear IV with homoskedastic errors alternative OIR tests to (8.39) have been proposed. Magdalinos (1988) contrasts a number of these tests. One can also use incremental OIR tests of a subset of overidentifying restrictions.

8.4.5. RESET Test

A common functional form misspecification may involve neglected nonlinearity in some of the regressors. Consider the regression $\mathbf{y} = \mathbf{x}'\beta + \mathbf{u}$, where we assume that the regressors enter linearly and are asymptotically uncorrelated with the error \mathbf{u} . To test for nonlinearity one straightforward approach is to enter power functions of exogenous

variables, most commonly squares, as additional independent regressors and test the statistical significance of these additional variables using a Wald test or an F -test. This requires the investigator to have specific reasons for considering nonlinearity, and clearly the technique will not work for categorical \mathbf{x} variables.

Ramsey (1969) suggested a test of omitted variables from the regression that can be formulated as a test of functional form. The proposal is to fit the initial regression and generate new regressors that are functions of fitted values $\hat{\mathbf{y}} = \mathbf{x}'\hat{\boldsymbol{\beta}}$, such as $\mathbf{w} = [(\mathbf{x}'\hat{\boldsymbol{\beta}})^2, (\mathbf{x}'\hat{\boldsymbol{\beta}})^3, \dots, (\mathbf{x}'\hat{\boldsymbol{\beta}})^p]$. Then estimate the model $\mathbf{y} = \mathbf{x}'\boldsymbol{\beta} + \mathbf{w}'\boldsymbol{\gamma} + \mathbf{u}$, and the test of nonlinearity is the Wald test of p restrictions, $H_0 : \boldsymbol{\gamma} = \mathbf{0}$ against $H_a : \boldsymbol{\gamma} \neq \mathbf{0}$. Typically a low value of p such as 2 or 3 is used. This test can be made robust to heteroskedasticity.

8.5. Discriminating between Nonnested Models

Two models are **nested** if one is a special case of the other; they are **nonnested** if neither can be represented as a special case of the other. Discriminating between nested models is possible using a standard hypothesis test of the parametric restrictions that reduce one model to the other. In the nonnested case, however, alternative methods need to be developed.

The presentation focuses on nonnested model discrimination within the likelihood framework, where results are well developed. A brief discussion of the nonlikelihood case is given in Section 8.5.4. Bayesian methods for model discrimination are presented in Section 13.8.

8.5.1. Information Criteria

Information criteria are log-likelihood criteria with degrees of freedom adjustment. The model with the smallest information criterion is preferred.

The essential intuition is that there exists a tension between model fit, as measured by the maximized log-likelihood value, and the principle of parsimony that favors a simple model. The fit of the model can be improved by increasing model complexity. However, parameters are only added if the resulting improvement in fit sufficiently compensates for loss of parsimony. Note that in this viewpoint it is not necessary that the set of models under consideration should include the “true dgp.” Different information criteria vary in how steeply they penalize model complexity.

Akaike (1973) originally proposed the **Akaike information criterion**

$$\text{AIC} = -2 \ln L + 2q, \quad (8.40)$$

where q is the number of parameters, with the model with lowest AIC preferred. The term *information criterion* is used because the underlying theory, presented more simply in Amemiya (1980), discriminates among models using the Kullback–Liebler information criterion (KLIC).

A considerable number of modifications to AIC have been proposed, all of the form $-2 \ln L + g(q, N)$ for specified penalty function $g(\cdot)$ that exceeds $2q$. The most popular

variation is the **Bayesian information criterion**

$$\text{BIC} = -2 \ln L + (\ln N)q, \quad (8.41)$$

proposed by Schwarz (1978). Schwarz assumed y has density in the exponential family with parameter θ , the j th model has parameter θ_j with $\dim[\theta_j] = q_j < \dim[\theta]$, and the prior across models is a weighted sum of the prior for each θ_j . He showed that under these assumptions maximizing the posterior probability (see Chapter 13) is asymptotically equivalent to choosing the model for which $\ln L - (\ln N)q_j/2$ is largest. Since this is equivalent to minimizing (8.41), the procedure of Schwarz has been labeled the Bayesian information criterion. A refinement of AIC based on minimization of KLIC that is similar to BIC is the **consistent AIC**, $\text{CAIC} = -2 \ln L + (1 + \ln N)q$. Some authors define criteria such as AIC and BIC by additionally dividing by N in the right-hand sides of (8.40) and (8.41).

If model parsimony is important, then BIC is more widely used as the model-size penalty for AIC is relatively low. Consider two nested models with q_1 and q_2 parameters, respectively, where $q_2 = q_1 + h$. An LR test is then possible and favors the larger model at significance level 5% if $2 \ln L$ increases by $\chi^2_{.05}(h)$. AIC favors the larger model if $2 \ln L$ increases by more than $2h$, a lesser penalty for model size than the LR test if $h < 7$. In particular for $h = 1$, that is, one restriction, the LR test uses a 5% critical value of 3.84 whereas AIC uses a much lower value of 2. The BIC favors the larger model if $2 \ln L$ increases by $h \ln N$, a much larger penalty than either AIC or an LR test of size 0.05 (unless N is exceptionally small).

The Bayesian information criterion increases the penalty as sample size increases, whereas traditional hypothesis tests at a significance level such as 5% do not. For nested models with $q_2 = q_1 + 1$ choosing the larger model on the basis of lower BIC is equivalent to using a two-sided t -test critical value of $\sqrt{\ln N}$, which equals 2.15, 3.03, and 3.72, respectively, for $N = 10^2, 10^4$, and 10^6 . By comparison traditional hypothesis tests with size 0.05 use an unchanging critical value of 1.96. More generally, for a $\chi^2(h)$ distributed test statistic the BIC suggests using a critical value of $h \ln N$ rather than the customary $\chi^2_{.05}(h)$.

Given their simplicity, penalized likelihood criteria are often used for selecting “the best model.” However, there is no clear answer as to which criterion, if any, should be preferred. Considerable approximation is involved in deriving the formulas for AIC and related measures, and loss functions other than minimization of KLIC, or maximization of the posterior probability in the case of BIC, might be much more appropriate. From a decision-theoretic viewpoint, the choice of the model from a set of models should depend on the intended use of that model. For example, the purpose of the model may be to summarize the main features of a complex reality, or to predict some outcome, or to test some important hypothesis. In applied work it is quite rare to see an explicit statement of the intended use of an econometric model.

8.5.2. Cox Likelihood Ratio Test of Nonnested Models

Consider choosing between two parametric models. Let model F_θ have density $f(y|\mathbf{x}, \theta)$ and model G_γ have density $g(y|\mathbf{x}, \gamma)$.

A likelihood ratio test of the model F_θ against G_γ is based on

$$\text{LR}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}) \equiv \mathcal{L}_f(\hat{\boldsymbol{\theta}}) - \mathcal{L}_g(\hat{\boldsymbol{\gamma}}) = \sum_{i=1}^N \ln \frac{f(y_i | \mathbf{x}_i, \hat{\boldsymbol{\theta}})}{g(y_i | \mathbf{x}_i, \hat{\boldsymbol{\gamma}})}. \quad (8.42)$$

If G_γ is nested in F_θ then, from Section 7.3.1, $2\text{LR}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}})$ is chi-square distributed under the null hypothesis that $F_\theta = G_\gamma$. However, this result no longer holds if the models are nonnested.

Cox (1961, 1962b) proposed solving this problem in the special case that F_θ is the true model but the models are not nested, by applying a central limit theorem under the assumption that F_θ is the true model.

This approach is computationally awkward to implement if one cannot analytically obtain $E_f[\ln(f(y|\mathbf{x}, \boldsymbol{\theta})/g(y|\mathbf{x}, \boldsymbol{\gamma}))]$, where E_f denotes expectation with respect to the density $f(y|\mathbf{x}, \boldsymbol{\theta})$. Furthermore, if a similar test statistic is obtained with the roles of F_θ and G_γ reversed it is possible to find both that model F_θ is rejected in favor of G_γ and that model G_γ is rejected in favor of F_θ . The test is therefore not necessarily one of model selection as it does not necessarily select one or the other; instead it is a model specification test that zero, one, or two of the models can pass.

The Cox statistic has been obtained analytically in some cases. For nonnested linear regression models $y = \mathbf{x}'\beta + u$ and $y = \mathbf{z}'\gamma + v$ with homoskedastic normally distributed errors (see Pesaran, 1974). For nonnested transformation models $h(y) = \mathbf{x}'\beta + u$ and $g(y) = \mathbf{z}'\gamma + v$, where $h(y)$ and $g(y)$ are known transformations; see Pesaran and Pesaran (1995), who use a simulation-based approach. This permits, for example, discrimination between linear and log-linear parametric models, with $h(\cdot)$ the identity transformation and $g(\cdot)$ the log transformation. Pesaran and Pesaran (1995) apply the idea to choosing between logit and probit models presented in Chapter 14.

8.5.3. Vuong Likelihood Ratio Test of Nonnested Models

Vuong (1989) provided a very general distribution theory for the LR test statistic that covers both nested and nonnested models and more remarkably permits the dgp to be an unknown density that differs from both $f(\cdot)$ and $g(\cdot)$.

The asymptotic results of Vuong, presented here to aid understanding of the variety of tests presented in Vuong's paper, are relatively complex as in some cases the test statistic is a weighted sum of chi-squares with weights that can be difficult to compute.

Vuong proposed a test of

$$H_0 : E_0 \left[\ln \frac{f(y|\mathbf{x}, \boldsymbol{\theta})}{g(y|\mathbf{x}, \boldsymbol{\gamma})} \right] = 0, \quad (8.43)$$

where E_0 denotes expectation with respect to the true dgp $h(y|\mathbf{x})$, which may be unknown. This is equivalent to testing $E_h[\ln(h/g)] - E_h[\ln(h/f)] = 0$, or testing whether the two densities f and g have the same Kullback–Liebler information criterion (see Section 5.7.2). One-sided alternatives are possible with $H_f : E_0[\ln(f/g)] > 0$ and $H_g : E_0[\ln(f/g)] < 0$.

An obvious test of H_0 is an m-test of whether the sample analogue $\text{LR}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\gamma}})$ defined in (8.42) differs from zero. Here the distribution of the test statistic is to be obtained with possibly unknown dgp. This is possible because from Section 5.7.1 the quasi-MLE $\widehat{\boldsymbol{\theta}}$ converges to the pseudo-true value $\boldsymbol{\theta}^*$ and $\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ has a limit normal distribution, with a similar result for the quasi-MLE $\widehat{\boldsymbol{\gamma}}$.

General Result

The resulting distribution of $\text{LR}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\gamma}})$ varies according to whether or not the two models, both possibly incorrect, are equivalent in the sense that $f(y|\mathbf{x}, \boldsymbol{\theta}_*) = g(y|\mathbf{x}, \boldsymbol{\gamma}_*)$, where $\boldsymbol{\theta}_*$ and $\boldsymbol{\gamma}_*$ are the pseudo-true values of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$.

If $f(y|\mathbf{x}, \boldsymbol{\theta}_*) = g(y|\mathbf{x}, \boldsymbol{\gamma}_*)$ then

$$\text{LR}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\gamma}}) \xrightarrow{d} M_{p+q}(\boldsymbol{\lambda}_*), \quad (8.44)$$

where p and q are the dimensions of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ and $M_{p+q}(\boldsymbol{\lambda}_*)$ denotes the cdf of the weighted sum of chi-squared variables $\sum_{j=1}^{p+q} \lambda_{*j} Z_j^2$. The Z_j^2 are iid $\chi^2(1)$ and λ_{*j} are the eigenvalues of the $(p+q) \times (p+q)$ matrix

$$\mathbf{W} = \begin{bmatrix} -\mathbf{B}_f(\boldsymbol{\theta}_*) \mathbf{A}_f(\boldsymbol{\theta}_*)^{-1} & -\mathbf{B}_{fg}(\boldsymbol{\theta}_*, \boldsymbol{\gamma}_*) \mathbf{A}_g(\boldsymbol{\gamma}_*)^{-1} \\ -\mathbf{B}_{gf}(\boldsymbol{\gamma}_*, \boldsymbol{\theta}_*) \mathbf{A}_f(\boldsymbol{\theta}_*)^{-1} & -\mathbf{B}_g(\boldsymbol{\gamma}_*) \mathbf{A}_g(\boldsymbol{\gamma}_*)^{-1} \end{bmatrix}, \quad (8.45)$$

where $\mathbf{A}_f(\boldsymbol{\theta}_*) = E_0[\partial^2 \ln f / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}']$, $\mathbf{B}_f(\boldsymbol{\theta}_*) = E_0[(\partial \ln f / \partial \boldsymbol{\theta})(\partial \ln f / \partial \boldsymbol{\theta}')$, the matrices $\mathbf{A}_g(\boldsymbol{\gamma}_*)$ and $\mathbf{B}_g(\boldsymbol{\gamma}_*)$ are similarly defined for the density $g(\cdot)$, the cross-matrix $\mathbf{B}_{fg}(\boldsymbol{\theta}_*, \boldsymbol{\gamma}_*) = E_0[(\partial \ln f / \partial \boldsymbol{\theta})(\partial \ln g / \partial \boldsymbol{\gamma}')]$, and expectations are with respect to the true dgp. For explanation and derivation of these results see Vuong (1989).

If instead $f(y|\mathbf{x}, \boldsymbol{\theta}_*) \neq g(y|\mathbf{x}, \boldsymbol{\gamma}_*)$, then under H_0

$$N^{-1/2} \text{LR}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\gamma}}) \xrightarrow{d} \mathcal{N}[0, \omega_*^2], \quad (8.46)$$

where

$$\omega_*^2 = V_0 \left[\ln \frac{f(y|\mathbf{x}, \boldsymbol{\theta}_*)}{g(y|\mathbf{x}, \boldsymbol{\gamma}_*)} \right], \quad (8.47)$$

and the variance is with respect to the true dgp. For derivation again see Vuong (1989).

Use of these results varies with whether or not one model is assumed to be correctly specified and with the nesting relationship between the two models.

Vuong differentiated among three types of model comparisons. The models F_θ and G_γ are (1) **nested** with G_γ nested in F_θ if $G_\gamma \subset F_\theta$; (2) **strictly nonnested models** if and only if $F_\theta \cap G_\gamma = \emptyset$ so that neither model can specialize to the other; and (3) **overlapping** if $F_\theta \cap G_\gamma \neq \emptyset$ and $F_\theta \not\subseteq G_\gamma$ and $G_\gamma \not\subseteq F_\theta$. Similar distinctions are made by Pesaran and Pesaran (1995).

Both (2) and (3) are nonnested models, but they require different testing procedures. Examples of strictly nonnested models are linear models with different error distributions and nonlinear regression models with the same error distributions but different functional forms for the conditional mean. For overlapping models some specializations of the two models are equal. An example is linear models with some regressors in common and some regressors not in common.

Nested Models

For nested models it is necessarily the case that $f(y|\mathbf{x}, \boldsymbol{\theta}_*) = g(y|\mathbf{x}, \boldsymbol{\gamma}_*)$. For G_γ nested in $F_{\boldsymbol{\theta}}$, H_0 is tested against $H_f : E_0[\ln(f/g)] > 0$.

For density possibly misspecified the weighted chi-square result (8.44) is appropriate, using the eigenvalues $\widehat{\lambda}_j$ of the sample analogue of \mathbf{W} in (8.45). Alternatively, one can use eigenvalues $\widetilde{\lambda}_j$ of the sample analogue of the smaller matrix

$$\underline{\mathbf{W}} = \mathbf{B}_f(\boldsymbol{\theta}_*)[\mathbf{D}(\boldsymbol{\gamma}_*)\mathbf{A}_g(\boldsymbol{\gamma}_*)^{-1}\mathbf{D}(\boldsymbol{\gamma}_*)' - \mathbf{A}_f(\boldsymbol{\theta}_*)^{-1}],$$

where $\mathbf{D}(\boldsymbol{\gamma}_*) = \partial\phi(\boldsymbol{\gamma}_*)/\partial\boldsymbol{\gamma}$ and the constrained quasi-MLE $\widetilde{\boldsymbol{\theta}} = \phi(\widetilde{\boldsymbol{\gamma}})$, see Vuong (1989). This result provides a robustified version of the standard LR test for nested models.

If the density $f(\cdot)$ is actually correctly specified, or more generally satisfies the IM equality, we get the expected result that $2\text{LR}(\widehat{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\gamma}}) \xrightarrow{d} \chi^2(p-q)$ as then $(p-q)$ of the eigenvalues of \mathbf{W} or $\underline{\mathbf{W}}$ equal one whereas the others equal zero.

Strictly Nonnested Models

For strictly nonnested models it is necessarily the case that $f(y|\mathbf{x}, \boldsymbol{\theta}_*) \neq g(y|\mathbf{x}, \boldsymbol{\gamma}_*)$. The normal distribution result (8.46) is applicable, and a consistent estimate of ω_*^2 is

$$\widehat{\omega}^2 = \frac{1}{N} \sum_{i=1}^N \left(\ln \frac{f(y_i|\mathbf{x}_i, \widehat{\boldsymbol{\theta}})}{g(y_i|\mathbf{x}_i, \widehat{\boldsymbol{\gamma}})} \right)^2 - \left(\frac{1}{N} \sum_{i=1}^N \ln \frac{f(y_i|\mathbf{x}_i, \widehat{\boldsymbol{\theta}})}{g(y_i|\mathbf{x}_i, \widehat{\boldsymbol{\gamma}})} \right)^2. \quad (8.48)$$

Thus form

$$T_{\text{LR}} = N^{-1/2} \text{LR}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\gamma}}) / \widehat{\omega} \xrightarrow{d} \mathcal{N}[0, 1]. \quad (8.49)$$

For tests with critical value c , H_0 is rejected in favor of $H_f : E_0[\ln(f/g)] > 0$ if $T_{\text{LR}} > c$, H_0 is rejected in favor of $H_g : E_0[\ln(f/g)] < 0$ if $T_{\text{LR}} < -c$, and discrimination between the two models is not possible if $|T_{\text{LR}}| < c$. The test can be modified to permit log-likelihood penalties similar to AIC and BIC; see Vuong (1989, p. 316). An asymptotically equivalent statistic to (8.49) replaces $\widehat{\omega}^2$ by $\widetilde{\omega}^2$ equal to just the first term in the right-hand side of (8.48).

This test assumes that both models are misspecified. If instead one of the models is assumed to be correctly specified, the Cox test approach of Section 8.5.2 needs to be used.

Overlapping Models

For overlapping models it is not clear a priori as to whether or not $f(y|\mathbf{x}, \boldsymbol{\theta}_*) = g(y|\mathbf{x}, \boldsymbol{\gamma}_*)$, and one needs to first test this condition.

Vuong (1989) proposes testing whether or not the variance ω_*^2 defined in (8.47) equals zero, since $\omega_*^2 = 0$ if and only if $f(\cdot) = g(\cdot)$. Thus compute $\widehat{\omega}^2$ in (8.48). Under $H_0^\omega : \omega_*^2 = 0$

$$N\widehat{\omega}^2 \xrightarrow{d} M_{p+q}(\boldsymbol{\lambda}_*), \quad (8.50)$$

where the $M_{p+q}(\boldsymbol{\lambda}_*)$ distribution is defined after (8.44). Hypothesis H_0^ω is rejected at level α if $N\widehat{\omega}^2$ exceeds the upper α percentile of the $M_{p+q}(\widehat{\boldsymbol{\lambda}})$ distribution, using the eigenvalues $\widehat{\lambda}_j$ of the sample analogue of \mathbf{W} in (8.45). Alternatively, and more simply, one can test the conditions that $\boldsymbol{\theta}_*$ and $\boldsymbol{\gamma}_*$ must satisfy for $f(\cdot) = g(\cdot)$. Examples are given in Lien and Vuong (1987).

If H_0^ω is not rejected, or the conditions for $f(\cdot) = g(\cdot)$ are not rejected, conclude that it is not possible to discriminate between the two models given the data. If H_0^ω is rejected, or the conditions for $f(\cdot) = g(\cdot)$ are rejected, then test H_0 against H_f or H_g using T_{LR} as detailed in the strictly nonnested case. In this latter case the significance level is at most the maximum of the significance levels for each of the two tests.

This test assumes that both models are misspecified. If instead one of the models is assumed to be correctly specified, then the other model must also be correctly specified for the two models to be equivalent. Thus $f(y|\mathbf{x}, \boldsymbol{\theta}_*) = g(y|\mathbf{x}, \boldsymbol{\gamma}_*)$ under H_0 , and one can directly move to the LR test using the weighted chi-square result (8.44). Let c_1 and c_2 be upper tail and lower tail critical values, respectively. If $2LR(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\gamma}}) > c_1$ then H_0 is rejected in favor of H_f ; if $2LR(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\gamma}}) < c_2$ then H_0 is rejected in favor of H_g ; and the test is otherwise inconclusive.

8.5.4. Other Nonnested Model Comparisons

The preceding methods are restricted to fully parametric models. Methods for discriminating between models that are only partially parameterized, such as linear regression without the assumption of normality, are less clear-cut.

The information criteria of Section 8.5.1 can be replaced by criteria developed using loss functions other than KLIC. A variety of measures corresponding to different loss functions are presented in Amemiya (1980). These measures are often motivated for nested models but may also be applicable to nonnested models.

A simple approach is to compare predictive ability, selecting the model with lowest value of mean-squared error $(N - q)^{-1} \sum_i (y_i - \widehat{y}_i)^2$. For linear regression this is equivalent to choosing the model with highest adjusted R^2 , which is generally viewed as providing too small a penalty for model complexity. An adaptation for nonparametric regression is leave-one-out cross-validation (see Section 9.5.3).

Formal tests to discriminate between nonnested models in the nonlikelihood case often take one of two approaches. **Artificial nesting**, proposed by Davidson and MacKinnon (1984), embeds the two nonnested models into a more general artificial model and leads to so-called J tests and P tests and related tests. The **encompassing principle**, proposed by Mizon and Richard (1986), leads to a quite general framework for testing one model against a competing nonnested model. White (1994) links this approach with CM tests. For a summary of this literature see Davidson and MacKinnon (1993, chapter 11).

8.5.5. Nonnested Models Example

A sample of 100 observations is generated from a Poisson model with mean $E[y|\mathbf{x}] = \exp(\beta_1 + \beta_2 x_2 + \beta_3 x_3)$, where $x_2, x_3 \sim \mathcal{N}[0, 1]$, and $(\beta_1, \beta_2, \beta_3) = (0.5, 0.5, 0.5)$.

Table 8.2. Nonnested Model Comparisons for Poisson Regression Example^a

Test Type	Model 1	Model 2	Conclusion
-2ln L	366.86	352.18	Model 2 preferred
AIC	370.86	358.18	Model 2 preferred
BIC	376.07	366.00	Model 2 preferred
$N\hat{\omega}^2$	7.84 with $p = 0.000$		Can discriminate
$T_{LR} = N^{-1/2}LR/\hat{\omega}$	-0.883 with $p = 0.377$		No model favored

^a $N = 100$. Model 1 is Poisson regression of y on intercept and x_2 . Model 2 is Poisson regression of y on intercept, x_3 , and x_3^2 . The final two rows are for the Vuong test for nonoverlapping models (see the text).

The dependent variable y has sample mean 1.92 and standard deviation 1.84. Two incorrect nonnested models were estimated by Poisson regression:

$$\text{Model 1: } \widehat{E}[y|\mathbf{x}] = \exp(0.608 + 0.291x_2), \quad (8.08) \quad (4.03)$$

$$\text{Model 2: } \widehat{E}[y|\mathbf{x}] = \exp(0.493 + 0.359x_3 + 0.091x_3^2), \quad (5.14) \quad (5.10) \quad (1.78)$$

where t -statistics are given in parentheses.

The first three rows of Table 8.2 give various information criteria, with the model with smallest value preferred. The first does not penalize number of parameters and favors model 2. The second and third measures defined in (8.40) and (8.41) give larger penalty to model 2, which has an additional parameter, but still lead to the larger model 2 being favored.

The final two rows of the Table 8.2 summarize Vuong's test, here a test of overlapping models.

First, test the condition of equality of the densities when evaluated at the pseudotrue values. The statistic $\hat{\omega}^2$ in (8.48) is easily computed given expressions for the densities. The difficult part is computing an estimate of the matrix \mathbf{W} in (8.45). For the Poisson density we can use $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$ defined at the end of Section 5.2.3 and $\widehat{\mathbf{B}}_{fg} = N^{-1} \sum_i (y_i - \widehat{\mu}_{fi}) \mathbf{x}_{fi} \times (y_i - \widehat{\mu}_{gi}) \mathbf{x}_{gi}'$. The eigenvalues of $\widehat{\mathbf{W}}$ are $\lambda_1 = 0.29$, $\lambda_2 = 1.00$, $\lambda_3 = 1.06$, $\lambda_4 = 1.48$, and $\lambda_5 = 2.75$. The p -value for the test statistic $N\hat{\omega}^2$ with distribution given in (8.44) is obtained as the proportion of draws of $\sum_{j=1}^5 \lambda_j z_j^2$, say 10,000 draws, which exceed $N\hat{\omega}^2 = 69.14$. Here $p = 0.000 < 0.05$ and we conclude that it is possible to discriminate between the models. The critical value at level 0.05 in this example equals 16.10, quite a bit higher than $\chi^2_{.05}(5) = 11.07$.

Given discrimination is possible, then the second test can be applied. Here $T_{LR} = -0.883$ favors the second model, since it is negative. However, using a standard normal two-tail test at 5% the difference is not statistically significant. In this example $\hat{\omega}^2$ is quite large, which means the first test statistic $N\hat{\omega}^2$ is large but the second test statistic $N^{-1/2}LR(\widehat{\theta}, \widehat{\gamma})/\widehat{\omega}$ is small.

8.6. Consequences of Testing

In practice more than one test is performed before one reaches a preferred model. This leads to several complications that practitioners usually ignore.

8.6.1. Pretest Estimation

The use of specification tests to choose a model complicates the distribution of an estimator. For example, suppose we choose between two estimators $\hat{\theta}$ and $\tilde{\theta}$ on the basis of a statistical test at 5%. For instance, $\hat{\theta}$ and $\tilde{\theta}$ may be estimators in unrestricted and restricted models. Then the actual estimator is $\theta^+ = w\hat{\theta} + (1 - w)\tilde{\theta}$, where the random variable w takes value 1 if the test favors $\hat{\theta}$ and 0 if the test favors $\tilde{\theta}$. In short, the estimator depends on the restricted and unrestricted estimators and on a random variable w , which in turn depends on the significance level of the test. Hence θ^+ is an estimator with complex properties. This is called a **pretest estimator**, as the estimator is based on an initial test. The distribution of θ^+ has been obtained for the linear regression model under normality and is nonstandard.

In theory statistical inference should be based on the distribution of θ^+ . In practice inference is based on the distribution of $\hat{\theta}$ if $w = 1$ or of $\tilde{\theta}$ if $w = 0$, ignoring the randomness in w . This is done for simplicity, as even in the simplest models the distribution of the estimator becomes intractable when several such tests are performed.

8.6.2. Order of Testing

Different conclusions can be drawn according to the **order** in which tests are conducted.

One possible ordering is from **general to specific** model. For example, one may estimate a general model for demand before testing restrictions from consumer demand theory such as homogeneity and symmetry. Or the cycle may go from **specific to general** model, with regressors added as needed and additional complications such as endogeneity controlled for if present. Such orderings are natural when choosing which regressors to include in a model, but when specification tests are also being performed it is not uncommon to use both general to specific and specific to general orderings in the same study.

A related issue is that of **joint versus separate tests**. For example, the significance of two regressors can be tested by either two individual t -tests of significance or a joint F -test or $\chi^2(2)$ test of significance. A general discussion was given in Section 7.2.7 and an example is given later in Section 18.7.

8.6.3. Data Mining

Taken to its extreme, the extensive use of tests to select a model has been called **data mining** (Lovell, 1983). For example, one may search among several hundred possible

predictors of y and choose just those predictors that are significant at 5% on a two-sided test. Computer programs exist that automate such searches and are commonly used in some branches of applied statistics. Unfortunately, such broad searches will lead to discovery of spurious relationships, since a test with size 0.05 leads to erroneous findings of statistical significance 5% of the time. Lovell pointed out that the application of such a methodology tends to overestimate the goodness-of-fit measures (e.g., R^2) and underestimate the sampling variances of regression coefficients, even when it succeeds in uncovering the variables that feature in the data-generating process. Using standard tests and reporting p -values without taking account of the model-search procedure is misleading because nominal and actual p -values are not the same. White (2001b) and Sullivan, Timmermann, and White (2001) show how to use bootstrap methods to calculate the true statistical significance of regressors. See also P. Hansen (2003).

The motivation for data mining is sometimes to conserve degrees of freedom or to avoid overparameterization (“clutter”). More importantly, many aspects of specification, such as the functional form of covariates, are left unresolved by underlying theory. Given specification uncertainty, justification exists for specification searching (Sargan, 2001). However, care needs to be taken especially if small samples are analyzed and the number of specification searches is large relative to the sample size. When the specification search is sequential, with a large number of steps, and with each step determined by a previous test outcome, the statistical properties of the procedure as a whole are complex and analytically intractable.

8.6.4. A Practical Approach

Applied microeconomics research generally minimizes the problem of pretest estimation by making judicious use of hypothesis tests. Economic theory is used to guide the selection of regressors, to greatly reduce the number of potential regressors. If the sample size is large there is little purpose served by dropping “insignificant” variables. Final results often use regressions that include statistically insignificant regressors for control variables, such as region, industry, and occupation dummies in an earnings regression. Clutter can be avoided by not reporting unimportant coefficients in a full model specification but noting that fact in an appropriate place. This can lead to some loss of precision in estimating the key regressors of interest, such as years of schooling in an earnings regression, but guards against bias caused by erroneously dropping variables that should be included.

Good practice is to use only part of the sample (“training sample”) for specification searches and model selection, and then report results using the preferred model estimated using a completely separate part of the sample (“estimation sample”). In such circumstances pretesting does not affect the distribution of the estimator, if the subsamples are independent. This procedure is usually only implemented when sample sizes are very large, because using less than the full sample in final estimation leads to a loss in estimator precision.

8.7. Model Diagnostics

In this section we discuss goodness-of-fit measures and definitions of residuals in non-linear models. Useful measures are those that reveal model deficiency in some particular dimension.

8.7.1. Pseudo- R^2 Measures

Goodness of fit is interpreted as closeness of fitted values to sample values of the dependent variable.

For linear models with K regressors the most direct measure is the **standard error of the regression**, which is the estimated standard deviation of the error term,

$$s = \left[\frac{1}{N - K} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \right]^{1/2}.$$

For example, a standard error of regression of 0.10 in a log-earnings regression means that approximately 95% of the fitted values are within 0.20 of the actual value of log-earnings, or within 22% of actual earnings using $e^{0.2} \approx 1.22$. This measure is the same as the in-sample root mean squared error where \hat{y}_i is viewed as a forecast of y_i , aside from a degrees of freedom correction. Alternatively, one can use the **mean absolute error** $(N - K)^{-1} \sum_i |y_i - \hat{y}_i|$. The same measures can be used for nonlinear regression models, provided the nonlinear models lead to a predicted value \hat{y}_i of the dependent variable.

A related measure in linear models is R^2 , the **coefficient of multiple determination**. This explains the fraction of variation of the dependent variable explained by the regressors. The statistic R^2 is more commonly reported than s , even though s may be more informative in evaluating the goodness of fit.

A **pseudo- R^2** is an extension of R^2 to nonlinear regression model. There are several interpretations of R^2 in the linear model. These lead to several possible pseudo- R^2 measures that in nonlinear models differ and do not necessarily have the properties of lying between zero and one and increasing as regressors are added. We present several of these measures that, for simplicity, are not adjusted for degrees of freedom.

One approach bases R^2 on decomposition of the total sum of squares (TSS), with

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 + 2 \sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}).$$

The first sum in the right-hand side is the residual sum of squares (RSS) and the second term is the explained sum of squares (ESS). This leads to two possible measures:

$$\begin{aligned} R_{\text{RES}}^2 &= 1 - \text{RSS/TSS}, \\ R_{\text{EXP}}^2 &= \text{ESS/TSS}. \end{aligned}$$

For OLS regression in the linear model with intercept the third sum equals zero, so $R_{\text{RES}}^2 = R_{\text{EXP}}^2$. However, this simplification does not occur in other models and in general $R_{\text{RES}}^2 \neq R_{\text{EXP}}^2$ in nonlinear models. The measure R_{RES}^2 can be less than zero, R_{EXP}^2

can exceed one, and both measures may decrease as regressors are added though R_{RES}^2 will increase for NLS regression of the nonlinear model as then the estimator is minimizing RSS.

A closely related measure uses

$$R_{\text{COR}}^2 = \widehat{\text{Cor}}^2 [y_i, \hat{y}_i],$$

the squared correlation between actual and fitted values. The measure R_{COR}^2 lies between zero and one and equals R^2 in OLS regression for the linear model with intercept. In nonlinear models R_{COR}^2 can decrease as regressors are added.

A third approach uses weighted sums of squares that control for the intrinsic heteroskedasticity of cross-section data. Let $\hat{\sigma}_i^2$ be the fitted conditional variance of y_i , where it is assumed that heteroskedasticity is explicitly modeled as is the case for FGLS and for models such as logit and Poisson. Then we can use

$$R_{\text{WSS}}^2 = 1 - \text{WRSS}/\text{WTSS},$$

where the weighted residual sum of squares $\text{WRSS} = \sum_i (y_i - \hat{y}_i)^2 / \hat{\sigma}_i^2$, $\text{WTSS} = \sum_i (y_i - \hat{\mu})^2 / \hat{\sigma}^2$, and $\hat{\mu}$ and $\hat{\sigma}^2$ are the estimated mean and variance in the intercept-only model. This can be called a Pearson R^2 because WRSS equals the Pearson statistic, which, aside from any finite-sample corrections, should equal N if heteroskedasticity is correctly modeled. Note that R_{WSS}^2 can be less than zero and decrease as regressors are added.

A fourth approach is a generalization of R^2 to objective functions other than the sum of squared residuals. Let $Q_N(\theta)$ denote the objective function being maximized, Q_0 denote its value in the intercept-only model, Q_{fit} denote the value in the fitted model, and Q_{\max} denote the largest possible value of $Q_N(\theta)$. Then the maximum potential gain in the objective function resulting from inclusion of regressors is $Q_{\max} - Q_0$ and the actual gain is $Q_{\text{fit}} - Q_0$. This suggests the measure

$$R_{\text{RG}}^2 = \frac{Q_{\text{fit}} - Q_0}{Q_{\max} - Q_0} = 1 - \frac{Q_{\max} - Q_{\text{fit}}}{Q_{\max} - Q_0},$$

where the subscript RG means **relative gain**. For least-squares estimation the loss function maximized is minus the residual sum of squares. Then $Q_0 = -\text{TSS}$, $Q_{\text{fit}} = -\text{RSS}$, and $Q_{\max} = 0$, so $R_{\text{RG}}^2 = \text{ESS/TSS}$ for OLS or NLS regression. The measure R_{RG}^2 has the advantage of lying between zero and one and increasing as regressors are added. For ML estimation the loss function is $Q_N(\theta) = \ln L_N(\theta)$. Then R_{RG}^2 cannot always be used as in some models there may be no bound on Q_{\max} . For example, for the linear model under normality $L_N(\beta, \sigma^2) \rightarrow \infty$ as $\sigma^2 \rightarrow 0$. For ML and quasi-ML estimation of linear exponential family models, such as logit and Poisson, Q_{\max} is usually known and R_{RG}^2 can be shown to be an R^2 based on the deviance residuals defined in the next section.

A related measure to R_{RG}^2 is $R_Q^2 = 1 - Q_{\text{fit}}/Q_0$. This measure increases as regressors are added. It equals R_{RG}^2 if $Q_{\max} = 0$, which is the case for OLS regression and for binary and multinomial models. Otherwise, for discrete data this measure may have upper bound less than one, whereas for continuous data the measure

may not be bounded between zero and one as the log-likelihood can be negative or positive. For example, for ML estimation with continuous density it is possible that $Q_0 = 1$ and $Q_{\text{fit}} = 4$, leading to $R_Q^2 = -3$, or that $Q_0 = -1$ and $Q_{\text{fit}} = 4$, leading to $R_Q^2 = 5$.

For nonlinear models there is therefore no universal pseudo- R^2 . The most useful measures may be R_{COR}^2 , as correlation coefficients are easily interpreted, and R_{RG}^2 in special cases that Q_{\max} is known. Cameron and Windmeijer (1997) analyze many of the measures and Cameron and Windmeijer (1996) apply these measures to count data models.

8.7.2. Residual Analysis

Microeconomics analysis actually places little emphasis on residual analysis, compared to some other areas of statistics. If data sets are small then there is concern that residual analysis may lead to overfitting of the model. If the data set is large then there is a belief that residual analysis may be unnecessary as a single observation will have little impact on the analysis. We therefore give a brief summary. A more extensive discussion is given in, for example, McCullagh and Nelder (1989) and Cameron and Trivedi (1998, chapter 5). Econometricians have had particular interest in defining residuals in censored and truncated models.

A wide range of residuals have been proposed for nonlinear regression models. Consider a scalar dependent variable y_i with fitted value $\hat{y}_i = \hat{\mu}_i = \mu(\mathbf{x}_i, \hat{\theta})$. The **raw residual** is $r_i = y_i - \hat{\mu}_i$. The **Pearson residual** is the obvious correction for heteroskedasticity $p_i = (y_i - \hat{\mu}_i)/\hat{\sigma}_i$, where $\hat{\sigma}_i$ is an estimate of the conditional variance of y_i . This requires a specification of the variance for y_i , which is done for models such as the Poisson. For an LEF density (see Section 5.7.3) the **deviance residual** is $d_i = \text{sign}(y_i - \hat{\mu}_i)\sqrt{2[l(y_i) - l(\hat{\mu}_i)]}$, where $l(y)$ denotes the log-density of $y|\mu$ evaluated at $\mu = y$ and $l(\hat{\mu})$ denotes evaluation at $\mu = \hat{\mu}$. A motivation for the deviance residual is that the sum of squares of these residuals is the deviance statistic that is the generalization for LEF models of the sum of raw residuals in the linear model. The **Anscombe residual** is defined to be the transformation of y that is closest to normality, then standardized to mean zero and variance 1. This transformation has been obtained for LEF densities.

Small-sample corrections to residuals have been proposed to account for estimation error in $\hat{\mu}_i$. For the linear model this entails division of residuals by $\sqrt{1 - h_{ii}}$, where h_{ii} is the i th diagonal entry in the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$. These residuals are felt to have better finite-sample performance. Since \mathbf{H} has rank K , the number of regressors, the average value of h_{ii} is K/N and values of h_{ii} in excess of $2K/N$ are viewed as having high leverage. These results extend to LEF models with $\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}\mathbf{W}^{1/2}$, where $\mathbf{W} = \text{Diag}[w_{ii}]$ and $w_{ii} = g'(\mathbf{x}_i'\boldsymbol{\beta})/\sigma_i^2$ with $g(\mathbf{x}_i'\boldsymbol{\beta})$ and σ_i^2 the specified conditional mean and variance, respectively. McCullagh and Nelder (1989) provide a summary.

More generally, Cox and Snell (1968) define a **generalized residual** to be any scalar function $r_i = r(y_i, \mathbf{x}_i, \hat{\theta})$ that satisfies some relatively weak conditions. One way that such residuals arise is that many estimators have first-order conditions of the form

$\sum_i \mathbf{g}(x_i, \theta)r(y_i, x_i, \hat{\theta}) = \mathbf{0}$, where y_i appears in the scalar $r(\cdot)$ but not in the vector $\mathbf{g}(\cdot)$. See also White (1994).

For regression models based on a normal latent variable (see Chapters 14 and 16) Chesher and Irish (1987) propose using $E[\varepsilon_i^*|y_i]$ as the residual, where $y_i^* = \mu_i + \varepsilon_i^*$ is the unobserved latent variable and $y_i = g(y_i^*)$ is the observed dependent variable. Particular choices of $g(\cdot)$ correspond to the probit and Tobit models. Gouriéroux et al. (1987) generalize this approach to LEF densities. A natural approach in this context is to treat residuals as missing data, along the lines of the expectation maximum algorithm in Section 10.3.

A common use of residuals is in plots against other variables of interest. Plots of residuals against fitted values can reveal poor model fit; plots of residuals against omitted variables can suggest further regressors to include in the model; and plots of residuals against included regressors can suggest need for a different functional form. It can be helpful to include a nonparametric regression line in such plots, (see Chapter 9). If data take only a few discrete values the plots can be difficult to interpret because of clustering at just a few values, and it can be helpful to use a so-called jitter feature that adds some random noise to the data to reduce the clustering.

Some parametric models imply that an appropriately defined residual should be normally distributed. This can be checked by a normal scores plot that orders residuals r_i from smallest to largest and plots them against the values predicted if the residuals were exactly normally distributed. Thus plot ordered r_i against $\bar{r} + s_r \Phi^{-1}((i - 0.5)/N)$, where \bar{r} and s_r are the sample mean and standard deviation of r and $\Phi^{-1}(\cdot)$ is the inverse of the standard normal cdf.

8.7.3. Diagnostics Example

Table 8.3 uses the same data-generating process as in Section 8.5.5. The dependent variable y has sample mean 1.92 and standard deviation 1.84. Poisson regression of y on x_3 and of y on x_3 and x_3^2 yields

$$\text{Model 1: } \widehat{E}[y|\mathbf{x}] = \exp(0.586 + 0.389x_3), \quad (5.20)$$

$$\text{Model 2: } \widehat{E}[y|\mathbf{x}] = \exp(0.493 + 0.359x_3 + 0.091x_3^2), \quad (5.10)$$

where t -statistics are given in parentheses.

In this example all R^2 measures increase with addition of x_3^2 as regressor, though by quite different amounts given that in this example all but the last R^2 have similar values. More generally the first three R^2 are scaled similarly and R_{RES}^2 and R_{COR}^2 can be quite close, but the remaining three measures are scaled quite differently. Only the last two R^2 measures are guaranteed to increase as a regressor is added, unless the objective function is the sum of squared errors. The measure R_{RG}^2 can be constructed here, as the Poisson log-likelihood is maximized if the fitted mean $\widehat{\mu}_i = y_i$ for all i , leading to $Q_{\max} = \sum_i [y_i \ln y_i - y_i - \ln y_i!]$, where $y \ln y = 0$ when $y = 0$.

Additionally, three residuals were calculated for the second model. The sample mean and standard deviation of residuals were, respectively, 0 and 1.65 for the raw

Table 8.3. *Pseudo R²s: Poisson Regression Example^a*

Diagnostic	Model 1	Model 2	Difference
s where $s^2 = \text{RSS}/(N-K)$	0.1662	0.1661	0.0001
$R_{\text{RES}}^2 = 1 - \text{RSS}/\text{TSS}$	0.1885	0.1962	+0.0077
$R_{\text{EXP}}^2 = \text{ESS}/\text{TSS}$	0.1667	0.2087	+0.0402
$R_{\text{COR}}^2 = \widehat{\text{Cor}}^2 [y_i, \widehat{y}_i]$	0.1893	0.1964	+0.0067
$R_{\text{WSS}}^2 = 1 - \text{WRSS}/\text{WTSS}$	0.1562	0.1695	+0.0233
$R_{\text{RG}}^2 = (Q_{\text{fit}} - Q_0)/(Q_{\text{max}} - Q_0)$	0.1552	0.1712	+0.0160
$R_Q^2 = 1 - Q_{\text{fit}}/Q_0$	0.0733	0.0808	+0.0075

^a $N = 100$. Model 1 is Poisson regression of y on intercept and x_3 . Model 2 is Poisson regression of y on intercept, x_3 , and x_3^2 . RSS is residual sum of squares (SS), ESS is explained SS, TSS is total sum of squares, WRSS is weighted RSS, WTSS is weighted TSS, Q_{fit} is fitted value of objective function, Q_0 is fitted value in intercept-only model, and Q_{max} is the maximum possible value of the objective function given the data and exists only for some objective functions.

residuals, 0.01 and 1.97 for the Pearson residuals, and -0.21 and 1.22 for the deviance residuals. The zero mean for the raw residual is a property of Poisson regression with intercept included that is shared by very few other models. The larger standard deviation of the raw residuals reflects the lack of scaling and the fact that here the standard deviation of y exceeds 1. The correlations between pairs of these residuals all exceed 0.96. This is likely to happen when R^2 is low so that $\widehat{y}_i \simeq \bar{y}$.

8.8. Practical Considerations

m-Tests and Hausman tests are most easily implemented by use of auxiliary regressions. One should be aware that these auxiliary regressions may be valid only under distributional assumptions that are stronger than those made to obtain the usual robust standard errors of regression coefficients. Some robust tests have been presented in Section 8.4.

With a large enough data set and fixed significance level such as 5% the sample moment conditions implied by a model will be rejected, except in the unrealistic case that all aspects of the model–functional form, regressors, and distribution – are correctly specified. In classical testing situations this is often a desired result. In particular, with a large enough sample, regression coefficients will always be significantly different from zero and many studies seek such a result. However, for specification tests the desire is usually to not reject, so that one can say that the model is correctly specified. Perhaps for this reason specification tests are under-utilized.

As an illustration, consider tests of correct specification of life-cycle models of consumption. Unless samples are small a dedicated specification tester is likely to reject the model at 5%. For example, suppose a model specification test statistic is $\chi^2(12)$ distributed when applied to a sample with $N = 3,000$ has a p -value of 0.02. It is not clear that the life-cycle model is providing a poor explanation of the

data, even though it would be formally rejected at the 5% significance level. One possibility is to increase the critical value as sample size increases using BIC (see Section 8.5.1).

Another reason for underutilization of specification tests is difficulty in computation and poor size property of tests when more convenient auxiliary regressions are used to implement an asymptotically equivalent version of a test. These drawbacks can be greatly reduced by use of the bootstrap. Chapter 11 presents bootstrap methods to implement many of the tests given in this chapter.

8.9. Bibliographic Notes

- 8.2** The conditional moment test, due to Newey (1985) and Tauchen (1985), is a generalization of the information matrix test of White (1982). For ML estimation, the computation of the m-test by auxiliary regression generalizes methods of Lancaster (1984) and Chesher (1984) for the IM test. A good overview of m-tests is given in Pagan and Vella (1989). The m-test provides a very general framework for viewing testing. It can be shown to nest all tests, such as Wald, LM, LR, and Hausman tests. This unifying element is emphasized in White (1994).
- 8.3** The Hausman test was proposed by Hausman (1978), with earlier references already given in Section 8.3 and a good survey provided by Ruud (1984).
- 8.4** The econometrics texts by Greene (2003), Davidson and McKinnon (1993) and Wooldridge (2002) present many of the standard specification tests.
- 8.5** Pesaran and Pesaran (1993) discuss how the Cox (1961, 1962b) nonnested test can be implemented when an analytical expression for the expectation of the log-likelihood is not available. Alternatively, the test of Vuong (1989) can be used.
- 8.7** Model diagnostics for nonlinear models are often obtained by extension of results for the linear regression model to generalized linear models such as logit and Poisson models. A detailed discussion with references to the literature is given in Cameron and Trivedi (1998, Chapter 5).

Exercises

- 8–1** Suppose $y = \mathbf{x}'\beta + u$, where $u \sim \mathcal{N}[0, \sigma^2]$, with parameter vector $\theta = [\beta', \sigma^2]'$ and density $f(y|\theta) = (1/\sqrt{2\pi}\sigma) \exp[-(y - \mathbf{x}'\beta)^2/2\sigma^2]$. We have a sample of N independent observations.
 - (a) Explain why a test of the moment condition $E[\mathbf{x}(y - \mathbf{x}'\beta)^3]$ is a test of the assumption of normally distributed errors.
 - (b) Give the expressions for $\hat{\mathbf{m}}_i$ and $\hat{\mathbf{s}}_i$ given in (8.5) necessary to implement the m-test based on the moment condition in part (a).
 - (c) Suppose $\text{dim}[\mathbf{x}] = 10$, $N = 100$, and the auxiliary regression in (8.5) yields an uncentered R^2 of 0.2. What do you conclude at level 0.05?
 - (d) For this example give the moment conditions tested by White's information matrix test.
- 8–2** Consider the multinomial version of the PCGF test given in (8.23) with p_j replaced by $\hat{p}_j = N^{-1} \sum_i F_j(\mathbf{x}_i, \hat{\theta})$. Show that PCGF can be expressed as CGF in (8.27)

with $\widehat{\mathbf{V}} = \text{Diag}[N\widehat{p}_j]$. [Conclude that in the multinomial case Andrew's test statistic simplifies to Pearson's statistic.]

- 8-3** (Adapted from Amemiya, 1985). For the Hausman test given in Section 8.4.1 let $V_{11} = V[\widehat{\theta}]$, $V_{22} = V[\widetilde{\theta}]$, and $V_{12} = \text{Cov}[\widehat{\theta}, \widetilde{\theta}]$.

- (a) Show that the estimator $\bar{\theta} = \widehat{\theta} + [V_{11} + V_{22} - 2V_{12}]^{-1}(\widetilde{\theta} - \widehat{\theta})$ has asymptotic variance matrix $V[\bar{\theta}] = V_{11} - [V_{11} - V_{12}][V_{11} + V_{22} - 2V_{12}]^{-1}[V_{11} - V_{12}]$.
- (b) Hence show that $V[\bar{\theta}]$ is less than $V[\widehat{\theta}]$ in the matrix sense unless $\text{Cov}[\widehat{\theta}, \widetilde{\theta}] = V[\widehat{\theta}]$.
- (c) Now suppose that $\widehat{\theta}$ is fully efficient. Can $V[\bar{\theta}]$ be less than $V[\widehat{\theta}]$? What do you conclude?

- 8-4** Suppose that two models are non-nested and there are $N = 200$ observations. For model 1, the number of parameters $q = 10$ and $\ln L = -400$. For model 3, $q = 10$ and $\ln L = -380$.

- (a) Which model is favored using AIC?
- (b) Which model is favored using BIC?
- (c) Which model would be favored if the models were actually nested and we used a likelihood ratio test at level 0.05?

- 8-5** Use the health expenditure data of Section 16.6. The model is a probit regression of DMED, an indicator variable for positive health expenditures, against the 17 regressors listed in the second paragraph of Section 16.6. You should obtain the estimates given in the first column of Table 16.1.

- (a) Test the joint statistical significance of the self-rated health indicators HLTHG, HLTHF, and HLTHP at level 0.05 using a Hausman test. [This may require some additional coding, depending on the package used.]
- (b) Is the Hausman test the best test to use here?
- (c) Does an information matrix test at level 0.05 support the restrictions of this model? [This will require some additional coding.]
- (d) Discriminate between a model that drops HLTHG, HLTHF, and HLTHP and a model that drops LC, IDP, and LPI on the basis of R^2_{RES} , R^2_{EXP} , R^2_{COR} , and R^2_{RG} .

CHAPTER 9

Semiparametric Methods

9.1. Introduction

In this chapter we present methods for data analysis that require less model specification than the methods of the preceding chapters.

We begin with nonparametric estimation. This makes very minimal assumptions regarding the process that generated the data. One leading example is estimation of a continuous density using a kernel density estimate. This has the attraction of providing a smoother estimate than the familiar histogram. A second leading example is nonparametric regression, such as kernel regression, on a scalar regressor. This places a flexible curve on an (x, y) scatterplot with no parametric restrictions on the form of the curve. Nonparametric estimates have numerous uses, including data description, exploratory analysis of data and of fitted residuals from a regression model, and summary across simulations of parameter estimates obtained from a Monte Carlo study.

Econometric analysis emphasizes multivariate regression of a scalar y on a vector of regressors \mathbf{x} . However, nonparametric methods, although theoretically possible with an infinitely large sample, break down in practice because the data need to be sliced in several dimensions, leading to too few data points in each slice.

As a result econometricians have focused on semiparametric methods. These combine a parametric component, greatly reducing the dimensionality, with a nonparametric component. One important application is to permit more flexible models of the conditional mean. For example, the conditional mean $E[y|\mathbf{x}]$ may be parameterized to be of the single-index form $g(\mathbf{x}'\boldsymbol{\beta})$, where the functional form for $g(\cdot)$ is not specified but is instead nonparametrically estimated, along with the unknown parameters $\boldsymbol{\beta}$. Another important application relaxes distributional assumptions that if misspecified lead to inconsistent parameter estimates. For example, we may wish to obtain consistent estimates of $\boldsymbol{\beta}$ in a linear regression model $y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon$ when data on y are truncated or censored (see Chapter 16), without having to correctly specify the particular distribution of the error term ε .

The asymptotic theory for nonparametric methods differs from that for more parametric methods. Estimates are obtained by cutting the data into ever smaller slices as $N \rightarrow \infty$ and estimating local behavior within each slice. Since less than N observations are being used in estimating each slice the convergence rate is slower than that obtained in the preceding chapters. Nonetheless, in the simplest cases nonparametric estimates are still asymptotically normally distributed. In some leading cases of semiparametric regression the estimators of parameters β have the usual property of converging at rate $N^{-1/2}$, so that scaling by \sqrt{N} leads to a limit normal distribution, whereas the nonparametric component of the model converges at a slower rate N^{-r} , $r < 1/2$.

Because nonparametric methods are local averaging methods, different choices of localness lead to different finite-sample results. In some restrictive cases there are rules or methods to determine the bandwidth or window width used in local averaging, just as there are rules for determining the number of bins in a histogram given the number of observations. In addition, it is common practice to use the nonscientific method of choosing the bandwidth that gives a graph that to the eye looks reasonably smooth yet is still capable of picking up details in the relationship of interest.

Nonparametric methods form the bulk of this chapter, both because they are of intrinsic interest and because they are an essential input for semiparametric methods, presented most notably in the chapters on discrete and censored dependent-variable models. Kernel methods are emphasized as they are relatively simple to present and because “It is argued that all smoothing methods are in an asymptotic sense essentially equivalent to kernel smoothing” (Härdle, 1990, p. xi).

Section 9.2 provides examples of nonparametric density estimation and nonparametric regression applied to data. Kernel density estimation is presented in Section 9.3. Local regression is discussed in Section 9.4, to provide motivation for the formal treatment of kernel regression given in Section 9.5. Section 9.6 presents nonparametric regression methods other than kernel methods. The vast topic of semiparametric regression is then introduced in Section 9.7.

9.2. Nonparametric Example: Hourly Wage

As an example we consider the hourly wage and education for 175 women aged 36 years who worked in 1993. The data are from the Michigan Panel Survey of Income Dynamics. It is easily established that the distribution of the hourly wage is right-skewed and so we model $\ln \text{wage}$, the natural logarithm of the hourly wage.

We give just one example of nonparametric density estimation and one of nonparametric regression and illustrate the important role of bandwidth selection. Sections 9.3 to 9.6 then provide the underlying theory.

9.2.1. Nonparametric Density Estimate

A histogram of the natural logarithm of wage is given in Figure 9.1. To provide detail the bin width is chosen so that there are 30 bins, each of width about 0.20. This is an

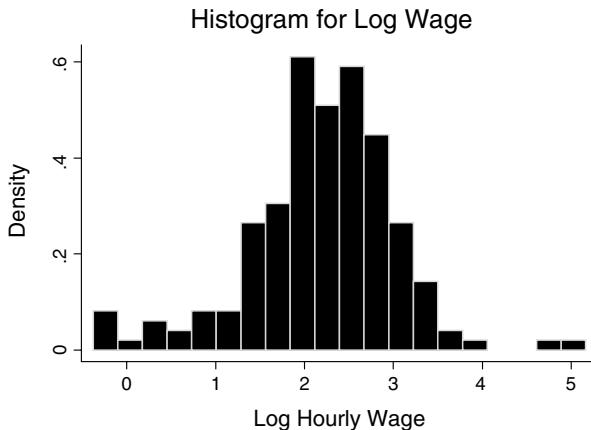


Figure 9.1: Histogram for natural logarithm of hourly wage. Data for 175 U.S. women aged 36 years who worked in 1993.

unusually narrow bin width for only 175 observations, but many details are lost with a larger bin width. The log-wage data seem to be reasonably symmetric, though they are possibly slightly left-skewed.

The standard smoothed nonparametric density estimate is the kernel density estimate defined in (9.3). Here we use the Epanechnikov kernel defined in Table 9.1.

The essential decision in implementation is the choice of bandwidth. For this example Silverman's plug-in estimate defined in (9.13) yields bandwidth of $h = 0.545$. Then the kernel estimate is a weighted average of those observations that have log wage within 0.21 units of the log wage at the current point of evaluation, with greatest weight placed on data closest to the current point of evaluation. Figure 9.2 presents three kernel density estimates, with bandwidths of 0.273, 0.545 and 1.091, respectively

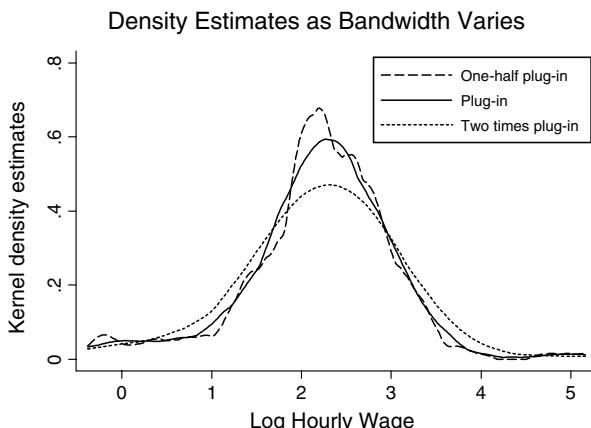


Figure 9.2: Kernel density estimates for log wage for three different bandwidths using the Epanechnikov kernel. The plug-in bandwidth is $h = 0.545$. Same data as Figure 9.1.

corresponding to one-half the plug-in, the plug-in, and two times the plug-in bandwidth. Clearly the smallest bandwidth is too small as it leads to too jagged a density estimate. The largest bandwidth oversmooths the data. The middle bandwidth, the plug-in value of 0.545, seems the best choice. It gives a reasonably smooth density estimate.

What might we do with this kernel density estimate? One possibility is to compare the density to the normal, by superimposing a normal density with mean equal to the sample mean and variance equal to the sample variance. The graph is not reproduced here but reveals that the kernel density estimate with preferred bandwidth 0.545 is considerably more peaked than the normal. A second possibility is to compare log-wage kernel density estimates for different subgroups, such as by educational attainment or by full-time or part-time work status.

9.2.2. Nonparametric Regression

We consider the relationship between log wage and education. The nonparametric method used here is the Lowess local regression method, a local weighted average estimator (see Equation (9.16) and Section 9.6.2).

A local weighted regression line at each point x is fitted using centered subsets that include the closest $0.8N$ observations, the program default, where N is the sample size, and the weights decline as we move away from x . For values of x near the end points, smaller uncentered subsets are used.

Figure 9.3 gives a scatter plot of log wage against education and three Lowess regression curves for bandwidths of 0.8, 0.4 and 0.1. The first two bandwidths give similar curves. The relationship appears to be quadratic, but this may be speculative as the data are relatively sparse at low education levels, with less than 10% of the sample having less than 10 years of schooling. For the majority of the data a linear relationship may also work well. For simplicity we have not presented 95% confidence intervals or bands that might also be provided.

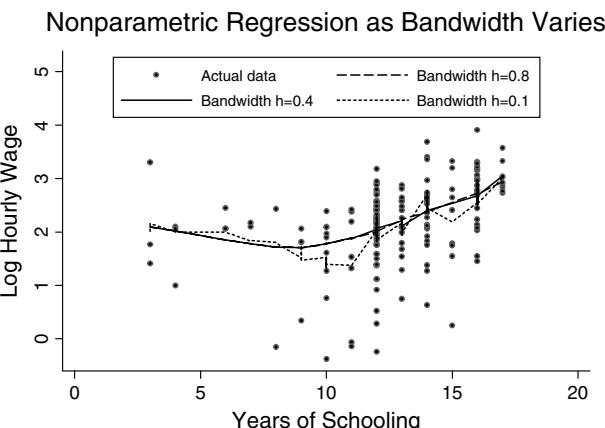


Figure 9.3: Nonparametric regression of log wage on education for three different bandwidths using Lowess regression. Same sample as Figure 9.1.

9.3. Kernel Density Estimation

Nonparametric density estimates are useful for comparison across different groups and for comparison to a benchmark density such as the normal. Compared to a histogram they have the advantage of providing a smoother density estimate. A key decision, analogous to choosing the number of bins in a histogram, is bandwidth choice. We focus on the standard nonparametric density estimator, the kernel density estimator. A detailed presentation is given as results also relevant for regression are more simply obtained for density estimation.

9.3.1. Histogram

A **histogram** is an estimate of the density formed by splitting the range of x into equally spaced intervals and calculating the fraction of the sample in each interval.

We give a more formal presentation of the histogram, one that extends naturally to the smoother kernel density estimator. Consider estimation of the density $f(x_0)$ of a scalar continuous random variable x evaluated at x_0 . Since the density is the derivative of the cdf $F(x_0)$ (i.e., $f(x_0) = dF(x_0)/dx$) we have

$$\begin{aligned} f(x_0) &= \lim_{h \rightarrow 0} \frac{F(x_0 + h) - F(x_0 - h)}{2h} \\ &= \lim_{h \rightarrow 0} \frac{\Pr[x_0 - h < x < x_0 + h]}{2h}. \end{aligned}$$

For a sample $\{x_i, i = 1, \dots, N\}$ of size N , this suggests using the estimator

$$\widehat{f}_{\text{HIST}}(x_0) = \frac{1}{N} \sum_{i=1}^N \frac{\mathbf{1}(x_0 - h < x_i < x_0 + h)}{2h}, \quad (9.1)$$

where the indicator function

$$\mathbf{1}(A) = \begin{cases} 1 & \text{if event } A \text{ occurs,} \\ 0 & \text{otherwise.} \end{cases}$$

The estimator $\widehat{f}_{\text{HIST}}(x_0)$ is a histogram estimate centered at x_0 with bin width $2h$, since it equals the fraction of the sample that lies between $x_0 - h$ and $x_0 + h$ divided by the bin width $2h$. If $\widehat{f}_{\text{HIST}}$ is evaluated over the range of x at equally spaced values of x , each $2h$ units apart, it yields a histogram.

The estimator $\widehat{f}_{\text{HIST}}(x_0)$ gives all observations in $x_0 \pm h$ equal weight as is clear from rewriting (9.1) as

$$\widehat{f}_{\text{HIST}}(x_0) = \frac{1}{Nh} \sum_{i=1}^N \frac{1}{2} \times \mathbf{1}\left(\left|\frac{x_i - x_0}{h}\right| < 1\right). \quad (9.2)$$

This leads to a density estimate that is a step function, even if the underlying density is continuous. Smoother estimates can be obtained by using weighting functions other than the indicator function chosen here.

9.3.2. Kernel Density Estimator

The **kernel density estimator**, introduced by Rosenblatt (1956), generalizes the histogram estimate (9.2) by using an alternative weighting function, so

$$\hat{f}(x_0) = \frac{1}{Nh} \sum_{i=1}^N K\left(\frac{x_i - x_0}{h}\right). \quad (9.3)$$

The weighting function $K(\cdot)$ is called a **kernel function** and satisfies restrictions given in the next section. The parameter h is a smoothing parameter called the **bandwidth**, and two times h is the **window width**. The density is estimated by evaluating $\hat{f}(x_0)$ at a wider range of values of x_0 than used in forming a histogram; usually evaluation is at the sample values x_1, \dots, x_N . This also helps provide a density estimate smoother than a histogram.

9.3.3. Kernel Functions

The kernel function $K(\cdot)$ is a continuous function, symmetric around zero, that integrates to unity and satisfies additional boundedness conditions. Following Lee (1996) we assume that the kernel satisfies the following conditions:

- (i) $K(z)$ is symmetric around 0 and is continuous.
- (ii) $\int K(z)dz = 1$, $\int zK(z)dz = 0$, and $\int |K(z)|dz < \infty$.
- (iii) Either (a) $K(z) = 0$ if $|z| \geq z_0$ for some z_0 or (b) $|z|K(z) \rightarrow 0$ as $|z| \rightarrow \infty$.
- (iv) $\int z^2 K(z)dz = \kappa$, where κ is a constant.

In practice kernel functions work better if they satisfy condition (iiia) rather than just the weaker condition (iiib). Then restricting attention to the interval $[-1, 1]$ rather than $[-z_0, z_0]$ is simply a normalization for convenience, and usually $K(z)$ is restricted to $z \in [-1, 1]$.

Some commonly used kernel functions are given in Table 9.1. The uniform kernel uses the same weights as a histogram of bin width $2h$, except that it produces a running histogram that is evaluated at a series of points x_0 rather than using fixed bins. The Gaussian kernel satisfies (iiib) rather than (iiia) because it does not restrict $z \in [-1, 1]$. A p th-order kernel is one whose first nonzero moment is the p th moment. The first seven kernels are of second order and satisfy the second condition in (ii). The last two kernels are fourth-order kernels. Such **higher order kernels** can increase rates of convergence if $f(x)$ is more than twice differentiable (see Section 9.3.10), though they can take negative values. Table 9.1 also gives the parameter δ , defined in (9.11) and used in Section 9.3.6 to aid bandwidth choice, for some of the kernels.

Given $K(\cdot)$ and h the estimator is very simple to implement. If the kernel estimator is evaluated at r distinct values of x_0 then computation of the kernel estimator requires at most Nr operations, when the kernel has unbounded support. Considerable computational savings on this are possible; see, for example, Härdle (1990, p. 35).

Table 9.1. Kernel Functions: Commonly Used Examples^a

Kernel	Kernel Function $K(z)$	δ
Uniform (or box or rectangular)	$\frac{1}{2} \times \mathbf{1}(z < 1)$	1.3510
Triangular (or triangle)	$(1 - z) \times \mathbf{1}(z < 1)$	–
Epanechnikov (or quadratic)	$\frac{3}{4}(1 - z^2) \times \mathbf{1}(z < 1)$	1.7188
Quartic (or biweight)	$\frac{15}{16}(1 - z^2)^2 \times \mathbf{1}(z < 1)$	2.0362
Triweight	$\frac{35}{32}(1 - z^2)^3 \times \mathbf{1}(z < 1)$	2.3122
Tricubic	$\frac{70}{81}(1 - z ^3)^3 \times \mathbf{1}(z < 1)$	–
Gaussian (or normal)	$(2\pi)^{-1/2} \exp(-z^2/2)$	0.7764
Fourth-order Gaussian	$\frac{1}{2}(3 - z^2)(2\pi)^{-1/2} \exp(-z^2/2)$	–
Fourth-order quartic	$\frac{15}{32}(3 - 10z^2 + 7z^4) \times \mathbf{1}(z < 1)$	–

^a The constant δ is defined in (9.11) and is used to obtain Silverman's plug-in estimate given in (9.13).

9.3.4. Kernel Density Example

The key choice of bandwidth h has already been illustrated in Figure 9.2.

Here we illustrate the choice of kernel using generated data, a random sample of size 100 drawn from the $\mathcal{N}[0, 25^2]$ distribution. For the particular sample drawn the sample mean is 2.81 and the sample standard deviation is 25.27.

Figure 9.4 shows the effect of using different kernels. For Epanechnikov, Gaussian, quartic and uniform kernels, Silverman's plug-in estimate given in (9.13) yields bandwidths of, respectively, 0.545, 0.246, 0.246, and 0.214. The resulting kernel density estimates are very similar, even for the uniform kernel which produces a running histogram. The variation in density estimate with kernel choice is much less than the variation with bandwidth choice evident in Figure 9.2.

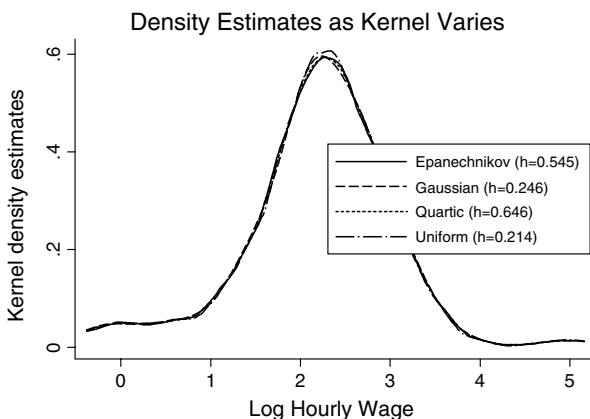


Figure 9.4: Kernel density estimates for log wage for four different kernels using the corresponding Silverman's plug-in estimate for bandwidth. Same data as Figure 9.1.

9.3.5. Statistical Inference

We present the distribution of the kernel density estimator $\hat{f}(x)$ for given choice of $K(\cdot)$ and h , assuming the data x are iid. The estimate $\hat{f}(x)$ is biased. This bias goes to zero asymptotically if the bandwidth $h \rightarrow 0$ as $N \rightarrow \infty$, so $\hat{f}(x)$ is consistent. However, the bias term does not necessarily disappear in the asymptotic normal distribution for $\hat{f}(x)$, complicating statistical inference.

Mean and Variance

The mean and variance of $\hat{f}(x_0)$ are obtained in Section 9.8.1, assuming that the second derivative of $f(x)$ exists and is bounded and that the kernel satisfies $\int z K(z) dz = 0$, as assumed in property (ii) of Section 9.3.3.

The kernel density estimator is biased with **bias term** $b(x_0)$ that depends on the bandwidth, the curvature of the true density, and the kernel used according to

$$b(x_0) = E[\hat{f}(x_0)] - f(x_0) = \frac{1}{2} h^2 f''(x_0) \int z^2 K(z) dz. \quad (9.4)$$

The kernel estimator is biased of size $O(h^2)$, where we use the order of magnitude notation that a function $a(h)$ is $O(h^k)$ if $a(h)/h^k$ is finite. The bias disappears asymptotically if $h \rightarrow 0$ as $N \rightarrow \infty$.

Assuming $h \rightarrow 0$ and $N \rightarrow \infty$, the **variance** of the kernel density estimator is

$$V[\hat{f}(x_0)] = \frac{1}{Nh} f(x_0) \int K(z)^2 dz + o\left(\frac{1}{Nh}\right), \quad (9.5)$$

where a function $a(h)$ is $o(h^k)$ if $a(h)/h^k \rightarrow 0$. The variance depends on the sample size, bandwidth, the true density, and the kernel. The variance disappears if $Nh \rightarrow \infty$, which requires that while $h \rightarrow 0$ it must do so at a slower rate than $N \rightarrow \infty$.

Consistency

The kernel estimator is **pointwise consistent**, that is, consistent at a particular point $x = x_0$, if both the bias and variance disappear. This is the case if $h \rightarrow 0$ and $Nh \rightarrow \infty$.

For estimation of $f(x)$ at all values of x the stronger condition of **uniform convergence**, that is, $\sup_{x_0} |\hat{f}(x_0) - f(x_0)| \xrightarrow{P} 0$, can be shown to occur if $Nh/\ln N \rightarrow \infty$. This requires h larger than for pointwise convergence.

Asymptotic Normality

The preceding results show that asymptotically $\hat{f}(x_0)$ has mean $f(x_0) + b(x_0)$ and variance $(Nh)^{-1} f(x_0) \int K(z)^2 dz$. It follows that if a central limit theorem can be

applied, the kernel density estimator has **limit distribution**

$$\sqrt{Nh}(\widehat{f}(x_0) - f(x_0) - b(x_0)) \xrightarrow{d} \mathcal{N}\left[0, f(x_0) \int K(z)^2 dz\right]. \quad (9.6)$$

The central limit theorem applied is a nonstandard one and requires condition (iv); see, for example, Lee (1996, p. 139) or Pagan and Ullah (1999, p. 40).

It is important to note the presence of the bias term $b(x_0)$, defined in (9.4). For typical choices of bandwidth this term does not disappear, complicating computation of confidence intervals (presented in Section 9.3.7).

9.3.6. Bandwidth Choice

The choice of bandwidth h is much more important than choice of kernel function $K(\cdot)$. There is a tension between setting h small to reduce bias and setting h large to ensure smoothness. A natural metric to use is therefore **mean-squared error (MSE)**, the sum of bias squared and variance.

From (9.4) the bias is $O(h^2)$ and from (9.5) the variance is $O((Nh)^{-1})$. Intuitively MSE is minimized by choosing h so that bias squared and variance are of the same order, so $h^4 = (Nh)^{-1}$, which implies the optimal bandwidth $h = O(N^{-0.2})$ and $\sqrt{Nh} = O(N^{0.4})$. We now give a more formal treatment that includes a practical plug-in estimate for h .

Mean Integrated Squared Error

A **local** measure of the performance of the kernel density estimate at x_0 is the MSE

$$\text{MSE}[\widehat{f}(x_0)] = E[(\widehat{f}(x_0) - f(x_0))^2], \quad (9.7)$$

where the expectation is with respect to the density $f(x)$. Since MSE equals variance plus squared bias, (9.4) and (9.5) yield the MSE of the kernel density estimate

$$\text{MSE}[\widehat{f}(x_0)] \simeq \frac{1}{Nh} f(x_0) \int K(z)^2 dz + \left\{ \frac{1}{2} h^2 f''(x_0) \int z^2 K(z) dz \right\}^2. \quad (9.8)$$

To obtain a **global** measure of performance at all values of x_0 we begin by defining the **integrated squared error (ISE)**

$$\text{ISE}(h) = \int (\widehat{f}(x_0) - f(x_0))^2 dx_0, \quad (9.9)$$

the continuous analogue of summing squared error over all x_0 in the discrete case. This is written as a function of h to emphasize dependence on the bandwidth. We then eliminate the dependence of $\widehat{f}(x_0)$ on x values other than x_0 by taking the expected

9.3. KERNEL DENSITY ESTIMATION

value of the ISE with respect to the density $f(x)$. This yields the **mean integrated squared error (MISE)**,

$$\begin{aligned}\text{MISE}(h) &= \mathbb{E} [\text{ISE}(h)] \\ &= \mathbb{E} \left[\int (\hat{f}(x_0) - f(x_0))^2 dx_0 \right] \\ &= \int \mathbb{E}[(\hat{f}(x_0) - f(x_0))^2] dx_0 \\ &= \int \text{MSE}[\hat{f}(x_0)] dx_0,\end{aligned}$$

where $\text{MSE}[\hat{f}(x)]$ is defined in (9.8). From the preceding algebra MISE equals the **integrated mean-squared error (IMSE)**.

Optimal Bandwidth

The optimal bandwidth minimizes MISE. Differentiating $\text{MISE}(h)$ with respect to h and setting the derivative to zero yields the **optimal bandwidth**

$$h^* = \delta \left(\int f''(x_0)^2 dx_0 \right)^{-0.2} N^{-0.2}, \quad (9.10)$$

where δ depends on the kernel function used, with

$$\delta = \left(\frac{\int K(z)^2 dz}{\left(\int z^2 K(z) dz \right)^2} \right)^{0.2}. \quad (9.11)$$

This result is due to Silverman (1986).

Since $h^* = O(N^{-0.2})$, we have $h^* \rightarrow 0$ as $N \rightarrow \infty$ and $Nh^* = O(N^{0.8}) \rightarrow \infty$ as required for consistency. The bias in $\hat{f}(x_0)$ is $O(h^{*2}) = O(N^{-0.4})$, which disappears as $N \rightarrow \infty$. For a histogram estimate it can be shown that $h^* = O(N^{-0.2})$ and $\text{MISE}(h^*) = O(N^{-2/3})$, inferior to $\text{MISE}(h^*) = O(N^{-4/5})$ for the kernel density estimate.

The optimal bandwidth depends on the curvature of the density, with h^* lower if $f(x)$ is highly variable.

Optimal Kernel

The optimal bandwidth varies with the kernel (see (9.10) and (9.11)). It can be shown that $\text{MISE}(h^*)$ varies little across kernels, provided different optimal h^* are used for different kernels (Figure 9.4 provides an illustration). It can be shown that the **optimal kernel** is the Epanechnikov, though this advantage is slight.

Bandwidth choice is much more important than kernel choice and from (9.10) this varies with the kernel.

Plug-in Bandwidth Estimate

A **plug-in estimate** for the bandwidth is a simple formula for h that depends on the sample size N and the sample standard deviation s .

A useful starting point is to assume that the data are normally distributed. Then $\int f''(x_0)^2 dx_0 = 3/(8\sqrt{\pi}\sigma^5) = 0.2116/\sigma^5$, in which case (9.10) specializes to

$$h^* = 1.3643\delta N^{-0.2}, \quad (9.12)$$

where s is the sample standard deviation of x and δ is given in Table 9.1 for several kernels. For the Epanechnikov kernel $h^* = 2.345N^{-0.2}s$, and for the Gaussian kernel $h^* = 1.059N^{-0.2}s$. The considerably lower bandwidth for the normal kernel arises because, unlike most kernels, the normal kernel gives some weight to x_i even if $|x_i - x_0| > h$. In practice one uses **Silverman's plug-in estimate**

$$h^* = 1.3643\delta N^{-0.2} \min(s, iqr/1.349), \quad (9.13)$$

where iqr is the sample interquartile range. This uses $iqr/1.349$ as an alternative estimate of σ that protects against outliers, which can increase s and lead to too large an h .

These **plug-in estimates** for h work well in practice, especially for symmetric unimodal densities, even if $f(x)$ is not the normal density. Nonetheless, one should also check by using variations such as twice and half the plug-in estimate.

For the example in Figures 9.2 and 9.4 we have $177^{-0.2} = 0.3551$, $s = 0.8282$, and $iqr/1.349 = 0.6459$, so (9.13) yields $h^* = 0.3173\delta$. For the Epanechnikov kernel, for example, this yields $h^* = 0.545$ since $\delta = 1.7188$ from Table 9.1.

Cross-Validation

From (9.9), $\text{ISE}(h) = \int \widehat{f}^2(x_0)dx_0 - 2 \int \widehat{f}(x_0)f(x_0)dx_0 + \int f^2(x_0)dx_0$. The third term does not depend on h . An alternative data-driven approach estimates the first two terms in $\text{ISE}(h)$ by

$$\text{CV}(h) = \frac{1}{N^2h} \sum_i \sum_j K^{(2)}\left(\frac{x_i - x_j}{h}\right) - \frac{2}{N} \sum_{i=1}^N \widehat{f}_{-i}(x_i), \quad (9.14)$$

where $K^{(2)}(u) = \int K(u-t)K(t)dt$ is the convolution of K with itself, and $\widehat{f}_{-i}(x_i)$ is the leave-one-out kernel estimator of $f(x_i)$. See Lee (1996, p. 137) or Pagan and Ullah (1999, p. 51) for a derivation. The **cross-validation estimate** h_{CV} is chosen to minimize $\widehat{\text{CV}}(h)$. It can be shown that $h_{\text{CV}} \xrightarrow{P} h^*$ as $N \rightarrow \infty$, but the rate of convergence is very slow.

Obtaining h_{CV} is computationally burdensome because $\widehat{\text{ISE}}(h)$ needs to be computed for a range of values of h . It is often not necessary to cross-validate for kernel density estimation as the plug-in estimate usually provides a good starting point.

9.3.7. Confidence Intervals

Kernel density estimates are usually presented without confidence intervals, but it is possible to construct pointwise confidence intervals for $f(x_0)$, where pointwise means evaluated at a particular value of x_0 . A simple procedure is to obtain confidence intervals at a small number of evaluation points x_0 , say 10, that are evenly distributed over the range of x and plot these along with the estimated density curves.

The result (9.6) yields the following 95% **confidence interval** for $f(x_0)$:

$$f(x_0) \in \widehat{f}(x_0) - b(x_0) \pm 1.96 \times \sqrt{\frac{1}{Nh} \widehat{f}(x_0) \int K(z)^2 dz}.$$

For most kernels $\int K(z)^2 dz$ is easily obtained by analytical methods.

The situation is complicated by the bias term, which should not be ignored in finite samples, even though asymptotically $b(x_0) \xrightarrow{P} 0$. This is because with optimal bandwidth $h^* = O(N^{-0.2})$ the bias of the rescaled random variable $\sqrt{Nh}(\widehat{f}(x_0) - f(x_0))$ given in (9.6) does not disappear, since $\sqrt{Nh^*}$ times $O(h^{*2}) = O(1)$. The bias can be estimated using (9.4) and a kernel estimate of $f''(x_0)$, but in practice the estimate of $f''(x_0)$ is noisy. Instead, the usual method is to reduce the bias in computing the confidence interval, but not $\widehat{f}(x_0)$ itself, by undersmoothing, that is, by choosing $h < h^*$ so that $h^* = o(N^{-0.2})$. Other approaches include using a higher order kernel, such as the fourth-order kernels given in Table 9.1, or bootstrapping (see Section 11.6.5).

One can also compute confidence bands for $f(x)$ over all possible values of x . These are wider than the pointwise confidence intervals for each value x_0 .

9.3.8. Estimation of Derivatives of a Density

In some cases estimates of the **derivatives** of a density need to be made. For example, estimation of the bias term of $\widehat{f}(x_0)$ given in (9.4) requires an estimate of $f''(x_0)$.

For simplicity we present estimates of the first derivative. A finite-difference approach uses $\widehat{f}'(x_0) = [\widehat{f}(x_0 + \Delta) - \widehat{f}(x_0 - \Delta)]/2\Delta$. A calculus approach instead takes the first derivative of $\widehat{f}(x_0)$ in (9.3), yielding $\widehat{f}'(x_0) = -(Nh^2)^{-1} \sum_i K'((x_i - x_0)/h)$.

Intuitively, a larger bandwidth should be used to estimate derivatives, which can be more variable than $f(x_0)$. The bias of $\widehat{f}^{(s)}(x_0)$ is as before but the variance converges more slowly, leading to optimal bandwidth $h^* = O(N^{-1/(2s+2p+1)})$ if $f(x_0)$ is p times differentiable. For kernel estimation of the first derivative we need $p \geq 3$.

9.3.9. Multivariate Kernel Density Estimate

The preceding discussion considered kernel density estimation for scalar x . For the density of the k -dimensional random variable \mathbf{x} , the **multivariate kernel density estimator** is

$$\widehat{f}(\mathbf{x}_0) = \frac{1}{Nh^k} \sum_{i=1}^N K\left(\frac{\mathbf{x}_i - \mathbf{x}_0}{h}\right),$$

where $K(\cdot)$ is now a k -dimensional kernel. Usually $K(\cdot)$ is a **product kernel**, the product of one-dimensional kernels. Multivariate kernels such as the multivariate normal density or spherical kernels proportionate to $K(\mathbf{z}'\mathbf{z})$ can also be used. The kernel $K(\cdot)$ satisfies properties similar to properties given in the one-dimensional case; see Lee (1996, p. 125).

The analytical results and expressions are similar to those before, except that the variance of $\hat{f}(\mathbf{x}_0)$ declines at rate $O(Nh^k)$, which for $k > 1$ is slower than $O(Nh)$ in the one-dimensional case. Then

$$\sqrt{Nh^k}(\hat{f}(\mathbf{x}_0) - f(\mathbf{x}_0) - b(\mathbf{x}_0)) \xrightarrow{d} \mathcal{N}\left[0, f(\mathbf{x}_0) \int K(\mathbf{z})^2 d\mathbf{z}\right].$$

The optimal bandwidth choice is $h = O(N^{-1/(k+4)})$, which is larger than $O(N^{-0.2})$ in the one-dimensional case, and implies $\sqrt{Nh^k} = O(N^{2/(4+k)})$. The plug-in and cross-validation methods can be extended to the multivariate case. For the product normal kernel Scott's plug-in estimate for the j th component of \mathbf{x} is $h_j = N^{-1/(k+4)}s_j$, where s_j is the sample standard deviation of x_j .

Problems of **sparseness** of data are more likely to arise with a multivariate kernel. There is a curse of dimensionality, as fewer observations in the vicinity of \mathbf{x}_0 receive substantial weight when \mathbf{x} is of higher dimension. Even when this is not a problem, plotting even a bivariate kernel density estimate requires a three-dimensional plot that can be difficult to read and interpret.

One use of a multivariate kernel density estimate is to permit estimation of a conditional density. Since $f(y|x) = f(x, y)/f(x)$, an obvious estimator is $\hat{f}(y|x) = \hat{f}(x, y)/\hat{f}(x)$, where $\hat{f}(x, y)$ and $\hat{f}(x)$ are bivariate and univariate kernel density estimates.

9.3.10. Higher Order Kernels

The preceding analysis assumes $f(x)$ is twice differentiable, a necessary assumption to obtain the bias term in (9.4). If $f(x)$ is more than twice differentiable then using higher order kernels (see Section 9.3.3 for fourth-order examples) reduces the order of the bias, leading to smaller h^* and faster rates of convergence. A general statement is that if \mathbf{x} is k dimensional and $f(\mathbf{x})$ is p times differentiable and a p th-order kernel is used, then the kernel estimate $\hat{f}(\mathbf{x}_0)$ of $f(\mathbf{x})$ has optimal rate of convergence $N^{-p/(2p+k)}$ when $h^* = O(N^{-1/(2p+k)})$.

9.3.11. Alternative Nonparametric Density Estimates

The kernel density estimate is the standard nonparametric estimate. Other density estimates are presented, for example, in Pagan and Ullah (1999). These often use approaches such as nearest-neighbors methods that are more commonly used in nonparametric regression and are presented briefly in Section 9.6.

9.4. Nonparametric Local Regression

We consider regression of scalar dependent variable y on a scalar regressor variable x . The regression model is

$$\begin{aligned} y_i &= m(x_i) + \varepsilon_i, \quad i = 1, \dots, N, \\ \varepsilon_i &\sim \text{iid } [0, \sigma_\varepsilon^2]. \end{aligned} \tag{9.15}$$

The complication is that the functional form $m(\cdot)$ is not specified, so NLS estimation is not possible.

This section provides a simple general treatment of **nonparametric regression** using **local weighted averages**. Specialization to kernel regression is given in Section 9.5 and other commonly used local weighted methods are presented in Section 9.6.

9.4.1. Local Weighted Averages

Suppose that for a distinct value of the regressor, say x_0 , there are multiple observations on y , say N_0 observations. Then an obvious simple estimator for $m(x_0)$ is the sample average of these N_0 values of y , which we denote $\tilde{m}(x_0)$. It follows that $\tilde{m}(x_0) \sim [m(x_0), N_0^{-1}\sigma_\varepsilon^2]$, since it is the average of N_0 observations that by (9.15) are iid with mean $m(x_0)$ and variance σ_ε^2 .

The estimator $\tilde{m}(x_0)$ is unbiased but not necessarily consistent. Consistency requires $N_0 \rightarrow \infty$ as $N \rightarrow \infty$, so that $V[\tilde{m}(x_0)] \rightarrow 0$. With discrete regressors this estimator may be very noisy in finite samples because N_0 may be small. Even worse, for continuous regressors there may be only one observation for which x_i takes the particular value x_0 , even as $N \rightarrow \infty$.

The problem of sparseness in data can be overcome by averaging observed values of y when x is close to x_0 , in addition to when x exactly equals x_0 . We begin by noting that the estimator $\tilde{m}(x_0)$ can be expressed as a weighted average of the dependent variable, with $\tilde{m}(x_0) = \sum_i w_{i0} y_i$, where the weights w_{i0} equal $1/N_0$ if $x_i = x_0$ and equal 0 if $x_i \neq x_0$. Thus the weights vary with both the evaluation point x_0 and the sample values of the regressors.

More generally we consider the **local weighted average estimator**

$$\hat{m}(x_0) = \sum_{i=1}^N w_{i0,h} y_i, \tag{9.16}$$

where the weights

$$w_{i0,h} = w(x_i, x_0, h)$$

sum to one, so $\sum_i w_{i0,h} = 1$. The weights are specified to increase as x_i becomes closer to x_0 .

The additional parameter h is generic notation for a **window width parameter**. It is defined so that smaller values of h lead to a smaller window and more weight being placed on those observations with x_i close to x_0 . In the specific example of kernel regression, h is the **bandwidth**. Other methods given in Section 9.6 have alternative **smoothing parameters** that play a similar role to h here. As h becomes smaller $\hat{m}(x_0)$

becomes less biased, as only observations close to x_0 are being used, but more variable, as fewer observations are being used.

The OLS predictor for the linear regression model is a weighted average of y_i , since some algebra yields

$$\hat{m}_{\text{OLS}}(x_0) = \sum_{i=1}^N \left\{ \frac{1}{N} + \frac{(x_0 - \bar{x})(x_i - \bar{x})}{\sum_j (x_j - \bar{x})^2} \right\} y_i.$$

The OLS weights, however, can actually increase with increasing distance between x_0 and x_i if, for example, $x_i > x_0 > \bar{x}$. **Local regression** instead uses weights that are decreasing in $|x_i - x_0|$.

9.4.2. K-Nearest Neighbors Example

We consider a simple example, the unweighted average of the y values corresponding to the closest $(k - 1)/2$ observations on x less than x_0 and the closest $(k - 1)/2$ observations on x greater than x_0 .

Order the observations by increasing x values. Then evaluation at $x_0 = x_i$ yields

$$\hat{m}_k(x_i) = \frac{1}{k}(y_{i-(k-1)/2} + \cdots + y_{i+(k-1)/2}),$$

where for simplicity k is odd, and potential modifications caused by ties and values of x_0 close to the end points x_1 or x_N are ignored. This estimator can be expressed as a special case of (9.16) with weight

$$w_{i0,k} = \frac{1}{k} \times \mathbf{1}\left(|i - 0| < \frac{k-1}{2}\right), \quad x_1 < x_2 < \cdots < x_0 < \cdots < x_N.$$

This estimator has many names. We refer to it as a (symmetrized) **k-nearest neighbors estimator** (k -NN), defined in Section 9.6.1. It is also a standard **local running average** or **running mean** or **moving average** of length k centered at x_0 that is used, for example, to plot a time series y against time x . The parameter k plays the role of the window width h in Section 9.4.1, with small k corresponding to small h .

As an illustration, consider data generated from the model

$$\begin{aligned} y_i &= 150 + 6.5x_i - 0.15x_i^2 + 0.001x_i^3 + \varepsilon_i, \quad i = 1, \dots, 100, \\ x_i &= i, \\ \varepsilon_i &\sim \mathcal{N}[0, 25^2]. \end{aligned} \tag{9.17}$$

The mean of y is a cubic in x , with x taking values 1, 2, ..., 100, with turning points at $x = 20$ and $x = 80$. To this is added a normally distributed error term with standard deviation 25.

Figure 9.5 plots the symmetrized k -NN estimator with $k = 5$ and 25. Both moving averages suggest a cubic relationship. The second is smoother than the first but is still quite jagged despite one-quarter of the sample being used to form the average. The OLS regression line is also given on the diagram.

K-Nearest Neighbors Regression as k Varies

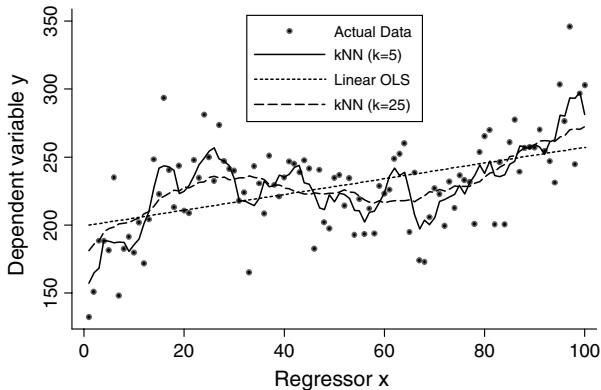


Figure 9.5: k -nearest neighbors regression curve for two different choices of k , as well as OLS regression line. The data are generated from a cubic polynomial model.

The slope of $\hat{m}_k(x)$ is flatter at the end points when $k = 25$ rather than $k = 5$. This illustrates a **boundary problem** in estimating $m(x)$ at the end points. For example, for the smallest regressor value x_1 there are no lower valued observations on x to be included, and the average becomes a one-sided average $\hat{m}_k(x_1) = (y_1 + \dots + y_{1+(k-1)/2})/[(k+1)/2]$. Since for these data $m_k(x)$ is increasing in x in this region, this leads to $\hat{m}_k(x_1)$ being an overestimate and the overstatement is increasing in k . Such boundary problems are reduced by instead using methods given in Section 9.6.2.

9.4.3. Lowess Regression Example

Using alternative weights to those used to form the symmetrized k -NN estimator can lead to better estimates of $m(x)$.

An example is the Lowess estimator, defined in Section 9.6.2. This provides a smoother estimate of $m(x)$ as it uses kernel weights rather than an indicator function, analogous to a kernel density estimate being smoother than a running histogram. It also has smaller bias (see Section 9.6.2), which is especially beneficial in estimating $m(x)$ at the end points.

Figure 9.6 plots, for data generated by (9.17), the Lowess estimate with $k = 25$. This local regression estimate is quite close to the true cubic conditional mean function, which is also drawn. Comparing Figure 9.6 to Figure 9.5 for symmetrized k -NN with $k = 25$, we see that Lowess regression leads to a much smoother regression function estimate and more precise estimation at the boundaries.

9.4.4. Statistical Inference

When the error term is normally distributed and analysis is conditional on x_1, \dots, x_N , the exact small-sample distribution of $\hat{m}(x_0)$ in (9.16) can be easily obtained.

Lowess Nonparametric Regression

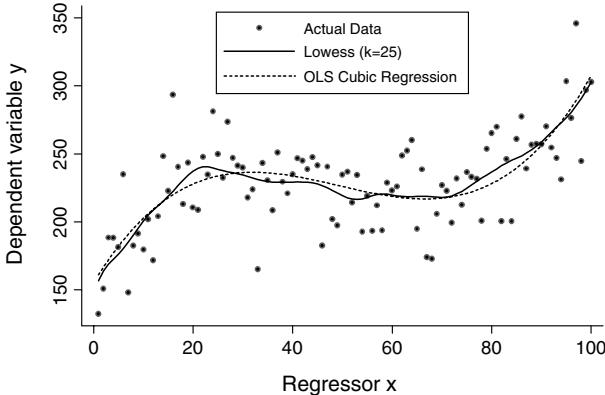


Figure 9.6: Nonparametric regression curve using Lowess, as well as a cubic regression curve. Same generated data as Figure 9.5.

Substituting $y_i = m(x_i) + \varepsilon_i$ into the definition of $\hat{m}(x_0)$ leads directly to

$$\hat{m}(x_0) - \sum_{i=1}^N w_{i0,h} m(x_i) = \sum_{i=1}^N w_{i0,h} \varepsilon_i,$$

which implies with fixed regressors, and if ε_i are iid $\mathcal{N}[0, \sigma_\varepsilon^2]$, that

$$\hat{m}(x_0) \sim \mathcal{N} \left[\sum_{i=1}^N w_{i0,h} m(x_i), \sigma_\varepsilon^2 \sum_{i=1}^N w_{i0,h}^2 \right]. \quad (9.18)$$

Note that in general $\hat{m}(x_0)$ is biased and the distribution is not necessarily centered around $m(x_0)$.

With stochastic regressors and nonnormal errors, we condition on x_1, \dots, x_N and apply a central limit theorem for U-statistics that is appropriate for double summations (see, for example, Pagan and Ullah, 1999, p. 359). Then for ε_i iid $[0, \sigma_\varepsilon^2]$,

$$c(N) \sum_{i=1}^N w_{i0,h} \varepsilon_i \xrightarrow{d} \mathcal{N} \left[0, \sigma_\varepsilon^2 \lim c(N)^2 \sum_{i=1}^N w_{i0,h}^2 \right], \quad (9.19)$$

where $c(N)$ is a function of the sample size with $O(c(N)) < N^{1/2}$ that can vary with the local estimator. For example, $c(N) = \sqrt{Nh}$ for kernel regression and $c(N) = N^{0.4}$ for kernel regression with optimal bandwidth. Then

$$c(N)(\hat{m}(x_0) - m(x_0) - b(x_0)) \xrightarrow{d} \mathcal{N} \left[0, \sigma_\varepsilon^2 \lim c(N)^2 \sum_{i=1}^N w_{i0,h}^2 \right], \quad (9.20)$$

where $b(x_0) = m(x_0) - \sum_i w_{i0,h} m(x_i)$. Note that (9.20) yields (9.18) for the asymptotic distribution of $\hat{m}(x_0)$.

Clearly, the distribution of $\hat{m}(x_0)$, a simple weighted average, can be obtained under alternative distributional assumptions. For example, for **heteroskedastic errors**

the variance in (9.19) and (9.20) is replaced by $\lim c(N)^2 \sum_i \sigma_{\varepsilon,i}^2 w_{i0,h}^2$, which can be consistently estimated by replacing $\sigma_{\varepsilon,i}^2$ by the squared residual $(y_i - \hat{m}(x_i))^2$. Alternatively, one can bootstrap (see Section 11.6.5).

9.4.5. Bandwidth Choice

Throughout this chapter we follow the nonparametric terminology that an estimator $\hat{\theta}$ of θ_0 has **convergence rate** N^{-r} if $\hat{\theta} = \theta_0 + O_p(N^{-r})$, so that $N^r(\hat{\theta} - \theta_0) = O_p(1)$ and ideally $N^r(\hat{\theta} - \theta_0)$ has a limit normal distribution. Note in particular that an estimator that is commonly called a \sqrt{N} -consistent estimator is converging at rate $N^{-1/2}$. Nonparametric estimators typically have a slower rate of convergence than this, with $r < 1/2$, because small bandwidth h is needed to eliminate bias but then less than N observations are being used to estimate $\hat{m}(x_0)$.

As an example, consider the k -NN example of Section 9.4.2. Suppose $k = N^{4/5}$, so that for example $k = 251$ if $N = 1,000$. Then the estimator is consistent as the moving average uses $N^{4/5}/N = N^{-1/5}$ of the sample and is therefore collapsing around x_0 as $N \rightarrow \infty$. Using (9.18), the variance of the moving average estimator is $\sigma_{\varepsilon}^2 \sum_i w_{i0,k}^2 = \sigma_{\varepsilon}^2 \times k \times (1/k)^2 = \sigma_{\varepsilon}^2 \times 1/k = \sigma_{\varepsilon}^2 N^{-4/5}$, so in (9.19) $c(N) = \sqrt{k} = \sqrt{N^{4/5}} = N^{0.4}$, which is less than $N^{1/2}$. Other values of k also ensure consistency, provided $k < O(N)$.

More generally, a range of values of the bandwidth parameter eliminates asymptotic bias, but smaller bandwidth increases variability. In this literature this trade-off is accounted for by minimizing mean-squared error, the sum of variance and bias squared.

Stone (1980) showed that if \mathbf{x} is k dimensional and $m(\mathbf{x})$ is p times differentiable then the **fastest possible rate of convergence** for a nonparametric estimator of an s th-order derivative of $m(\mathbf{x})$ is N^{-r} , where $r = (p-s)/(2p+k)$. This rate decreases as the order of the derivative increases and as the dimension of \mathbf{x} increases. It increases the more differentiable $m(\mathbf{x})$ is assumed to be, approaching $N^{-1/2}$ if $m(\mathbf{x})$ has derivatives of order approaching infinity. For scalar regression estimation of $m(x)$ it is customary to assume existence of $m''(x)$, in which case $r = 2/5$ and the fastest convergence rate is $N^{-0.4}$.

9.5. Kernel Regression

Kernel regression is a weighted average estimator using kernel weights. Issues such as bias and choice of bandwidth presented for kernel density estimation are also relevant here. However, there is less guidance for choice of bandwidth than in the regression case. Also, while we present kernel regression for pedagogical reasons, kernel local regression estimators are often used in practice (see Section 9.6).

9.5.1. Kernel Regression Estimator

The goal in kernel regression is to estimate the regression function $m(x)$ in the model $y = m(x) + \varepsilon$ defined in (9.15).

From Section 9.4.1, an obvious estimator of $m(x_0)$ is the average of the sample values y_i of the dependent variable corresponding to the x_i s close to x_0 . A variation on this is to find the average of the y_i s for all observations with x_i within distance h of x_0 . This can be formally expressed as

$$\widehat{m}(x_0) \equiv \frac{\sum_{i=1}^N \mathbf{1}\left(\left|\frac{x_i-x_0}{h}\right| < 1\right) y_i}{\sum_{i=1}^N \mathbf{1}\left(\left|\frac{x_i-x_0}{h}\right| < 1\right)},$$

where as before $\mathbf{1}(A) = 1$ if event A occurs and equals 0 otherwise. The numerator sums the y values and the denominator gives the number of y values that are summed.

This expression gives equal weights to all observations close to x_0 , but it may be preferable to give the greatest weight at x_0 and decrease the weight as we move away. Thus more generally we consider a kernel weighting function $K(\cdot)$, introduced in Section 9.3.2. This yields the **kernel regression estimator**

$$\widehat{m}(x_0) \equiv \frac{\frac{1}{Nh} \sum_{i=1}^N K\left(\frac{x_i-x_0}{h}\right) y_i}{\frac{1}{Nh} \sum_{i=1}^N K\left(\frac{x_i-x_0}{h}\right)}. \quad (9.21)$$

Several common kernel functions – uniform, Gaussian, Epanechnikov, and quartic – have already been given in Table 9.1.

The constant h is called the **bandwidth**, and $2h$ is called the **window width**. The bandwidth plays the same role as k in the k -NN example of Section 9.4.2.

The estimator (9.21) was proposed by Nadaraya (1964) and Watson (1964), who gave an alternative derivation. The conditional mean $m(x) = \int y f(y|x) dy = \int y [f(y, x)/f(x)] dy$, which can be estimated by $\widehat{m}(x) = \int y [\widehat{f}(y, x)/f(x)] dy$, where $f(y, x)$ and $f(x)$ are bivariate and univariate kernel density estimators. It can be shown that this equals the estimator in (9.21). The statistics literature also considers kernel regression in the **fixed design** or fixed regressors case where $f(x)$ is known and need not be estimated, whereas we consider only the case of **stochastic regressors** that arises with observational data.

The kernel regression estimator is a special case of the weighted average (9.16), with weights

$$w_{i0,h} = \frac{\frac{1}{Nh} K\left(\frac{x_i-x_0}{h}\right)}{\frac{1}{Nh} \sum_{i=1}^N K\left(\frac{x_i-x_0}{h}\right)}, \quad (9.22)$$

which by construction sum over i to one. The general results of Section 9.4 are relevant, but we give a more detailed analysis.

9.5.2. Statistical Inference

We present the distribution of the kernel regression estimator $\widehat{m}(x)$ for given choice of $K(\cdot)$ and h , assuming the data x are iid. We implicitly assume that regressors are continuous. With discrete regressors $\widehat{m}(x_0)$ will still collapse on $m(x_0)$, and both $\widehat{m}(x_0)$ in the limit and $m(x_0)$ are step functions.

Consistency

Consistency of $\widehat{m}(x_0)$ for the conditional mean function $m(x_0)$ requires $h \rightarrow 0$, so that substantial weight is given only to x_i very close to x_0 . At the same time we need many x_i close to x_0 , so that many observations are used in forming the weighted average. Formally, $\widehat{m}(x_0) \xrightarrow{P} m(x_0)$ if $h \rightarrow 0$ and $Nh \rightarrow \infty$ as $N \rightarrow \infty$.

Bias

The kernel regression estimator is biased of size $O(h^2)$, with **bias** term

$$b(x_0) = h^2 \left(m'(x_0) \frac{f'(x_0)}{f(x_0)} + \frac{1}{2} m''(x_0) \right) \int z^2 K(z) dz \quad (9.23)$$

(see Section 9.8.2) assuming $m(x)$ is twice differentiable. As for kernel density estimation, the bias varies with the kernel function used. More importantly, the bias depends on the slope and curvature of the regression function $m(x_0)$ and the slope of the density $f(x_0)$ of the regressors, whereas for density estimation the bias depended only on the second derivatives of $f(x_0)$. The bias can be particularly large at the end points, as illustrated in Section 9.4.2.

The bias can be reduced by using higher order kernels, defined in Section 9.3.3, and boundary modifications such as specific boundary kernels. Local polynomial regression and modifications such as Lowess (see Section 9.6.2) have the attraction that the term in (9.23) depending on $m'(x_0)$ drops out and perform well at the boundaries.

Asymptotic Normality

In Section 9.8.2 it is shown that, for x_i iid with density $f(x_i)$, the kernel regression estimator has **limit distribution**

$$\sqrt{Nh}(\widehat{m}(x_0) - m(x_0) - b(x_0)) \xrightarrow{d} \mathcal{N}\left[0, \frac{\sigma_\varepsilon^2}{f(x_0)} \int K(z)^2 dz\right]. \quad (9.24)$$

The **variance** term in (9.24) is larger for small $f(x_0)$, so as expected the variance of $\widehat{m}(x_0)$ is larger in regions where x is sparse.

9.5.3. Bandwidth Choice

Incorporating values of y_i for which $x_i \neq x_0$ into the weighted average introduces bias, since $E[y_i|x_i] = m(x_i) \neq m(x_0)$ for $x_i \neq x_0$. However, using these additional points reduces the variance of the estimator, since we are averaging over more data. The optimal bandwidth balances the trade-off between increased bias and decreased variance, using squared error loss. Unlike kernel density estimation, plug-in approaches are impractical and cross-validation is used more extensively.

For simplicity most studies focus on choosing one bandwidth for all values of x_0 . Some methods with variable bandwidths, notably k -NN and Lowess, are given in Section 9.6.

Mean Integrated Squared Error

The local performance of $\hat{m}(\cdot)$ at x_0 is measured by the **mean-squared error**, given by

$$\text{MSE}[\hat{m}(x_0)] = \mathbb{E}[(\hat{m}(x_0) - m(x_0))^2],$$

where the expectation eliminates dependence of $\hat{m}(x_0)$ on x . Since MSE equals variance plus squared bias, the MSE can be obtained using (9.23) and (9.24).

Similar to Section 9.3.6, the **integrated square error** is

$$\text{ISE}(h) = \int (\hat{m}(x_0) - m(x_0))^2 f(x_0) dx_0,$$

where $f(x)$ denotes the density of the regressors x , and the **mean integrated square error**, or equivalently the integrated mean-squared error, is

$$\text{MISE}(h) = \int \text{MSE}[\hat{m}(x_0)] f(x_0) dx_0.$$

Optimal Bandwidth

The **optimal bandwidth** h^* minimizes $\text{MISE}(h)$. This yields $h^* = O(N^{-0.2})$ since the bias is $O(h^2)$ from (9.23); the variance is $O((Nh)^{-1})$ from (9.24) since an $O(1)$ variance is obtained after scaling $\hat{m}(x_0)$ by \sqrt{Nh} ; and for bias squared and variance to be of the same order $(h^2)^2 = (Nh)^{-1}$ or $h = N^{-0.2}$. The kernel estimate then converges to $m(x_0)$ at rate $(Nh^*)^{-1/2} = N^{-0.4}$ rather than the usual $N^{-0.5}$ for parametric analysis.

Plug-in Bandwidth Estimate

One can obtain an exact expression for h^* that minimizes $\text{MISE}(h)$, using calculus methods similar to those in Section 9.3.5 for the kernel density estimator. Then h^* depends on the bias and variance expressions in (9.23) and (9.24).

A **plug-in approach** calculates h^* using estimates of these unknowns. However, estimation of $m''(x)$, for example, requires nonparametric methods that in turn require an initial bandwidth choice, but h^* also depends on unknowns such as $m''(x)$. Given these complications one should be wary of plug-in estimates. More common is to use cross-validation, presented in the following.

It can also be shown that $\text{MISE}(h^*)$ is minimized if the Epanichnikov kernel is used (see Härdle, 1990, p. 186, or Härdle and Linton, 1994, p. 2321), though as in the kernel regression case $\text{MISE}(h^*)$ is not much larger for other kernels. The key issue is determination of h^* , which will vary with kernel and the data.

Cross-Validation

An empirical estimate of the optimal h can be obtained by the leave-one-out **cross-validation** procedure. This chooses \hat{h}^* that minimizes

$$\text{CV}(h) = \sum_{i=1}^N (y_i - \hat{m}_{-i}(x_i))^2 \pi(x_i), \quad (9.25)$$

where $\pi(x_i)$ is a weighting function (discussed in the following) and

$$\widehat{m}_{-i}(x_i) = \sum_{j \neq i} w_{ji,h} y_j / \sum_{j \neq i} w_{ji,h} \quad (9.26)$$

is a **leave-one-out estimate** of $m(x_i)$ obtained by the kernel formula (9.21), or more generally by a weighted procedure (9.16), with the modification that y_i is dropped.

Cross-validation is not as computationally intensive as it first appears. It can be shown that

$$y_i - \widehat{m}_{-i}(x_i) = \frac{y_i - \widehat{m}(x_i)}{1 - [w_{ii,h} / \sum_j w_{ji,h}]}, \quad (9.27)$$

so that for each value of h cross-validation requires only one computation of the weighted averages $\widehat{m}(x_i)$, $i = 1, \dots, N$.

The weights $\pi(x_i)$ are introduced to potentially downweight the end points, which otherwise may receive too much importance since local weighted estimates can be quite highly biased at the end points as illustrated in Section 9.4.2. For example, observations with x_i outside the 5th to 95th percentiles may not be used in calculating $\text{CV}(h)$, in which case $\pi(x_i) = 0$ for these observations and $\pi(x_i) = 1$ otherwise. The term cross-validation is used as it validates the ability to predict the i th observation using all the other observations in the data set. The i th observation is dropped because if instead it was additionally used in the prediction, then $\text{CV}(h)$ would be trivially minimized when $\widehat{m}_h(x_i) = y_i$, $i = 1, \dots, N$. $\text{CV}(h)$ is also called the **estimated prediction error**.

Härdle and Marron (1985) showed that minimizing $\text{CV}(h)$ is asymptotically equivalent to minimizing a modification of $\text{ISE}(h)$ and $\text{MISE}(h)$. The modification includes weight function $\pi(x_0)$ in the integrand, as well as the **averaged squared error** (ASE) $N^{-1} \sum_i (\widehat{m}(x_i) - m(x_i))^2 \pi(x_i)$, which is a discrete sample approximation to $\text{ISE}(h)$. The measure $\text{CV}(h)$ converges at the slow rate of $O(N^{-0.1})$ however, so $\text{CV}(h)$ can be quite variable in finite samples.

Generalized Cross-Validation

An alternative to leave-one-out cross validation is to use a measure similar to $\text{CV}(h)$ but one that more simply uses $\widehat{m}(x_i)$ rather than $\widehat{m}_{-i}(x_i)$ and then adds a model complexity penalty that increases as the bandwidth h decreases. This leads to

$$\text{PV}(h) = \sum_{i=1}^N (y_i - \widehat{m}(x_i))^2 \pi(x_i) p(w_{ii,h}),$$

where $p(\cdot)$ is the penalty function and $w_{ii,h}$ is the weight given to the i th observation in $\widehat{m}(x_i) = \sum_j w_{ji,h} y_j$.

A popular example is the **generalized cross-validation measure** that uses the penalty function $p(w_{ii,h}) = (1 - w_{ii,h})^2$. Other penalties are given in Härdle (1990, p. 167) and Härdle and Linton (1994, p. 2323).

Cross-Validation Example

For the local running average example in Section 9.4.2, $\text{CV}(k) = 54,811, 56,666, 63,456, 65,605$, and $69,939$ for $k = 3, 5, 7, 9$, and 25 , respectively. In this case all observations were used to calculate $\text{CV}(k)$, with $\pi(x_i) = 1$, despite possible end-point problems. There is no real gain after $k = 5$, though from Figure 9.5 this value produced too rough an estimate and in practice one would choose a higher value of k to get a smoother curve.

More generally cross-validation is by no means perfect and it is common to “eyeball” fitted nonparametric curves to select h to achieve a desired degree of smoothness.

Trimming

The denominator of the kernel estimator in (9.21) is $\widehat{f}(x_0)$, the kernel estimate of the density of the regressor at x_0 . At some evaluation points $\widehat{f}(x_i)$ can be very small, leading to a very large estimate $\widehat{m}(x_i)$. **Trimming** eliminates or greatly downweights all points with $\widehat{f}(x_i) < b$, say, where $b \rightarrow 0$ at an appropriate rate as $N \rightarrow \infty$. Such problems are most likely to occur in the tails of the distribution. For nonparametric estimation one can just focus on estimation of $m(x_i)$ for more central values of x_i , and values in the tails may be downweighted in cross-validation. However, the semiparametric methods of Section 9.7 can entail computation of $\widehat{m}(x_i)$ at all values of x_i , in which case it is not unusual to trim. Ideally, the trimming function should make no difference asymptotically, though it will make a difference in finite samples.

9.5.4. Confidence Intervals

Kernel regression estimates should generally be presented with pointwise confidence intervals. A simple procedure is to present pointwise confidence intervals for $f(x_0)$ evaluated at, for example, x_0 equal to the first through ninth deciles of x .

If the bias $b(x_0)$ in $\widehat{m}(x_0)$ is ignored, (9.24) yields the following 95% **confidence interval**:

$$m(x_0) \in \widehat{m}(x_0) \pm 1.96 \sqrt{\frac{1}{Nh} \frac{\widehat{\sigma}_\varepsilon^2}{\widehat{f}(x_0)} \int K(z)^2 dz},$$

where $\widehat{\sigma}_\varepsilon^2 = \sum_i w_{i0,h} \widehat{\varepsilon}_i^2$ and $w_{i0,h}$ is defined in (9.22) and $\widehat{f}(x_0)$ is the kernel density estimate at x_0 . This estimate assumes homoskedastic errors, though is likely to be somewhat robust to heteroskedasticity since observations close to x_0 are given the greatest weight. Alternatively, from the discussion after (9.20) a heteroskedastic robust 95% confidence interval is $\widehat{m}(x_0) \pm 1.96 \widehat{s}_0$, where $\widehat{s}_0^2 = \sum_i w_{i0,h}^2 \widehat{\varepsilon}_i^2$.

As in the kernel density case, the bias in $\widehat{m}(x_0)$ should not be ignored. As already noted, estimation of the bias is difficult. Instead, the standard procedure is to undersmooth, with smaller bandwidth h satisfying $h = o(N^{-0.2})$ rather than the optimal $h^* = O(N^{-0.2})$.

Härdle (1990) gives a detailed presentation of confidence intervals, including uniform confidence bands rather than pointwise intervals, and the bootstrap methods given in Section 11.6.5.

9.5.5. Derivative Estimation

In regression we are often interested in how the conditional mean of y changes with changes in x , the **marginal effect**, rather than the conditional mean per se.

Kernel estimates can be easily used to form the derivative. The general result is that the s th derivative of the kernel regression estimate, $\hat{m}^{(s)}(x_0)$, is consistent for $m^{(s)}(x_0)$, the s th derivative of the conditional mean $m(x_0)$. Either calculus or finite-difference approaches can be taken.

As an example, consider estimation of the first derivative in the generated-data example of the previous section. Let z_1, \dots, z_N denote the ordered points at which the kernel regression function is evaluated and $\hat{m}(z_1), \dots, \hat{m}(z_N)$ denote the estimates at these points. A finite-difference estimate is $\hat{m}'(z_i) = [\hat{m}(z_i) - \hat{m}(z_{i-1})]/[z_i - z_{i-1}]$. This is plotted in Figure 9.7, along with the true derivative, which for the dgp given in (9.17) is the quadratic $m'(z_i) = 6.5 - 0.30z_i + 0.003z_i^2$. As expected the derivative estimate is somewhat noisy, but it picks up the essentials. Derivative estimates should be based on oversmoothed estimates of the conditional mean. For further details see Pagan and Ullah (1999, chapter 4). Härdle (1990, p. 160) presents adaptation of cross-validation to derivative estimation.

In addition to the local derivative $m'(x_0)$ we may also be interested in the average derivative $E[m'(x)]$. The average derivative estimator given in Section 9.7.4 provides a \sqrt{N} -consistent and asymptotically normal estimate of $E[m'(x)]$.

9.5.6. Conditional Moment Estimation

The kernel regression methods for the conditional mean $E[y|x] = m(x)$ can be extended to nonparametric estimation of other conditional moments.

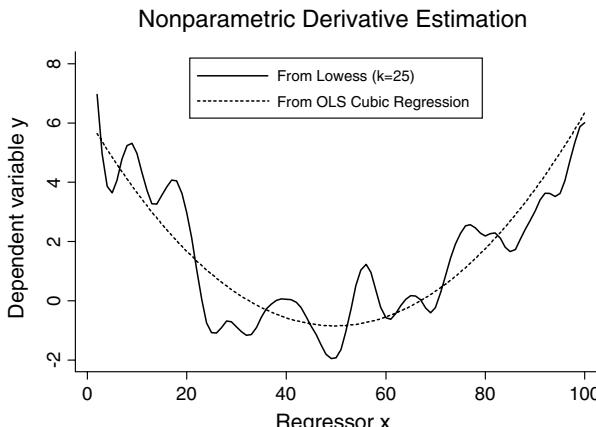


Figure 9.7: Nonparametric derivative estimate using previously estimated Lowess regression curve, as well as using a cubic regression curve. Same generated data as Figure 9.5.

For raw **conditional moments** such as $E[y^k|x]$ we use the weighted average

$$\widehat{E}[y^k|x_0] = \sum_{i=1}^N w_{i0,h} y_i^k, \quad (9.28)$$

where the weights $w_{i0,h}$ may be the same weights as used for estimation of $m(x_0)$.

Central conditional moments can then be computed by reexpressing them as weighted sums of raw moments. For example, since $V[y|x] = E[y^2|x] - (E[y|x])^2$, the conditional variance can be estimated by $\widehat{E}[y^2|x_0] - \widehat{m}(x_0)^2$. One expects that higher order conditional moments will be estimated with more noise than will be the conditional mean.

9.5.7. Multivariate Kernel Regression

We have focused on kernel regression on a single regressor. For regression of scalar y on k -dimensional vector \mathbf{x} , that is, $y_i = m(\mathbf{x}_i) + \varepsilon_i = m(x_{1i}, \dots, x_{ki}) + \varepsilon_i$, the kernel estimator of $m(\mathbf{x}_0)$ becomes

$$\widehat{m}(\mathbf{x}_0) \equiv \frac{\frac{1}{Nh^k} \sum_{i=1}^N K\left(\frac{\mathbf{x}_i - \mathbf{x}_0}{h}\right) y_i}{\frac{1}{Nh^k} \sum_{i=1}^N K\left(\frac{\mathbf{x}_i - \mathbf{x}_0}{h}\right)},$$

where $K(\cdot)$ is now a **multivariate kernel**. Often $K(\cdot)$ is the product of k one-dimensional kernels, though multivariate kernels such as the multivariate normal density can be used.

If a product kernel is used the regressors should be transformed to a common scale by dividing by the standard deviation. Then the cross-validation measure (9.25) can be used to determine a common optimal bandwidth h^* , though determining which \mathbf{x}_i should be downweighted as the result of closeness to the end points is more complicated when \mathbf{x} is multivariate. Alternatively, regressors need not be rescaled, but then different bandwidths should be used for each regressor.

The asymptotic results and expressions are similar to those considered before, as the estimate is again a local average of the y_i . The bias $b(\mathbf{x}_0)$ is again $O(h^2)$ as before, but the variance of $\widehat{m}(\mathbf{x}_0)$ declines at a rate $O(Nh^k)$, slower than in the one-dimensional case since essentially a smaller fraction of the sample is being used to form $\widehat{m}(\mathbf{x}_0)$. Then

$$\sqrt{Nh^k}(\widehat{m}(\mathbf{x}_0) - m(\mathbf{x}_0) - b(\mathbf{x}_0)) \xrightarrow{d} \mathcal{N}\left[0, \frac{\sigma_\varepsilon^2}{f(\mathbf{x}_0)} \int K(\mathbf{z})^2 d\mathbf{z}\right].$$

The optimal bandwidth choice is $h^* = O(N^{-1/(k+4)})$, which is larger than $O(N^{-0.2})$ in the one-dimensional case. The corresponding optimal rate of convergence of $\widehat{m}(\mathbf{x}_0)$ is $N^{-2/(k+4)}$.

This result and the earlier scalar result assumes that $m(x)$ is twice differentiable, a necessary assumption to obtain the bias term in (9.23). If $m(x)$ is instead p times differentiable then kernel estimation using a p th order kernel (see Section 9.3.3) reduces the order of the bias, leading to smaller h^* and faster rates of convergence that attain Stone's bound given in Section 9.4.5; see Härdle (1990, p. 93) for further details. Other nonparametric estimators given in the next section can also attain Stone's bound.

The convergence rate decreases as the number of regressors increases, approaching N^0 as the number of regressors approaches infinity. This **curse of dimensionality** greatly restricts the use of nonparametric methods in regression models with several regressors. Semiparametric models (see Section 9.7) place additional structure so that the nonparametric components are of low dimension.

9.5.8. Tests of Parametric Models

An obvious test of correct specification of a parametric model of the conditional mean is to compare the fitted mean with that obtained from a nonparametric model.

Let $\hat{m}_\theta(\mathbf{x})$ denote a parametric estimator of $E[y|\mathbf{x}]$ and $\hat{m}_h(\mathbf{x})$ denote a nonparametric estimator such as a kernel estimator. One approach is to compare $\hat{m}_\theta(\mathbf{x})$ with $\hat{m}_h(\mathbf{x})$ at a range of values of \mathbf{x} . This is complicated by the need to correct for asymptotic bias in $\hat{m}_h(\mathbf{x})$ (see Härdle and Mammen, 1993). A second approach is to consider conditional moment tests of the form $N^{-1} \sum_i w_i(y_i - \hat{m}_\theta(\mathbf{x}_i))$, where different weights, based in part on kernel regression, test failure of $E[y|\mathbf{x}] = m_\theta(\mathbf{x})$ in different directions. For example, Horowitz and Härdle (1994) use $w_i = \hat{m}_h(\mathbf{x}_i) - \hat{m}_\theta(\mathbf{x}_i)$. Pagan and Ullah (1999, pp. 141–150) and Yatchew (2003, pp. 119–124) survey some of the methods used.

9.6. Alternative Nonparametric Regression Estimators

Section 9.4 introduced local regression methods that estimate the regression function $m(x_0)$ by a local weighted average $\hat{m}(x_0) = \sum_i w_{i0,h} y_i$, where the weights $w_{i0,h} = w(x_i, x_0, h)$ differ with the point of evaluation x_0 and the sample value of x_i . Section 9.5 presented detailed results when the weights are kernel weights.

Here we consider other commonly used local estimators that correspond to other weights. Many of the results of Section 9.5 carry through, with similar optimal rates of convergence and use of cross-validation for bandwidth selection, though the exact expressions for bias and variance differ from those in (9.23) and (9.24). The estimators given in Section 9.6.2 are especially popular.

9.6.1. Nearest Neighbors Estimator

The **k -nearest neighbor estimator** is the equally weighted average of the y values for the k observations of x_i closest to x_0 . Define $N_k(x_0)$ to be the set of k observations of x_i closest to x_0 . Then

$$\hat{m}_{k-NN}(x_0) = \frac{1}{k} \sum_{i=1}^N \mathbf{1}(x_i \in N_k(x_0)) y_i. \quad (9.29)$$

This estimator is a kernel estimator with uniform weights (see Table 9.1) except that the bandwidth is variable. Here the bandwidth h_0 at x_0 equals the distance between x_0 and the furthest of the k nearest neighbors, and more formally $h_0 \simeq k/(2Nf(x_0))$.

The quantity k/N is called the **span**. Smoother curves can be obtained by using kernel weights in (9.29).

The estimator has the attraction of providing a simple rule for variable bandwidth selection. It is computationally faster to use a **symmetrized** version that uses the $k/2$ nearest neighbors to the left and a similar number to the right, which is the local running average method used in Section 9.4.2. Then one can use an updating formula on observations ordered by increasing x_i , as then one observation leaves the data and one enters as x_0 increases.

9.6.2. Local Linear Regression and Lowess

The kernel regression estimator is a **local constant estimator** because it assumes that $m(x)$ equals a constant in the local neighborhood of x_0 . Instead, one can let $m(x)$ be linear in the neighborhood of x_0 , so that $m(x) = a_0 + b_0(x - x_0)$ in the neighborhood of x_0 .

To implement this idea, note that the kernel regression estimator $\hat{m}(x_0)$ can be obtained by minimizing $\sum_i K((x_i - x_0)/h)(y_i - m_0)^2$ with respect to m_0 . The **local linear regression estimator** minimizes

$$\sum_{i=1}^N K\left(\frac{x_i - x_0}{h}\right)(y_i - a_0 - b_0(x_i - x_0))^2, \quad (9.30)$$

with respect to a_0 and b_0 , where $K(\cdot)$ is a kernel weighting function. Then $\hat{m}(x) = \hat{a}_0 + \hat{b}_0(x - x_0)$ in the neighborhood of x_0 . The estimate at exactly x_0 is then $\hat{m}(x) = \hat{a}_0$, and \hat{b}_0 provides an estimate of the first derivative $\hat{m}'(x_0)$. More generally, a **local polynomial estimator of degree p** minimizes

$$\sum_{i=1}^N K\left(\frac{x_i - x_0}{h}\right)(y_i - a_{0,0} - a_{0,1}(x_i - x_0) - \cdots - a_{0,p}\frac{(x_i - x_0)^p}{p!})^2, \quad (9.31)$$

yielding $\hat{m}^{(s)}(x_0) = \hat{a}_{0,s}$.

Fan and Gijbels (1996) list many properties and attractions of this method. Estimation entails only weighted least-squares regression at each evaluation point x_0 . The estimators can be expressed as a weighted average of y_i , since they are LS estimators. The local linear estimator has bias term $b(x_0) = h^2 (\frac{1}{2}m''(x_0)) \int z^2 K(z) dz$, which, unlike the bias for kernel regression given in (9.23), does not depend on $m'(x_0)$. This is especially beneficial for overcoming the boundary problems illustrated in Section 9.4.2. For estimating an s th-order derivative a good choice of p is $p = s + 1$ so that, for example, one uses a local quadratic estimator to estimate the first derivative.

A standard local regression estimator is the **locally weighted scatterplot smoothing** or **Lowess estimator** of Cleveland (1979). This is a variant of local polynomial estimation that in (9.31) uses a variable bandwidth $h_{0,k}$ determined by the distance from x_0 to its k th nearest neighbor; uses the tricubic kernel $K(z) = (70/81)(1 - |z|^3)^3 \mathbf{1}(|z| < 1)$; and downweights observations with large residuals $y_i - \hat{m}(x_i)$, which requires passing through the data N times. For a summary see Fan and Gijbels (1996, p. 24). Lowess is attractive compared to kernel regression as it uses a variable bandwidth, robustifies

against outliers, and uses a local polynomial estimator to minimize boundary problems. However, it is computationally intensive.

Another popular variation is the **supersmooth**er of Friedman (1984) (see Härdle, 1990, p. 181). The starting point is symmetrized k -NN, using local linear fit rather than local constant fit for better fit at the boundary. Rather than use a fixed span or fixed k , however, the supersmooth is a variable span smoother where the variable span is determined by local cross-validation that entails nine passes over the data. Compared to Lowess the supersmooth does not robustify against outliers, but it permits the span to vary and is fast to compute.

9.6.3. Smoothing Spline Estimator

The **cubic smoothing spline estimator** $\hat{m}_\lambda(x)$ minimizes the penalized residual sum of squares

$$\text{PRSS}(\lambda) = \sum_{i=1}^N (y_i - m(x_i))^2 + \lambda \int (m''(x))^2 dx, \quad (9.32)$$

where λ is a smoothing parameter. As elsewhere in this chapter squared error loss is used. The first term alone leads to a very rough fit since then $\hat{m}(x_i) = y_i$. The second term is introduced to penalize roughness. The cross-validation methods of Section 9.5.3 can be used to determine λ , with larger values of λ leading to a smoother curve.

Härdle (1990, pp. 56–65) shows that $\hat{m}_\lambda(x)$ is a cubic polynomial between successive x -values and that the estimator can be expressed as a local weighted average of the y s and is asymptotically equivalent to a kernel estimator with a particular variable kernel. In microeconomics smoothing splines are used less frequently than the other methods presented here. The approach can be adapted to other roughness penalties and other loss functions.

9.6.4. Series Estimators

Series estimators approximate a regression function by a weighted sum of K functions $z_1(x), \dots, z_K(x)$,

$$\hat{m}_K(x) = \sum_{j=1}^K \hat{\beta}_j z_j(x), \quad (9.33)$$

where the coefficients $\hat{\beta}_1, \dots, \hat{\beta}_K$ are simply obtained by OLS regression of y on $z_1(x), \dots, z_K(x)$. The functions $z_1(x), \dots, z_K(x)$ form a truncated series. Examples include a $(K-1)$ th-order polynomial approximation or power series with $z_j(x) = x^{j-1}$, $j = 1, \dots, K$; orthogonal and orthonormal polynomial variants (see Section 12.3.1); truncated Fourier series where the regressor is rescaled so that $x \in [0, 2\pi]$; the Fourier flexible functional form of Gallant (1981), which is a truncated Fourier series plus the terms x and x^2 ; and regression splines that approximate the regression function $m(x)$ by polynomial functions between a given number of knots that are joined at the knots.

The approach differs from that in Section 9.4 as it is a global approximation approach to estimation of $m(x)$, rather than a local approach to estimation of $m(x_0)$. Nonetheless, $\widehat{m}_K(x) \xrightarrow{P} m(x_0)$ if $K \rightarrow \infty$ at an appropriate rate as $N \rightarrow \infty$. From Newey (1997) if \mathbf{x} is k dimensional and $m(\mathbf{x})$ is p times differentiable the mean integrated squared error (see Section 9.5.3) $\text{MISE}(h) = O(K^{-2p/k} + K/N)$, where the first term reflects bias and the second term variance. Equating these gives the optimal $K^* = N^{k/(2p+k)}$, so K grows but at slower rate than the sample size. The convergence rate of $\widehat{m}_{K^*}(x)$ equals the fastest possible rate of Stone (1980), given in Section 9.4.5. Intuitively, series estimators may not be robust as outliers may have a global rather than merely local impact on $\widehat{m}(x)$, but this conjecture is not tested in typical examples given in texts.

Andrews (1991) and Newey (1997) give a very general treatment that includes the multivariate case, estimation of functionals other than the conditional mean, and extensions to semiparametric models where series methods are most often used.

9.7. Semiparametric Regression

The preceding analysis has emphasized regression models without any structure. In microeconomics some structure is usually placed on the regression model.

First, economic theory may place some structure, such as symmetry and homogeneity restrictions, in a demand function. Such information may be incorporated into nonparametric regression; see, for example, Matzkin (1994).

Second, and more frequently, econometric models include so many potential regressors that the curse of dimensionality makes fully nonparametric analysis impractical. Instead, it is common to estimate a **semiparametric model** that loosely speaking combines a parametric component with a nonparametric component; see Powell (1994) for a careful discussion of the term semiparametric.

There are many different semiparametric models and myriad methods are often available to consistently estimate these models. In this section we present just a few leading examples. Applications are given elsewhere in this book, including the binary outcome models and censored regression models given in Chapters 14 and 16.

9.7.1. Examples

Table 9.2 presents several leading examples of semiparametric regression. The first two examples, detailed in the following, generalize the linear model $\mathbf{x}'\beta$ by adding an unspecified component $\lambda(\mathbf{z})$ or by permitting an unspecified transformation $g(\mathbf{x}'\beta)$, whereas the third combines the first two. The next three models, used more in applied statistics than econometrics, reduce the dimensionality by assuming additivity or separability of the regressors but are otherwise nonparametric. We detail the generalized additive model. Related to these are **neural network models**; see Kuan and White (1994). The last example, also detailed in the following, is a flexible model of the conditional variance. Care needs to be taken to ensure that semiparametric models

Table 9.2. Semiparametric Models: Leading Examples

Name	Model	Parametric	Nonparametric
Partially linear	$E[y \mathbf{x}, \mathbf{z}] = \mathbf{x}'\beta + \lambda(\mathbf{z})$	β	$\lambda(\cdot)$
Single index	$E[y \mathbf{x}] = g(\mathbf{x}'\beta)$	β	$g(\cdot)$
Generalized partial linear	$E[y \mathbf{x}, \mathbf{z}] = g(\mathbf{x}'\beta + \lambda(\mathbf{z}))$	β	$g(\cdot), \lambda(\cdot)$
Generalized additive	$E[y \mathbf{x}] = c + \sum_{j=1}^k g_j(x_j)$	—	$g_j(\cdot)$
Partial additive	$E[y \mathbf{x}, \mathbf{z}] = \mathbf{x}'\beta + c + \sum_{j=1}^k g_j(z_j)$	β	$g_j(\cdot)$
Projection pursuit	$E[y \mathbf{x}] = \sum_{j=1}^M g_j(\mathbf{x}'_j \beta_j)$	β_j	$g_j(\cdot)$
Heteroskedastic linear	$E[y \mathbf{x}] = \mathbf{x}'\beta; V[y \mathbf{x}] = \sigma^2(\mathbf{x})$	β	$\sigma^2(\cdot)$

are identified. For example, see the discussion of single-index models. In addition to estimation of β , interest also lies in the marginal effects such as $\partial E[y|\mathbf{x}, \mathbf{z}]/\partial \mathbf{x}$.

9.7.2. Efficiency of Semiparametric Estimators

We consider loss of efficiency in estimating by semiparametric rather than parametric methods, ahead of presenting results for several leading semiparametric models.

Our summary follows Robinson (1988b), who considers a semiparametric model with parametric component denoted β and nonparametric component denoted G that depends on infinitely many nuisance parameters. Examples of G include the shape of the distribution of a symmetrically distributed iid error and the single-index function $g(\cdot)$ given in (9.37) in Section 9.7.4. The estimator $\widehat{\beta} = \beta(\widehat{G})$, where \widehat{G} is a nonparametric estimator of G .

Ideally, the estimator $\widehat{\beta}$ is **adaptive**, meaning that there is no efficiency loss in having to estimate G by nonparametric methods, so that

$$\sqrt{N}(\widehat{\beta} - \beta) \xrightarrow{d} \mathcal{N}[\mathbf{0}, V_G],$$

where V_G is the covariance matrix for any shape function G in the particular class being considered. Within the likelihood framework V_G is the Cramer–Rao lower bound. In the second-moment context V_G is given by the Gauss–Markov theorem or a generalization such as to GMM. A leading example of an adaptive estimator is estimation with specified conditional mean function but with unknown functional form for heteroskedasticity (see Section 9.7.6).

If the estimator $\widehat{\beta}$ is not adaptive then the next best optimality property is for the estimator to attain the **semiparametric efficiency bound** V_G^* , so that

$$\sqrt{N}(\widehat{\beta} - \beta) \xrightarrow{d} \mathcal{N}[\mathbf{0}, V_G^*],$$

where V_G^* is a generalization of the Cramer–Rao lower bound or its second-moment analogue that provides the smallest variance matrix possible given the specified semiparametric model. For an adaptive estimator $V_G^* = V_G$, but usually V_G^* exceeds V_G . Semiparametric efficiency bounds are introduced in Section 9.7.8. They can be

obtained only in some semiparametric settings, and even when they are known no estimator may exist that attains the bound. An example that attains the bound is the binary choice model estimator of Klein and Spady (1993) (see Section 14.7.4).

If the semiparametric efficiency bound is not attained or is not known, then the next best property is that $\sqrt{N}(\widehat{\beta} - \beta) \xrightarrow{d} \mathcal{N}[\mathbf{0}, V_G^{**}]$ for V_G^{**} greater than V_G^* , which permits the usual statistical inference. More generally, $\sqrt{N}(\widehat{\beta} - \beta) = O_p(1)$ but is not necessarily normally distributed. Finally, consistent but less than \sqrt{N} -consistent estimators have the property that $N^r(\widehat{\beta} - \beta) = O_p(1)$, where $r < 0.5$. Often asymptotic normality cannot be established. This often arises when the parametric and nonparametric parts are treated equally, so that maximization occurs jointly over β and G . There are many examples, particularly in discrete and truncated choice models.

Despite their potential inefficiency, semiparametric estimators are attractive because they can retain consistency in settings where a fully parametric estimator is inconsistent. Powell (1994, p. 2513) presents a table that summarizes the existence of consistent and \sqrt{N} -consistent asymptotic normal estimators for a range of semiparametric models.

9.7.3. Partially Linear Model

The **partially linear model** specifies the conditional mean to be the usual linear regression function plus an unspecified nonlinear component, so

$$E[y|\mathbf{x}, \mathbf{z}] = \mathbf{x}'\beta + \lambda(\mathbf{z}), \quad (9.34)$$

where the scalar function $\lambda(\cdot)$ is unspecified.

An example is the estimation of a demand function for electricity, where \mathbf{z} reflects time-of-day or weather indicators such as temperature. A second example is the sample selection model given in Section 16.5. Ignoring $\lambda(\mathbf{z})$ leads to inconsistent β owing to omitted variables bias, unless $\text{Cov}[\mathbf{x}, \lambda(\mathbf{z})] = \mathbf{0}$. In applications interest may lie in β , $\lambda(\mathbf{z})$ or both. Fully nonparametric estimation of $E[y|\mathbf{x}, \mathbf{z}]$ is possible but leads to less than \sqrt{N} -consistent estimation of β .

Robinson Difference Estimator

Instead, Robinson (1988a) proposed the following method. The regression model implies

$$y = \mathbf{x}'\beta + \lambda(\mathbf{z}) + u,$$

where the error $u = y - E[y|\mathbf{x}, \mathbf{z}]$. This in turn implies

$$E[y|\mathbf{z}] = E[\mathbf{x}|\mathbf{z}]\beta + \lambda(\mathbf{z})$$

since $E[u|\mathbf{x}, \mathbf{z}] = 0$ implies $E[u|\mathbf{z}] = 0$. Subtracting the two equations yields

$$y - E[y|\mathbf{z}] = (\mathbf{x} - E[\mathbf{x}|\mathbf{z}])'\beta + u. \quad (9.35)$$

The conditional moments in (9.35) are unknown, but they can be replaced by nonparametric estimates.

Thus Robinson proposed the OLS regression estimation of

$$y_i - \hat{m}_{yi} = (\mathbf{x} - \hat{\mathbf{m}}_{xi})'\beta + v, \quad (9.36)$$

where \hat{m}_{yi} and $\hat{\mathbf{m}}_{xi}$ are predictions from nonparametric regression of, respectively, y_i and \mathbf{x}_i on \mathbf{z}_i . Given independence over i , the OLS estimator of β in (9.36) is \sqrt{N} consistent and asymptotically normal with

$$\sqrt{N}(\hat{\beta}_{PL} - \beta) \xrightarrow{d} \mathcal{N}\left[\mathbf{0}, \sigma^2 \left(\text{plim} \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - E[\mathbf{x}_i | \mathbf{z}_i])(\mathbf{x}_i - E[\mathbf{x}_i | \mathbf{z}_i])' \right)^{-1} \right],$$

assuming u_i is iid $[0, \sigma^2]$. Not specifying $\lambda(\mathbf{z})$ generally leads to an efficiency loss, though there is no loss if $E[\mathbf{x}|\mathbf{z}]$ is linear in \mathbf{z} . To estimate $V[\hat{\beta}_{PL}]$ simply replace $(\mathbf{x}_i - E[\mathbf{x}_i | \mathbf{z}_i])$ by $(\mathbf{x}_i - \hat{\mathbf{m}}_{xi})$. The asymptotic result generalizes to heteroskedastic errors, in which case one just uses the usual Eicker–White standard errors from the OLS regression (9.36). Since $\lambda(\mathbf{z}) = E[y|\mathbf{z}] - E[\mathbf{x}|\mathbf{z}]'\beta$ it can be consistently estimated by $\hat{\lambda}(\mathbf{z}) = \hat{m}_{yi} - \hat{\mathbf{m}}_{xi}'\hat{\beta}$.

A variety of nonparametric estimators \hat{m}_{yi} and $\hat{\mathbf{m}}_{xi}$ can be used. Robinson (1988a) used kernel estimates that require convergence at rate no slower than $N^{-1/4}$ so that oversmoothing or higher order kernels are needed if the dimension of \mathbf{z} is large; see Pagan and Ullah (1999, p. 205). Note also that the kernel estimators may be trimmed (see Section 9.5.3).

Other Estimators

Several other methods lead to \sqrt{N} -consistent estimates of β in the partially linear model. Speckman (1988) also used kernels. Engle et al. (1986) used a generalization of the cubic smoothing spline estimator. Andrews (1991) presented regression of y on \mathbf{x} and a series approximation for $\lambda(\mathbf{z})$ given in Section 9.6.4. Yatchew (1997) presents a simple differencing estimator.

9.7.4. Single-Index Models

A **single-index model** specifies the conditional mean to be an unknown scalar function of a linear combination of the regressors, with

$$E[y|\mathbf{x}] = g(\mathbf{x}'\beta), \quad (9.37)$$

where the scalar function $g(\cdot)$ is unspecified. The advantages of single-index models have been presented in Section 5.2.4. Here the function $g(\cdot)$ is obtained from the data, whereas previous examples specified, for example, $E[y|\mathbf{x}] = \exp(\mathbf{x}'\beta)$.

Identification

Ichimura (1993) presents **identification conditions** for the single-index model. For unknown function $g(\cdot)$ the single-index model β is only identified up to location and scale. To see this note that for scalar v the function $g^*(a + bv)$ can always be expressed

as $g(v)$, so the function $g^*(a + b\mathbf{x}'\beta)$ is equivalent to $g(\mathbf{x}'\beta)$. Additionally, $g(\cdot)$ must be differentiable. In the simplest case all regressors are continuous. If instead some regressors are discrete, then at least one regressor must be continuous and if $g(\cdot)$ is monotonic then bounds can be obtained for β .

Average Derivative Estimator

For continuous regressors, Stoker (1986) observed that if the conditional mean is single index then the vector of average derivatives of the conditional mean determines β up to scale, since for $m(\mathbf{x}_i) = g(\mathbf{x}'_i/\beta)$

$$\delta \equiv E \left[\frac{\partial m(\mathbf{x})}{\partial \mathbf{x}} \right] = E[g'(\mathbf{x}'\beta)]\beta, \quad (9.38)$$

and $E[g'(\mathbf{x}'_i/\beta)]$ is a scalar. Furthermore, by the generalized information matrix equality given in Section 5.6.3, for any function $h(\mathbf{x})$, $E[\partial h(\mathbf{x})/\partial \mathbf{x}] = -E[h(\mathbf{x})s(\mathbf{x})]$, where $s(\mathbf{x}) = \partial \ln f(\mathbf{x})/\partial \mathbf{x} = f'(\mathbf{x})/f(\mathbf{x})$ and $f(\mathbf{x})$ is the density of \mathbf{x} . Thus

$$\delta = -E[m(\mathbf{x})s(\mathbf{x})] = -E[E[y|\mathbf{x}]s(\mathbf{x})]. \quad (9.39)$$

It follows that δ , and hence β up to scale, can be estimated by the **average derivative (AD) estimator**

$$\hat{\delta}_{AD} = -\frac{1}{N} \sum_{i=1}^N y_i \hat{s}(\mathbf{x}_i), \quad (9.40)$$

where $\hat{s}(\mathbf{x}_i) = \hat{f}'(\mathbf{x}_i)/\hat{f}(\mathbf{x}_i)$ can be obtained by kernel estimation of the density of \mathbf{x}_i and its first derivative. The estimator $\hat{\delta}$ is \sqrt{N} consistent and its asymptotic normal distribution was derived by Härdle and Stoker (1989). The function $g(\cdot)$ can be estimated by nonparametric regression of y_i on $\mathbf{x}'_i \hat{\delta}$. Note that $\hat{\delta}_{AD}$ provides an estimate of $E[m'(\mathbf{x})]$ regardless of whether a single-index model is relevant.

A weakness of $\hat{\delta}_{AD}$ is that $\hat{s}(\mathbf{x}_i)$ can be very large if $\hat{f}(\mathbf{x}_i)$ is small. One possibility is to trim when $\hat{f}(\mathbf{x}_i)$ is small. Powell, Stock, and Stoker (1989) instead observed that the result (9.38) extends to weighted derivatives with $\delta \equiv E[w(\mathbf{x})m'(\mathbf{x})]$. Especially convenient is to choose $w(\mathbf{x}) = f(\mathbf{x})$, which yields the **density weighted average derivative (DWAD) estimator**

$$\hat{\delta}_{DWAD} = -\frac{1}{N} \sum_{i=1}^N y_i \hat{f}'(\mathbf{x}_i), \quad (9.41)$$

which no longer divides by $\hat{f}(\mathbf{x}_i)$. This yields a \sqrt{N} -consistent and asymptotically normal estimate of β up to scale. For example, if the first component of β is normalized to one then $\hat{\beta}_1 = 1$ and $\hat{\beta}_j = \hat{\delta}_j/\hat{\delta}_1$ for $j > 1$.

These methods require continuous regressors so that the derivatives exist. Horowitz and Härdle (1996) present extension to discrete regressors.

Semiparametric Least Squares

An alternative estimator of the single-index model was proposed by Ichimura (1993). Begin by assuming that $g(\cdot)$ is known, in which case the WLS estimator of β minimizes

$$S_N(\beta) = \frac{1}{N} \sum_{i=1}^N w_i(x)(y_i - g(\mathbf{x}'_i \beta))^2.$$

For unknown $g(\cdot)$ Ichimura proposed replacing $g(\mathbf{x}'_i \beta)$ by a nonparametric estimate $\widehat{g}(\mathbf{x}'_i \beta)$, leading to the **weighted semiparametric least-squares (WSLS) estimator** $\widehat{\beta}_{\text{WSLS}}$ that minimizes

$$Q_N(\beta) = \frac{1}{N} \sum_{i=1}^N \pi(x_i) w_i(x) (y_i - \widehat{g}(\mathbf{x}'_i \beta))^2,$$

where $\pi(x_i)$ is a trimming function that drops observations if the kernel regression estimate of the scalar $\mathbf{x}'_i \beta$ is small, and $\widehat{g}(\mathbf{x}'_i \beta)$ is a leave-one-out kernel estimator from regression of y_i on $\mathbf{x}'_i \beta$. This is a \sqrt{N} -consistent and asymptotically normal estimate of β up to scale that is generally more efficient than the DWAD estimator. For heteroskedastic data the most efficient estimator is the analogue of feasible GLS that uses estimated weight function $\widehat{w}_i(x) = 1/\widehat{\sigma}_i^2$, where $\widehat{\sigma}_i^2$ is the kernel estimate given in (9.43) of Section 9.7.6 and where $\widehat{u}_i = y_i - \widehat{g}(\mathbf{x}'_i \widehat{\beta})$ and $\widehat{\beta}$ is obtained from initial minimization of $Q_N(\beta)$ with $w_i(x) = 1$.

The WSLS estimator is computed by iterative methods. Begin with an initial estimator $\widehat{\beta}^{(1)}$, such as the DWAD estimator with first component normalized to one. Form the kernel estimate $\widehat{g}(\mathbf{x}'_i \widehat{\beta}^{(1)})$ and hence $Q_N(\widehat{\beta}^{(1)})$, perturb $\widehat{\beta}^{(1)}$ to obtain the gradient $g_N(\widehat{\beta}^{(1)}) = \partial Q_N(\beta)/\partial \beta|_{\widehat{\beta}^{(1)}}$ and hence an update $\widehat{\beta}^{(2)} = \widehat{\beta}^{(1)} + \mathbf{A}_N g_N(\widehat{\beta}^{(1)})$, and so on. This estimator is considerably more difficult to calculate than the DWAD estimator, especially as $Q_N(\beta)$ can be nonconvex and multimodal.

9.7.5. Generalized Additive Models

Generalized additive models specify $E[y|\mathbf{x}] = g_1(x_1) + \dots + g_k(x_k)$, a specialization of the fully nonparametric model $E[y|\mathbf{x}] = g(x_1, \dots, x_k)$. This specialization results in the estimated subfunctions $\widehat{g}_j(x_j)$ converging at the rate for a one-dimensional nonparametric regression rather than the slower rate of a k -dimensional nonparametric regression.

A well-developed methodology exists for estimating such models (see Hastie and Tibsharani, 1990). This is automated in some statistical packages such as S-Plus. Plots of the estimated subfunctions $\widehat{g}_j(x_j)$ on x_j trace out the marginal effects of x_j on $E[y|\mathbf{x}]$, so the additive model can provide a useful tool for exploratory data analysis. The model sees little use in microeconomics in part because many applications such as censoring, truncation, and discrete outcomes lead naturally to single-index and partially linear models.

9.7.6. Heteroskedastic Linear Model

The **heteroskedastic linear model** specifies

$$\begin{aligned} E[y|\mathbf{x}] &= \mathbf{x}'\boldsymbol{\beta}, \\ V[y|\mathbf{x}] &= \sigma^2(\mathbf{x}), \end{aligned}$$

where the variance function $\sigma^2(\cdot)$ is unspecified.

The assumption that errors are heteroskedastic is the standard cross-section data assumption in modern microeconomics. One can obtain consistent but inefficient estimates of $\boldsymbol{\beta}$ by doing OLS and using the Eicker–White heteroskedastic-consistent estimate of the variance matrix of the OLS estimator. Cragg (1983) and Amemiya (1983) proposed an IV estimator that is more efficient than OLS but still not fully efficient. Feasible GLS provides a fully efficient second-moment estimator but is not attractive as it requires specification of a functional form for $\sigma^2(\mathbf{x})$ such as $\sigma^2(\mathbf{x}) = \exp(\mathbf{x}'\boldsymbol{\gamma})$.

Robinson (1987) proposed a variant of FGLS using a nonparametric estimator of $\sigma_i^2 = \sigma^2(\mathbf{x}_i)$. Then

$$\widehat{\boldsymbol{\beta}}_{HLM} = \left(\sum_{i=1}^N \widehat{\sigma}_i^{-2} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\sum_{i=1}^N \widehat{\sigma}_i^{-2} \mathbf{x}_i y_i \right), \quad (9.42)$$

where Robinson (1987) used a k -NN estimator of σ_i^2 with uniform weight, so

$$\widehat{\sigma}_i^2 = \frac{1}{k} \sum_{j=1}^N \mathbf{1}(\mathbf{x}_j \in N_k(\mathbf{x}_i)) \widehat{u}_j^2, \quad (9.43)$$

where $\widehat{u}_i = y_i - \mathbf{x}_i' \widehat{\boldsymbol{\beta}}_{OLS}$ is the residual from first-stage OLS regression of y_i on x_i and $N_k(\mathbf{x}_i)$ is the set of k observations of \mathbf{x}_j closest to \mathbf{x}_i in weighted Euclidean norm. Then

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{HLM} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathcal{N} \left[\mathbf{0}, \left(\text{plim} \frac{1}{N} \sum_{i=1}^N \sigma_i^{-2} (\mathbf{x}_i) \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \right]],$$

assuming u_i is iid $[0, \sigma^2(\mathbf{x}_i)]$. This estimator is **adaptive** as it attains the Gauss–Markov bound so is as efficient as the GLS estimator when σ_i^2 is known. The variance matrix is consistently estimated by $(N^{-1} \sum_i \widehat{\sigma}_i^{-2} \mathbf{x}_i \mathbf{x}_i')^{-1}$.

In principle other nonparametric estimators of $\sigma^2(\mathbf{x}_i)$ might be used, but Carroll (1982) and others originally proposed use of a kernel estimator of σ_i^2 and found that proof of efficiency was possible only under very restrictive assumptions on \mathbf{x}_i . The Robinson method extends to models with nonlinear mean function.

9.7.7. Seminonparametric MLE

Suppose y_i is iid with specified density $f(y_i|\mathbf{x}_i, \boldsymbol{\beta})$. In general, misspecification of the density leads to inconsistent parameter estimates. Gallant and Nychka (1987) proposed approximating the unknown true density by a power-series expansion around the density $f(y|\mathbf{x}, \boldsymbol{\beta})$. To ensure a positive density they actually use a **squared power-series**

expansion around $f(y|\mathbf{x}, \beta)$, yielding

$$h_p(y|\mathbf{x}, \beta, \alpha) = \frac{(p(y|\alpha))^2 f(y|\mathbf{x}, \beta)}{\int (p(z|\alpha))^2 f(y|z, \beta) dz}, \quad (9.44)$$

where $p(y|\alpha)$ is a p th order polynomial in y , α is the vector of coefficients of the polynomial, and division by the denominator ensures that probabilities integrate or sum to one. The estimator of β and α maximizes the log-likelihood $\sum_{i=1}^N \ln h_p(y_i|\mathbf{x}_i, \beta, \alpha)$. The approach generalizes immediately to multivariate \mathbf{y}_i . The estimator is called the **seminonparametric maximum likelihood estimator** because it is a nonparametric estimator that can be estimated in the same way as a maximum likelihood estimator. Gallant and Nychka (1987) showed that under fairly general conditions the estimator yields consistent estimates of the density if the order p of the polynomial increases with sample size N at an appropriate rate.

This result provides a strong basis for using (9.44) to obtain a class of flexible distributions for any particular data. The method is particularly simple if the polynomial series $p(y|\alpha)$ is the orthogonal or orthonormal polynomial series (see Section 12.3.1) for the baseline density $f(y|\mathbf{x}, \beta)$, as then the normalizing factor in the denominator can be simply constructed. The order of the polynomial can be chosen using information criteria, with measures that penalize model complexity more than AIC used in practice. Regular ML statistical inference is possible if one ignores the data-dependent selection of the polynomial order and assumes that the resulting density $h_p(y|\mathbf{x}, \beta, \alpha)$ is correctly specified. An example of this approach for count data regression is given in Cameron and Johansson (1997).

9.7.8. Semiparametric Efficiency Bounds

Semiparametric efficiency bounds extend efficiency bounds such as Cramer–Rao or the Gauss–Markov theorem to cases where the dgp has a nonparametric component. The best semiparametric methods achieve this efficiency bound.

We use β to denote parameters we wish to estimate, which may include variance components such as σ^2 , and η to denote nuisance parameters. For simplicity we consider ML estimation with a nonparametric component.

We begin with the fully parametric case. The MLE $(\widehat{\beta}, \widehat{\eta})$ maximizes $\mathcal{L}(\beta, \eta) = \ln L(\beta, \eta)$. Let $\theta = (\beta, \eta)$ and let $\mathcal{I}_{\theta\theta}$ be the information matrix defined in (5.43). Then $\sqrt{N}(\widehat{\theta} - \theta) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathcal{I}_{\theta\theta}^{-1}]$. For $\sqrt{N}(\widehat{\beta} - \beta)$, partitioned inversion of $\mathcal{I}_{\theta\theta}$ leads to

$$\mathbf{V}^* = (\mathcal{I}_{\beta\beta} - \mathcal{I}_{\beta\eta}\mathcal{I}_{\eta\eta}^{-1}\mathcal{I}_{\eta\beta})^{-1} \quad (9.45)$$

as the efficiency bound for estimation of β when η is unknown. There is an efficiency loss when η is unknown, unless the information matrix is block diagonal so that $\mathcal{I}_{\beta\eta} = \mathbf{0}$ and the variance reduces to $\mathcal{I}_{\beta\beta}^{-1}$.

Now consider extension to the nonparametric case. Suppose we have a parametric submodel, say $\mathcal{L}_0(\beta)$, that involves β alone. Consider the family of all possible parametric models $\mathcal{L}(\beta, \eta)$ that nest $\mathcal{L}_0(\beta)$ for some value of η . The semiparametric

efficiency bound is the largest value of V^* given in (9.45) over all possible parametric models $\mathcal{L}(\beta, \eta)$, but this is difficult to obtain.

Simplification is possible by considering

$$\tilde{\mathbf{s}}_\beta = \mathbf{s}_\beta - E[\mathbf{s}_\beta | \mathbf{s}_\eta],$$

where \mathbf{s}_θ denotes the score $\partial \mathcal{L} / \partial \theta$, and $\tilde{\mathbf{s}}_\beta$ is the score for β after concentrating out η . For finite-dimensional η it can be shown that $E[N^{-1}\tilde{\mathbf{s}}_\beta \tilde{\mathbf{s}}'_\beta] = V^*$. Here η is instead infinite dimensional. Assume iid data and let \mathbf{s}_{η_i} denote the i th component in the sum that leads to the score \mathbf{s}_θ . Begun et al. (1983) define the tangent set to be the set of all linear combinations of \mathbf{s}_{η_i} . When this tangent set is linear and closed the largest value of V^* in (9.45) equals

$$\Omega = (\text{plim } N^{-1}\tilde{\mathbf{s}}_\beta \tilde{\mathbf{s}}'_\beta)^{-1} = (E[\tilde{\mathbf{s}}_\beta \tilde{\mathbf{s}}'_\beta])^{-1}.$$

The matrix Ω is then the semiparametric efficiency bound.

In applications one first obtains $\mathbf{s}_\eta = \sum_i \mathbf{s}_{\eta_i}$. Then obtain $E[\mathbf{s}_{\beta_i} | \mathbf{s}_\eta]$, which may entail assumptions such as symmetry of errors that place restrictions on the class of semiparametric models being considered. This yields $\tilde{\mathbf{s}}_{\beta_i}$ and hence Ω . For more details and applications see Newey (1990b), Pagan and Ullah (1999), and Severini and Tripathi (2001).

9.8. Derivations of Mean and Variance of Kernel Estimators

Nonparametric estimation entails a balance between smoothness (variance) and bias (mean). Here we derive the mean and variance of kernel density and kernel regression estimators. The derivations follow those of M. J. Lee (1996).

9.8.1. Mean and Variance of Kernel Density Estimator

Since x_i are iid each term in the summation has the same expected value and

$$\begin{aligned} E[\hat{f}(x_0)] &= E\left[\frac{1}{h} K\left(\frac{x-x_0}{h}\right)\right] \\ &= \int \frac{1}{h} K\left(\frac{x-x_0}{h}\right) f(x) dx. \end{aligned}$$

By change of variable to $z = (x - x_0)/h$ so that $x = x_0 + hz$ and $dx/dz = h$ we obtain

$$E[\hat{f}(x_0)] = \int K(z) f(x_0 + hz) dz.$$

A second-order Taylor series expansion of $f(x_0 + hz)$ around $f(x_0)$ yields

$$\begin{aligned} E[\hat{f}(x_0)] &= \int K(z) \{f(x_0) + f'(x_0)hz + \frac{1}{2}f''(x_0)(hz)^2\} dz \\ &= f(x_0) \int K(z) dz + hf'(x_0) \int z K(z) dz + \frac{1}{2}h^2 f''(x_0) \int z^2 K(z) dz. \end{aligned}$$

Since the kernel $K(z)$ integrates to unity this simplifies to

$$E[\widehat{f}(x_0)] - f(x_0) = hf'(x_0) \int z K(z) dz + \frac{1}{2} h^2 f''(x_0) \int z^2 K(z) dz.$$

If additionally the kernel satisfies $\int z K(z) dz = 0$, assumed in condition (ii) in Section 9.3.3, and second derivatives of f are bounded, then the first term on the right-hand side disappears, yielding $E[\widehat{f}(x_0)] - f(x_0) = b(x_0)$, where $b(x_0)$ is defined in (9.4).

To obtain the variance of $\widehat{f}(x_0)$, begin by noting that if y_i are iid then $V[\bar{y}] = N^{-1}V[y] = N^{-1}E[y^2] - N^{-1}(E[y])^2$. Thus

$$V[\widehat{f}(x_0)] = \frac{1}{N} E \left[\left(\frac{1}{h} K \left(\frac{x-x_0}{h} \right) \right)^2 \right] - \frac{1}{N} (E \left[\frac{1}{h} K \left(\frac{x-x_0}{h} \right) \right])^2.$$

Now by change of variables and first-order Taylor series expansion

$$\begin{aligned} E \left[\left(\frac{1}{h} K \left(\frac{x-x_0}{h} \right) \right)^2 \right] &= \int \frac{1}{h} K(z)^2 \{f(x_0) + f'(x_0)hz\} dz \\ &= \frac{1}{h} f(x_0) \int K(z)^2 dz + f'(x_0) \int z K(z)^2 dz. \end{aligned}$$

It follows that

$$\begin{aligned} V[\widehat{f}(x_0)] &= \frac{1}{Nh} f(x_0) \int K(z)^2 dz + \frac{1}{N} f'(x) \int z K(z)^2 dz \\ &\quad - \frac{1}{N} [f(x_0) + \frac{h^2}{2} f''(x_0) [\int z^2 K(z) dz]]^2. \end{aligned}$$

For $h \rightarrow 0$ and $N \rightarrow \infty$ this is dominated by the first term, leading to Equation (9.5).

9.8.2. Distribution of Kernel Regression Estimator

We obtain the distribution for regressors x_i that are iid with density $f(x)$. From Section 9.5.1 the kernel estimator is a weighted average $\widehat{m}(x_0) = \sum_i w_{i0,h} y_i$, where the kernel weights $w_{i0,h}$ are given in (9.22). Since the weights sum to unity we have $\widehat{m}(x_0) - m(x_0) = \sum_i w_{i0,h} (y_i - m(x_0))$. Substituting (9.15) for y_i , and normalizing by \sqrt{Nh} as in the kernel density estimator case we have

$$\sqrt{Nh}(\widehat{m}(x_0) - m(x_0)) = \sqrt{Nh} \sum_{i=1}^N w_{i0,h} (m(x_i) - m(x_0) + \varepsilon_i). \quad (9.46)$$

One approach to obtaining the limit distribution of (9.46) is to take a second-order Taylor series expansion of $m(x_i)$ around x_0 . This approach is not always taken because the weights $w_{i0,h}$ are complicated by the normalization that they sum to one (see (9.22)).

Instead, we take the approach of Lee (1996, pp. 148–151) following Bierens (1987, pp. 106–108). Note that the denominator of the weight function is the kernel estimate of the density of x_0 , since $\widehat{f}(x_0) = (Nh)^{-1} \sum_i K((x_i - x_0)/h)$. Then (9.46) yields

$$\sqrt{Nh}(\widehat{m}(x_0) - m(x_0)) = \frac{1}{\sqrt{Nh}} \sum_{i=1}^N K \left(\frac{x_i - x_0}{h} \right) (m(x_i) - m(x_0) + \varepsilon_i) \Bigg/ \widehat{f}(x_0). \quad (9.47)$$

We apply the Transformation Theorem (Theorem A.12) to (9.47), using $\widehat{f}(x_0) \xrightarrow{P} f(x_0)$ for the denominator, while several steps are needed to obtain a limit normal

distribution for the numerator:

$$\begin{aligned} & \frac{1}{\sqrt{Nh}} \sum_{i=1}^N K\left(\frac{x_i - x_0}{h}\right) (m(x_i) - m(x_0) + \varepsilon_i) \\ &= \frac{1}{\sqrt{Nh}} \sum_{i=1}^N K\left(\frac{x_i - x_0}{h}\right) (m(x_i) - m(x_0)) + \frac{1}{\sqrt{Nh}} \sum_{i=1}^N K\left(\frac{x_i - x_0}{h}\right) \varepsilon_i. \end{aligned} \quad (9.48)$$

Consider the first sum in (9.48); if a law of large numbers can be applied it converges in probability to its mean

$$\begin{aligned} & E\left[\frac{1}{\sqrt{Nh}} \sum_{i=1}^N K\left(\frac{x_i - x_0}{h}\right) (m(x_i) - m(x_0))\right] \\ &= \frac{\sqrt{N}}{\sqrt{h}} \int K\left(\frac{x - x_0}{h}\right) (m(x) - m(x_0)) f(x) dx \\ &= \sqrt{Nh} \int K(z) (m(x_0 + hz) - m(x_0)) f(x_0 + hz) dz \\ &= \sqrt{Nh} \int K(z) \left(hz m'(x_0) + \frac{1}{2} h^2 z^2 m''(x_0)\right) (f(x_0) + hz f'(x_0)) dz \\ &= \sqrt{Nh} \left\{ \int K(z) h^2 z^2 m'(x_0) f'(x_0) dz + \int K(z) \frac{1}{2} h^2 z^2 m''(x_0) f(x_0) dz \right\} \\ &= \sqrt{Nh} h^2 \left(m'(x_0) f'(x_0) + \frac{1}{2} m''(x_0) f(x_0)\right) \int z^2 K(z) dz \\ &= \sqrt{Nh} f(x_0) b(x_0), \end{aligned} \quad (9.49)$$

where $b(x_0)$ is defined in (9.23). The first equality uses x_i iid; the second equality is change of variables to $z = (x - x_0)/h$; the third equality applies a second-order Taylor series expansion to $m(x_0 + hz)$ and a first-order Taylor series expansion to $f(x_0 + hz)$; the fourth equality follows because upon expanding the product to four terms, the two terms given dominate the others (see, e.g., Lee, 1996, p. 150).

Now consider the second sum in (9.48); the terms in the sum clearly have mean zero, and the variance of each term, dropping subscript i , is

$$\begin{aligned} V\left[K\left(\frac{x - x_0}{h}\right) \varepsilon\right] &= E\left[K^2\left(\frac{x - x_0}{h}\right) \varepsilon^2\right] \\ &= \int K^2\left(\frac{x - x_0}{h}\right) V[\varepsilon|x] f(x) dx \\ &= h \int K^2(z) V[\varepsilon|x_0 + hz] f(x_0 + hz) dz \\ &= h V[\varepsilon|x_0] f(x_0) \int K^2(z) dz, \end{aligned} \quad (9.50)$$

by change of variables to $z = (x - x_0)/h$ with $dx = hdz$ in the third-line term, and letting $h \rightarrow 0$ to get the last line. It follows upon applying a central limit theorem that

$$\frac{1}{\sqrt{Nh}} \sum_{i=1}^N K\left(\frac{x_i - x_0}{h}\right) \varepsilon_i \xrightarrow{d} \mathcal{N}\left[0, V[\varepsilon|x_0] f(x_0) \int K^2(z) dz\right]. \quad (9.51)$$

Combining (9.49) and (9.51), we have that $\sqrt{Nh}(\hat{m}(x_0) - m(x_0))$ defined in (9.47) converges to $1/f(x_0)$ times $\mathcal{N}[\sqrt{Nh}f(x_0)b(x_0), V[\varepsilon|x_0]f(x_0)\int K^2(z)dz]$. Division of the mean by $f(x_0)$ and the variance by $f(x_0)^2$ leads to the limit distribution given in (9.24).

9.9. Practical Considerations

All-purpose regression packages increasingly offer adequate methods for univariate nonparametric density estimation and regression. The programming language XPlore emphasizes nonparametric and graphical methods; details on many of the methods are provided at its Web site.

Nonparametric univariate density estimation is straightforward, using a kernel density estimate based on a kernel such as the Gaussian or Epanechnikov. Easily computed plug-in estimates for the bandwidth provide a useful starting point that one may then, say, halve or double to see if there is an improvement.

Nonparametric univariate regression is also straightforward, aside from bandwidth selection. If relatively unbiased estimates of the regression function at the end points are desired, then local linear regression or Lowess estimates are better than kernel regression. Plug-in estimates for the bandwidth are more difficult to obtain and cross-validation is instead used (see Section 9.5.3) along with eyeballing the scatterplot with a fitted line. The degree of desired smoothness can vary with application. For nonparametric multivariate regression such eyeballing may be impossible.

Semiparametric regression is more complicated. It can entail subtleties such as trimming and undersmoothing the nonparametric component since typically estimation of the parametric component involves averaging the nonparametric component. For such purposes one generally uses specialized code written in languages such as Gauss, Matlab, Splus, or XPlore. For the nonparametric estimation component considerable computational savings can be obtained through use of fast computing algorithms such as binning and updating; see, for example, Fan and Gijbels (1996) and Härdle and Linton (1994).

All methods require at some stage specification of a bandwidth or window width. Different choices lead to different estimates in finite samples, and the differences can be quite large as illustrated in many of the figures in this chapter. By contrast, within a fully parametric framework different researchers estimating the same model by ML will all obtain the same parameter estimates. This indeterminateness is a detraction of nonparametric methods, though the hope is that in semiparametric methods at least the spillover effects to the parametric component of the model may be small.

9.10. Bibliographic Notes

Nonparametric estimation is well presented in many statistics texts, including Fan and Gijbels (1996). Ruppert, Wand, and Carroll (2003) present application of many semiparametric methods. The econometrics books by Härdle (1990), M. J. Lee (1996), Horowitz (1998b), Pagan and Ullah (1999), and Yatchew (2003) cover both nonparametric and semiparametric estimation.

Pagan and Ullah (1999) is particularly comprehensive. Yatchew (2003) is oriented to the applied econometrician. He emphasizes the partial linear and single-index models and practical aspects of their implementation such as computation of confidence intervals.

- 9.3** Key early references for kernel density estimation are Rosenblatt (1956) and Parzen (1962). Silverman's (1986) is a classic book on nonparametric density estimation.
- 9.4** A quite general statement of optimal rates of convergence for nonparametric estimators is given in Stone (1980).
- 9.5** Kernel regression estimation was proposed by Nadaraya (1964) and Watson (1964). A very helpful and relatively simple survey of kernel and nearest-neighbors regression is by Altman (1992). There are many other surveys in the statistics literature. Härdle (1990, chapter 5) has a lengthy discussion of bandwidth choice and confidence intervals.
- 9.6** Many approaches to nonparametric local regression are contained in Stone (1977). For series estimators see Andrews (1991) and Newey (1997).
- 9.6** For semiparametric efficiency bounds see the survey by Newey (1990b) and the more recent paper by Severini and Tripathi (2001). An early econometrics application was given by Chamberlain (1987).
- 9.7** The econometrics literature focuses on semiparametric regression. Survey papers include those by Powell (1994), Robinson (1988b), and, at a more introductory level, Yatchew (1998). Additional references are given in elsewhere in this book, notably in Sections 14.7, 15.11, 16.9, 20.5, and 23.8. The applied study by Bellemare, Melenberg, and Van Soest (2002) illustrates several semiparametric methods.

Exercises

- 9–1** Suppose we obtain a kernel density estimate using the uniform kernel (see Table 9.1) with $h = 1$ and a sample of size $N = 100$. Suppose in fact the data $x \sim \mathcal{N}[0, 1]$.
- (a) Calculate the bias of the kernel density estimate at $x_0 = 1$ using (9.4).
 - (b) Is the bias large relative to the true value $\phi(1)$, where $\phi(\cdot)$ is the standard normal pdf?
 - (c) Calculate the variance of the kernel density estimate at $x_0 = 1$ using (9.5).
 - (d) Which is making a bigger contribution to MSE at $x_0 = 1$, variance or bias squared?
 - (e) Using results in Section 9.3.7, give a 95% confidence interval for the density at $x_0 = 1$ based on the kernel density estimate $\hat{f}(1)$.
 - (f) For this example, what is the optimal bandwidth h^* from (9.10).
- 9–2** Suppose we obtain a kernel regression estimate using a uniform kernel (see Table 9.1) with $h = 1$ and a sample of size $N = 100$. Suppose in fact the data $x \sim \mathcal{N}[0, 1]$ and the conditional mean function is $m(x) = x^2$.
- (a) Calculate the bias of the kernel regression estimate at $x_0 = 1$ using (9.23).
 - (b) Is the bias large relative to the true value $m(1) = 1$?
 - (c) Calculate the variance of the kernel regression estimate at $x_0 = 1$ using (9.24).
 - (d) Which is making a bigger contribution to MSE at $x_0 = 1$, variance or bias squared?

- (e) Using results in Section 9.5.4, give a 95% confidence interval for $E[y|x_0 = 1]$ based on the kernel regression estimate $\hat{m}(1)$.

9–3 This question assumes access to a nonparametric density estimation program. Use the Section 4.6.4 data on health expenditure. Use a kernel density estimate with Gaussian kernel (if available).

- (a) Obtain the kernel density estimate for health expenditure, choosing a suitable bandwidth by eyeballing and trial and error. State the bandwidth chosen.
- (b) Obtain the kernel density estimate for natural logarithm of health expenditure, choosing a suitable bandwidth by eyeballing and trial and error. State the bandwidth chosen.
- (c) Compare your answer in part (b) to an appropriate histogram.
- (d) If possible superimpose a fitted normal density on the same graph as the kernel density estimate from part (b). Do health expenditures appear to be log-normally distributed?

9–4 This question assumes access to a kernel regression program or other nonparametric smoother. Use the complete sample of the Section 4.6.4 data on natural logarithm of health expenditure (y) and natural logarithm of total expenditure (x).

- (a) Obtain the kernel regression density estimate for health expenditure, choosing a good bandwidth by eyeballing and trial and error. State the bandwidth chosen.
- (b) Given part (a), does health appear to be a normal good?
- (c) Given part (a), does health appear to be a superior good?
- (d) Compare your nonparametric estimates with predictions from linear and quadratic regression.

Numerical Optimization

10.1. Introduction

Theoretical results on consistency and the asymptotic distribution of an estimator defined as the solution to an optimization problem were presented in Chapters 5 and 6. The more practical issue of how to numerically obtain the optimum, that is, how to calculate the parameter estimates, when there is no explicit formula for the estimator, comprises the subject of this chapter.

For the applied researcher estimation of standard nonlinear models, such as logit, probit, Tobit, proportional hazards, and Poisson, is seemingly no different from estimation of an OLS model. A statistical package obtains the estimates and reports coefficients, standard errors, t -statistics, and p -values. Computational problems generally only arise for the same reasons that OLS may fail, such as multicollinearity or incorrect data input.

Estimation of less standard nonlinear models, including minor variants of a standard model, may require writing a program. This may be possible within a standard statistical package. If not, then a programming language is used. Especially in the latter case a knowledge of optimization methods becomes necessary.

General considerations for optimization are presented in Section 10.2. Various iterative methods, including the Newton–Raphson and Gauss–Newton gradient methods, are described in Section 10.3. Practical issues, including some common pitfalls, are presented in Section 10.4. These issues become especially relevant when the optimization method fails to produce parameter estimates.

10.2. General Considerations

Microeconometric analysis is often based on an estimator $\widehat{\boldsymbol{\theta}}$ that maximizes a stochastic objective function $Q_N(\boldsymbol{\theta})$, where usually $\widehat{\boldsymbol{\theta}}$ solves the first-order conditions $\partial Q_N(\boldsymbol{\theta})/\partial \boldsymbol{\theta} = \mathbf{0}$. A minimization problem can be recast as a maximization by multiplying the objective function by minus one. In nonlinear applications there will

generally be no explicit solution to the first-order conditions, a nonlinear system of q equations in the q unknowns θ .

A grid search procedure is usually impractical and iterative methods, usually gradient methods, are employed.

10.2.1. Grid Search

In **grid search methods**, the procedure is to select many values of θ along a grid, compute $Q_N(\theta)$ for each of these values, and choose as the estimator $\hat{\theta}$ the value that provides the largest (locally or globally depending on the application) value of $Q_N(\theta)$.

If a fine enough grid can be chosen this method will always work. It is generally impractical, however, to choose a fine enough grid without further restrictions. For example, if 10 parameters need to be estimated and the grid evaluates each parameter at just 10 points, a very sparse grid, there are 10^{10} or 10 billion evaluations.

Grid search methods are nonetheless useful in applications where the grid search need only be performed among a subset of the parameters. They also permit viewing the response surface to verify that in using iterative methods one need not be concerned about multiple maxima. For example, many time-series packages do this for the scalar AR(1) coefficient in a regression model with AR(1) error. A second example is doing a grid search for the scalar inclusive parameter in a nested logit model (see Section 15.6). Of course, grid search methods may have to be used if nothing else works.

10.2.2. Iterative Methods

Virtually all microeconometric applications instead use **iterative methods**. These update the current estimate of θ using a particular rule. Given an s th-round estimate $\hat{\theta}_s$ the iterative method provides a rule that yields a new estimate $\hat{\theta}_{s+1}$, where $\hat{\theta}_s$ denotes the s th-round estimate rather than the s th component of $\hat{\theta}$. Ideally, the new estimate is a move toward the maximum, so that $Q_N(\hat{\theta}_{s+1}) > Q_N(\hat{\theta}_s)$, but in general this cannot be guaranteed. Also, gradient estimates may find a local maximum but not necessarily the global maximum.

10.2.3. Gradient Methods

Most iterative methods are **gradient methods** that change $\hat{\theta}_s$ in a direction determined by the gradient. The update formula is a matrix weighted average of the gradient

$$\hat{\theta}_{s+1} = \hat{\theta}_s + \mathbf{A}_s \mathbf{g}_s, \quad s = 1, \dots, S, \quad (10.1)$$

where \mathbf{A}_s is a $q \times q$ matrix that depends on $\hat{\theta}_s$, and

$$\mathbf{g}_s = \left. \frac{\partial Q_N(\theta)}{\partial \theta} \right|_{\hat{\theta}_s} \quad (10.2)$$

is the $q \times 1$ **gradient** vector evaluated at $\hat{\theta}_s$. Different gradient methods use different matrices \mathbf{A}_s , detailed in Section 10.3. A leading example is the Newton–Raphson method, which sets $\mathbf{A}_s = -\mathbf{H}_s^{-1}$, where \mathbf{H}_s is the Hessian matrix defined later in (10.6).

Note that in this chapter \mathbf{A} and \mathbf{g} denote quantities that differ from those in other chapters. Here \mathbf{A} is not the matrix that appears in the limit distribution of an estimator and \mathbf{g} is not the conditional mean of \mathbf{y} in the nonlinear regression model.

Ideally, the matrix \mathbf{A}_s is **positive definite** for a maximum (or negative definite for a minimum), as then it is likely that $Q_N(\widehat{\boldsymbol{\theta}}_{s+1}) > Q_N(\widehat{\boldsymbol{\theta}}_s)$. This follows from the first-order Taylor series expansion $Q_N(\widehat{\boldsymbol{\theta}}_{s+1}) = Q_N(\widehat{\boldsymbol{\theta}}_s) + \mathbf{g}'_s(\widehat{\boldsymbol{\theta}}_{s+1} - \widehat{\boldsymbol{\theta}}_s) + R$, where R is a remainder. Substituting in the update formula (10.1) yields

$$Q_N(\widehat{\boldsymbol{\theta}}_{s+1}) - Q_N(\widehat{\boldsymbol{\theta}}_s) = \mathbf{g}'_s \mathbf{A}_s \mathbf{g}_s + R,$$

which is greater than zero if \mathbf{A}_s is positive definite and the remainder R is sufficiently small, since for a positive definite square matrix \mathbf{A} the quadratic form $\mathbf{x}'\mathbf{A}\mathbf{x} > 0$ for all column vectors $\mathbf{x} \neq \mathbf{0}$. Too small a value of \mathbf{A}_s leads to an iterative procedure that is too slow; however, too large a value of \mathbf{A}_s may lead to overshooting, even if \mathbf{A}_s is positive definite, as the remainder term cannot be ignored for large changes.

A common modification to gradient methods is to add a **step-size adjustment** to prevent possible overshooting or undershooting, so

$$\widehat{\boldsymbol{\theta}}_{s+1} = \widehat{\boldsymbol{\theta}}_s + \widehat{\lambda}_s \mathbf{A}_s \mathbf{g}_s, \quad (10.3)$$

where the stepsize $\widehat{\lambda}_s$ is a scalar chosen to maximize $Q_N(\widehat{\boldsymbol{\theta}}_{s+1})$. At the s th round first calculate $\mathbf{A}_s \mathbf{g}_s$, which may involve considerable computation. Then calculate $Q_N(\widehat{\boldsymbol{\theta}})$, where $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_s + \lambda \mathbf{A}_s \mathbf{g}_s$ for a range of values of λ (called a **line search**), and choose $\widehat{\lambda}_s$ as that λ that maximizes $Q_N(\widehat{\boldsymbol{\theta}})$. Considerable computational savings are possible because the gradient and \mathbf{A}_s are not recomputed along the line search.

A second modification is sometimes made when the matrix \mathbf{A}_s is defined as the inverse of a matrix \mathbf{B}_s , say, so that $\mathbf{A}_s = \mathbf{B}_s^{-1}$. Then if \mathbf{B}_s is close to singular a matrix of constants, say \mathbf{C} , is added or subtracted to permit inversion, so $\mathbf{A}_s = (\mathbf{B}_s + \mathbf{C})^{-1}$. Similar adjustments can be made if \mathbf{A}_s is not positive definite. Further discussion of computation of \mathbf{A}_s is given in Section 10.3.

Gradient methods are most likely to converge to the local maximum nearest the starting values. If the objective function has multiple local optima then a range of starting values should be used to increase the chance of finding the global maximum.

10.2.4. Gradient Method Example

Consider calculation of the NLS estimator in the exponential regression model when the only regressor is the intercept. Then $E[y] = e^\beta$ and a little algebra yields the gradient $g = N^{-1} \sum_i (y_i - e^\beta) e^\beta = (\bar{y} - e^\beta) e^\beta$. Suppose in (10.1) we use $\mathbf{A}_s = e^{-2\widehat{\beta}_s}$, which corresponds to the method of scoring variant of the Newton–Raphson algorithm presented later in Section 10.3.2. The iterative method simplifies to $\widehat{\beta}_{s+1} = \widehat{\beta}_s + (\bar{y} - e^{\widehat{\beta}_s}) / e^{\widehat{\beta}_s}$.

As an example of the performance of this algorithm, suppose $\bar{y} = 2$ and the starting value is $\widehat{\beta}_1 = 0$. This leads to the iterations listed in Table 10.1. There is very rapid convergence to the NLS estimate, which for this simple example can be analytically obtained as $\widehat{\beta} = \ln \bar{y} = \ln 2 = 0.693147$. The objective function increases throughout,

Table 10.1. Gradient Method Results

Round	Estimate	Gradient	Objective Function
s	$\hat{\beta}_s$	g_s	$Q_N(\hat{\beta}_s) = -\frac{1}{2N} \sum_i (y_i - e^{\beta})^2$
1	0.000000	1.000000	$1.500000 - \sum_i y_i^2 / 2N$
2	1.000000	-1.952492	$1.742036 - \sum_i y_i^2 / 2N$
3	0.735758	-0.181711	$1.996210 - \sum_i y_i^2 / 2N$
4	0.694042	-0.003585	$1.999998 - \sum_i y_i^2 / 2N$
5	0.693147	-0.000002	$2.000000 - \sum_i y_i^2 / 2N$

a consequence of use of the NR algorithm with globally concave objective function. Note that overshooting occurs in the first iteration, from $\hat{\beta}_1 = 0.0$ to $\hat{\beta}_2 = 1.0$, greater than $\hat{\beta} = 0.693$.

Quick convergence usually occurs when the NR algorithm is used and the objective function is globally concave. The challenge in practice is that nonstandard nonlinear models often have objective functions that are not globally concave.

10.2.5. Method of Moments and GMM Estimators

For m-estimators $Q_N(\theta) = N^{-1} \sum_i q_i(\theta)$ and the gradient $\mathbf{g}(\theta) = N^{-1} \sum_i \partial q_i(\theta) / \partial \theta$.

For GMM estimators $Q_N(\theta)$ is a quadratic form (see Section 6.3.2) and the gradient takes the more complicated form

$$\mathbf{g}(\theta) = \left[N^{-1} \sum_i \partial \mathbf{h}_i(\theta)' / \partial \theta \right] \times \mathbf{W}_N \times \left[N^{-1} \sum_i \mathbf{h}_i(\theta) \right].$$

Some gradient methods can then no longer be used as they work only for averages. Methods given in Section 10.3 that can still be used include Newton-Raphson, steepest ascent, DFP, BFG, and simulated annealing.

Method of moments and estimating equations estimators are defined as solving a system of equations, but they can be converted to a numerical optimization problem similar to GMM. The estimator $\hat{\theta}$ that solves the q equations $N^{-1} \sum_i \mathbf{h}_i(\theta) = \mathbf{0}$ can be obtained by minimizing $Q_N(\theta) = [N^{-1} \sum_i \mathbf{h}_i(\theta)]' [N^{-1} \sum_i \mathbf{h}_i(\theta)]$.

10.2.6. Convergence Criteria

Iterations continue until there is virtually no change. Programs ideally stop when all of the following occur: (1) A small relative change occurs in the objective function $Q_N(\hat{\theta}_s)$; (2) a small change of the gradient vector \mathbf{g}_s occurs relative to the Hessian; and (3) a small relative change occurs in the parameter estimates $\hat{\theta}_s$. Statistical packages typically choose default threshold values for these three changes, called **convergence criteria**. These values can often be changed by the user. A conservative value is 10^{-6} .

In addition there is usually a **maximum number of iterations** that will be attempted. If this maximum is reached estimates are typically reported. The estimates should not be used, however, unless convergence has been achieved.

If convergence is achieved then a local maximum has been obtained. However, there is no guarantee that the global maximum is obtained, unless the objective function is globally concave.

10.2.7. Starting Values

The number of iterations is considerably reduced if the initial **starting values** $\hat{\theta}_1$ are close to $\hat{\theta}$. Consistent parameter estimates are obviously good estimates to use as starting values. A poor choice of starting values can lead to failure of iterative methods. In particular, for some estimators and gradient methods it may not be possible to compute \mathbf{g}_1 or \mathbf{A}_1 if the starting value is $\hat{\theta}_1 = \mathbf{0}$.

If the objective function is not globally concave it is good practice to use a range of starting values to increase the chance of obtaining a global maximum.

10.2.8. Numerical and Analytical Derivatives

Any gradient method by definition uses derivatives of the objective function. Either numerical derivatives or analytical derivatives may be used.

Numerical derivatives are computed using

$$\frac{\Delta Q_N(\hat{\theta}_s)}{\Delta \theta_j} = \frac{Q_N(\hat{\theta}_s + h\mathbf{e}_j) - Q_N(\hat{\theta}_s - h\mathbf{e}_j)}{2h}, \quad j = 1, \dots, q, \quad (10.4)$$

where h is small and $\mathbf{e}_j = (0 \dots 0 \ 1 \ 0 \dots 0)'$ is a vector with unity in the j th row and zeros elsewhere.

In theory h should be very small, as formally $\partial Q_N(\boldsymbol{\theta})/\partial \theta_j$ equals the limit of $\Delta Q_N(\boldsymbol{\theta})/\Delta \theta_j$ as $h \rightarrow 0$. In practice too small a value of h leads to inaccuracy owing to rounding error. For this reason calculations using numerical derivatives should always be done in double precision or quadruple precision rather than single precision. Although a program may use a default value such as $h = 10^{-6}$, other values will be better for any particular problem. For example, a smaller value of h is appropriate if the dependent variable y in NLS regression is measured in thousands of dollars rather than dollars (with regressors not rescaled), since then $\boldsymbol{\theta}$ will be one-thousandth the size.

A drawback of using numerical derivatives is that these derivatives have to be computed many times – for each of the q parameters, for each of the N observations, and for each of the S iterations. This requires $2qNS$ evaluations of the objective function, where each evaluation itself may be computationally burdensome.

An alternative is to use **analytical derivatives**. These will be more accurate than numerical derivatives and may be much quicker to compute, especially if the analytical derivatives are simpler than the objective function itself. Moreover, only qNS function evaluations are needed.

For methods that additionally require calculation of second derivatives to form \mathbf{A}_s there is even greater benefit to providing analytical derivatives. Even if just analytical first derivatives are given, the second derivative may then be more quickly and

accurately obtained as the numerical first derivative of the analytical first derivative. Statistical packages often provide the user with the option of providing analytical first and second derivatives.

Numerical derivatives have the advantage of requiring no coding beyond providing the objective function. This saves coding time and eliminates one possible source of user error, though some packages have the ability to take analytical derivatives.

If computational time is a factor or if there is concern about accuracy of calculations, however, it is worthwhile going to the trouble of providing analytical derivatives. It is still good practice then to check that the analytical derivatives have been correctly coded by obtaining parameter estimates using numerical derivatives, with starting values the estimates obtained using analytical derivatives.

10.2.9. Nongradient Methods

Gradient methods presume the objective function is sufficiently smooth to ensure existence of the gradient. For some examples, notably least absolute deviations (LAD), quantile regression, and maximum score estimation, there is no gradient and alternative iterative methods are used.

For example, for LAD the objective function $Q_N(\boldsymbol{\theta}_s) = N^{-1} \sum_i |y_i - \mathbf{x}_i \boldsymbol{\beta}|$ has no derivative and **linear programming methods** are used in place of gradient methods. Such examples are sufficiently rare in microeconomics that we focus almost exclusively on gradient methods.

For objective functions that are difficult to maximize, particularly because of multiple local optima, use can be made of nongradient methods such as simulated annealing (presented in Section 10.3.8) and **genetic algorithms** (see Dorsey and Mayer, 1995).

10.3. Specific Methods

The leading method for obtaining a globally concave objective function is the Newton–Raphson iterative method. The other methods, such as steepest descent and DFP, are usually learnt and employed when the Newton–Raphson method fails. Another common method is the Gauss–Newton method for the NLS estimator. This method is not as universal as the Newton–Raphson method, as it is applicable only to least-squares problems, and it can be obtained as a minor adaptation of the Newton–Raphson method. These various methods are designed to obtain a local optimum given some starting values for the parameters.

This section also presents the expectation method, which is particularly useful in missing data problems, and the method of simulated annealing, which is an example of a nongradient method and is more likely to yield a global rather than local maximum.

10.3.1. Newton–Raphson Method

The **Newton–Raphson (NR) method** is a popular gradient method that works especially well if the objective function is globally concave in $\boldsymbol{\theta}$. In this method

$$\widehat{\boldsymbol{\theta}}_{s+1} = \widehat{\boldsymbol{\theta}}_s - \mathbf{H}_s^{-1} \mathbf{g}_s, \quad (10.5)$$

where \mathbf{g}_s is defined in (10.2) and

$$\mathbf{H}_s = \left. \frac{\partial^2 Q_N(\theta)}{\partial \theta \partial \theta'} \right|_{\hat{\theta}_s} \quad (10.6)$$

is the $q \times q$ Hessian matrix evaluated at $\hat{\theta}_s$. These formulas apply to both maximization and minimization of $Q_N(\theta)$ since premultiplying $Q_N(\theta)$ by minus one changes the sign of both \mathbf{H}_s^{-1} and \mathbf{g}_s .

To motivate the NR method, begin with the s th-round estimate $\hat{\theta}_s$ for θ . Then by second-order Taylor series expansion around $\hat{\theta}_s$

$$Q_N(\theta) = Q_N(\hat{\theta}_s) + \left. \frac{\partial Q_N(\theta)}{\partial \theta'} \right|_{\hat{\theta}_s} (\theta - \hat{\theta}_s) + \frac{1}{2} (\theta - \hat{\theta}_s)' \left. \frac{\partial^2 Q_N(\theta)}{\partial \theta \partial \theta'} \right|_{\hat{\theta}_s} (\theta - \hat{\theta}_s) + R.$$

Ignoring the remainder term R and using more compact notation, we approximate $Q_N(\theta)$ by

$$Q_N^*(\theta) = Q_N(\hat{\theta}_s) + \mathbf{g}'_s (\theta - \hat{\theta}_s) + \frac{1}{2} (\theta - \hat{\theta}_s)' \mathbf{H}_s (\theta - \hat{\theta}_s),$$

where \mathbf{g}_s and \mathbf{H}_s are defined in (10.2) and (10.6). To maximize the approximation $Q_N^*(\theta)$ with respect to θ we set the derivative to zero. Then $\mathbf{g}_s + \mathbf{H}_s(\theta - \hat{\theta}_s) = \mathbf{0}$, and solving for θ yields $\hat{\theta}_{s+1} = \hat{\theta}_s - \mathbf{H}_s^{-1} \mathbf{g}_s$, which is (10.5). The NR update therefore maximizes a second-order Taylor series approximation to $Q_N(\theta)$ evaluated at $\hat{\theta}_s$.

To see whether NR iterations will necessarily increase $Q_N(\theta)$, substitute the $(s+1)$ th-round estimate back into the Taylor series approximation to obtain

$$Q_N(\hat{\theta}_{s+1}) = Q_N(\hat{\theta}_s) - \frac{1}{2} (\hat{\theta}_{s+1} - \hat{\theta}_s)' \mathbf{H}_s (\hat{\theta}_{s+1} - \hat{\theta}_s) + R.$$

Ignoring the remainder term, we see that this increases (or decreases) if \mathbf{H}_s is negative (or positive) definite. At a local maximum the Hessian is negative semi-definite, but away from the maximum this may not be the case even for well-defined problems. If the NR method strays into such territory it may not necessarily move toward the maximum. Furthermore the Hessian is then singular, in which case \mathbf{H}_s^{-1} in (10.5) cannot be computed. Clearly, the NR method works best for maximization (or minimization) problems if the objective function is globally concave (or convex), as then \mathbf{H}_s is always negative (or positive) definite. In such cases convergence often occurs within 10 iterations.

An additional attraction of the NR method arises if the starting value $\hat{\theta}_1$ is root- N consistent, that is, if $\sqrt{N}(\hat{\theta}_1 - \theta_0)$ has a proper limiting distribution. Then the second-round estimator $\hat{\theta}_2$ can be shown to have the same asymptotic distribution as the estimator obtained by iterating to convergence. There is therefore no theoretical gain to further iteration. An example is feasible GLS, where initial OLS leads to consistent regression parameter estimates, and these in turn are used to obtain consistent variance parameter estimates, which are then used to obtain efficient GLS. A second example is use of easily obtained consistent estimates as starting values before maximizing a complicated likelihood function. Although there is no need to iterate further, in practice most researchers still prefer to iterate to convergence unless this is computationally too

time consuming. One advantage of iterating to convergence is that different researchers should obtain the same parameter estimates, whereas different initial root- N consistent estimates lead to second-round parameter estimates that will differ even though they are asymptotically equivalent.

10.3.2. Method of Scoring

A common modification of the NR method is the **method of scoring** (MS). In this method the Hessian matrix is replaced by its expected value

$$\mathbf{H}_{\text{MS},s} = \mathbb{E} \left[\frac{\partial^2 Q_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right] \Bigg|_{\hat{\boldsymbol{\theta}}_s}. \quad (10.7)$$

This substitution is especially advantageous when applied to the MLE (i.e., $Q_N(\boldsymbol{\theta}) = N^{-1}\mathcal{L}_N(\boldsymbol{\theta})$), because the expected value should be negative definite, since by the information matrix equality (see Section 5.6.3), $\mathbf{H}_{\text{MS},s} = \mathbb{E} [\partial \mathcal{L}_N / \partial \boldsymbol{\theta} \times \partial \mathcal{L}_N / \partial \boldsymbol{\theta}']$, which is positive definite since it is a covariance matrix. Obtaining the expectation in (10.7) is possible only for m-estimators and even then may be analytically difficult.

The method of scoring algorithm for the MLE of generalized linear models, such as the Poisson, probit, and logit, can be shown to be implementable using iteratively reweighted least squares (see McCullagh and Nelder, 1989). This was advantageous to early adopters of these models who only had access to an OLS program.

The method of scoring can also be applied to m-estimators other than the MLE, though then $\mathbf{H}_{\text{MS},s}$ may not be negative definite.

10.3.3. BHHH Method

The **BHHH method** of Berndt, Hall, Hall, and Hausman (1974) uses (10.1) with weighting matrix $\mathbf{A}_s = -\mathbf{H}_{\text{BHHH},s}^{-1}$ where the matrix

$$\mathbf{H}_{\text{BHHH},s} = - \sum_{i=1}^N \frac{\partial q_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial q_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \Bigg|_{\hat{\boldsymbol{\theta}}_s}, \quad (10.8)$$

and $Q_N(\boldsymbol{\theta}) = \sum_i q_i(\boldsymbol{\theta})$. Compared to NR, this has the advantage of requiring evaluation of first derivatives only, offering considerable computational savings.

To justify this method, begin with the method of scoring for the MLE, in which case $Q_N(\boldsymbol{\theta}) = \sum_i \ln f_i(\boldsymbol{\theta})$, where $f_i(\boldsymbol{\theta})$ is the log-density. The information matrix equality can be expressed as

$$\mathbb{E} \left[\frac{\partial^2 \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right] = -\mathbb{E} \left[\sum_{i=1}^N \frac{\partial \ln f_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \sum_{j=1}^N \frac{\partial \ln f_j(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right],$$

and independence over i implies

$$\mathbb{E} \left[\frac{\partial^2 \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right] = - \sum_{i=1}^N \mathbb{E} \left[\frac{\partial \ln f_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln f_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right].$$

Dropping the expectation leads to (10.8).

The BHHH method can also be applied to estimators other than the MLE, in which case it is viewed as simply another choice of matrix \mathbf{A}_s in (10.1) rather than as an estimate of the Hessian matrix \mathbf{H}_s .

The BHHH method is used for many cross-section m-estimators as it can work well and requires only first derivatives.

10.3.4. Method of Steepest Ascent

The **method of steepest ascent** sets $\mathbf{A}_s = \mathbf{I}_q$, the simplest choice of weighting matrix. A line search is then done (see (10.3)) to scale \mathbf{I}_q by a constant λ_s .

The line search can be done manually. In practice it is common to use the optimal λ for the line search, which can be shown to be $\lambda_s = -\mathbf{g}'_s \mathbf{g}_s / \mathbf{g}'_s \mathbf{H}_s \mathbf{g}_s$, where \mathbf{H}_s is the Hessian matrix. This optimal λ_s requires computation of the Hessian, in which case one might instead use NR. The advantage of steepest ascent rather than NR is that \mathbf{H}_s can be singular, though \mathbf{H}_s still needs to be negative definite to ensure $\lambda_s < 0$ so that $\lambda_s \mathbf{I}_q$ is negative definite.

10.3.5. DFP and BFGS Methods

The **DFP algorithm** due to Davidon, Fletcher, and Powell is a gradient method with weighting matrix \mathbf{A}_s that is positive definite and requires computation of only first derivatives, unlike NR, which requires computation of the Hessian. Here the method is presented without derivation.

The weighting matrix \mathbf{A}_s is computed by the recursion

$$\mathbf{A}_s = \mathbf{A}_{s-1} + \frac{\boldsymbol{\delta}_{s-1} \boldsymbol{\delta}'_{s-1}}{\boldsymbol{\delta}'_{s-1} \boldsymbol{\gamma}_{s-1}} + \frac{\mathbf{A}_{s-1} \boldsymbol{\gamma}_{s-1} \boldsymbol{\gamma}'_{s-1} \mathbf{A}_{s-1}}{\boldsymbol{\gamma}'_{s-1} \mathbf{A}_{s-1} \boldsymbol{\gamma}_{s-1}}, \quad (10.9)$$

where $\boldsymbol{\delta}_{s-1} = \mathbf{A}_{s-1} \mathbf{g}_{s-1}$ and $\boldsymbol{\gamma}_{s-1} = \mathbf{g}_s - \mathbf{g}_{s-1}$. By inspection of the right-hand side of (10.9), \mathbf{A}_s will be positive definite provided the initial \mathbf{A}_0 is positive definite (e.g., $\mathbf{A}_0 = \mathbf{I}_q$).

The procedure converges quite well in many statistical applications. Eventually \mathbf{A}_s goes to the theoretically preferred $-\mathbf{H}_s^{-1}$. In principle this method can also provide an approximate estimate of the inverse of the Hessian for use in computation of standard errors, without needing either second derivatives or matrix inversion. In practice, however, this estimate can be a poor one.

A refinement of the DFP algorithm is the **BFGS algorithm** of Boyden, Fletcher, Goldfarb, and Shannon with

$$\mathbf{A}_s = \mathbf{A}_{s-1} + \frac{\boldsymbol{\delta}_{s-1} \boldsymbol{\delta}'_{s-1}}{\boldsymbol{\delta}'_{s-1} \boldsymbol{\gamma}_{s-1}} + \frac{\mathbf{A}_{s-1} \boldsymbol{\gamma}_{s-1} \boldsymbol{\gamma}'_{s-1} \mathbf{A}_{s-1}}{\boldsymbol{\gamma}'_{s-1} \mathbf{A}_{s-1} \boldsymbol{\gamma}_{s-1}} - (\boldsymbol{\gamma}'_{s-1} \mathbf{A}_{s-1} \boldsymbol{\gamma}_{s-1}) \boldsymbol{\eta}_{s-1} \boldsymbol{\eta}'_{s-1}, \quad (10.10)$$

where $\boldsymbol{\eta}_{s-1} = (\boldsymbol{\delta}_{s-1} / \boldsymbol{\delta}'_{s-1} \boldsymbol{\gamma}_{s-1}) - (\mathbf{A}_{s-1} \boldsymbol{\gamma}_{s-1} / \boldsymbol{\gamma}'_{s-1} \mathbf{A}_{s-1} \boldsymbol{\gamma}_{s-1})$.

10.3.6. Gauss–Newton Method

The **Gauss–Newton (GN) method** is an iterative method for the NLS estimator that can be implemented by iterative OLS.

Specifically, for NLS with conditional mean function $g(\mathbf{x}_i, \boldsymbol{\beta})$, the GN method sets the parameter change vector $(\widehat{\boldsymbol{\beta}}_{s+1} - \widehat{\boldsymbol{\beta}}_s)$ equal to the OLS coefficient estimates from the artificial regression

$$y_i - g(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_s) = \frac{\partial g_i}{\partial \boldsymbol{\beta}'} \Big|_{\widehat{\boldsymbol{\beta}}_s} \boldsymbol{\beta} + v_i. \quad (10.11)$$

Equivalently, $\widehat{\boldsymbol{\beta}}_{s+1}$ equals the OLS coefficient estimates from the artificial regression

$$y_i - g(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_s) - \frac{\partial g_i}{\partial \boldsymbol{\beta}'} \Big|_{\widehat{\boldsymbol{\beta}}_s} \widehat{\boldsymbol{\beta}}_s = \frac{\partial g_i}{\partial \boldsymbol{\beta}'} \Big|_{\widehat{\boldsymbol{\beta}}_s} \boldsymbol{\beta} + v_i. \quad (10.12)$$

To derive this method, let $\widehat{\boldsymbol{\beta}}_s$ be a starting value, approximate $g(\mathbf{x}_i, \boldsymbol{\beta})$ by a first-order Taylor series expansion

$$g(\mathbf{x}_i, \boldsymbol{\beta}) = g(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_s) + \frac{\partial g_i}{\partial \boldsymbol{\beta}} \Big|_{\widehat{\boldsymbol{\beta}}_s} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_s),$$

and substitute this in the least-squares objective function $Q_N(\boldsymbol{\beta})$ to obtain the approximation

$$Q_N^*(\boldsymbol{\beta}) = \sum_{i=1}^N \left(y_i - g(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_s) - \frac{\partial g_i}{\partial \boldsymbol{\beta}} \Big|_{\widehat{\boldsymbol{\beta}}_s} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_s) \right)^2.$$

But this is the sum of squared residuals for OLS regression of $y_i - g(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_s)$ on $\partial g_i / \partial \boldsymbol{\beta}' \Big|_{\widehat{\boldsymbol{\beta}}_s}$ with parameter vector $(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_s)$, leading to (10.11). More formally,

$$\widehat{\boldsymbol{\beta}}_{s+1} = \widehat{\boldsymbol{\beta}}_s + \left[\sum_i \frac{\partial g_i}{\partial \boldsymbol{\beta}} \Big|_{\widehat{\boldsymbol{\beta}}_s} \frac{\partial g_i}{\partial \boldsymbol{\beta}'} \Big|_{\widehat{\boldsymbol{\beta}}_s} \right]^{-1} \sum_i \frac{\partial g_i}{\partial \boldsymbol{\beta}} \Big|_{\widehat{\boldsymbol{\beta}}_s} (y_i - g(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_s)). \quad (10.13)$$

This is the gradient method (10.1) with vector $\mathbf{g}_s = \sum_i \partial g_i / \partial \boldsymbol{\beta} \Big|_{\widehat{\boldsymbol{\beta}}_s} (y_i - g(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_s))$ weighted by matrix $\mathbf{A}_s = [\sum_i \partial g_i / \partial \boldsymbol{\beta} \times \partial g_i / \partial \boldsymbol{\beta}']_{\widehat{\boldsymbol{\beta}}_s}^{-1}$.

The iterative method (10.13) equals the method of scoring variant of the Newton–Raphson algorithm for NLS estimation since, from Section 5.8, the second sum on the right-hand side is the gradient vector and the first sum is minus the expected value of the Hessian (see also Section 10.3.9). The Gauss–Newton algorithm is therefore a special case of the Newton–Raphson, and NR is emphasized more here as it can be applied to a much wider range of problems than can GN.

10.3.7. Expectation Maximization

There are a number of data and model formulations considered in this book that can be thought of as involving incomplete or missing data. For example, outcome variables of interest (e.g., expenditure or the length of a spell in some state) may be right-censored. That is, for some cases we may observe the actual expenditure or spell length, whereas

in other cases we may only know that the outcome exceeded some specific value, say c^* . A second example involves a multiple regression in which the data matrix looks as follows:

$$\begin{bmatrix} \mathbf{y}_1 & \mathbf{X}_1 \\ ? & \mathbf{X}_2 \end{bmatrix},$$

where ? stands for missing data. Here we envisage a situation in which we wish to estimate a linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, where $\mathbf{y}' = [\mathbf{y}_1 ?]$, $\mathbf{X}' = [\mathbf{X}_1 \mathbf{X}_2]$, but a subset of variables \mathbf{y} is missing. A third example involves estimating the parameters $(\theta_1, \theta_2, \dots, \theta_C, \pi_1, \dots, \pi_C)$ of a C -component mixture distribution, also called a latent class model, $h(\mathbf{y}|\mathbf{X}) = \sum_{j=1}^C \pi_j f_j(\mathbf{y}_j|\mathbf{X}_j, \theta_j)$, where $f_j(\mathbf{y}_j|\mathbf{X}_j, \theta_j)$ are well-defined pdfs. Here π_j ($j = 1, \dots, C$) are unknown sampling fractions corresponding to the C latent densities from which the observations are sampled. It is convenient to think of this problem also as a missing data problem in the sense that if the sampling fractions were known constants then estimation would be simpler.

The **expectation maximization (EM)** framework provides a unifying framework for developing algorithms for problems that can be interpreted as involving missing data. Although particular solutions to this type of estimation problem have long been found in the literature, Dempster, Laird, and Rubin (1977) provided a definitive treatment.

Let \mathbf{y} denote the vector dependent variable of interest, determined by the underlying latent variable vector \mathbf{y}^* . Let $f^*(\mathbf{y}^*|\mathbf{X}, \boldsymbol{\theta})$ denote the joint density of the latent variables, conditional on regressors \mathbf{X} , and let $f(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$ denote the joint density of the observed variables. Let there be a many-to-one mapping from the sample space of \mathbf{y} to that of \mathbf{y}^* ; that is, the value of the latent variable \mathbf{y}^* uniquely determines \mathbf{y} , but the value of \mathbf{y} does not uniquely determine \mathbf{y}^* . It follows that $f(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = f^*(\mathbf{y}^*|\mathbf{X}, \boldsymbol{\theta})/f(\mathbf{y}^*|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta})$, since from Bayes rule the conditional density $f(\mathbf{y}^*|\mathbf{y}) = f(\mathbf{y}, \mathbf{y}^*)/f(\mathbf{y}) = f^*(\mathbf{y}^*)/f(\mathbf{y})$, where the final equality uses $f(\mathbf{y}^*, \mathbf{y}) = f^*(\mathbf{y}^*)$ as \mathbf{y}^* uniquely determines \mathbf{y} . Rearranging gives $f(\mathbf{y}) = f^*(\mathbf{y}^*)/f(\mathbf{y}^*|\mathbf{y})$.

The MLE maximizes

$$Q_N(\boldsymbol{\theta}) = \frac{1}{N} \mathcal{L}_N(\boldsymbol{\theta}) = \frac{1}{N} \ln f^*(\mathbf{y}^*|\mathbf{X}, \boldsymbol{\theta}) - \frac{1}{N} \ln f(\mathbf{y}^*|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}). \quad (10.14)$$

Because \mathbf{y}^* is unobserved the first term in the log-likelihood is ignored. The second term is replaced by its expected value, which will not involve \mathbf{y}^* , where at the s th round this expectation is evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_s$.

The **expectation (E)** part of the **EM algorithm** calculates

$$Q_N(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}_s) = -E \left[\frac{1}{N} \ln f(\mathbf{y}^*|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}) | \mathbf{y}, \mathbf{X}, \widehat{\boldsymbol{\theta}}_s \right], \quad (10.15)$$

where expectation is with respect to the density $f(\mathbf{y}^*|\mathbf{y}, \mathbf{X}, \widehat{\boldsymbol{\theta}}_s)$. The **maximization (M)** part of the EM algorithm maximizes $Q_N(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}_s)$ to obtain $\widehat{\boldsymbol{\theta}}_{s+1}$.

The full EM algorithm is iterative. The likelihood is maximized, given the expected value of the latent variable; the expected value is evaluated afresh given the current value of $\boldsymbol{\theta}$. The iterative process continues until convergence is achieved. The EM algorithm has the advantage of always leading to an increase or constancy in $Q_N(\boldsymbol{\theta})$;

see Amemiya (1985, p. 376). The EM algorithm is applied to a latent class model in Section 18.5.3 and to missing data in Section 27.5.

There is a very extensive literature on situations where the EM algorithm can be usefully applied, even though it can be applied to only a subset of optimization problems. The EM algorithm is easy to program in many cases and its use was further encouraged by considerations of limited computing power and storage that are no longer paramount. Despite these attractions, for censored data models and latent class models direct estimation using Newton–Raphson type iterative procedures is often found to be faster and more efficient computationally.

10.3.8. Simulated Annealing

Simulated annealing (SA) is a nongradient iterative method reviewed by Goffe, Ferrier, and Rogers (1994). It differs from gradient methods in permitting movements that decrease rather than increase the objective function to be maximized, so that one is not locked in to moving steadily toward one particular local maximum.

Given a value $\hat{\theta}_s$ at the s th round we perturb the j th component of $\hat{\theta}_s$ to obtain a new trial value of

$$\theta_s^* = \hat{\theta}_s + [0 \cdots 0 \ (\lambda_j r_j) 0 \cdots 0]', \quad (10.16)$$

where λ_j is a prespecified step length and r_j is a draw from a uniform distribution on $(-1, 1)$. The new trial value is used, that is, the method sets $\hat{\theta}_{s+1} = \theta_s^*$, if it increases the objective function, or if it does not increase the value of the objective function but does pass the Metropolis criterion that

$$\exp((Q_N(\theta_s^*) - Q_N(\hat{\theta}_s))/T_s) > u, \quad (10.17)$$

where u is a drawing from a uniform $(0, 1)$ distribution and T_s is a scaling parameter called the **temperature**. Thus not only uphill moves are accepted, but downhill moves are also accepted with a probability that decreases with the difference between $Q_N(\theta_s^*)$ and $Q_N(\hat{\theta}_s)$ and that increases with the temperature. The terms simulated annealing and temperature come from analogy with minimizing thermal energy by slowly cooling (annealing) a molten metal.

The user needs to set the step-size parameter λ_j . Goffe et al. (1994) suggest periodically adjusting λ_j so that 50% of all moves over a number of iterations are accepted. The temperature also needs to be chosen and reduced during the course of iterations. Then the algorithm initially is searching over a wide range of parameter values before steadily locking in on a particular region.

Fast simulated annealing (FSA), proposed by Szu and Hartley (1987), is a faster method. It replaces the uniform $(-1, 1)$ random number r_j by a Cauchy random variable r_j scaled by the temperature and permits a fixed step length v_j . The method also uses a simpler adjustment of the temperature over iterations with T_s equal to the initial temperature divided by the number of FSA iterations, where one iteration is a full cycle over the q components of θ .

Cameron and Johansson (1997) discuss and use simulated annealing, following the methods of Horowitz (1992). This begins with FSA but on grounds of computational

savings switches to gradient methods (BFGS) when relatively little change in $Q_N(\cdot)$ occurs over a number of iterations or after many (250) FSA iterations. In a simulation they find that NR with a number of different starting values offers a considerable improvement over NR with just one set of starting values, but even better is FSA with a number of different starting values.

10.3.9. Example: Exponential Regression

Consider the nonlinear regression model with exponential conditional mean

$$E[y_i | \mathbf{x}_i] = \exp(\mathbf{x}'_i \boldsymbol{\beta}), \quad (10.18)$$

where \mathbf{x}_i and $\boldsymbol{\beta}$ are $K \times 1$ vectors. The NLS estimator $\widehat{\boldsymbol{\beta}}$ minimizes

$$Q_N(\boldsymbol{\beta}) = \sum_i (y_i - \exp(\mathbf{x}'_i \boldsymbol{\beta}))^2, \quad (10.19)$$

where for notational simplicity scaling by $2/N$ is ignored. The first-order conditions are nonlinear in $\boldsymbol{\beta}$ and there is no explicit solution for $\boldsymbol{\beta}$. Instead, gradient methods need to be used.

For this example the gradient and Hessian are, respectively,

$$\mathbf{g} = -2 \sum_i (y_i - e^{\mathbf{x}'_i \boldsymbol{\beta}}) e^{\mathbf{x}'_i \boldsymbol{\beta}} \mathbf{x}_i \quad (10.20)$$

and

$$\mathbf{H} = 2 \sum_i \left\{ e^{\mathbf{x}'_i \boldsymbol{\beta}} e^{\mathbf{x}'_i \boldsymbol{\beta}} \mathbf{x}_i \mathbf{x}'_i - 2(y_i - e^{\mathbf{x}'_i \boldsymbol{\beta}}) e^{\mathbf{x}'_i \boldsymbol{\beta}} \mathbf{x}_i \mathbf{x}'_i \right\}. \quad (10.21)$$

The NR iterative method (10.5) uses \mathbf{g}_s and \mathbf{H}_s equal to (10.20) and (10.21) evaluated at $\widehat{\boldsymbol{\beta}}_s$.

A simpler method of scoring variation of NR notes that (10.18) implies

$$E[\mathbf{H}] = 2 \sum_i e^{\mathbf{x}'_i \boldsymbol{\beta}} e^{\mathbf{x}'_i \boldsymbol{\beta}} \mathbf{x}_i \mathbf{x}'_i. \quad (10.22)$$

Using $E[\mathbf{H}_s]$ in place of \mathbf{H}_s yields

$$\widehat{\boldsymbol{\beta}}_{s+1} - \widehat{\boldsymbol{\beta}}_s = \left[\sum_i e^{\mathbf{x}'_i \widehat{\boldsymbol{\beta}}_s} e^{\mathbf{x}'_i \widehat{\boldsymbol{\beta}}_s} \mathbf{x}_i \mathbf{x}'_i \right]^{-1} \sum_i e^{\mathbf{x}'_i \widehat{\boldsymbol{\beta}}_s} \mathbf{x}_i (y_i - e^{\mathbf{x}'_i \widehat{\boldsymbol{\beta}}_s}).$$

It follows that $\widehat{\boldsymbol{\beta}}_{s+1} - \widehat{\boldsymbol{\beta}}_s$ can be computed from OLS regression of $(y_i - e^{\mathbf{x}'_i \widehat{\boldsymbol{\beta}}_s})$ on $e^{\mathbf{x}'_i \widehat{\boldsymbol{\beta}}_s} \mathbf{x}_i$. This is also the Gauss–Newton regression (10.11), since $\partial g(\mathbf{x}_i, \boldsymbol{\beta}) / \partial \boldsymbol{\beta} = \exp(\mathbf{x}'_i \widehat{\boldsymbol{\beta}}_s) \mathbf{x}_i$ for the exponential conditional mean (10.18). Specialization to $\exp(\mathbf{x}'_i \boldsymbol{\beta}) = \exp(\boldsymbol{\beta})$ gives the iterative procedure presented in Section 10.2.4.

10.4. Practical Considerations

Some practical issues have already been presented in Section 10.2, notably convergence criteria, modifications such as step-size adjustment, and the use of numerical rather than analytical derivatives. In this section a brief overview of statistical packages

is given, followed by a discussion of common pitfalls that can arise in computation of a nonlinear estimator.

10.4.1. Statistical Packages

All standard microeconometric packages such as Limdep, Stata, PCTSP, and SAS have built-in procedures to estimate basic nonlinear models such as logit and probit. These packages are simple to use, requiring no knowledge of iterative methods or even of the model being used. For example, the command for logit regression might be “logit y x” rather than the command “ols y x” for OLS. Nonlinear least squares requires some code to convey to the package the particular functional form for $g(\mathbf{x}, \boldsymbol{\beta})$ one wishes to specify. Estimation should be quick and accurate as the program should exploit the structure of the particular model. For example, if the objective function is globally concave then the method of scoring might be used.

If a statistical package does not contain a particular model then one needs to write one's own code. This situation can arise with even minor variation of standard models, such as imposing restrictions on parameters or using parameterizations that are not of single-index form. The code may be written using one's own favorite statistical package or using other more specialized programming languages. Possibilities include (1) built-in optimization procedures within the statistical package that require specification of the objective function and possibly its derivatives; (2) matrix commands within the statistical package to compute \mathbf{A}_s and \mathbf{g}_s and iterate; (3) a matrix programming language such as Gauss, Matlab, OX, SAS/IML, or S-Plus, and possibly add-on optimization routines; (4) a programming language such as Fortran or C++; and (5) an optimization package such as those in GAMS, GQOPT, or NAGLIB.

The first and second methods are attractive because they do not force the user to learn a new program. The first method is particularly simple for m-estimation as it can require merely specification of the subfunction $q_i(\boldsymbol{\theta})$ for the i th observation rather than specification of $Q_N(\boldsymbol{\theta})$. In practice, however, the optimization procedures for user-defined functions in the standard packages are more likely to encounter numerical problems than if more specialized programs are used. Moreover, for some packages the second method can require learning arcane forms of matrix commands.

For nonlinear problems, the third method is the best, although this might require the user to learn a matrix programming language from scratch. One then is set up to handle virtually any econometric problem encountered, and the optimization routines that come with matrix programming languages are usually adequate. Also, many authors make available the code used in specific papers.

The fourth and fifth methods generally require a higher level of programming sophistication than the third method. The fourth method can lead to much faster computation and the fifth method can solve the most numerically challenging optimization problems.

Other practical issues include cost of software; the software used by colleagues; and whether the software has clear error messages and useful debugging features, such as a trace program that tracks line-by-line program execution. The value of using software similar to that used by other colleagues cannot be underestimated.

Table 10.2. Computational Difficulties: A Partial Checklist

Problem	Check
Data read incorrectly	Print full descriptive statistics.
Inprecise calculation	Use analytical derivatives or numerical with different step size h .
Multicollinearity	Check condition number of $\mathbf{X}'\mathbf{X}$. Try subset of regressors.
Singular matrix in iterations	Try method not requiring matrix inversion such as DFP.
Poor starting values	Try a range of different starting values.
Model not identified	Difficult to check. Obvious are dummy variable traps.
Strange parameter values	Constant included/excluded? Iterations actually converged?
Different standard errors	Which method was used to calculate variance matrix?

10.4.2. Computational Difficulties

Computational difficulties are, in practice, situations where it is not possible to obtain an estimate of the parameters. For example, an error message may indicate that the estimator cannot be calculated because the Hessian is singular. There are many possible reasons for this, as detailed in the following and summarized in Table 10.2. These reasons may also provide explanation for another common situation of parameter estimates that are obtained but are seemingly in error.

First, the data may not have been read in correctly. This is a remarkably common oversight. With large data sets it is not practical to print out all the data. However, at a minimum one should always obtain descriptive statistics and check for anomalies such as incorrect range for a variable, unusually large or small sample mean, and unusually large or small standard deviation (including a value of zero, which indicates no variation). See Section 3.5.4 for further details.

Second, there may be calculation errors. To minimize these all calculations should be done in double precision or even quadruple precision rather than single precision. It is helpful to rescale the data so that the regressors have similar means and variances. For example, it may be better to use annual income in thousands of dollars rather than in dollars. If numerical derivatives are used it may be necessary to alter the change value h in (10.4). Care needs to be paid to how functions are evaluated. For example, the function $\ln \Gamma(y)$, where $\Gamma(\cdot)$ is the gamma function, is best evaluated using the log-gamma function. If instead one evaluates the gamma function followed by the log function considerable numerical error arises even for moderate sized y .

Third, multicollinearity may be a problem. In single-index models (see Section 5.2.4) the usual checks for multicollinearity will carry over. The correlation matrix for the regressors can be printed, though this only considers pairwise correlation. Better is to use the condition number of $\mathbf{X}'\mathbf{X}$, that is, the square root of the ratio of the largest to smallest eigenvalue of $\mathbf{X}'\mathbf{X}$. If this exceeds 100 then problems may arise. For more highly nonlinear models than single-index ones it is possible to have problems even if the condition number is not large. If one suspects multicollinearity is causing

numerical problems then see whether it is possible to estimate the model with a subset of the variables that are less likely to be collinear.

Fourth, a noninvertible Hessian during iterations does not necessarily imply singularity at the true maximum. It is worthwhile trying a range of iterative methods such as steepest ascent with line search and DFP, not just Newton–Raphson. This problem may also result from multicollinearity.

Fifth, try different starting values. The iterative gradient methods are designed to obtain a local maximum rather than the global maximum. One way to guard against this is to begin iterations at a wide range of starting values. A second way is to perform a grid search. Both of these approaches theoretically require evaluations at many different points if the dimension of θ is large, but it may be sufficient to do a detailed analysis for a stripped-down version of the model that includes just the few regressors thought to be most statistically significant.

Lastly, the model may not be identified. Indeed a standard necessary condition for model identification is that the Hessian be invertible. As with linear models, simple checks include avoiding dummy variable traps and, if a subset of data is being used in initial analysis, determining that all variables in the subset of the data have some variation. For example, if data are ordered by gender or by age or by region then problems can arise if these appear as indicator variables and the chosen subset is of individuals of a particular gender, age, or region. For nonlinear models it can be difficult to theoretically determine that the model is not identified. Often one first eliminates all other potential causes before returning to a careful analysis of model identification.

Even after parameter estimates are successfully obtained computational problems can still arise, as it may not be possible to obtain estimates of the variance matrix $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}'^{-1}$. This situation can arise when the iterative method used, such as DFP, does not use the Hessian matrix \mathbf{A}^{-1} as the weighting matrix in the iterations. First check that the iterative method has indeed converged rather than, for example, stopping at a default maximum number of iterations. If convergence has occurred, try alternative estimates of \mathbf{A} , using the expected Hessian or using more accurate numerical computations by, for example, using analytical rather than numerical derivatives. If such solutions still fail it is possible that the model is not identified, with this nonidentification being finessed at the parameter estimation stage by using an iterative method that did not compute the Hessian.

Other perceived computational problems are parameter and variance estimates that do not accord with prior beliefs. For parameter estimates obvious checks include ensuring correct treatment of an intercept term (inclusion or exclusion, depending on the context), that convergence has been achieved, and that a global maximum is obtained (by trying a range of starting values). If standard errors of parameter estimates differ across statistical packages that give the same parameter estimates, the most likely cause is that a different method has been used to construct the variance matrix estimate (see Section 5.5.2).

A good computational strategy is to start with a small subset of the data and regressors, say one regressor and 100 observations. This simplifies detailed tracing of the program either manually, such as by printing out key output along the way, or using

a built-in trace facility if the program has one. If the program passes this test then computational problems with the full model and data are less likely to be due to incorrect data input or coding errors and are more likely due to genuine computational difficulties such as multicollinearity or poor starting values.

A good way to test program validity is to construct a simulated data set where the true parameters are known. For a large sample size, say $N = 10,000$, the estimated parameter values should be close to the true values.

Finally, note that obtaining reasonable computational results from estimation of a nonlinear model does not guarantee correct results. For example, many early published applications of multinomial probit models reported apparently sensible results, yet the models estimated have subsequently been determined to be not identified (see Section 15.8.1).

10.5. Bibliographic Notes

Numerical problems can arise even in linear models, and it is instructive to read Davidson and MacKinnon (1993, Section 1.5) and Greene (2003, appendix E). Standard references for statistical computation are Kennedy and Gentle (1980) and especially Press et al. (1993) and related co-authored books by Press. For evaluation of functions the standard reference is Abramowitz and Stegun (1971). Quandt (1983) presents many computational issues, including optimization.

5.3 Summaries of iterative methods are given in Amemiya (1985, Section 4.4), Davidson and MacKinnon (1993, Section 6.7), Maddala (1977, Section 9.8), and especially Greene (2003, appendix E.6). Harvey (1990) gives many applications of the GN algorithm, which, owing to its simplicity, is the usual iterative method for NLS estimation. For the EM algorithm see especially Amemiya (1985, pp. 375–378). For SA see Goffe et al. (1994).

Exercises

10–1 Consider calculation of the MLE in the logit regression model when the only regressor is the intercept. Then $E[y] = 1/(1 + e^{-\beta})$ and the gradient of the scaled log-likelihood function $g(\beta) = (y - 1/(1 + e^{-\beta}))$. Suppose a sample yields $\bar{y} = 0.8$ and the starting value is $\beta = 0.0$.

- (a) Calculate β for the first six iterations of the Newton–Raphson algorithm.
- (b) Calculate the first six iterations of a gradient algorithm that sets $A_s = 1$ in (10.1), so $\hat{\beta}_{s+1} = \hat{\beta}_s + g_s$.
- (c) Compare the performance of the methods in parts (a) and (b).

10–2 Consider the nonlinear regression model $y = \alpha x_1 + \gamma/(x_2 - \delta) + u$, where x_1 and x_2 are exogenous regressors independent of the iid error $u \sim \mathcal{N}[0, \sigma^2]$.

- (a) Derive the equation for the Gauss–Newton algorithm for estimating (α, γ, δ) .
- (b) Derive the equation for the Newton–Raphson algorithm for estimating (α, γ, δ) .
- (c) Explain the importance of not arbitrarily choosing the starting values of the algorithm.

10–3 Suppose that the pdf of y has a C -component mixture form, $f(y|\pi) = \sum_{j=1}^C \pi_j f_j(y)$, where $\pi = (\pi_1, \dots, \pi_C)$, $\pi_j > 0$, $\sum_{j=1}^C \pi_j = 1$. The π_j are

unknown mixing proportions whereas the parameters of the densities $f_j(y)$ are presumed known.

- (a) Given a random sample on y_i , $i = 1, \dots, N$, write the general log-likelihood function and obtain the first-order conditions for $\hat{\pi}_{\text{ML}}$. Verify that there is no explicit solution for $\hat{\pi}_{\text{ML}}$.
- (b) Let \mathbf{z}_i be a $C \times 1$ vector of latent categorical variables, $i = 1, \dots, N$, such that $z_{ji} = 1$ if y comes from the j th component of the mixture and $z_{ji} = 0$ otherwise. Write down the likelihood function in terms of the observed and latent variables as if the latent variable were observed.
- (c) Devise an EM algorithm for estimating π . [Hint: If z_{ji} were observable the MLE of $\hat{\pi}_j = N^{-1} \sum_i z_{ji}$. The E step requires calculation of $E[z_{ji}|y_i]$; the M step requires replacing z_{ji} by $E[z_{ji}|y_i]$ and then solving for π .]

- 10–4** Let (y_{1i}, y_{2i}) , $i = 1, \dots, N$, have a bivariate normal distribution with mean (μ_1, μ_2) and covariance parameters $(\sigma_{11}, \sigma_{12}, \sigma_{22})$ and correlation coefficient ρ . Suppose that all N observations on y_1 are available but there are $m < N$ missing observations on y_2 . Using the fact that the marginal distribution of y_j is $\mathcal{N}[\mu_j, \sigma_{jj}]$, and that conditionally $y_2|y_1 \sim \mathcal{N}[\mu_{2.1}, \sigma_{22.1}]$, where $\mu_{2.1} = \mu_2 + \sigma_{12}/\sigma_{22}(y_1 - \mu_1)$, $\sigma_{22.1} = (1 - \rho^2)\sigma_{22}$, devise an EM algorithm for imputing the missing observations on y_1 .

PART THREE

Simulation-Based Methods

Part 1 emphasized that microeconometric models are frequently nonlinear models estimated using large and heterogeneous data sets drawn from surveys that are complex and subject to a variety of sampling biases. A realistic depiction of the economic phenomena in such settings often requires the use of models for which estimation and subsequent statistical inference are difficult. Advances in computing hardware and software now make it feasible to tackle such tasks. Part 3 presents modern, computer-intensive, simulation-based methods of estimation and inference that mitigate some of these difficulties. The background required to cover this material varies somewhat with the chapter, but the essential base is least squares and maximum likelihood estimation.

Chapter 11 presents bootstrap methods for statistical inference. These methods have the attraction of providing a simple way to obtain standard errors when the formulae from asymptotic theory are complex, as is the case, for example, for some two-step estimators. Furthermore, if implemented appropriately, a bootstrap can lead to a more refined asymptotic theory that may then lead to better statistical inference in small samples.

Chapter 12 presents simulation-based estimation methods. These methods permit estimation in situations where standard computational methods may not permit calculation of an estimator, because of the presence of an integral over a probability distribution that leads to no closed-form solution.

Chapter 13 surveys Bayesian methods that provide an approach to estimation and inference that is quite different from the classical approach used in other chapters of this book. Despite this different approach, in practice in large sample settings the Bayesian approach produces similar results to those from classical methods. Further, they often do so in a computationally more efficient manner.

Bootstrap Methods

11.1. Introduction

Exact finite-sample results are unavailable for most microeconomics estimators and related test statistics. The statistical inference methods presented in preceding chapters rely on asymptotic theory that usually leads to limit normal and chi-square distributions.

An alternative approximation is provided by the bootstrap, due to Efron (1979, 1982). This approximates the distribution of a statistic by a Monte Carlo simulation, with sampling done from the empirical distribution or the fitted distribution of the observed data. The additional computation required is usually feasible given advances in computing power. Like conventional methods, however, bootstrap methods rely on asymptotic theory and are only exact in infinitely large samples.

The wide range of bootstrap methods can be classified into two broad approaches. First, the simplest bootstrap methods can permit statistical inference when conventional methods such as standard error computation are difficult to implement. Second, more complicated bootstraps can have the additional advantage of providing asymptotic refinements that can lead to a better approximation in-finite samples. Applied researchers are most often interested in the first aspect of the bootstrap. Theoreticians emphasize the second, especially in settings where the usual asymptotic methods work poorly in finite samples.

The econometrics literature focuses on use of the bootstrap in hypothesis testing, which relies on approximation of probabilities in the tails of the distributions of statistics. Other applications are to confidence intervals, estimation of standard errors, and bias reduction. The bootstrap is straightforward to implement for smooth \sqrt{N} -consistent estimators based on iid samples, though bootstraps with asymptotic refinements are underutilized. Caution is needed in other settings, including nonsmooth estimators such as the median, nonparametric estimators, and inference for data that are not iid.

A reasonably self-contained summary of the bootstrap is provided in Section 11.2, an example is given in Section 11.3, and some theory is provided in Section 11.4.

Further variations of the bootstrap are presented in Section 11.5. Section 11.6 presents use of the bootstrap for specific types of data and specific methods used often in microeconomics.

11.2. Bootstrap Summary

We summarize key bootstrap methods for estimator $\hat{\theta}$ and associated statistics based on an iid sample $\{\mathbf{w}_1, \dots, \mathbf{w}_N\}$, where usually $\mathbf{w}_i = (y_i, \mathbf{x}_i)$ and $\hat{\theta}$ is a smooth estimator that is \sqrt{N} consistent and asymptotically normally distributed. For notational simplicity we generally present results for scalar θ . For vector θ in most instances replace θ by θ_j , the j th component of θ .

Statistics of interest include the usual regression output: the estimate $\hat{\theta}$; standard errors $s_{\hat{\theta}}$; t -statistic $t = (\hat{\theta} - \theta_0)/s_{\hat{\theta}}$, where θ_0 is the null hypothesis value; the associated critical value or p -value for this statistic; and a confidence interval.

This section presents bootstraps for each of these statistics. Some motivation is also provided, with the underlying theory sketched in Section 11.4.

11.2.1. Bootstrap without Refinement

Consider estimation of the variance of the sample mean $\hat{\mu} = \bar{y} = N^{-1} \sum_{i=1}^N y_i$, where the scalar random variable y_i is iid $[\mu, \sigma^2]$, when it is not known that $V[\hat{\mu}] = \sigma^2/N$.

The variance of $\hat{\mu}$ could be obtained by obtaining S such samples of size N from the population, leading to S sample means and hence S estimates $\hat{\mu}_s = \bar{y}_s$, $s = 1, \dots, S$. Then we could estimate $V[\hat{\mu}]$ by $(S-1)^{-1} \sum_{s=1}^S (\hat{\mu}_s - \bar{\mu})^2$, where $\bar{\mu} = S^{-1} \sum_{s=1}^S \hat{\mu}_s$.

Of course this approach is not possible, as we only have one sample. A bootstrap can implement this approach by viewing the sample as the population. Then the finite population is now the actual data y_1, \dots, y_N . The distribution of $\hat{\mu}$ can be obtained by drawing B bootstrap samples from this population of size N , where each bootstrap sample of size N is obtained by sampling from y_1, \dots, y_N with replacement. This leads to B sample means and hence B estimates $\hat{\mu}_b = \bar{y}_b$, $b = 1, \dots, B$. Then estimate $V[\hat{\mu}]$ by $(B-1)^{-1} \sum_{b=1}^B (\hat{\mu}_b - \bar{\mu})^2$, where $\bar{\mu} = B^{-1} \sum_{b=1}^B \hat{\mu}_b$. Sampling with replacement may seem to be a departure from usual sampling methods, but in fact standard sampling theory assumes sampling with replacement rather than without replacement (see Section 24.2.2).

With additional information other ways to obtain bootstrap samples may be possible. For example, if it is known that $y_i \sim \mathcal{N}[\mu, \sigma^2]$ then we could obtain B bootstrap samples of size N by drawing from the $\mathcal{N}[\hat{\mu}, s^2]$ distribution. This bootstrap is an example of a parametric bootstrap, whereas the preceding bootstrap was from the empirical distribution.

More generally, for estimator $\hat{\theta}$ similar bootstraps can be used to, for example, estimate $V[\hat{\theta}]$ and hence standard errors when analytical formulas for $V[\hat{\theta}]$ are complex. Such bootstraps are usually valid for observations \mathbf{w}_i that are iid over i , and they have similar properties to estimates obtained using the usual asymptotic theory.

11.2.2. Asymptotic Refinements

In some settings it is possible to improve on the preceding bootstrap and obtain estimates that are equivalent to those obtained using a more refined asymptotic theory that may better approximate the finite-sample distribution of $\widehat{\theta}$. Much of this chapter is directed to such **asymptotic refinements**.

Usual asymptotic theory uses the result that $\sqrt{N}(\widehat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}[0, \sigma^2]$. Thus

$$\Pr[\sqrt{N}(\widehat{\theta} - \theta_0)/\sigma \leq z] = \Phi(z) + R_1, \quad (11.1)$$

where $\Phi(\cdot)$ is the standard normal cdf and R_1 is a remainder term that disappears as $N \rightarrow \infty$.

This result is based on asymptotic theory detailed in Section 5.3 that includes application of a central limit theorem. The CLT is based on a truncated power-series expansion. The Edgeworth expansion, detailed in Section 11.4.3, includes additional terms in the expansion. With one extra term this yields

$$\Pr[\sqrt{N}(\widehat{\theta} - \theta_0)/\sigma \leq z] = \Phi(z) + \frac{g_1(z)\phi(z)}{\sqrt{N}} + R_2, \quad (11.2)$$

where $\phi(\cdot)$ is the standard normal density, $g_1(\cdot)$ is a bounded function given after (11.13) in Section 11.4.3 and R_2 is a remainder term that disappears as $N \rightarrow \infty$.

The Edgeworth expansion is difficult to implement theoretically as the function $g_1(\cdot)$ is data dependent in a complicated way. A bootstrap *with asymptotic refinement* provides a simple computational method to implement the Edgeworth expansion. The theory is given in Section 11.4.4.

Since $R_1 = O(N^{-1/2})$ and $R_2 = O(N^{-1})$, asymptotically $R_2 < R_1$, leading to a better approximation as $N \rightarrow \infty$. However, in finite samples it is possible that $R_2 > R_1$. A bootstrap with asymptotic refinement provides a better approximation asymptotically that hopefully leads to a better approximation in samples of the finite sizes typically used. Nevertheless, there is no guarantee and simulation studies are frequently used to verify that finite-sample gains do indeed occur.

11.2.3. Asymptotically Pivotal Statistic

For asymptotic refinement to occur, the statistic being bootstrapped must be an **asymptotically pivotal statistic**, meaning a statistic whose limit distribution does not depend on unknown parameters. This result is explained in Section 11.4.4.

As an example, consider sampling from $y_i \sim [\mu, \sigma^2]$. Then the estimate $\widehat{\mu} = \bar{y} \stackrel{a}{\sim} \mathcal{N}[\mu, \sigma^2/N]$ is not asymptotically pivotal even given a null hypothesis value $\mu = \mu_0$ since its distribution depends on the unknown parameter σ^2 . However, the **studentized statistic** $t = (\widehat{\mu} - \mu_0)/s_{\widehat{\mu}} \stackrel{a}{\sim} \mathcal{N}[0, 1]$ is asymptotically pivotal.

Estimators are usually not asymptotically pivotal. However, conventional asymptotically standard normal or chi-squared distributed test statistics, including Wald, Lagrange multiplier, and likelihood ratio tests, and related confidence intervals, are asymptotically pivotal.

11.2.4. The Bootstrap

In this section we provide a broad description of the bootstrap, with further details given in subsequent sections.

Bootstrap Algorithm

A general **bootstrap algorithm** is as follows:

1. Given data $\mathbf{w}_1, \dots, \mathbf{w}_N$, draw a bootstrap sample of size N using a method given in the following and denote this new sample $\mathbf{w}_1^*, \dots, \mathbf{w}_N^*$.
2. Calculate an appropriate statistic using the bootstrap sample. Examples include (a) the estimate $\hat{\theta}^*$ of θ , (b) the standard error $s_{\hat{\theta}}^*$ of the estimate $\hat{\theta}^*$, and (c) a t -statistic $t^* = (\hat{\theta}^* - \hat{\theta})/s_{\hat{\theta}}^*$ centered at the original estimate $\hat{\theta}$. Here $\hat{\theta}^*$ and $s_{\hat{\theta}}^*$ are calculated in the usual way but using the new bootstrap sample rather than the original sample.
3. Repeat steps 1 and 2 B independent times, where B is a large number, obtaining B bootstrap replications of the statistic of interest, such as $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$ or t_1^*, \dots, t_B^* .
4. Use these B bootstrap replications to obtain a bootstrapped version of the statistic, as detailed in the following subsections.

Implementation can vary according to how bootstrap samples are obtained, how many bootstraps are performed, what statistic is being bootstrapped, and whether or not that statistic is asymptotically pivotal.

Bootstrap Sampling Methods

The bootstrap dgp in step 1 is used to approximate the true unknown dgp.

The simplest bootstrapping method is to use the empirical distribution of the data, which treats the sample as being the population. Then $\mathbf{w}_1^*, \dots, \mathbf{w}_N^*$ are obtained by sampling with replacement from $\mathbf{w}_1, \dots, \mathbf{w}_N$. In each bootstrap sample so obtained, some of the original data points will appear multiple times whereas others will not appear at all. This method is an **empirical distribution function (EDF) bootstrap** or **nonparametric bootstrap**. It is also called a **paired bootstrap** since in single-equation regression models $\mathbf{w}_i = (y_i, \mathbf{x}_i)$, so here both y_i and \mathbf{x}_i are resampled.

Suppose the conditional distribution of the data is specified, say $y|\mathbf{x} \sim F(\mathbf{x}, \boldsymbol{\theta}_0)$, and an estimate $\hat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}_0$ is available. Then in step 1 we can instead form a bootstrap sample by using the original \mathbf{x}_i while generating y_i by random draws from $F(\mathbf{x}_i, \hat{\boldsymbol{\theta}})$. This corresponds to regressors fixed in repeated samples (see Section 4.4.5). Alternatively, we may first resample \mathbf{x}_i^* from $\mathbf{x}_1, \dots, \mathbf{x}_N$ and then generate y_i from $F(\mathbf{x}_i^*, \hat{\boldsymbol{\theta}})$, $i = 1, \dots, N$. Both are examples of a **parametric bootstrap** that can be applied in fully parametric models.

For regression model with additive iid error, say $y_i = g(\mathbf{x}_i, \boldsymbol{\beta}) + u_i$, we can form fitted residuals $\hat{u}_1, \dots, \hat{u}_N$, where $\hat{u}_i = y_i - g(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$. Then in step 1 bootstrap from these residuals to get a new draw of residuals, say $(\hat{u}_1^*, \dots, \hat{u}_N^*)$, leading to a bootstrap sample $(y_1^*, \mathbf{x}_1), \dots, (y_N^*, \mathbf{x}_N)$, where $y_i^* = g(\mathbf{x}_i, \hat{\boldsymbol{\beta}}) + u_i^*$. This bootstrap is called a

residual bootstrap. It uses information intermediate between the nonparametric and parametric bootstrap. It can be applied if the error term has distribution that does not depend on unknown parameters.

We emphasize the paired bootstrap on grounds of its simplicity, applicability to a wide range of nonlinear models, and reliance on weak distributional assumptions. However, the other bootstraps generally provide a better approximation (see Horowitz, 2001, p. 3185) and should be used if the stronger model assumptions they entail are warranted.

The Number of Bootstraps

The bootstrap asymptotics rely on $N \rightarrow \infty$ and so the bootstrap can be asymptotically valid even for low B . However, clearly the bootstrap is more accurate as $B \rightarrow \infty$. A sufficiently large value of B varies with one's tolerance for bootstrap-induced simulation error and with the purpose of the bootstrap.

Andrews and Buchinsky (2000) present an application-specific numerical method to determine the number of replications B needed to ensure a given level of accuracy or, equivalently, the level of accuracy obtained for a given value of B . Let λ denote the quantity of interest, such as a standard error or a critical value, $\widehat{\lambda}_\infty$ denote the ideal bootstrap estimate with $B = \infty$, and $\widehat{\lambda}_B$ denote the estimate with B bootstraps. Then Andrews and Buchinsky (2000) show that

$$\sqrt{B}(\widehat{\lambda}_B - \widehat{\lambda}_\infty)/\widehat{\lambda}_\infty \xrightarrow{d} \mathcal{N}[0, \omega],$$

where ω varies with the application and is defined in Table III of Andrews and Buchinsky (2000). It follows that $\Pr[\delta \leq z_{\tau/2}\sqrt{\omega/B}] = 1 - \tau$, where $\delta = |\widehat{\lambda}_B - \widehat{\lambda}_\infty|/\widehat{\lambda}_\infty$ denotes the relative discrepancy caused by only B replications. Thus $B \geq \omega z_{\tau/2}^2/\delta^2$ ensures the relative discrepancy is less than δ with probability at least $1 - \tau$. Alternatively, given B replications the relative discrepancy is less than $\delta = z_{\tau/2}\sqrt{\omega/B}$.

To provide concrete guidelines we propose the rule of thumb that

$$B = 384\omega.$$

This ensures that the relative discrepancy is less than 10% with probability at least 0.95, since $z_{0.025}^2/0.1^2 = 384$. The only difficult part in implementation is estimation of ω , which varies with the application.

For standard error estimation $\omega = (2 + \gamma_4)/4$, where γ_4 is the coefficient of excess kurtosis for the bootstrap estimator $\widehat{\theta}^*$. Intuitively, fatter tails in the distribution of the estimator mean outliers are more likely, contaminating standard error estimation. It follows that $B = 384 \times (1/2) = 192$ is enough if $\gamma_4 = 0$ whereas $B = 960$ is needed if $\gamma_4 = 8$. These values are higher than those proposed by Efron and Tibsharani (1993, p. 52), who state that $B = 200$ is almost always enough.

For a symmetric two-sided test or confidence interval at level α , $\omega = \alpha(1 - \alpha)/[2z_{\alpha/2}\phi(z_{\alpha/2})]^2$. This leads to $B = 348$ for $\alpha = 0.05$ and $B = 685$ for $\alpha = 0.01$. As expected more bootstraps are needed the further one goes into the tails of the distribution.

For a one-sided test or nonsymmetric two-sided test or confidence interval at level α , $\omega = \alpha(1 - \alpha)/[z_\alpha \phi(z_\alpha)]^2$. This leads to $B = 634$ for $\alpha = 0.05$ and $B = 989$ for $\alpha = 0.01$. More bootstraps are needed when testing in one tail. For chi-squared tests with h degrees of freedom $\omega = \alpha(1 - \alpha)/[\chi_\alpha^2(h)f(\chi_\alpha^2(h))]^2$, where $f(\cdot)$ is the $\chi^2(h)$ density.

For test p -values $\omega = (1 - p)/p$. For example, if $p = 0.05$ then $\omega = 19$ and $B = 7,296$. Many more bootstraps are needed for precise calculation of the test p -value compared to hypothesis rejection if a critical value is exceeded.

For bias-corrected estimation of θ a simple rule uses $\widehat{\omega} = \widehat{\sigma}^2/\widehat{\theta}^2$, where the estimator $\widehat{\theta}$ has standard error $\widehat{\sigma}$. For example, if the usual t -statistic $t = \widehat{\theta}/\widehat{\sigma} = 2$ then $\widehat{\omega} = 1/4$ and $B = 96$. Andrews and Buchinsky (2000) provide many more details and refinements of these results.

For hypothesis testing, Davidson and MacKinnon (2000) provide an alternative approach. They focus on the loss of power caused by bootstrapping with finite B . (Note that there is no power loss if $B = \infty$.) On the basis of simulations they recommend at least $B = 399$ for tests at level 0.05, and at least $B = 1,499$ for tests at level 0.01. They argue that for testing their approach is superior to that of Andrews and Buchinsky.

Several other papers by Davidson and MacKinnon, summarized in MacKinnon (2002), emphasize practical considerations in bootstrap inference. For hypothesis testing at level α choose B so that $\alpha(B + 1)$ is an integer. For example, at $\alpha = 0.05$ let $B = 399$ rather than 400. If instead $B = 400$ it is unclear on an upper one-sided alternative test whether the 20th or 21st largest bootstrap t -statistic is the critical value. For nonlinear models computation can be reduced by performing only a few Newton–Raphson iterations in each bootstrap sample from starting values equal to the initial parameter estimates.

11.2.5. Standard Error Estimation

The **bootstrap estimate of variance** of an estimator is the usual formula for estimating a variance, applied to the B bootstrap replications $\widehat{\theta}_1^*, \dots, \widehat{\theta}_B^*$:

$$s_{\widehat{\theta}, \text{Boot}}^2 = \frac{1}{B-1} \sum_{b=1}^B (\widehat{\theta}_b^* - \overline{\widehat{\theta}}^*)^2, \quad (11.3)$$

where

$$\overline{\widehat{\theta}}^* = B^{-1} \sum_{b=1}^B \widehat{\theta}_b^*. \quad (11.4)$$

Taking the square root yields $s_{\widehat{\theta}, \text{Boot}}$, the **bootstrap estimate of the standard error**.

This bootstrap provides no asymptotic refinement. Nonetheless, it can be extraordinarily useful when it is difficult to obtain standard errors using conventional methods. There are many examples. The estimate $\widehat{\theta}$ may be a **sequential two-step m-estimator** whose standard error is difficult to compute using the results given in Section 6.8. The estimate $\widehat{\theta}$ may be a 2SLS estimator estimated using a package that

only reports standard errors assuming homoskedastic errors but the errors are actually **heteroskedastic**. The estimate $\hat{\theta}$ may be a **function of other parameters** that are actually estimated, for example, $\hat{\theta} = \hat{\alpha}/\hat{\beta}$, and the bootstrap can be used instead of the delta method. For **clustered data** with many small clusters, such as short panels, cluster-robust standard errors can be obtained by resampling the clusters.

Since the bootstrap estimate $s_{\hat{\theta},\text{Boot}}$ is consistent, it can be used in place of $s_{\hat{\theta}}$ in the usual asymptotic formula to form confidence intervals and hypothesis tests that are asymptotically valid. Thus asymptotic statistical inference is possible in settings where it is difficult to obtain standard errors by other methods. However, there will be **no improvement** in finite-sample performance. To obtain an asymptotic refinement the methods of the next section are needed.

11.2.6. Hypothesis Testing

Here we consider tests on an individual coefficient, denoted θ . The test may be either an upper one-tailed alternative of $H_0 : \theta \leq \theta_0$ against $H_a : \theta > \theta_0$ or a two-sided test of $H_0 : \theta = \theta_0$ against $H_a : \theta \neq \theta_0$. Other tests are deferred to Section 11.6.3.

Tests with Asymptotic Refinement

The usual test statistic $T_N = (\hat{\theta} - \theta_0)/s_{\hat{\theta}}$ provides the potential for asymptotic refinement, as it is asymptotically pivotal since its asymptotic standard normal distribution does not depend on unknown parameters. We perform B bootstrap replications producing B test statistics t_1^*, \dots, t_B^* , where

$$t_b^* = (\hat{\theta}_b^* - \hat{\theta})/s_{\hat{\theta}_b^*}. \quad (11.5)$$

The estimates t_b^* are centered around the original estimate $\hat{\theta}$ since resampling is from a distribution centered around $\hat{\theta}$. The empirical distribution of t_1^*, \dots, t_B^* , ordered from smallest to largest, is then used to approximate the distribution of T_N as follows.

For an upper one-tailed alternative test the **bootstrap critical value** (at level α) is the upper α quantile of the B ordered test statistics. For example, if $B = 999$ and $\alpha = 0.05$ then the critical value is the 950th highest value of t^* , since then $(B + 1)(1 - \alpha) = 950$. For a similar lower tail one-sided test the critical value is the 50th smallest value of t^* .

One can also compute a **bootstrap p-value** in the obvious way. For example, if the original statistic t lies between the 914th and 915th largest values of 999 bootstrap replicates then the p -value for a upper one-tailed alternative test is $1 - 914/(B + 1) = 0.086$.

For a two-sided test a distinction needs to be made between symmetrical and nonsymmetrical tests. For a **nonsymmetrical test** or **equal-tailed test** the bootstrap **critical values** (at level α) are the lower $\alpha/2$ and upper $\alpha/2$ quantiles of the ordered test statistics t^* , and the null hypothesis is rejected at level α if the original t -statistic lies outside this range. For a **symmetrical test** we instead order $|t^*|$ and the bootstrap

critical value (at level α) is the upper α quantile of the ordered $|t^*|$. The null hypothesis is rejected at level α if $|t|$ exceeds this critical value.

These tests, using the **percentile- t method**, provide asymptotic refinements. For a one-sided t -test and for a nonsymmetrical two-sided t -test the true size of the test is $\alpha + O(N^{-1/2})$ with standard asymptotic critical values and $\alpha + O(N^{-1})$ with bootstrap critical values. For a two-sided symmetrical t -test or for an asymptotic chi-square test the asymptotic approximations work better, and the true size of the test is $\alpha + O(N^{-1})$ using standard asymptotic critical values and $\alpha + O(N^{-2})$ using bootstrap critical values.

Tests without Asymptotic Refinement

Alternative bootstrap methods can be used that although asymptotically valid do not provide an asymptotic refinement.

One approach already mentioned at the end of Section 11.2.5 is to compute $t = (\hat{\theta} - \theta_0)/s_{\hat{\theta},\text{boot}}$, where the bootstrap estimate $s_{\hat{\theta},\text{boot}}$ given in (11.3) replaces the usual estimate $s_{\hat{\theta}}$, and compare this test statistic to critical values from the standard normal distribution.

A second approach, exposited here for a two-sided test of $H_0 : \theta = \theta_0$ against $H_a : \theta \neq \theta_0$, finds the lower $\alpha/2$ and upper $\alpha/2$ quantiles of the bootstrap estimates $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$ and rejects H_0 if θ_0 falls outside this region. This is called the **percentile method**. Asymptotic refinement is obtained by using t_b^* in (11.5) that centers around $\hat{\theta}$ rather than θ_0 and using a different standard error $s_{\hat{\theta}}^*$ in each bootstrap.

These two bootstraps have the attraction of not requiring computation of $s_{\hat{\theta}}$, the usual standard error estimate based on asymptotic theory.

11.2.7. Confidence Intervals

Much of the statistics literature considers confidence interval estimation rather than its flip side of hypothesis tests. Here instead we began with hypothesis tests, so only a brief presentation of confidence intervals is necessary.

An asymptotic refinement is based on the t -statistic, which is asymptotically pivotal. Thus from steps 1–3 in Section 11.2.4 we obtain bootstrap replication t -statistics t_1^*, \dots, t_B^* . Then let $t_{[1-\alpha/2]}^*$ and $t_{[\alpha/2]}^*$ denote the lower and upper $\alpha/2$ quantiles of these t -statistics. The **percentile- t method** 100(1 – α) percent confidence interval is

$$(\hat{\theta} - t_{[1-\alpha/2]}^* \times s_{\hat{\theta}}, \hat{\theta} + t_{[\alpha/2]}^* \times s_{\hat{\theta}}), \quad (11.6)$$

where $\hat{\theta}$ and $s_{\hat{\theta}}$ are the estimate and standard error from the original sample.

An alternative is the **bias-corrected and accelerated (BC_a) method** detailed in Efron (1987). This offers an asymptotic refinement in a wider class of problems than the percentile- t method.

Other methods provide an asymptotically valid confidence interval, but without asymptotic refinement. First, one can use the bootstrap estimate of the standard

error in the usual confidence interval formula, leading to interval $(\widehat{\theta} - z_{[1-\alpha/2]} \times s_{\widehat{\theta}, \text{boot}}, \widehat{\theta} + z_{[\alpha/2]} \times s_{\widehat{\theta}, \text{boot}})$. Second, the **percentile method** confidence interval is the distance between the lower $\alpha/2$ and upper $\alpha/2$ quantiles of the B bootstrap estimates $\widehat{\theta}_1^*, \dots, \widehat{\theta}_B^*$ of θ .

11.2.8. Bias Reduction

Nonlinear estimators are usually biased in finite samples, though this bias goes to zero asymptotically if the estimator is consistent. For example, if μ^3 is estimated by $\widehat{\theta} = \bar{y}^3$, where y_i is iid $[\mu, \sigma^2]$, then $E[\widehat{\theta} - \mu^3] = 3\mu\sigma^2/N + E[(y - \mu)^3]/N^2$.

More generally, for a \sqrt{N} -consistent estimator

$$E[\widehat{\theta} - \theta_0] = \frac{a_N}{N} + \frac{b_N}{N^2} + \frac{c_N}{N^3} + \dots, \quad (11.7)$$

where a_N , b_N , and c_N are bounded constants that vary with the data and estimator (see Hall, 1992, p. 53). An alternative estimator $\widetilde{\theta}$ provides an asymptotic refinement if

$$E[\widetilde{\theta} - \theta_0] = \frac{B_N}{N^2} + \frac{C_N}{N^3} + \dots, \quad (11.8)$$

where B_N and C_N are bounded constants. For both estimators the bias disappears as $N \rightarrow \infty$. The latter estimator has the attraction that the bias goes to zero at a faster rate, and hence it is an asymptotic refinement, though in finite samples it is possible that $(B_N/N^2) > (a_N/N + b_N/N^2)$.

We wish to estimate the bias $E[\widehat{\theta}] - \theta$. This is the distance between the expected value or population average value of the parameter and the parameter value generating the data. The bootstrap replaces the population with the sample, so that the bootstrap samples are generated by parameter $\widehat{\theta}$, which has average value $\overline{\widehat{\theta}}^*$ over the bootstraps. The **bootstrap estimate of the bias** is then

$$\text{Bias}_{\widehat{\theta}} = (\overline{\widehat{\theta}}^* - \widehat{\theta}), \quad (11.9)$$

where $\overline{\widehat{\theta}}^*$ is defined in (11.4).

Suppose, for example, that $\widehat{\theta} = 4$ and $\overline{\widehat{\theta}}^* = 5$. Then the estimated bias is $(5 - 4) = 1$, an upward bias of 1. Since $\widehat{\theta}$ overestimates by 1, bias correction requires *subtracting* 1 from $\widehat{\theta}$, giving a bias-corrected estimate of 3. More generally, the **bootstrap bias-corrected estimator** of θ is

$$\begin{aligned} \widehat{\theta}_{\text{Boot}} &= \widehat{\theta} - (\overline{\widehat{\theta}}^* - \widehat{\theta}) \\ &= 2\widehat{\theta} - \overline{\widehat{\theta}}^*. \end{aligned} \quad (11.10)$$

Note that $\overline{\widehat{\theta}}^*$ itself is not the bias-corrected estimate. For more details on the direction of the correction, which may seem puzzling, see Efron and Tibsharani (1993, p. 138). For typical \sqrt{N} -consistent estimators the asymptotic bias of $\widehat{\theta}$ is $O(N^{-1})$ whereas the asymptotic bias of $\widehat{\theta}_{\text{Boot}}$ is instead $O(N^{-2})$.

In practice bias correction is seldom used for \sqrt{N} -consistent estimators, as the bootstrap estimate can be more variable than the original estimate $\widehat{\theta}$ and the bias is often

small relative to the standard error of the estimate. Bootstrap bias correction is used for estimators that converge at rate less than \sqrt{N} , notably nonparametric regression and density estimators.

11.3. Bootstrap Example

As a bootstrap example, consider the exponential regression model introduced in Section 5.9. Here the data are generated from an exponential distribution with an exponential mean with two regressors:

$$\begin{aligned} y_i | \mathbf{x}_i &\sim \text{exponential}(\lambda_i), \quad i = 1, \dots, 50, \\ \lambda_i &= \exp(\beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i}), \\ (x_{2i}, x_{3i}) &\sim \mathcal{N}[0.1, 0.1; 0.1^2, 0.1^2, 0.005], \\ (\beta_1, \beta_2, \beta_3) &= (-2, 2, 2). \end{aligned}$$

Maximum likelihood estimation on a sample of 50 observations yields $\hat{\beta}_1 = -2.192$; $\hat{\beta}_2 = 0.267$, $s_2 = 1.417$, and $t_2 = 0.188$; and $\hat{\beta}_3 = 4.664$, $s_3 = 1.741$, and $t_3 = 2.679$. For this ML example the standard errors were based on $-\hat{\mathbf{A}}^{-1}$, minus the inverse of the estimated Hessian matrix.

We concentrate on statistical inference for β_3 and demonstrate the bootstrap for standard error computation, test of statistical significance, confidence intervals, and bias correction. The differences between bootstrap and usual asymptotic estimates are relatively small in this example and can be much larger in other examples.

The results reported here are based on the paired bootstrap (see Section 11.2.4) with (y_i, x_{2i}, x_{3i}) jointly resampled with replacement $B = 999$ times. From Table 11.1, the 999 bootstrap replication estimates $\hat{\beta}_{3,b}^*, b = 1, \dots, 999$, had mean 4.716 and standard deviation of 1.939. Table 11.1 also gives key percentiles for $\hat{\beta}_3^*$ and t_3^* (defined in the following).

A parametric bootstrap could have been used instead. Then bootstrap samples would be obtained by drawing y_i from the exponential distribution with parameter $\exp(\hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i})$. In the case of tests of $H_0 : \beta_3 = 0$ the exponential parameter could instead be $\exp(\tilde{\beta}_1 + \tilde{\beta}_2 x_{2i})$, where $\tilde{\beta}_1$ and $\tilde{\beta}_2$ are then the restricted ML estimates from the original sample.

Standard errors: From (11.3) the bootstrap estimate of standard error is computed using the usual standard deviation formula for the 999 bootstrap replication estimates of β_3 . This yields estimate 1.939 compared to the usual asymptotic standard error estimate of 1.741. Note that this bootstrap offers no refinement and would only be used as a check or if finding the standard error by other means proved difficult.

Hypothesis testing with asymptotic refinement: We consider test of $H_0 : \beta_3 = 0$ against $H_a : \beta_3 \neq 0$ at level 0.05. A test with asymptotic refinement is based on the t -statistic, which is asymptotically pivotal. From Section 11.2.6 for each bootstrap we compute $t_3^* = (\hat{\beta}_3^* - 4.664)/s_{\hat{\beta}_3^*}$, which is centered on the estimate $\hat{\beta}_3 = 4.664$ from the original sample. For a nonsymmetrical test the bootstrap critical values

Table 11.1. Bootstrap Statistical Inference on a Slope Coefficient: Example^a

	$\hat{\beta}_3^*$	t_3^*	$z = t(\infty)$	$t(47)$
Mean	4.716	0.026	1.021	1.000
SD ^b	1.939	1.047	1.000	1.021
1%	−.336	−2.664	−2.326	−2.408
2.5%	0.501	−2.183	−1.960	−2.012
5%	1.545	−1.728	−1.645	−1.678
25%	3.570	−0.621	−0.675	−0.680
50%	4.772	0.062	0.000	0.000
75%	5.971	0.703	0.675	0.680
95%	7.811	1.706	1.645	1.678
97.5%	8.484	2.066	1.960	2.012
99.0%	9.427	2.529	2.326	2.408

^a Summary statistics and percentiles based on 999 paired bootstrap resamples for (1) estimate $\hat{\beta}_3^*$; (2) the associated statistics $t_3^* = (\hat{\beta}_3^* - \hat{\beta}_3)/s_{\hat{\beta}_3^*}$; (3) student t -distribution with 47 degrees of freedom; (4) standard normal distribution. Original dgp is one draw from the exponential distribution given in the text; the sample size is 50.

^b SD, standard deviation.

equal the lower and upper 2.5 percentiles of the 999 values of t_3^* , the 25th lowest and 25th highest values. From Table 11.1 these are −2.183 and 2.066. Since the t -statistic computed from the original sample $t_3 = (4.664 - 0)/1.741 = 2.679 > 2.066$, the null hypothesis is rejected. A symmetrical test that instead uses the upper 5 percentile of $|t_3^*|$ yields bootstrap critical value 2.078 that again leads to rejection of H_0 at level 0.05.

The bootstrap critical values in this example exceed those using the asymptotic approximation of either standard normal or $t(47)$, an ad hoc finite-sample adjustment motivated by the exact result for linear regression under normality. So the usual asymptotic results in this example lead to overrejection and have actual size that exceeds the nominal size. For example, at 5% the z critical region values of (−1.960, 1.960) are smaller than the bootstrap critical values (−2.183, 2.066). Figure 11.1 plots the bootstrap estimate based on t_3^* of the density of the t -test, smoothed using kernel methods, and compares it to the standard normal. The two densities appear close, though the left tail is notably fatter for the bootstrap estimate. Table 11.1 makes clearer the difference in the tails.

Hypothesis testing without asymptotic refinement: Alternative bootstrap testing methods can be used but do not offer an asymptotic refinement. First, using the bootstrap standard error estimate of 1.939, rather than the asymptotic standard error estimate of 1.741, yields $t_3 = (4.664 - 0)/1.939 = 2.405$. This leads to rejection at level 0.05 using either standard normal or $t(47)$ critical values. Second, from Table 11.1, 95% of the bootstrap estimates $\hat{\beta}_3^*$ lie in the range (0.501, 8.484), which does not include the hypothesized value of 0, so again we reject $H_0 : \beta_3 = 0$.

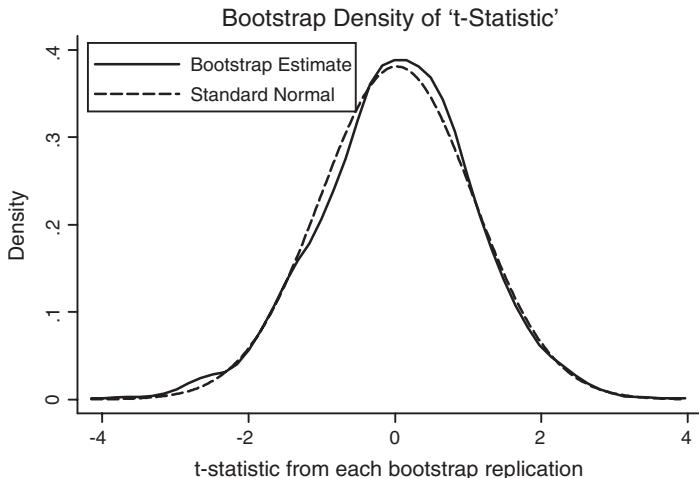


Figure 11.1: Bootstrap density of t -test statistic for slope equal to zero obtained from 999 bootstrap replications with standard normal density plotted for comparison. Data are generated from an exponential distribution regression model.

Confidence intervals: An asymptotic refinement is obtained using the 95% percentile- t confidence interval. Applying (11.6) yields $(4.664 - 2.183 \times 1.741, 4.664 + 2.066 \times 1.741)$ or $(0.864, 8.260)$. This compares to a conventional 95% asymptotic confidence interval of $4.664 \pm 1.960 \times 1.741$ or $(1.25, 8.08)$.

Other confidence intervals can be constructed, but these do not have an asymptotic refinement. Using the bootstrap standard error estimate leads to a 95% confidence interval $4.664 \pm 1.960 \times 1.939 = (0.864, 8.464)$. The percentile method uses the lower and upper 2.5 percentiles of the 999 bootstrap coefficient estimates, leading to a 95% confidence interval of $(0.501, 8.484)$.

Bias correction: The mean of the 999 bootstrap replication estimates of β_3 is 4.716, compared to the original estimate of 4.664. The estimated bias of $(4.716 - 4.664) = 0.052$ is quite small, especially compared to the standard error of $s_3 = 1.741$. The estimated bias is upward and (11.10) yields a bias-corrected estimate of β_3 equal to $4.664 - 0.052 = 4.612$.

The bootstrap relies on asymptotic theory and may actually provide a finite-sample approximation worse than that of conventional methods. To determine that the bootstrap is really an improvement here we need a full Monte Carlo analysis with, say, 1,000 samples of size 50 drawn from the exponential dgp, with each of these samples then bootstrapped, say, 999 times.

11.4. Bootstrap Theory

The exposition here follows the comprehensive survey of Horowitz (2001). Key results are consistency of the bootstrap and, if the bootstrap is applied to an asymptotically pivotal statistic, asymptotic refinement.

11.4.1. The Bootstrap

We use X_1, \dots, X_N as generic notation for the data, where for notational simplicity bold is not used for X_i even though it is usually a vector, such as (y_i, \mathbf{x}_i) . The data are assumed to be independent draws from distribution with cdf $F_0(x) = \Pr[X \leq x]$. In the simplest applications F_0 is in a finite-dimensional family, with $F_0 = F_0(x, \theta_0)$.

The statistic being considered is denoted $T_N = T_N(X_1, \dots, X_N)$. The exact finite-sample distribution of T_N is $G_N = G_N(t, F_0) = \Pr[T_N \leq t]$. The problem is to find a good approximation to G_N .

Conventional asymptotic theory uses the asymptotic distribution of T_N , denoted $G_\infty = G_\infty(t, F_0)$. This may theoretically depend on unknown F_0 , in which case we use a consistent estimate of F_0 . For example, use $\hat{F}_0 = F_0(\cdot, \hat{\theta})$, where $\hat{\theta}$ is consistent for θ_0 .

The empirical bootstrap takes a quite different approach to approximating $G_N(\cdot, F_0)$. Rather than replace G_N by G_∞ , the population cdf F_0 is replaced by a consistent estimator F_N of F_0 , such as the empirical distribution of the sample.

$G_N(\cdot, F_N)$ cannot be determined analytically but can be approximated by bootstrapping. One bootstrap resample with replacement yields the statistic $T_N^* = T_N(X_1^*, \dots, X_N^*)$. Repeating this step B independent times yields replications $T_{N,1}^*, \dots, T_{N,B}^*$. The empirical cdf of $T_{N,1}^*, \dots, T_{N,B}^*$ is the bootstrap estimate of the distribution of T , yielding

$$\hat{G}_N(t, F_N) = \frac{1}{B} \sum_{b=1}^B \mathbf{1}(T_{N,b}^* \leq t), \quad (11.11)$$

where $\mathbf{1}(A)$ equals one if event A occurs and equals zero otherwise. This is just the proportion of the bootstrap resamples for which the realized $T_N^* \leq t$.

The notation is summarized in Table 11.2.

11.4.2. Consistency of the Bootstrap

The bootstrap estimate $\hat{G}_N(t, F_N)$ clearly converges to $G_N(t, F_N)$ as the number of bootstraps $B \rightarrow \infty$. Consistency of the bootstrap estimate $\hat{G}_N(t, F_N)$ for $G_N(t, F_0)$

Table 11.2. *Bootstrap Theory Notation*

Quantity	Notation
Sample (iid)	X_1, \dots, X_N , where X_i is usually a vector
Population cdf of X	$F_0 = F_0(x, \theta_0) = \Pr[X \leq x]$
Statistic of interest	$T_N = T_N(X_1, \dots, X_N)$
Finite sample cdf of T_N	$G_N = G_N(t, F_0) = \Pr[T_N \leq t]$
Limit cdf of T_N	$G_\infty = G_\infty(t, F_0)$
Asymptotic cdf of T_N	$\hat{G}_\infty = G_\infty(t, \hat{F}_0)$, where $\hat{F}_0 = F_0(x, \hat{\theta})$
Bootstrap cdf of T_N	$\hat{G}_N(t, F_N) = B^{-1} \sum_{b=1}^B \mathbf{1}(T_{N,b}^* \leq t)$

therefore requires that

$$G_N(t, F_N) \xrightarrow{P} G_N(t, F_0),$$

uniformly in the statistic t and for all F_0 in the space of permitted cdfs.

Clearly, F_N must be consistent for F_0 . Additionally, *smoothness* in the dgp $F_0(x)$ is needed, so that $F_N(x)$ and $F_0(x)$ are close to each other uniformly in the observations x for large N . Moreover, *smoothness* in $G_N(\cdot, F)$, the cdf of the statistic considered as a functional of F , is required so that $G_N(\cdot, F_N)$ is close to $G_N(\cdot, F_0)$ when N is large.

Horowitz (2001, pp. 3166–3168) gives two formal theorems, one general and one for iid data, and provides examples of potential failure of the bootstrap, including estimation of the median and estimation with boundary constraints on parameters.

Subject to consistency of F_N for F_0 and smoothness requirements on F_0 and G_N , the bootstrap leads to consistent estimates and asymptotically valid inference. The bootstrap is consistent in a very wide range of settings.

11.4.3. Edgeworth Expansions

An additional attraction of the bootstrap is that it allows for asymptotic refinement. Singh (1981) provided a proof using Edgeworth expansions, which we now introduce.

Consider the asymptotic behavior of $Z_N = \sum_i X_i / \sqrt{N}$, where for simplicity X_i are standardized scalar random variables that are iid $[0, 1]$. Then application of a central limit theorem leads to a limit standard normal distribution for Z_N . More precisely, Z_N has cdf

$$G_N(z) = \Pr[Z_N \leq z] = \Phi(z) + O(N^{-1/2}), \quad (11.12)$$

where $\Phi(\cdot)$ is the standard normal cdf. The remainder term is ignored and regular asymptotic theory approximates $G_N(z)$ by $G_\infty(z) = \Phi(z)$.

The CLT leading to (11.12) is formally derived by a simple approximation of the **characteristic function** of Z_N , $E[e^{isZ_N}]$, where $i = -\sqrt{1}$. A better approximation expands this characteristic function in powers of $N^{-1/2}$. The usual **Edgeworth expansion** adds two additional terms, leading to

$$G_N(z) = \Pr[Z_N \leq z] = \Phi(z) + \frac{g_1(z)}{\sqrt{N}} + \frac{g_2(z)}{N} + O(N^{-3/2}), \quad (11.13)$$

where $g_1(z) = -(z^2 - 1)\phi(z)\kappa_3/6$, $\phi(\cdot)$ denotes the standard normal density, κ_3 is the third cumulant of Z_N , and the lengthy expression for $g_2(\cdot)$ is given in Rothenberg (1984, p. 895) or Amemiya (1985, p. 93). In general the r th **cumulant** κ_r is the r th coefficient in the series expansion $\ln(E[e^{isZ_N}]) = \sum_{r=0}^{\infty} \kappa_r(is)^r/r!$ of the log characteristic function or cumulant generating function.

The remainder term in (11.13) is ignored and an Edgeworth expansion approximates $G_N(z, F_0)$ by $G_\infty(z, F_0) = \Phi(z) + N^{-1/2}g_1(z) + N^{-1}g_2(z)$. If Z_N is a test statistic this can be used to compute p -values and critical values. Alternatively, (11.13) can be

inverted to

$$\Pr \left[Z_N + \frac{h_1(z)}{\sqrt{N}} + \frac{h_2(z)}{N} \leq z \right] \simeq \Phi(z), \quad (11.14)$$

for functions $h_1(z)$ and $h_2(z)$ given in Rothenberg (1984, p. 895). The left-hand side gives a modified statistic that will be better approximated by the standard normal than the original statistic Z_N .

The problem in application is that the cumulants of Z_N are needed to evaluate the functions $g_1(z)$ and $g_2(z)$ or $h_1(z)$ and $h_2(z)$. It can be very difficult to obtain analytical expressions for these cumulants (e.g., Sargan, 1980, and Phillips, 1983). The bootstrap provides a numerical method to implement the Edgeworth expansion without the need to calculate cumulants, as shown in the following.

11.4.4. Asymptotic Refinement via Bootstrap

We now return to the more general setting of Section 11.4.1, with the additional assumption that T_N has a limit normal distribution and usual \sqrt{N} asymptotics apply.

Conventional asymptotic methods use the limit cdf $G_\infty(t, F_0)$ as an approximation to the true cdf $G_N(t, F_0)$. For \sqrt{N} -consistent asymptotically normal estimators this has an error that in the limit behaves as a multiple of $N^{-1/2}$. We write this as

$$G_N(t, F_0) = G_\infty(t, F_0) + O(N^{-1/2}), \quad (11.15)$$

where in our example $G_\infty(t, F_0) = \Phi(t)$.

A better approximation is possible using an Edgeworth expansion. Then

$$G_N(t, F_0) = G_\infty(t, F_0) + \frac{g_1(t, F_0)}{\sqrt{N}} + \frac{g_2(t, F_0)}{N} + O(N^{-3/2}). \quad (11.16)$$

Unfortunately, as already noted, the functions $g_1(\cdot)$ and $g_2(\cdot)$ on the right-hand side can be difficult to construct.

Now consider the bootstrap estimator $G_N(t, F_N)$. An Edgeworth expansion yields

$$G_N(t, F_N) = G_\infty(t, F_N) + \frac{g_1(t, F_N)}{\sqrt{N}} + \frac{g_2(t, F_N)}{N} + O(N^{-3/2}); \quad (11.17)$$

see Hall (1992) for details. The bootstrap estimator $G_N(t, F_N)$ is used to approximate the finite-sample cdf $G_N(t, F_0)$. Subtracting (11.16) from (11.17), we get

$$\begin{aligned} G_N(t, F_N) - G_N(t, F_0) &= [G_\infty(t, F_N) - G_\infty(t, F_0)] \\ &\quad + \frac{[g_1(t, F_N) - g_1(t, F_0)]}{\sqrt{N}} + O(N^{-1}). \end{aligned} \quad (11.18)$$

Assume that F_N is \sqrt{N} consistent for the true cdf F_0 , so that $F_N - F_0 = O(N^{-1/2})$. For continuous function G_∞ the first term on the right-hand side of (11.18), $[G_\infty(t, F_N) - G_\infty(t, F_0)]$, is therefore $O(N^{-1/2})$, so $G_N(t, F_N) - G_N(t, F_0) = O(N^{-1/2})$.

The bootstrap approximation $G_N(t, F_N)$ is therefore in general no closer asymptotically to $G_N(t, F_0)$ than is the usual asymptotic approximation $G_\infty(t, F_0)$; see (11.15).

Now suppose the statistic T_N is *asymptotically pivotal*, so that its asymptotic distribution G_∞ does not depend on unknown parameters. Here this is the case if T_N is standardized so that its limit distribution is the standard normal. Then $G_\infty(t, F_N) = G_\infty(t, F_0)$, so (11.18) simplifies to

$$G_N(t, F_N) - G_N(t, F_0) = N^{-1/2}[g_1(t, F_N) - g_1(t, F_0)] + O(N^{-1}). \quad (11.19)$$

However, because $F_N - F_0 = O(N^{-1/2})$ we have that $[g_1(t, F_N) - g_1(t, F_0)] = O(N^{-1/2})$ for g_1 continuous in F . It follows upon simplification that $G_N(t, F_N) = G_N(t, F_0) + O(N^{-1})$. The bootstrap approximation $G_N(t, F_N)$ is now a better asymptotic approximation to $G_N(t, F_0)$ as the error is now $O(N^{-1})$.

In summary, for a bootstrap on an asymptotically pivotal statistic we have

$$G_N(t, F_0) = G_N(t, F_N) + O(N^{-1}), \quad (11.20)$$

an improvement on the conventional approximation $G_N(t, F_0) = G_\infty(t, F_0) + O(N^{-1/2})$.

The bootstrap on an asymptotically pivotal statistic therefore leads to an improved small-sample performance in the following sense. Let α be the nominal size for a test procedure. Usual asymptotic theory produces t -tests with actual size $\alpha + O(N^{-1/2})$, whereas the bootstrap produces t -tests with actual size $\alpha + O(N^{-1})$.

For symmetric two-sided hypothesis tests and confidence intervals the bootstrap on an asymptotically pivotal statistic can be shown to have approximation error $O(N^{-3/2})$ compared to error $O(N^{-1})$ using usual asymptotic theory.

The preceding results are restricted to asymptotically normal statistics. For chi-squared distributed test statistics the asymptotic gains are similar to those for symmetric two-sided hypothesis tests. For proof of bias reduction by bootstrapping, see Horowitz (2001, p. 3172).

The theoretical analysis leads to the following points. The bootstrap should be from distribution F_N consistent for F_0 . The bootstrap requires smoothness and continuity in F_0 and G_N , so that a modification of the standard bootstrap is needed if, for example, there is a discontinuity because of a boundary constraint on the parameters such as $\theta \geq 0$. The bootstrap assumes existence of low-order moments, as low-order cumulants appear in the function g_1 in the Edgeworth expansions. Asymptotic refinement requires use of an asymptotically pivotal statistic. The bootstrap refinement presented assumes iid data, so that modification is needed even for heteroskedastic errors. For more complete discussion see Horowitz (2001).

11.4.5. Power of Bootstrapped Tests

The analysis of the bootstrap has focused on getting tests with correct size in small samples. The size correction of the bootstrap will lead to changes in the power of tests, as will any size correction.

Intuitively, if the actual size of a test using first-order asymptotics exceeds the nominal size, then bootstrapping with asymptotic refinement will not only reduce the size toward the nominal size but, because of less frequent rejection, will also reduce the power of the test. Conversely, if the actual size is less than the nominal size then

bootstrapping will increase test power. This is observed in the simulation exercise of Horowitz (1994, p. 409). Interestingly, in his simulation he finds that although bootstrapping first-order asymptotically equivalent tests leads to tests with similar actual size (essentially equal to the nominal size) there can be considerable difference in test power across the bootstrapped tests.

11.5. Bootstrap Extensions

The bootstrap methods presented so far emphasize smooth \sqrt{N} -consistent asymptotically normal estimators based on iid data. The following extensions of the bootstrap permit for a wider range of applications a consistent bootstrap (Sections 11.5.1 and 11.5.2) or a consistent bootstrap with asymptotic refinement (Sections 11.5.3–11.5.5). The presentation of these more advanced methods is brief. Some are used in Section 11.6.

11.5.1. Subsampling Method

The **subsampling method** uses a sample of size m that is substantially smaller than the sample size N . The subsampling may be with replacement (Bickel, Gotze, and van Zwet, 1997) or without replacement (Politis and Romano, 1994).

Replacement subsampling provides subsamples that are random samples of the population, rather than random samples of an estimate of the distribution such as the sample in the case of a paired bootstrap. Replacement subsampling can then be consistent when failure of the smoothness conditions discussed in Section 11.4.2 leads to inconsistency of a full sample bootstrap. The associated asymptotic error for testing or confidence intervals, however, is of higher order of magnitude than the usual $O(N^{-1/2})$ obtained when a full sample bootstrap without refinement can be used.

Subsample bootstraps are useful when full sample bootstraps are invalid, or as a way to verify that a full sample bootstrap is valid. Results will differ with the choice of subsample size. And there is a considerable increase in sample error because a smaller fraction of the sample is being used. Indeed, we should have $(m/N) \rightarrow 0$ and $N \rightarrow \infty$. Politis, Romano, and Wolf (1999) and Horowitz (2001) provide further details.

11.5.2. Moving Blocks Bootstrap

The **moving blocks bootstrap** is used for data that are dependent rather than independent. This splits the sample into r nonoverlapping blocks of length l , where $rl \simeq N$. First, one samples with replacement from these blocks, to give r new blocks, which will have a different temporal ordering from the original r blocks. Then one estimates the parameters using this bootstrap sample.

The moving blocks method treats the randomly drawn blocks as being independent of each other, but allows dependence within the blocks. A similar blocking was actually used by Anderson (1971) to derive a central limit theorem for an m -dependent process. The moving blocks process requires $r \rightarrow \infty$ as $N \rightarrow \infty$ to ensure that we

are likely to draw consecutive blocks uncorrelated with each other. It also requires the block length $l \rightarrow \infty$ as $N \rightarrow \infty$. See, for example, Götze and Künsch (1996).

11.5.3. Nested Bootstrap

A **nested bootstrap**, introduced by Hall (1986), Beran (1987), and Loh (1987), is a bootstrap within a bootstrap. This method is especially useful if the bootstrap is on a statistic that is not asymptotically pivotal. In particular, if the standard error of the estimate is difficult to compute one can bootstrap the current bootstrap sample to obtain a bootstrap standard error estimate $s_{\hat{\theta}^*, \text{Boot}}$ and form $t^* = (\hat{\theta}^* - \hat{\theta})/s_{\hat{\theta}^*, \text{Boot}}$, and then apply the percentile- t method to the bootstrap replications t_1^*, \dots, t_B^* . This permits asymptotic refinements where a single round of bootstrap would not.

More generally, **iterated bootstrapping** is a way to improve the performance of the bootstrap by estimating the errors (i.e., bias) that arise from a single pass of the bootstrap, and correcting for these errors. In general each further iteration of the bootstrap reduces bias by a factor N^{-1} if the statistic is asymptotically pivotal and by a factor $N^{-1/2}$ otherwise. For a good exposition see Hall and Martin (1988). If B bootstraps are performed at each iteration then B^k bootstraps need to be performed if there are k iterations. For this reason at most two iterations, called a **double bootstrap** or **calibrated bootstrap**, are done.

Davison, Hinkley, and Schechtman (1986) proposed **balanced bootstrapping**. This method ensures that each sample observation is reused exactly the same number of times over all B bootstraps, leading to better bootstrap estimates. For implementation see Gleason (1988), whose algorithms add little to computational time compared to the usual unbalanced bootstrap.

11.5.4. Recentering and Rescaling

To yield an asymptotic refinement the bootstrap should be based on an estimate \widehat{F} of the dgp F_0 that imposes all the conditions of the model under consideration. A leading example arises with the residual bootstrap.

Least-squares residuals do not sum to zero in nonlinear models, or even in linear models if there is no intercept. The residual bootstrap (see Section 11.2.4) based on least-squares residuals will then fail to impose the restriction that $E[u_i] = 0$. The residual bootstrap should instead bootstrap the **recentered residual** $\widehat{u}_i - \bar{u}$, where $\bar{u} = N^{-1} \sum_{i=1}^N \widehat{u}_i$. Similar recentering should be done for paired bootstraps of GMM estimators in overidentified models (see Section 11.6.4).

Rescaling of residuals can also be useful. For example, in the linear regression model with iid errors resample from $(N/(N - K))^{1/2} \widehat{u}_i$ since these have variance s^2 . Other adjustments include using the standardized residual $\widehat{u}_i/\sqrt{(1 - h_{ii})s^2}$, where h_{ii} is the i th diagonal entry in the projection matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

11.5.5. The Jackknife

The bootstrap can be used for bias correction (see Section 11.2.8). An alternative resampling method is the jackknife, a precursor of the bootstrap. The jackknife uses N

deterministically defined subsamples of size $N - 1$ obtained by dropping in turn each of the N observations and recomputing the estimator.

To see how the jackknife works, let $\hat{\theta}_N$ denote the estimate of θ using all N observations, and let $\hat{\theta}_{N-1}$ denote the estimate of θ using the first $(N - 1)$ observations. If (11.7) holds then $E[\hat{\theta}_N] = \theta + a_N/N + b_N/N^2 + O(N^{-3})$ and $E[\hat{\theta}_{N-1}] = \theta + a_N/(N - 1) + b_N/(N - 1)^2 + O(N^{-3})$, which implies $E[N\hat{\theta}_N - (N - 1)\hat{\theta}_{N-1}] = \theta + O(N^{-2})$. Thus $N\hat{\theta}_N - (N - 1)\hat{\theta}_{N-1}$ has smaller bias than $\hat{\theta}_N$.

The estimator can be more variable, however, as it uses less of the data. As an extreme example, if $\hat{\theta} = \bar{y}$ then the new estimator is simply y_N , the N th observation. The variation can be reduced by dropping each observation in turn and averaging.

More formally then, consider the estimator $\hat{\theta}$ of a parameter vector θ based on a sample of size N from iid data. For $i = 1, \dots, N$ sequentially delete the i th observation and obtain N jackknife replication estimates $\hat{\theta}_{(-i)}$ from the N jackknife resamples of size $(N - 1)$. The **jackknife estimate of the bias** of $\hat{\theta}$ is $(N - 1)(\bar{\hat{\theta}} - \hat{\theta})$, where $\bar{\hat{\theta}} = N^{-1} \sum_i \hat{\theta}_{(-i)}$ is the average of the N jackknife replications $\hat{\theta}_{(-i)}$. The bias appears large because of multiplication by $(N - 1)$, but the differences $(\hat{\theta}_{(-i)} - \hat{\theta})$ are much smaller than in the bootstrap case since a jackknife resample differs from the original sample in only one observation.

This leads to the bias-corrected **jackknife estimate** of θ :

$$\begin{aligned}\hat{\theta}_{\text{Jack}} &= \hat{\theta} - (N - 1)(\bar{\hat{\theta}} - \hat{\theta}) \\ &= N\hat{\theta} - (N - 1)\bar{\hat{\theta}}.\end{aligned}\tag{11.21}$$

This reduces the bias from $O(N^{-1})$ to $O(N^{-2})$, which is the same order of bias reduction as for the bootstrap. It is assumed that, as for the bootstrap, the estimator is a smooth \sqrt{N} -consistent estimator. The jackknife estimate can have increased variance compared with $\hat{\theta}$, and examples where the jackknife fails are given in Miller (1974).

A simple example is estimation of σ^2 from an iid sample with $y_i \sim [\mu, \sigma^2]$. The estimate $\hat{\sigma}^2 = N^{-1} \sum_i (y_i - \bar{y})^2$, the MLE under normality, has $E[\hat{\sigma}^2] = \sigma^2(N - 1)/N$ so that the bias equals σ^2/N , which is $O(N^{-1})$. In this example the jackknife estimate can be shown to simplify to $\hat{\sigma}_{\text{Jack}}^2 = (N - 1)^{-1} \sum_i (y_i - \bar{y})^2$, so one does not need to compute N separate estimates $\hat{\sigma}_{(-i)}^2$. This is an unbiased estimate of σ^2 , so the bias is actually zero rather than the general result of $O(N^{-2})$.

The jackknife is due to Quenouille (1956). Tukey (1958) considered application to a wider range of statistics. In particular, the **jackknife estimate of the standard error** of an estimator $\hat{\theta}$ is

$$\widehat{\text{se}}_{\text{Jack}}[\hat{\theta}] = \left[\frac{N - 1}{N} \sum_{i=1}^N (\hat{\theta}_{(-i)} - \bar{\hat{\theta}})^2 \right]^{1/2}.\tag{11.22}$$

Tukey proposed the term jackknife by analogy to a Boy Scout jackknife that solves a variety of problems, each of which could be solved more efficiently by a specially constructed tool. The jackknife is a “rough and ready” method for bias reduction in many situations, but it is not the ideal method in any. The jackknife can be viewed as a linear approximation of the bootstrap (Efron and Tibsharani, 1993, p. 146). It requires

less computation than the bootstrap in small samples, as then $N < B$ is likely, but is outperformed by the bootstrap as $B \rightarrow \infty$.

Consider the linear regression model $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$, with $\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. An example of a biased estimator from OLS regression is a time-series model with lagged dependent variable as regressor. The regression estimator based on the i th jackknife sample $(\mathbf{X}_{(-i)}, \mathbf{y}_{(-i)})$ is given by

$$\begin{aligned}\widehat{\beta}_{(-i)} &= [\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)}]^{-1}\mathbf{X}'_{(-i)}\mathbf{y}_{(-i)} \\ &= [\mathbf{X}'\mathbf{X} - \mathbf{x}_i\mathbf{x}'_i]^{-1}(\mathbf{X}'\mathbf{y} - \mathbf{x}_i y_i) \\ &= \widehat{\beta} - [\mathbf{X}'\mathbf{X}]^{-1}\mathbf{x}_i(y_i - \mathbf{x}'_i\widehat{\beta}_{(-i)}).\end{aligned}$$

The third equality avoids the need to invert $\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)}$ for each i and is obtained using

$$[\mathbf{X}'\mathbf{X}]^{-1} = [\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)}]^{-1} - \frac{[\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)}]^{-1}\mathbf{x}_i\mathbf{x}'_i[\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)}]^{-1}}{1 + \mathbf{x}'_i[\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)}]^{-1}\mathbf{x}_i}.$$

Here the **pseudo-values** are given by $N\widehat{\beta} - (N - 1)\widehat{\beta}_{(-i)}$, and the jackknife estimator of $\widehat{\beta}$ is given by

$$\widehat{\beta}_{\text{Jack}} = N\widehat{\beta} - (N - 1)\frac{1}{N} \sum_{i=1}^N \widehat{\beta}_{(-i)}. \quad (11.23)$$

An interesting application of the jackknife to bias reduction is the jackknife IV estimator (see Section 6.4.4).

11.6. Bootstrap Applications

We consider application of the bootstrap taking into account typical microeconometric complications such as heteroskedasticity and clustering and more complicated estimators that can lead to failure of simple bootstraps.

11.6.1. Heteroskedastic Errors

For least squares in models with additive errors that are heteroskedastic, the standard procedure is to use White's heteroskedastic-consistent covariance matrix estimator (HCCME). This is well known to perform poorly in small samples. When done correctly, the bootstrap can provide an improvement.

The paired bootstrap leads to valid inference, since the essential assumption that (y_i, \mathbf{x}_i) is iid still permits $V[u_i | \mathbf{x}_i]$ to vary with \mathbf{x}_i (see Section 4.4.7). However, it does not offer an asymptotic refinement because it does not impose the condition that $E[u_i | \mathbf{x}_i] = 0$.

The usual residual bootstrap actually leads to invalid inference, since it assumes that $u_i | \mathbf{x}_i$ is iid and hence erroneously imposes the condition of homoskedastic errors. In terms of Section 11.4 theory, \widehat{F} is then inconsistent for F . One can specify a formal model for heteroskedasticity, say $u_i = \exp(\mathbf{z}'_i\alpha)\varepsilon_i$, where ε_i are iid, obtain estimate $\exp(\mathbf{z}'_i\widehat{\alpha})$, and then bootstrap the implied residuals $\widehat{\varepsilon}_i$. Consistency and asymptotic

refinement of this bootstrap requires correct specification of the functional form for the heteroskedasticity.

The **wild bootstrap**, introduced by Wu (1986) and Liu (1988) and studied further by Mammen (1993), provides asymptotic refinement without imposing such structure on the heteroskedasticity. This bootstrap replaces the OLS residual \hat{u}_i by the following residual:

$$\hat{u}_i^* = \begin{cases} \frac{1-\sqrt{5}}{2}\hat{u}_i \simeq -0.6180\hat{u}_i & \text{with probability } \frac{1+\sqrt{5}}{2\sqrt{5}} \simeq 0.7236, \\ [1 - \frac{1-\sqrt{5}}{2}]\hat{u}_i \simeq 1.6180\hat{u}_i & \text{with probability } 1 - \frac{1+\sqrt{5}}{2\sqrt{5}} \simeq 0.2764. \end{cases}$$

Taking expectations with respect to only this two-point distribution and performing some algebra yields $E[\hat{u}_i^{*2}] = 0$, $E[\hat{u}_i^{*3}] = \hat{u}_i^3$, and $E[\hat{u}_i^{*4}] = \hat{u}_i^4$. Thus \hat{u}_i^* leads to a residual with zero conditional mean as desired, since $E[\hat{u}_i^*|\hat{u}_i, \mathbf{x}_i] = 0$ implies $E[\hat{u}_i^*|\mathbf{x}_i] = 0$, while the second and third moments are unchanged.

The wild bootstrap resamples have i th observation (y_i^*, \mathbf{x}_i) , where $y_i^* = \mathbf{x}_i' \hat{\beta} + \hat{u}_i^*$. The resamples vary because of different realizations of \hat{u}_i^* . Simulations by Horowitz (1997, 2001) show that this bootstrap works much better than a paired bootstrap when there is heteroskedasticity and works well compared to other bootstrap methods even if there is no heteroskedasticity.

It seems surprising that this bootstrap should work because for the i th observation it draws from only two possible values for the residual, $-0.6180\hat{u}_i$ or $1.6180\hat{u}_i$. However, a similar draw is being made over all N observations and over B bootstrap iterations. Recall also that White's estimator replaces $E[u_i^2]$ by \hat{u}_i^2 , which, although incorrect for one observation, is valid when averaged over the sample. The wild bootstrap is instead drawing from a two-point distribution with mean 0 and variance \hat{u}_i^2 .

11.6.2. Panel Data and Clustered Data

Consider a linear panel regression model

$$\tilde{y}_{it} = \tilde{\mathbf{w}}_{it}' \theta + \tilde{u}_{it},$$

where i denotes individual and t denotes time period. Following the notation of Section 21.2.3, the tilda is added as the original data y_{it} and \mathbf{x}_{it} may first be transformed to eliminate fixed effects, for example. We assume that the errors \tilde{u}_{it} are independent over i , though they may be heteroskedastic and correlated over t for given i .

If the panel is short, so that T is finite and asymptotic theory relies on $N \rightarrow \infty$, then consistent standard errors for $\hat{\theta}$ can be obtained by a paired or EDF bootstrap that resamples over i but does not resample over t . In the preceding presentation \mathbf{w}_i becomes $[y_{i1}, \mathbf{x}_{i1}, \dots, y_{iT}, \mathbf{x}_{iT}]$ and we resample over i and obtain all T observations for the chosen i .

This **panel bootstrap**, also called a block bootstrap, can also be applied to the nonlinear panel models of Chapter 23. The key assumptions are that the panel is short and the data are independent over i . More generally, this bootstrap can be applied whenever data are clustered (see Section 24.5), provided cluster size is finite and the number of clusters goes to infinity.

The panel bootstrap produces standard errors that are asymptotically equivalent to panel robust sandwich standard errors (see Section 21.2.3). It does not provide an asymptotic refinement. However, it is quite simple to implement and is practically very useful as many packages do not automatically provide panel robust standard errors even for quite basic panel estimators such as the fixed effects estimator. Depending on the application, other bootstraps such as parametric and residual bootstraps may be possible, provided again that resampling is over i only.

Asymptotic refinement is straightforward if the errors are iid. More realistically, however, \tilde{u}_{it} will be heteroskedastic and correlated over t for given i . The wild bootstrap (see Section 11.6.1) should provide an asymptotic refinement in a linear model if the panel is short. Then wild bootstrap resamples have (i, t) th observation $(\tilde{y}_{it}^*, \tilde{\mathbf{w}}_{it})$, where $\tilde{y}_{it}^* = \tilde{\mathbf{w}}_{it}'\hat{\theta} + \tilde{u}_{it}^*$, $\tilde{u}_{it} = \tilde{y}_{it} - \tilde{\mathbf{w}}_{it}'\hat{\theta}$ and \tilde{u}_{it}^* is a draw from the two-point distribution given in Section 11.6.1.

11.6.3. Hypothesis and Specification Tests

Section 11.2.6 focused on tests of the hypothesis $\theta = \theta_0$. Here we consider more general tests. As in Section 11.2.6, the bootstrap can be used to perform hypothesis tests with or without asymptotic refinement.

Tests without Asymptotic Refinement

A leading example of the usefulness of the bootstrap is the Hausman test (see Section 8.3). Standard implementation of this test requires estimation of $V[\hat{\theta} - \tilde{\theta}]$, where $\hat{\theta}$ and $\tilde{\theta}$ are the two estimators being contrasted. Obtaining this estimate can be difficult unless the strong assumption is made that one of the estimators is fully efficient under H_0 . The paired bootstrap can be used instead, leading to consistent estimate

$$\widehat{V}_{\text{Boot}}[\hat{\theta} - \tilde{\theta}] = \frac{1}{B-1} \sum_{b=1}^B [(\hat{\theta}_b^* - \tilde{\theta}_b^*) - (\bar{\hat{\theta}}^* - \bar{\tilde{\theta}}^*)][(\hat{\theta}_b^* - \tilde{\theta}_b^*) - (\bar{\hat{\theta}}^* - \bar{\tilde{\theta}}^*)]',$$

where $\bar{\hat{\theta}}^* = B^{-1} \sum_b \hat{\theta}_b^*$ and $\bar{\tilde{\theta}}^* = B^{-1} \sum_b \tilde{\theta}_b^*$. Then compute

$$H = (\hat{\theta} - \tilde{\theta})' (\widehat{V}_{\text{Boot}}[\hat{\theta} - \tilde{\theta}])^{-1} (\hat{\theta} - \tilde{\theta}) \quad (11.24)$$

and compare to chi-square critical values. As mentioned in Chapter 8, a generalized inverse may need to be used and care may be needed to ensure chi-square critical values are obtained using the correct degrees of freedom.

More generally, this approach can be used for any standard normal test or chi-square distributed test where implementation is difficult because a variance matrix must be estimated. Examples include hypothesis tests based on a two-step estimator and the m-tests of Chapter 8.

Tests with Asymptotic Refinement

Many tests, especially those for fully parametric models such as the LM test and IM test, can be simply implemented using an auxiliary regression (see Sections 7.3.5 and

8.2.2). The resulting test statistics, however, perform poorly in finite samples as documented in many Monte Carlo studies. Such test statistics are easily computed and are asymptotically pivotal as the chi-square distribution does not depend on unknown parameters. They are therefore prime candidates for asymptotic refinement by bootstrap.

Consider the m-test of $H_0 : E[\mathbf{m}_i(y_i|\mathbf{x}_i, \theta)] = \mathbf{0}$ against $H_a : E[\mathbf{m}_i(y_i|\mathbf{x}_i, \theta)] \neq \mathbf{0}$ (see Section 8.2). From the original data estimate $\hat{\theta}$ by ML, and calculate the test statistic M. Using a parametric bootstrap, resample y_i^* from the fitted conditional density $f(y_i|\mathbf{x}_i, \hat{\theta})$, for fixed regressors in repeated samples, or from $f(y_i|\mathbf{x}_i^*, \hat{\theta})$. Compute $M_b^*, b = 1, \dots, B$, in the bootstrap resamples. Reject H_0 at level α if the original calculated statistic M exceeds the α quantile of $M_b^*, b = 1, \dots, B$.

Horowitz (1994) presented this bootstrap for the IM test and demonstrated with simulation examples that there are substantial finite-sample gains to this bootstrap. A detailed application by Drukker (2002) to specification tests for the tobit model suggests that conditional moment specification tests can be easily applied to fully parametric models, since any size distortion in the auxiliary regressions can be corrected through bootstrap.

Note that bootstrap tests without asymptotic refinement, such as the Hausman test given here, can be refined by use of the nested bootstrap given in Section 11.5.3.

11.6.4. GMM, Minimum Distance, and Empirical Likelihood in Overidentified Models

The GMM estimator is based on population moment conditions $E[\mathbf{h}(\mathbf{w}_i, \theta)] = \mathbf{0}$ (see Section 6.3.1). In a just-identified model a consistent estimator simply solves $N^{-1} \sum_i \mathbf{h}(\mathbf{w}_i, \hat{\theta}) = \mathbf{0}$. In overidentified models this estimator is no longer feasible. Instead, the GMM estimator is used (see Section 6.3.2).

Now consider bootstrapping, using the paired or EDF bootstrap. For GMM in an *overidentified* model $N^{-1} \sum_i \mathbf{h}(\mathbf{w}_i, \theta) \neq \mathbf{0}$, so this bootstrap does not impose on the bootstrap resamples the original population restriction that $E[\mathbf{h}(\mathbf{w}_i, \theta)] = \mathbf{0}$. As a result even if the asymptotically pivotal t -statistic is used there is no longer a bootstrap refinement, though bootstraps on $\hat{\theta}$ and related confidence intervals and t -test statistics remain consistent. More fundamentally, the bootstrap of the OIR test (see Section 6.3.8) can be shown to be inconsistent. We focus on cross-section data but similar issues arise for panel GMM estimators (see Chapter 22) in overidentified models.

Hall and Horowitz (1996) propose correcting this by **recentering**. Then the bootstrap is based on $\mathbf{h}^*(\mathbf{w}_i, \hat{\theta}) = \mathbf{h}(\mathbf{w}_i, \hat{\theta}) - N^{-1} \sum_i \mathbf{h}(\mathbf{w}_i, \hat{\theta})$ and asymptotic refinements can be obtained for statistics based on $\hat{\theta}$ including the OIR test.

Horowitz (1998) does similar recentering for the minimum distance estimator (see Section 6.7). He then applies the bootstrap to the covariance structure example of Altonji and Segal (1996) discussed in Section 6.3.5.

An alternative adjustment proposed by Brown and Newey (2002) is to not recenter but to instead resample the observations \mathbf{w}_i with probabilities that vary across observations rather than using equal weights $1/N$. Specifically, let $\Pr[\mathbf{w}^* = \mathbf{w}_i] = \hat{\pi}_i$, where $\hat{\pi}_i = (1 + \hat{\lambda} \hat{\mathbf{h}}_i)$, $\hat{\mathbf{h}}_i = \mathbf{h}(\mathbf{w}_i, \hat{\theta})$, and $\hat{\lambda}$ maximizes $\sum_i \ln(1 + \hat{\lambda} \hat{\mathbf{h}}_i)$. The motivation is that the probabilities $\hat{\pi}_i$ equivalently are the solution to an empirical likelihood (EL)

problem (see Section 6.8.2) of maximizing $\sum_i \ln \pi_i$ with respect to π_1, \dots, π_N subject to the constraints $\sum_i \pi_i \hat{\mathbf{h}}_i = \mathbf{0}$ and $\sum_i \pi_i = 1$. This **empirical likelihood bootstrap** of the GMM estimator therefore imposes the constraint $\sum_i \hat{\pi}_i \hat{\mathbf{h}}_i = \mathbf{0}$.

One could instead work directly with EL from the beginning, letting $\hat{\theta}$ be the EL estimator rather than the GMM estimator. The advantage of the Brown and Newey (2002) approach is that it avoids the more challenging computation of the EL estimator. Instead, one needs only the GMM estimator and solution of the concave programming problem of minimizing $\sum_i \ln(1 + \hat{\lambda}' \hat{\mathbf{h}}_i)$.

11.6.5. Nonparametric Regression

Nonparametric density and regression estimators converge at rate less than \sqrt{N} and are asymptotically biased. This complicates inference such as confidence intervals (see Sections 9.3.7 and 9.5.4).

We consider the kernel regression estimator $\hat{m}(x_0)$ of $m(x_0) = E[y|x=x_0]$ for observations (y, x) that are iid, though conditional heteroskedasticity is permitted. From Horowitz (2001, p. 3204), an asymptotically pivotal statistic is

$$t = \frac{\hat{m}(x_0) - m(x_0)}{s_{\hat{m}(x_0)}},$$

where $\hat{m}(x_0)$ is an undersmoothed kernel regression estimator with bandwidth $h = o(N^{-1/3})$ rather than the optimal $h^* = O(N^{-1/5})$ and

$$s_{\hat{m}(x_0)}^2 = \frac{1}{Nh[\hat{f}(x_0)]^2} \sum_{i=1}^N (y_i - \hat{m}(x_i))^2 K\left(\frac{x_i - x_0}{h}\right)^2,$$

where $\hat{f}(x_0)$ is a kernel estimate of the density $f(x)$ at $x = x_0$. A paired bootstrap resamples (y^*, x^*) and forms $t_b^* = [\hat{m}_b^*(x_0) - m(x_0)]/s_{\hat{m}(x_0), b}^*$, where $s_{\hat{m}(x_0), b}^*$ is computed using bootstrap sample kernel estimates $\hat{m}_b^*(x_i)$ and $\hat{f}_b^*(x_0)$. The percentile- t confidence interval of Section 11.2.7 then provides an asymptotic refinement. For a symmetrical confidence interval or symmetrical test at level α the error is $o((Nh^{-1}))$ rather than $O((Nh^{-1}))$ using first-order asymptotic approximation.

Several variations on this bootstrap are possible. Rather than using undersmoothing, bias can be eliminated by directly estimating the bias term given in Section 9.5.2. Also rather than using $s_{\hat{m}(x_0)}^2$, the variance term given in Section 9.5.2 can be directly estimated.

Yatchew (2003) provides considerable detail on implementing the bootstrap in nonparametric and semiparametric regression.

11.6.6. Nonsmooth Estimators

From Section 11.4.2 the bootstrap assumes smoothness in estimators and statistics. Otherwise the bootstrap may not offer an asymptotic refinement and may even be invalid.

As illustration we consider the LAD estimator and extension to binary data. The LAD estimator (see Section 4.6.2) has objective function $\sum_i |y_i - \mathbf{x}_i' \beta|$ that has

discontinuous first derivative. A bootstrap can provide a valid asymptotic approximation but does not provide an asymptotic refinement. For binary outcomes, the LAD estimator extends to the maximum score estimator of Manski (1975) (see Section 14.7.2). For this estimator the bootstrap is not even consistent.

In these examples bootstraps with asymptotic refinements can be obtained by using a smoothed version of the original objective function for the estimator. For example, the smoothed maximum score estimator of Horowitz (1992) is presented in Section 14.7.2.

11.6.7. Time Series

The bootstrap relies on resampling from an iid distribution. Time-series data therefore present obvious problems as the result of dependence.

The bootstrap is straightforward in the linear model with an ARMA error structure and resampling the underlying white noise error. As an example, suppose $y_t = \beta x_t + u_t$, where $u_t = \rho u_{t-1} + \varepsilon_t$ and ε_t is white noise. Then given estimates $\hat{\beta}$ and $\hat{\rho}$ we can recursively compute residuals as $\hat{\varepsilon}_t = \hat{u}_t - \hat{\rho} \hat{u}_{t-1} = y_t - x_t \hat{\beta} - \hat{\rho}(y_{t-1} - x_{t-1} \hat{\beta})$. Bootstrapping these residuals to give $\hat{\varepsilon}_t^*$, $t = 1, \dots, T$, we can then recursively compute $\hat{u}_t^* = \hat{\rho} \hat{u}_{t-1}^* + \hat{\varepsilon}_t^*$ and hence $y_t^* = \hat{\beta} x_t + \hat{u}_t^*$. Then regress y_t^* on x_t with AR(1) error. An early example was presented by Freedman (1984), who bootstrapped a dynamic linear simultaneous equations regression model estimated by 2SLS. Given linearity, simultaneity adds little problems. The dynamic nature of the model is handled by recursively constructing $y_t^* = f(y_{t-1}^*, \mathbf{x}_t, \mathbf{u}_t^*)$, where \mathbf{u}_t^* are obtained by resampling from the 2SLS structural equation residuals and $\mathbf{y}_0^* = \mathbf{y}_0$. Then perform 2SLS on each bootstrap sample.

This method assumes the underlying error is iid. For general dependent data without an ARMA specification, for example, nonstationary data, the moving blocks bootstrap presented in Section 11.5.2 can be used.

For testing unit roots or cointegration special care is needed in applying the bootstrap as the behavior of the test statistic changes discontinuously at the unit root. See, for example, Li and Maddala (1997). Although it is possible to implement a valid bootstrap in this situation, to date these bootstraps do not provide an asymptotic refinement.

11.7. Practical Considerations

The bootstrap without asymptotic refinement can be a very useful tool for the applied researcher in situations where it is difficult to perform inference by other means. This need can vary with available software and the practitioner's tool kit. The most common application of the bootstrap to date is computation of standard errors needed to conduct a Wald hypothesis test. Examples include heteroskedasticity-robust and panel-robust inference, inference for two-step estimators, and inference on transformations of estimators. Other potential applications include computation of m-test statistics such as the Hausman test.

The bootstrap can additionally provide an asymptotic refinement. Many Monte Carlo studies show that quite standard procedures can perform poorly in finite samples. There appears to be great potential for use of bootstrap refinements, currently unrealized. In some cases this could improve existing inference, such as use of the wild bootstrap in models with additive errors that are heteroskedastic. In other cases it should encourage increased use of methods that are currently under-utilized. In particular, model specification tests with good small-sample properties can be implemented by bootstrapping easily computed auxiliary regressions.

There are two barriers to the use of the bootstrap. First, the bootstrap is not always built into statistical packages. This will change over time, and for now constructing code for a bootstrap is not too difficult provided the package includes looping and the ability to save regression output. Second, there are subtleties involved. Asymptotic refinement requires use of an asymptotically pivotal statistic and the simplest bootstraps presume iid data and smoothness of estimators and statistics. This covers a wide class of applications but not all applications.

11.8. Bibliographic Notes

The bootstrap was proposed by Efron (1979) for the iid case. Singh (1981) and Bickel and Freedman (1981) provided early theory. A good introductory statistics treatment is by Efron and Tibsharani (1993), and a more advanced treatment is by Hall (1992). Extensions to the regression case were considered early on; see, for example, Freedman (1984). Most of the work by econometricians has occurred in the past 10 years. The survey of Horowitz (2001) is very comprehensive and is well complemented by the survey of Brownstone and Kazimi (1998), which considers many econometrics applications, and the paper by MacKinnon (2002).

Exercises

11–1 Consider the model $y = \alpha + \beta x + \varepsilon$, where α , β , and x are scalars and $\varepsilon \sim \mathcal{N}[0, \sigma^2]$. Generate a sample of size $N = 20$ with $\alpha = 2$, $\beta = 1$, and $\sigma^2 = 1$ and suppose that $x \sim \mathcal{N}[2, 2]$. We wish to test $H_0 : \beta = 1$ against $H_a : \beta \neq 1$ at level 0.05 using the t -statistic $t = (\hat{\beta} - 1)/\text{se}[\hat{\beta}]$. Do as much of the following as your software permits. Use $B = 499$ bootstrap replications.

- (a) Estimate the model by OLS, giving slope estimate $\hat{\beta}$.
- (b) Use a paired bootstrap to compute the standard error and compare this to the original sample estimate. Use the bootstrap standard error to test H_0 .
- (c) Use a paired bootstrap with asymptotic refinement to test H_0 .
- (d) Use a residual bootstrap to compute the standard error and compare this to the original sample estimate. Use the bootstrap standard error to test H_0 .
- (e) Use a residual bootstrap with asymptotic refinement to test H_0 .

11–2 Generate a sample of size 20 according from the following dgp. The two regressors are generated by $x_1 \sim \chi^2(4) - 4$ and $x_2 \sim 3.5 + \mathcal{U}[1, 2]$; the error is from a mixture of normals with $u \sim \mathcal{N}[0, 25]$ with probability 0.3 and $u \sim \mathcal{N}[0, 5]$ with

probability 0.7; and the dependent variable is $y = 1.3x_1 + 0.7x_2 + 0.5u$.

- (a) Estimate by OLS the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$.
- (b) Suppose we are interested in estimating the quantity $\gamma = \beta_1 + \beta_2^2$ from the data. Use the least-squares estimates to estimate this quantity. Use the delta method to obtain approximate standard error for this function.
- (c) Then estimate the standard error of $\hat{\gamma}$ using a paired bootstrap. Compare this to $se[\hat{\gamma}]$ from part (b) and explain the difference. For the bootstrap use $B = 25$ and $B = 200$.
- (d) Now test $H_0 : \gamma = 1.0$ at level 0.05 using a paired bootstrap with $B = 999$. Perform bootstrap tests without and with asymptotic refinement.

- 11–3** Use 200 observations from the Section 4.6.4 data on natural logarithm of health expenditure (y) and natural logarithm of total expenditure (x). Obtain OLS estimates of the model $y = \alpha + \beta x + u$. Use the paired bootstrap with $B = 999$.

- (a) Obtain a bootstrap estimate of the standard error of $\hat{\beta}$.
- (b) Use this standard error estimate to test $H_0 : \beta = 1$ against $H_a : \beta \neq 1$.
- (c) Do a bootstrap test with refinement of $H_0 : \beta = 1$ against $H_a : \beta \neq 1$ under the assumption that u is homoskedastic.
- (d) If u is heteroskedastic what happens to your method in (c)? Is the test still asymptotically valid, and if so does it offer an asymptotic refinement?
- (e) Do a bootstrap to obtain a bias-corrected estimate of β .

Simulation-Based Methods

12.1. Introduction

The nonlinear methods presented in the preceding chapters do not require closed-form solutions for the estimator. Nonetheless, they rely considerably on analytical tractability. In particular, the objective function for the estimator has been assumed to have a closed-form expression, and the asymptotic distribution of the estimator is based on a linearization of the estimating equations.

In the current chapter we present simulation-based estimation methods. The treatment of ML estimation in Chapter 5 presumed that the density $f(y|\mathbf{x}, \theta)$ has a closed-form expression. If there is no closed-form solution, ML estimation may still be possible if we instead use a good approximation $\hat{f}(y|\mathbf{x}, \theta)$ of $f(y|\mathbf{x}, \theta)$ to form the likelihood function. A common reason for lack of a closed-form expression for the density is the presence of an intractable expectation in the definition of $f(y|\mathbf{x}, \theta)$. For example, in a random coefficients model it may be difficult to integrate out the random parameters. If the expectation is replaced by a Monte Carlo approximation the resulting estimator is called a simulation-based estimator. A similar simulation approach can be applied to method of moments estimation based on a moment, such as the conditional mean, for which there is no closed-form solution. In the method of moments case it can be possible to obtain consistent parameter estimates with much less simulation than is necessary for consistency in the ML case.

These estimation methods are computer intensive because they make extensive use of Monte Carlo sampling methods. Their use raises questions of accuracy of approximations, efficiency of computation, and the sampling properties of the estimators that use such approximations.

Section 12.2 gives motivating examples for simulation-based estimation. Section 12.3 covers the basics of computing integrals, as an expectation with respect to a continuous random variable is an integral. Sections 12.4 and 12.5 present maximum simulated likelihood estimation and simulated moment-based estimation; Section 12.6 deals with indirect inference. These estimators require simulators, detailed in Section 12.7, and pseudo-random numbers, detailed in Section 12.8.

12.2. Examples

We consider examples where the conditional density of y given regressors \mathbf{x} and parameters θ is an integral

$$f(y|\mathbf{x}, \theta) = \int h(y|\mathbf{x}, \theta, \mathbf{u})g(\mathbf{u})d\mathbf{u}, \quad (12.1)$$

where the functional forms of $h(\cdot)$ and $g(\cdot)$ are known and \mathbf{u} denotes a random variable, not necessarily an error term, that needs to be integrated out. If there is no analytical solution for the integral, and hence no closed-form expression for the likelihood function, then simulation-based estimation methods are warranted.

12.2.1. Random Parameters Models

A **random parameters model** or **random coefficients model** permits regression parameters to vary across individuals according to some distribution. A fully parametric random parameters model specifies the dependent variable y_i conditional on regressors \mathbf{x}_i and given parameters γ_i to have conditional density $f(y_i|\mathbf{x}_i, \gamma_i)$, where γ_i are iid with density $g(\gamma_i|\theta)$. Inference is based on the density of y_i conditional on \mathbf{x}_i and given θ ,

$$f(y|\mathbf{x}, \theta) = \int f(y|\mathbf{x}, \gamma)g(\gamma|\theta)d\gamma. \quad (12.2)$$

This integral will not have a closed-form solution except in some special cases. A common specification is to assume normally distributed random parameters, with $\gamma_i \sim \mathcal{N}[\mu, \Sigma]$. Then $\gamma_i = \mu + \Sigma^{-1/2}\mathbf{u}_i$, where $\mathbf{u}_i \sim \mathcal{N}[\mathbf{0}, \mathbf{I}]$ and we can rewrite (12.2) in the form (12.1), where θ is a vector containing μ and the distinct components of Σ , and $g(\mathbf{u})$ is the $\mathcal{N}[\mathbf{0}, \mathbf{I}]$ density.

A simple example of a random parameters model is **neglected heterogeneity**. Then often just one parameter, usually the intercept, is assumed to be random and the integral is a one-dimensional integral that is easily approximated numerically. More generally, however, the dimension of the integral may be high.

Leading examples of random parameters and unobserved heterogeneity include (1) normally distributed random parameters in multinomial logit models (the random parameters logit model; see Chapter 15), (2) gamma distributed unobserved heterogeneity in Weibull duration models (see Chapter 19), (3) gamma distributed unobserved heterogeneity in Poisson count data models (see Chapter 20), and (4) individual-specific random effects in panel data models (see Chapter 21). Closed-form solutions for the resulting marginal density after integration over the distribution of heterogeneity are available in example 3 and for the linear model under normality in example 4. However, for examples 1 and 2 and many nonlinear applications of example 4 closed-form solutions are not available.

12.2.2. Limited Dependent Variable Models

A **limited dependent variable** (LDV) is a dependent variable that is observed only over part of its range, owing to censoring and truncation. Then the density of the observed variable involves integrals that may not have a closed-form expression.

A leading class of LDV models are **discrete choice models**, detailed in Chapters 14 and 15. We introduce discrete choice models here because they have been the focus of the econometrics literature on simulation-based estimation.

As an example, consider consumer choice among three mutually exclusive alternatives, such as among three different durable goods, only one of which is chosen by the individual. Suppose the consumer maximizes utility, and let the utilities of alternatives 1, 2, and 3 be given by U_1 , U_2 , and U_3 , respectively. The utilities U_1 , U_2 , and U_3 are not observed. Instead, we observe only a discrete outcome variable $y = 1, 2$, or 3 depending on which alternative is chosen.

Suppose alternative 1 is chosen, because it has the highest utility. Then the probability mass function is $p_1 = \Pr[y = 1]$, where

$$\begin{aligned} p_1 &= \Pr[U_1 - U_2 \geq 0, U_1 - U_3 \geq 0] \\ &= \Pr[(\mathbf{x}_1 - \mathbf{x}_2)' \boldsymbol{\beta} + \varepsilon_1 - \varepsilon_2 \geq 0, (\mathbf{x}_1 - \mathbf{x}_3)' \boldsymbol{\beta} + \varepsilon_1 - \varepsilon_3 \geq 0], \end{aligned}$$

if we make the common assumption (see Section 15.5.1) that $U_j = \mathbf{x}'_j \boldsymbol{\beta} + \varepsilon_j$, $j = 1, 2, 3$, where the regressor \mathbf{x} measures the different attributes of the three goods and the error ε can range over $(-\infty, \infty)$. Defining $u_1 = U_1 - U_2$ and $u_2 = U_1 - U_3$, we have that

$$p_1 = \int_0^\infty \int_0^\infty g(u_1, u_2) du_1 du_2, \quad (12.3)$$

where $g(u_1, u_2)$, or more formally $g(u_1, u_2 | \mathbf{x}, \boldsymbol{\theta})$, is the bivariate density of (u_1, u_2) , or equivalently

$$p_1 = \int_{-\infty}^\infty \int_{-\infty}^\infty 1[u_1 \geq 0, u_2 \geq 0] g(u_1, u_2) du_1 du_2, \quad (12.4)$$

where $1[A]$ is the indicator function equal to 1 if event A happens and equal to 0 otherwise.

The integral (12.4) is of the form (12.1). Because the integral is over only part of the range of (u_1, u_2) (see (12.3)) a closed-form solution may not exist, even though we know that $\int \int g(u_1, u_2) du_1 du_2 = 1$ if integration is over the entire range of (u_1, u_2) .

In particular, if the errors ε are normally distributed, as in the **multinomial probit model**, the integral (12.3) is over the positive orthant of a bivariate normal distribution. There is no closed-form solution for p , and hence no tractable expression for the density $f(y | \mathbf{x}, \boldsymbol{\theta})$ exists. In practice the dimension of the integral can be very high, making numerical approximation difficult, because for choice among m mutually exclusive alternatives the integral has dimension $m - 1$. Until simulation-based estimators were developed researchers either used models with $m \leq 4$ or chose other error distributions such as that leading to the much more restricted multinomial logit model.

12.2.3. ML Estimation

For simplicity consider the MLE. Assume independence over observations and that y has conditional density $f(y|\mathbf{x}, \boldsymbol{\theta})$.

The complication in the preceding two examples is that ML estimation is not practical as there is no closed-form expression for $f(y|\mathbf{x}, \boldsymbol{\theta})$, which is defined by an integral that does not simplify. Instead, we replace the integral by a numerical approximation $\widehat{f}(y|\mathbf{x}, \boldsymbol{\theta})$, and we maximize

$$\ln \widehat{L}_N(\boldsymbol{\theta}) = \sum_{i=1}^N \ln \widehat{f}(y_i|\mathbf{x}_i, \boldsymbol{\theta})$$

with respect to $\boldsymbol{\theta}$. The estimator will be consistent and have the same asymptotic distribution as the MLE if $\widehat{f}(y|\mathbf{x}, \boldsymbol{\theta})$ is a good approximation to $f(y|\mathbf{x}, \boldsymbol{\theta})$.

The resulting first-order conditions are usually nonlinear and are solved by iterative methods. Because $\widehat{f}(y_i|\mathbf{x}_i, \boldsymbol{\theta})$ varies with i and $\boldsymbol{\theta}$, evaluation of the gradient using numerical derivatives will require at least Nqr evaluations, where N is the sample size, q is the dimension of $\boldsymbol{\theta}$, and r is the number of iterations. For example, with 1,000 observations, 10 parameters, and 50 iterations there are at least 500,000 function evaluations.

This standard computational demand for nonlinear models now needs to be multiplied by the number of evaluations needed to compute an adequate approximation to the integral $f(y|\mathbf{x}, \boldsymbol{\theta})$. Clearly, methods that require relatively few evaluations are desired.

12.2.4. Bayesian Methods

Bayesian methods are given a separate treatment in Chapter 13. They involve computation of integrals that appear similar to (12.2), but they go one step further and obtain the (posterior) distribution of parameters rather than a point estimate such as the MLE.

12.3. Basics of Computing Integrals

We consider the **integral**

$$I = \int_a^b f(x) dx, \quad (12.5)$$

where $f(\cdot)$ is continuous on $[a, b]$, and the bounds of the integral need not be finite, so $a = -\infty$ and/or $b = \infty$ are possible. In this section x is initially a scalar and is used to denote the variable being integrated out. In regression applications integration is often with respect to a vector that is denoted \mathbf{u} since \mathbf{x} then denotes the regressors (see (12.1)). It is assumed that the integral exists, an important qualification that needs to be checked as approximation methods will yield a finite estimate of I even if the integral diverges.

We first present numerical integration or quadrature, useful for low-dimensional integrals. This is followed by Monte Carlo integration, which works better for high-dimensional integrals and is the focus of this chapter.

The material in this section pertains to the implementation phase of simulation-based estimation; therefore, some readers may prefer to read it after covering Sections 12.4–12.6.

12.3.1. Deterministic Numerical Integration

An integral can be interpreted as an area or a volume measure. **Deterministic numerical integration** or **quadrature** replaces the volume by a series of slices of smaller volumes that are then added up. Formally this involves evaluating the integrand at several points and taking a weighted sum of these values. The prefix deterministic is used to indicate that this method of approximation of an integral does not entail simulation.

Simpson's Rule

By the definition of an integral,

$$I = \lim_{\Delta x_i \rightarrow 0} \sum_{j=1}^n f(x_j) \Delta x_j, \quad (12.6)$$

where the range of $[a, b]$ of x is split into $(n + 1)$ points, $x_0 < x_1 < \dots < x_n$, and $n \rightarrow \infty$. Standard approximation methods are refinements of (12.6) that provide more accurate approximations for finite n . We present results for equally spaced points, though the methods can be generalized to evaluation at points that are not equally spaced. For simplicity we assume that $f(x)$ can be evaluated at the limit points a and b .

The **midpoint rule** evaluates at the midpoint $\bar{x}_j = \frac{1}{2}(x_{j-1} + x_j)$ of the interval $[x_{j-1}, x_j]$ and sums n rectangles that have base $(b - a)/n$ and height $f(\bar{x}_j)$. Thus I is approximated by

$$\hat{I}_M = \sum_{j=1}^n \frac{b - a}{n} f(\bar{x}_j). \quad (12.7)$$

The **trapezoidal rule** is an improvement that draws a straight line between $f(x_{j-1})$ and $f(x_j)$ and sums n trapezoids that have base $(b - a)/n$ and average height $(f(x_{j-1}) + f(x_j))/2$. Thus I is approximated by

$$\hat{I}_T = \sum_{j=1}^n \frac{(b - a)}{n} \frac{f(x_{j-1}) + f(x_j)}{2}. \quad (12.8)$$

Simpson's rule uses a quadratic curve among three successive points $f(x_{j-1})$, $f(x_j)$, and $f(x_{j+1})$, whereas the trapezoidal rule used a line between two successive points.

This leads to the approximation

$$\hat{I}_S = \sum_{j=0}^n \frac{(b-a)}{3n} w_j f(x_j), \quad (12.9)$$

where n is even, $w_j = 4$ if j is odd, and $w_j = 2$ if j is even, except $w_0 = w_n = 1$. Further generalization to permit a polynomial of degree p among $p+1$ successive points is possible.

Error bounds for these approximations increase as a power function of the range of integration, $b-a$, and decrease as a power function of the number of intervals. For Simpson's rule, $|I_S - I| \leq M_4(b-a)^5/180n^4$, where M_4 is the maximum absolute value of the fourth derivative of x on $[a, b]$. For the trapezoidal rule, $|I_T - I| \leq M_2(b-a)^3/12n^2$, where M_2 is the maximum absolute value of the second derivative of x on $[a, b]$. Clearly, the number of intervals needs to increase with the range of x , and one should test for sensitivity to the number of intervals.

Simpson's rule and related rules can work well for definite integrals over a bounded interval, but problems can clearly arise with indefinite integrals because of problems in evaluating in the tails. For example, suppose $[a, b] = [0, \infty)$. Then in choosing x_n there is a trade-off because the upper bound x_n should be large, but then the distance between evaluation points is large. At the least one should test for sensitivity to increases in x_n .

Gaussian Quadrature

Gaussian quadrature, where quadrature is an alternative name for numerical integration, was proposed by Gauss in 1814. It provides a rule for good choice of the evaluation points x_j , no longer equally spaced, and is especially useful for evaluating indefinite integrals.

We first reexpress the integral (12.5) as

$$I = \int_c^d w(x)r(x)dx, \quad (12.10)$$

where $w(x)$ is usually one of the following three functions, depending on the range of x : Gauss–Hermite quadrature sets $w(x) = e^{-x^2}$ and is used for $[c, d] = (-\infty, \infty)$, Gauss–Laguerre quadrature sets $w(x) = e^{-x}$ and is used when $[c, d] = (0, \infty)$, and Gauss–Legendre quadrature sets $w(x) = 1$ and is used when $[c, d] = [-1, 1]$.

In the simplest case (12.10) can be obtained from (12.5) by defining $r(x) = f(x)/w(x)$. More generally, a transformation of x may be needed so that, for example, the range $[2, \infty)$ in (12.5) becomes $[0, \infty)$ in (12.10). Some routines permit the user to simply provide $f(x)$ and the range of integration and automatically take care of any necessary transformations.

Gaussian quadrature approximates the integral (12.10) by the weighted sum

$$\hat{I}_G = \sum_{j=1}^m w_j r(x_j), \quad (12.11)$$

where the researcher chooses m ; the m points of evaluation x_j and the weights w_j are given in books such as Abramowitz and Stegun's (1971) or in computer code such as that provided in Press et al. (1993).

The theory behind the approximation is based on the **orthogonal polynomials** of $w(x)$, denoted $p_j(x)$, $j = 0, \dots, m$, that satisfy

$$\int_c^d w(x)p_j(x)p_k(x)dx = 0, \quad j \neq k, \quad j, k = 0, \dots, m.$$

If additionally $\int_c^d w(x)p_j^2(x)dx = 1$ then the polynomials are said to be orthonormal. The approximation (12.11) is exact if $r(x)$ is a polynomial of order $2m - 1$ or less, so the approximation works best if $r(x)$ in (12.10) is well approximated by a polynomial of order $2m - 1$. A good choice of the number of evaluation points m requires experimentation, but many applications use m no more than 20 or 30.

As an example consider **Gauss–Hermite quadrature**, commonly used in econometrics since integration is often over $(-\infty, \infty)$. For $w(x) = e^{-x^2}$ the orthogonal polynomials $p_j(x)$ are the Hermite polynomials $H_j(x)$, which in the orthonormal form are generated using the recursion $H_{j+1}(x) = \sqrt{2/(j+1)}xH_j(x) - \sqrt{j/(j+1)}H_{j-1}(x)$, $j = 1, \dots, m$, where $H_{-1} = 0$ and $H_0 = \pi^{-1/4}$. The m abscissas x_j are obtained as the m roots to $H_m(x) = 0$ and, for orthonormal Hermite polynomials, the weights $w_j = 1 / [jH_{j-1}(x_j)^2]$. As already noted x_j and w_j for given m are readily available in tables or computer code.

For definite integrals Gauss–Legendre quadrature usually performs better than Simpson's rule. The real advantage of Gaussian quadrature, however, is for indefinite integrals. Note that if integration is over $(-\infty, \infty)$ it may be possible by change of variable techniques to transform to an integral over $(0, \infty)$ and use Gauss–Laguerre quadrature rather than Gauss–Hermite quadrature.

There are many additional deterministic methods for computing integrals, including **Laplace approximation** (Tierney, Kass, and Kadane, 1989).

12.3.2. Integration by Direct Monte Carlo Sampling

Monte Carlo integration provides an alternative to deterministic numerical integration. In general the Monte Carlo integral estimate of $I = \int_a^b f(x) dx$ is

$$\hat{I}_{\text{MC}} = \sum_{s=1}^S f(x^s), \quad (12.12)$$

where x^1, \dots, x^S are S uniform draws of x in the range $[a, b]$. Compared to the midpoint rule we evaluate $f(x)$ at S randomly chosen points rather than n deterministic midpoints.

We focus on regression applications such as those given in Section 12.2. Then integration arises because we wish to obtain an expected value $E[h(x)]$, say, where the expectation is with respect to a random variable x that has, say, pdf $g(x)$. In the

continuous case we wish to evaluate

$$E[h(x)] = \int_a^b h(x)g(x)dx, \quad (12.13)$$

where throughout this chapter it is assumed that $E[h(x)] < \infty$, that is, the integral converges. Then $E[h(x)]$ can be estimated by the **direct Monte Carlo integral estimate**

$$\widehat{I}_{DMC} = \widehat{E}[h(x)] = S^{-1} \sum_{s=1}^S h(x^s), \quad (12.14)$$

where $\{x^s, s = 1, \dots, S\}$ is a Monte Carlo sample of S pseudo-random numbers from the density $g(x)$, obtained using methods given later in Section 12.8. The estimate (12.14) evaluates $h(x)$ using draws of x from the density $g(x)$, whereas the estimate (12.12) evaluates $h(x)g(x)$ using uniform draws of x as in (12.12). An advantage of (12.14) is that it can be applied to indefinite integrals, whereas obtaining uniform draws in (12.12) is problematic if the limits a or b are unbounded.

The estimate $\widehat{E}[h(x)]$ is an average of the function $f(\cdot)$ evaluated at each of the random draws x^s . Equivalently, $\widehat{E}[h(x)]$ is an average of the random variable $h(x_s)$, and its properties as $S \rightarrow \infty$ can be obtained if we can apply a law of large numbers and a central limit theorem. Here x^s is iid, so $h(x^s)$ is iid and we can apply Kolmogorov LLN (see Appendix A, Theorem A.8) since the existence of $E[h(x)]$ has already been assumed. It follows that

$$\widehat{E}[h(x)] \xrightarrow{P} E[h(x)] \text{ as } S \rightarrow \infty.$$

Also, since $h(x^s)$ is iid, the variance of $\widehat{E}[h(x)]$ equals $S^{-1}V[h(x)]$ assuming $V[h(x)]$ exists. The approximation is likely to be good for moderate size S if $S^{-1}V[h(x^s)]$ is small.

12.3.3. Integral Computation Example

Suppose $x \sim \mathcal{N}[0, 1]$, and we wish to compute the mean

$$E[x] = (\sqrt{2\pi})^{-1} \int_{-\infty}^{\infty} x \exp(-x^2/2) dx$$

and the moment $E[\exp(-\exp(x))]$, defined as the integral

$$E[\exp(-\exp(x))] = (\sqrt{2\pi})^{-1} \int_{-\infty}^{\infty} \exp(-\exp(x)) \exp(-x^2/2) dx.$$

An analytical expression for $E[x]$ exists and yields $E[x] = 0$. By contrast an analytical solution for $E[\exp(-\exp(x))]$ does not exist. Before seeking a numerical approximation, we first confirm that the integral does indeed converge. Since $\exp(-\exp(x))$ is strictly positive and monotonically decreasing with maximum value of 1 it follows that $|\exp(-\exp(x))| < 1$, so $E[\exp(-\exp(x))] < E[1] = 1$ and the integral converges.

These one-dimensional integrals are easily calculated using a deterministic numerical approximation. For example, consider using the midpoint rule with $n = 20$ equally

spaced evaluations between $x_0 = -5$ and $x_{20} = 5$. Then

$$\widehat{E}[x] = (\sqrt{2\pi})^{-1} \sum_{j=1}^{20} \frac{10}{20} \bar{x}_i \exp(-\bar{x}_j^2/2),$$

$$\widehat{E}[\exp(-\exp(x))] = (\sqrt{2\pi})^{-1} \sum_{j=1}^{20} \frac{10}{20} \exp(-\exp(\bar{x}_j)) \exp(-\bar{x}_j^2/2),$$

where $\bar{x}_j = -5.25 + j/2$. This yields $\widehat{E}[x] = 0$ to many decimal places, as expected, whereas $\widehat{E}[\exp(-\exp(x))] = 0.38175656$. The latter estimate changes little, not until the eighth decimal place, if instead we do $n = 200$ evaluations between -10 and 10 . Clearly deterministic numerical approximations work well here.

These integrals are also easily calculated using a Monte Carlo approximation, with

$$\widehat{E}[x] = \frac{1}{S} \sum_{s=1}^S x^s,$$

$$\widehat{E}[\exp(-\exp(x))] = \frac{1}{S} \sum_{s=1}^S \exp(-\exp(x^s)),$$

where x^s is the s th draw of S draws from the $\mathcal{N}[0, 1]$ distribution, and a method to make such draws is given in Appendix B. Table 12.1 gives estimates of $\widehat{E}[x]$ and $\widehat{E}[\exp(-\exp(x))]$ for various numbers of simulations S . Observe the tendency of the estimators to stabilize as $S \rightarrow \infty$, and to go to their respective true values of 0 and 0.38175656, where the latter is obtained by deterministic numerical approximation. However, even with $S = 10^6$ the estimate $\widehat{E}[x]$ still differs from zero in the fourth decimal place. Here $V[\widehat{E}[x]] = S^{-1}V[x^s] = 1/S$ since $V[x^s] = 1$, so even with $S = 10^6$ the standard deviation of $\widehat{E}[x]$ is a relatively large 0.001. Alternative methods that yield a Monte Carlo approximation with lower variance are given in Section 12.7.

Table 12.1. Monte Carlo Integration: Example for x
Standard Normal

S = Number of simulations	$\widehat{E}[x]$	$\widehat{E}[\exp(-\exp(x))]$
10	0.145	0.336
25	-0.209	0.435
50	0.050	0.369
100	-0.120	0.409
500	-0.059	0.398
1,000	0.005	0.382
10,000	-0.007	0.383
100,000	-0.000	0.382
1,000,000	-0.000	0.381

12.3.4. Higher Dimensional Integrals

Higher dimensional integrals can be evaluated using either deterministic or Monte Carlo integration, with the latter method preferred as the dimension increases.

Deterministic integration is best done using multivariate Gaussian quadrature or, if the limits of integration are not too complicated, by reducing an m -dimensional integral to a series of m one-dimensional integrals evaluated using, say, Gaussian quadrature. However, from the definition of the integral in (12.6) it is clear that the number of evaluations will have to go up by the power m . For example, if 20 function evaluations are needed for a one-dimensional integral, then a five-dimensional integral may require 5^{20} or 95 trillion function evaluations. Such high precision may not be needed in an estimation setting where similar computations are being done for each individual observation and then summed, but even then the number of evaluations will rise substantially with the dimension of the integral.

Performing Monte Carlo integration in higher dimensions is straightforward: Just define x in (12.13) and (12.14) be a vector, and make draws from the multivariate density $g(\mathbf{x})$. There is apparently no curse of dimensionality. One should bear in mind, however, that simple Monte Carlo integration will not work if the integrand is strongly peaked, and it is possible that such peaks may become more prominent in higher dimensions. In particular, for the discrete choice example in Section 12.2.2 the integrand in (12.4) may be nonzero over only a small part of the range of (u, v) , a complication pursued in Section 12.7. Moreover, drawing from a multivariate distribution can be more difficult than drawing from a univariate distribution.

12.4. Maximum Simulated Likelihood Estimation

We now consider application of these ideas to ML estimation when no analytical expression is available for the density. The key result is that simulation can lead to an estimator with the same distribution as the MLE, provided that the number of simulation draws made to compute the density for each observation goes to infinity.

12.4.1. Simulators

Suppose the conditional density $f(y|\mathbf{x}, \theta)$ for an observation involves an intractable integral. Specifically, suppose that, as in (12.1),

$$f(y_i|\mathbf{x}_i, \theta) = \int h(y_i|\mathbf{x}_i, \theta, \mathbf{u}_i)g(\mathbf{u}_i)d\mathbf{u}_i, \quad (12.15)$$

which needs to be estimated if there is no closed-form solution.

The **direct simulator** for $f(y_i|\mathbf{x}_i, \theta)$ is the obvious Monte Carlo integral estimate

$$\hat{f}(y_i|\mathbf{x}_i, \mathbf{u}_{iS}, \theta) = \frac{1}{S} \sum_{s=1}^S h(y_i|\mathbf{x}_i, \theta, \mathbf{u}_i^s), \quad (12.16)$$

where \mathbf{u}_{iS} is a vector of S draws \mathbf{u}_i^s , $s = 1, \dots, S$, that are independent draws from $g(\mathbf{u}_i)$. This simply averages $h(y_i|\mathbf{x}_i, \boldsymbol{\theta}, \mathbf{u}_i^s)$ over the S draws. From Section 12.3.2, \widehat{f}_i is unbiased for f_i and is consistent for f_i as the number of draws $S \rightarrow \infty$.

Simulators other than the direct simulator can be used, and these are detailed in Section 12.7. These can yield an estimate \widetilde{f}_i that better approximates f_i for a finite number of draws by, for example, permitting correlation among the draws provided they still have marginal distribution $g(\mathbf{u}_i)$. More generally, then, a **simulator** for $f(y_i|\mathbf{x}_i, \boldsymbol{\theta})$ is a Monte Carlo estimate

$$\widehat{f}(y_i|\mathbf{x}_i, \mathbf{u}_{iS}, \boldsymbol{\theta}) = \frac{1}{S} \sum_{s=1}^S \widetilde{f}(y_i|\mathbf{x}_i, \boldsymbol{\theta}, \mathbf{u}_i^s), \quad (12.17)$$

where \mathbf{u}_i^s , $s = 1, \dots, S$, are S draws with marginal density $g(\mathbf{u}_i)$ but not necessarily independent over s . To be useful the simulator $\widehat{f}_i \xrightarrow{P} f_i$ as $S \rightarrow \infty$. This is likely if the **subsimulator** $\widetilde{f}(\cdot)$ is an **unbiased simulator** with the property that

$$E[\widetilde{f}(y|\mathbf{x}, \boldsymbol{\theta}, \mathbf{u}^s)] = f(y|\mathbf{x}, \boldsymbol{\theta}). \quad (12.18)$$

A desirable property of a simulator is that \widehat{f}_i be differentiable in $\boldsymbol{\theta}$, so that standard iterative gradient methods can be used to compute the estimate of $\boldsymbol{\theta}$. To eliminate “chatter” caused by simulation and ensure numerical convergence, the underlying Monte Carlo draws used to construct \widehat{f}_i should not be redrawn as $\boldsymbol{\theta}$ changes across iterations.

12.4.2. MSL Estimator

Given independence over i , the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_{ML}$ maximizes $\ln L_N(\boldsymbol{\theta}) = \sum_{i=1}^N \ln f(y_i|\mathbf{x}_i, \boldsymbol{\theta})$. The **maximum simulated likelihood** (MSL) estimator $\widehat{\boldsymbol{\theta}}_{MSL}$ instead maximizes the log-likelihood based on a simulated estimate of the density, or

$$\ln \widehat{L}_N(\boldsymbol{\theta}) = \sum_{i=1}^N \ln \widehat{f}(y_i|\mathbf{x}_i, \mathbf{u}_{iS}, \boldsymbol{\theta}), \quad (12.19)$$

where the simulator $\widehat{f}(\cdot)$ is defined in (12.17). If $\widehat{f}(\cdot)$ is differentiable in $\boldsymbol{\theta}$ then $\widehat{\boldsymbol{\theta}}_{MSL}$ can be computed using the standard gradient methods of Chapter 10, with either analytical or numerical derivatives used.

12.4.3. Distribution of the MSL Estimator

From the general consistency proof method outlined in Section 5.3.2, the MSL estimator will have the same probability limit as the ML estimator if the approximating objective function $N^{-1} \ln \widehat{L}_N(\boldsymbol{\theta})$ has the same probability limit as the original objective function $N^{-1} \ln L_N(\boldsymbol{\theta})$. This occurs if $\ln \widehat{f}_i - \ln f_i \xrightarrow{P} 0$, which in turn happens if $\widehat{f}_i - f_i \xrightarrow{P} 0$ as $S \rightarrow \infty$.

Even if the MSL estimator is consistent, it is possible that simulation error will inflate the variance of the MSL estimator compared to the ML estimator. As an example

of a formal statement of conditions under which the MSL estimator is fully efficient we give the following proposition, which is a rephrasing of a theorem in Gouriéroux and Monfort (1991).

Proposition 12.1 (Distribution of MSL Estimator) (Gouriéroux and Monfort 1991): *Assume the following:*

- (i) *The data are from a simple random sample from a dgp with conditional density $f(y|\mathbf{x}, \boldsymbol{\theta}_0)$ that satisfies the regularity conditions so that the ML estimator is consistent and asymptotically normal with limit variance matrix $\mathbf{A}^{-1}(\boldsymbol{\theta}_0)$, where*

$$\mathbf{A}(\boldsymbol{\theta}_0) = -\text{plim} \left[N^{-1} \sum_{i=1}^N \frac{\partial^2 \ln f(y_i|\mathbf{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \Big|_{\boldsymbol{\theta}_0} \right].$$

- (ii) *The density f is estimated using the simulator \hat{f} in (12.17) with \tilde{f} unbiased for f .*

Then the **maximum simulated likelihood estimator** defined in (12.19) is asymptotically equivalent to the ML estimator if $S, N \rightarrow \infty$ and $\sqrt{N}/S \rightarrow 0$, and it has a limit normal distribution with

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{\text{MSL}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}[0, \mathbf{A}^{-1}(\boldsymbol{\theta}_0)]. \quad (12.20)$$

The MSL estimator is actually consistent under the weaker condition that $S, N \rightarrow \infty$. This is satisfied if, for example, $S = N^{0.4}/a$ for some constant a . However, then $\sqrt{N}/S = aN^{0.1} \rightarrow \infty$, so the MSL estimator is not fully efficient according to Proposition 12.1. By the usual first-order Taylor series expansion the limit distribution of $\sqrt{N}(\hat{\boldsymbol{\theta}}_{\text{MSL}} - \boldsymbol{\theta}_0)$ is a matrix multiple of $N^{-1/2} \sum_i \partial \ln \hat{f}_i / \partial \boldsymbol{\theta}|_{\boldsymbol{\theta}_0}$, which depends on both variability of $\partial \ln f_i / \partial \boldsymbol{\theta}$ and simulation error in the approximation \hat{f}_i . Proposition 12.1 says that for this simulation error to disappear asymptotically the number of draws S must increase with sample size at rate in excess of \sqrt{N} .

The variance matrix of the MSL estimator requires estimation of $\mathbf{A}(\boldsymbol{\theta}_0)$. It is easiest to use a simulated variant of the BHHH estimate defined in Section 5.5.2. Since $\partial \ln f_i / \partial \boldsymbol{\theta} = (\partial f_i / \partial \boldsymbol{\theta}) / f_i$, the BHHH estimate for the information matrix is

$$\widehat{\mathbf{B}} = \frac{1}{N} \sum_{i=1}^N \frac{\partial f_i(\widehat{\boldsymbol{\theta}})/\partial \boldsymbol{\theta}}{f_i(\widehat{\boldsymbol{\theta}})} \frac{\partial f_i(\widehat{\boldsymbol{\theta}})/\partial \boldsymbol{\theta}'}{f_i(\widehat{\boldsymbol{\theta}})}.$$

Because there is no closed-form solution for f_i and $\partial f_i / \partial \boldsymbol{\theta}$ this expression cannot be computed. So we replace f_i by the simulator \hat{f}_i defined in (12.17), yielding the simulated estimate of the asymptotic variance

$$\widehat{\mathbf{V}}[\widehat{\boldsymbol{\theta}}_{\text{MSL}}] = \left(\sum_{i=1}^N \left(\frac{\sum_{s=1}^S \partial \tilde{f}_i^s(\widehat{\boldsymbol{\theta}})/\partial \boldsymbol{\theta}}{\sum_{s=1}^S f_i^s(\widehat{\boldsymbol{\theta}})} \frac{\sum_{s=1}^S \partial \tilde{f}_i^s(\widehat{\boldsymbol{\theta}})/\partial \boldsymbol{\theta}'}{\sum_{s=1}^S f_i^s(\widehat{\boldsymbol{\theta}})} \right) \right)^{-1}, \quad (12.21)$$

where $\tilde{f}_i^s(\widehat{\boldsymbol{\theta}}) = \tilde{f}(y_i|\mathbf{x}_i, u_i^s, \widehat{\boldsymbol{\theta}}_{\text{MSL}})$. Alternative estimates of the variance matrix can be obtained by similar adaptation of the Hessian estimate and sandwich estimates defined in Section 5.5.2.

An important practical issue concerns the number of simulations. One can increase the number of simulations as the sample size increases, but the level or the absolute value of S remains indeterminate. If there is little difference in the estimates using 2,400 simulations, say, rather than 2,600, then we might take this as an indication that 2,400 simulations is an adequate number. Suppose now that the sample increases four fold. By how much should we increase the number of simulations? Proposition 12.1 suggests that we should more than double S to more than 4,800, so that the ratio \sqrt{N}/S decreases toward zero. However, notice that in this case we may not be sure if \sqrt{N}/S , here equal to 1/30 if $S = 2,400$ and $N = 6,400$, say, is sufficiently close to zero. So the question of whether one has done enough simulations is difficult to answer. Many practitioners rely on rough indicators of convergence of point estimates, informally based on checking the gradients of $L_N(\boldsymbol{\theta})$. A formal test-based approach to choosing S is discussed in Hajivassiliou (2000).

12.4.4. Asymptotic Bias-Adjusted MSL

The MSL estimator is inconsistent, or asymptotically biased, when the number of simulations $S < \infty$. This bias arises for finite S because $\ln \hat{f}_i$ is biased for $\ln f_i$ even if the simulator \hat{f}_i is unbiased for f_i , as the consequence of taking the natural logarithm. Thus $N^{-1} \ln \hat{L}_N(\boldsymbol{\theta})$ and $N^{-1} \ln L_N(\boldsymbol{\theta})$ have different probability limits for finite S . This motivates a search for alternative simulation-based estimators, since we can never set $S = \infty$ and it may be computationally expensive to set S to be large.

The obvious approach is to find an unbiased simulator for the log-density $\ln f_i$, rather than for f_i , but in practice this is not possible. Instead, in this section we present a bias-corrected version of MSL, and in the following section we present an alternative, less efficient estimator than MSL that is consistent for finite S .

Gouriéroux and Monfort (1991) give an expression for the bias of the MSL estimator. The inconsistency of the MSL estimator for fixed S comes from the fact that then $\ln \hat{f}$ is an inconsistent estimator of $\ln f$. A way of reducing the inconsistency is to use a bias-adjusted log-likelihood function. Write

$$\ln \hat{f} = \ln[f + (\hat{f} - f)].$$

Taking a second-order Taylor expansion around $\ln f$ yields

$$\ln \hat{f} \simeq \ln f + \frac{\hat{f} - f}{f} - \frac{1}{2} \frac{(\hat{f} - f)^2}{f^2}.$$

Integrating with respect to the density of \mathbf{u} , and solving for $\ln f$, yields

$$\ln f \simeq E_{\mathbf{u}}[\ln \hat{f}] + \frac{1}{2} \frac{E_{\mathbf{u}}[(\hat{f} - f)^2]}{f}, \quad (12.22)$$

assuming \hat{f} is an unbiased simulator so that $E_{\mathbf{u}}[\hat{f}] = f$. This expression makes it clear that a simulator \hat{f} with small variance leads to lower bias.

A bias-corrected estimator uses an *adjusted log-likelihood* based on the right-hand side of (12.22). For the simulator (12.17), \hat{f} equals $S^{-1} \sum_s \tilde{f}^s$ and $E_{\mathbf{u}}[(\hat{f} - f)^2]$ equals $S^{-1} \sum_s E_{\mathbf{u}}[(\tilde{f}^s - f)^2]$. Given draws independent over s the latter can be

approximated by $S^{-1} \sum_s (\tilde{f}^s - \hat{f})^2$. Then (12.22) yields the **first-order asymptotic bias-corrected MSL estimator**, $\hat{\theta}_{\text{BCMSL}}$, which maximizes

$$\ln \hat{L}_{B,N}(\theta) = \sum_{i=1}^N \left[\ln \hat{f}(y_i | \mathbf{x}_i, \mathbf{u}_{iS}, \theta) + \frac{1}{2S} \frac{\sum_{s=1}^S [\tilde{f}(y_i, \mathbf{x}_i, u_i^s, \theta) - \hat{f}(y_i | \mathbf{x}_i, \mathbf{u}_{iS}, \theta)]^2}{\hat{f}(y_i | \mathbf{x}_i, \mathbf{u}_{iS}, \theta)^2} \right],$$

where $\hat{f}(y_i | \mathbf{x}_i, \mathbf{u}_{iS}, \theta) = S^{-1} \sum_s \tilde{f}(y_i, \mathbf{x}_i, u_i^s, \theta)$. The usefulness of this bias-reduction technique will vary from case to case, as the assumption that bias is small may not always hold.

12.4.5. Unobserved Heterogeneity Example

Suppose that $y_i \sim \mathcal{N}[\theta_i, 1]$, where the scalar parameter θ_i varies across individuals with $\theta_i = \theta + u_i$, with u_i representing unobserved heterogeneity that is assumed to have a known distribution. The density of y conditional on u is simply

$$f(y|u, \theta) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -(y - \theta - u)^2 / 2 \right\}. \quad (12.23)$$

However, inference on θ needs to be based on the marginal density of y (i.e., marginal with respect to u), which requires integrating out u . Here we assume that u has density

$$g(u) = e^{-u} \exp(-e^{-u}), \quad (12.24)$$

a skewed distribution that has nonzero mean and for simplicity does not depend on unknown parameters.

Maximum likelihood estimation is not possible as the marginal density $f(y|\theta)$, which equals $\int f(y|\theta, u)g(u)du$, has no closed-form solution. We instead use the MSL estimator using the direct simulator in (12.16), so that $\hat{\theta}_{\text{MSL}}$ maximizes

$$\ln \hat{L}_N(\theta) = \frac{1}{N} \sum_{i=1}^N \ln \left(\frac{1}{S} \sum_{s=1}^S \frac{1}{\sqrt{2\pi}} \exp \left\{ -(y_i - \theta - u_i^s)^2 / 2 \right\} \right), \quad (12.25)$$

where u_i^s , $s = 1, \dots, S$, are draws from the extreme value density $g(u_i)$ in (12.24). The MSL estimator $\hat{\theta}_{\text{MSL}}$ is the solution to the first-order conditions

$$\frac{\partial \ln \hat{L}_N(\theta)}{\partial \theta} = \frac{1}{N} \sum_{i=1}^N \frac{\sum_{s=1}^S (y_i - \theta - u_i^s) \exp \left\{ -(y_i - \theta - u_i^s)^2 / 2 \right\}}{\sum_{s=1}^S \exp \left\{ -(y_i - \theta - u_i^s)^2 / 2 \right\}} = 0, \quad (12.26)$$

upon some simplification. There is no closed-form solution for θ , but standard iterative methods can be used to compute $\hat{\theta}_{\text{MSL}}$.

Consistency of the MSL estimator requires the number of draws $S \rightarrow \infty$, in addition to the usual sample size $N \rightarrow \infty$, so the method is potentially computationally intensive. The MSL estimator is then asymptotically normally distributed as usual, with asymptotic variance most easily estimated using the BHHH estimator (12.21),

Table 12.2. Maximum Simulated Likelihood Estimation: Example

Number of Simulations	$S = 1$	$S = 10$	$S = 100$	$S = 1,000$	$S = 10,000$
MSL estimate $\hat{\theta}$	1.0416	1.0594	1.1775	1.1845	1.1828
Standard error	(.0968)	(.1093)	(.1453)	(.1448)	(.0091)
$\ln \hat{L}(\hat{\theta})$	-136.31	-174.38	-190.44	-192.43	-192.35

which yields

$$\widehat{V}[\hat{\theta}_{\text{MSL}}] = \left(\sum_{i=1}^N \left[\frac{\sum_{s=1}^S (y_i - \hat{\theta}_{\text{MSL}} - u_i^s) \exp \{-(y_i - \hat{\theta}_{\text{MSL}} - u_i^s)^2/2\}}{\sum_{s=1}^S \exp \{-(y_i - \hat{\theta}_{\text{MSL}} - u_i^s)^2/2\}} \right]^2 \right)^{-1}. \quad (12.27)$$

This estimator is fully efficient.

To illustrate we consider a sample $\{y_1, \dots, y_{100}\}$ of size $N = 100$ generated from the model of (12.23) and (12.24) with $\theta = 1$. Table 12.2 gives estimates as the number of draws S increases. For small S the MSL estimator is inconsistent. By $S = 10,000$ the estimator $\hat{\theta}_{\text{MSL}}$ has stabilized, though the estimated standard error bounces around quite a bit. The simulated log-likelihood decreases as S increases but eventually stabilizes. This decrease is expected as the simulator is unbiased for $f(y|\theta)$ but is biased upward for $\ln f(y|\theta)$ since by Jensen's inequality $\ln E[\hat{f}(y|\theta)] > E[\ln \hat{f}(y|\theta)]$ because the natural logarithm function is globally concave; see Appendix A (Section A.8).

12.5. Moment-Based Simulation Estimation

The simulation approach to estimation when there is no closed-form expression for the objective function can be extended to estimators other than the MLE. Furthermore, in some cases it is possible to obtain consistent parameter estimates with only a few simulations per observation, though there is then an efficiency loss.

12.5.1. Simulated m-Estimators

Consider an m-estimator that has as its objective function (see Section 5.2.2)

$$Q_N(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^N q(y_i, \mathbf{x}_i, \boldsymbol{\theta}).$$

Maximum likelihood is the special case $q(y, \mathbf{x}, \boldsymbol{\theta}) = \ln f(y|\mathbf{x}, \boldsymbol{\theta})$.

Suppose there is no closed-form expression for $q(\cdot)$, but a simulated estimate is available. Then a **simulated m-estimator** minimizes

$$\widehat{Q}_N(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^N \widehat{q}(y_i, \mathbf{x}_i, \mathbf{u}_{iS}, \boldsymbol{\theta}), \quad (12.28)$$

where, similar to Section 12.4.1, \hat{q}_i is an estimate of q_i based on a vector \mathbf{u}_{iS} of S draws \mathbf{u}_i^s , $s = 1, \dots, S$, from an appropriate distribution. Usually, $\hat{q}(\cdot) = S^{-1} \sum_s \tilde{q}(y_i | \mathbf{x}_i, \boldsymbol{\theta}, \mathbf{u}_i^s)$, where \mathbf{u}_i^s is the s th draw.

The simulated m-estimator will be consistent if the m-estimator is consistent and additionally

$$\text{plim } \hat{Q}_N(\boldsymbol{\theta}) = \text{plim } Q_N(\boldsymbol{\theta}), \quad (12.29)$$

since from Section 5.3 the necessary condition for consistency of the original m-estimator is that $\text{plim } Q_N(\boldsymbol{\theta})$ is maximized at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Here the first plim is with respect to all stochastic variables, including the simulated draws \mathbf{u}_{iS} , whereas the second plim does not depend on \mathbf{u}_{iS} .

Condition (12.29) is satisfied if the simulator is such that $\hat{q}_i - q_i \xrightarrow{P} 0$ as $S \rightarrow \infty$, since then $N^{-1} \sum_i \hat{q}_i - N^{-1} \sum_i q_i \xrightarrow{P} 0$. This was the assumption made in Section 12.4. Furthermore, the simulated m-estimator should have the same limit distribution as the m-estimator if, as in Section 12.4, S increases with sample size so that $\sqrt{N}/S \rightarrow 0$. This requires many simulations.

12.5.2. Reducing the Number of Simulations

Now suppose the simulator \hat{q}_i is not only consistent but is unbiased. Then by application of a law of large numbers, and for simplicity suppressing stochastic variables other than the simulated draws, $\text{plim } \hat{Q}_N(\boldsymbol{\theta}) = \lim N^{-1} \sum_i E_{\mathbf{u}_{iS}}[\hat{q}_i] = \lim N^{-1} \sum_i q_i = \text{plim } Q_N(\boldsymbol{\theta})$ and condition (12.29) is satisfied. Thus the simulated m-estimator is consistent with as little as one draw of \mathbf{u}_i per observation, provided $E_{\mathbf{u}_{iS}}[\hat{q}_i] = q_i$.

Unfortunately, this result is difficult to implement, as in applications it is rarely possible to find an unbiased simulator for q_i . For example, with ML estimation it can be possible to find an unbiased simulator for the density f_i , but it is not possible to find an unbiased simulator for $\ln f_i$. Similarly, for NLS estimation it can be possible to find an unbiased estimator for the conditional mean, but it is not possible to find an unbiased simulator for the squared error, which involves the square of the conditional mean.

In some cases this result can be implemented, however, if the estimator is a method of moments or GMM estimator rather than an m-estimator.

12.5.3. Method of Simulated Moments

Suppose theory leads to a conditional moment condition

$$E[m(y_i, \mathbf{x}_i, \boldsymbol{\theta}_0) | \mathbf{x}_i] = 0, \quad (12.30)$$

where $m(\cdot)$ is a scalar for simplicity. Let \mathbf{w}_i denote instruments, a function of \mathbf{x}_i and possibly $\boldsymbol{\theta}_0$, that satisfy

$$E[\mathbf{w}_i m(y_i, \mathbf{x}_i, \boldsymbol{\theta}_0)] = 0. \quad (12.31)$$

The method of moments estimator $\widehat{\theta}_{MM}$ (see Chapter 6.3.1) minimizes

$$Q_N(\boldsymbol{\theta}) = \left[\frac{1}{N} \sum_{i=1}^N \mathbf{w}_i m(y_i, \mathbf{x}_i, \boldsymbol{\theta}) \right]' \left[\frac{1}{N} \sum_{i=1}^N \mathbf{w}_i m(y_i, \mathbf{x}_i, \boldsymbol{\theta}) \right], \quad (12.32)$$

where for simplicity the just-identified case that $\dim[\mathbf{w}_i] = \dim[\boldsymbol{\theta}]$ is assumed. Results do generalize to the overidentified case, but the notation is more cumbersome as a weighting matrix then needs to be introduced and estimation is by GMM.

The method of moments estimator is consistent and has limit normal distribution with variance matrix that depends in part on the choice of instruments \mathbf{w}_i . An example is nonlinear regression, where $m(y, \mathbf{x}, \boldsymbol{\theta}) = y - E[y|\mathbf{x}]$ is the error term and the conditional mean $E[y|\mathbf{x}]$ is a specified function of \mathbf{x} and $\boldsymbol{\theta}$. Then the best choice of instrument is $\mathbf{w} = \partial E[y|\mathbf{x}] / \partial \boldsymbol{\theta}|_{\boldsymbol{\theta}_0}$ if the error is homoskedastic, since then the method of moments estimator has the same first-order conditions as those for the NLS estimator.

Now suppose there is no closed-form expression for $m(y, \mathbf{x}, \boldsymbol{\theta})$. For example, a nonlinear regression model may lack a closed-form expression for the conditional mean. Instead, $m(y, \mathbf{x}, \boldsymbol{\theta})$ is an integral

$$m(y_i, \mathbf{x}_i, \boldsymbol{\theta}) = \int h(y_i, \mathbf{x}_i, \mathbf{u}_i, \boldsymbol{\theta}) g(\mathbf{u}_i) d\mathbf{u}_i, \quad (12.33)$$

for some functions $h(\cdot)$ and $g(\cdot)$, that has no closed-form solution. Obtaining a method of moments estimator is no longer feasible.

The **method of simulated moments** (MSM) estimator $\widehat{\theta}_{MSM}$ instead minimizes

$$\widehat{Q}_N(\boldsymbol{\theta}) = \left[\frac{1}{N} \sum_{i=1}^N \mathbf{w}_i \widehat{m}(y_i, \mathbf{x}_i, \mathbf{u}_{iS}, \boldsymbol{\theta}) \right]' \left[\frac{1}{N} \sum_{i=1}^N \mathbf{w}_i \widehat{m}(y_i, \mathbf{x}_i, \mathbf{u}_{iS}, \boldsymbol{\theta}) \right], \quad (12.34)$$

where $\widehat{m}(y_i, \mathbf{x}_i, \mathbf{u}_{iS}, \boldsymbol{\theta})$ is an **unbiased simulator** for $m(y_i, \mathbf{x}_i, \boldsymbol{\theta})$ that satisfies the condition

$$E[\widehat{m}(y_i, \mathbf{x}_i, \mathbf{u}_{iS}, \boldsymbol{\theta})] = m(y_i, \mathbf{x}_i, \boldsymbol{\theta}), \quad (12.35)$$

and \mathbf{u}_{iS} denotes S draws from the marginal density $g(\mathbf{u}_i)$ and $S \geq 1$. Examples of m_i and unbiased simulator \widehat{m}_i are given in the following.

12.5.4. Distribution of MSM Estimator

The MSM estimator was proposed by McFadden (1989), who proved the following properties for the estimator.

Proposition 12.2 (Distribution of MSM Estimator) (McFadden 1989): Assume the following:

- (i) The data are from a simple random sample from a dgp, where $m(y, \mathbf{x}, \boldsymbol{\theta}_0)$ has zero conditional expectation as in (12.30) and $\mathbf{w}_i m(y, \mathbf{x}, \boldsymbol{\theta}_0)$ has zero unconditional expectation as in (12.31) and assumptions are satisfied so that the MM estimator that minimizes (12.32) is consistent and asymptotically normal.
- (ii) The function $m(y, \mathbf{x}, \boldsymbol{\theta}_0)$ is defined by (12.33) and is estimated using the unbiased simulator $\widehat{m}(y, \mathbf{x}, \boldsymbol{\theta}_0)$ that satisfies (12.35).

Then with S fixed the **method of simulated moments estimator** that minimizes (12.34) is consistent and asymptotically normal as $N \rightarrow \infty$ and has a limit normal distribution with

$$\sqrt{N}(\widehat{\boldsymbol{\theta}}_{\text{MSM}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}\left[\mathbf{0}, \mathbf{A}^{-1}(\boldsymbol{\theta}_0)\mathbf{B}(\boldsymbol{\theta}_0)\mathbf{A}^{-1}(\boldsymbol{\theta}_0)'\right], \quad (12.36)$$

where

$$\mathbf{A}(\boldsymbol{\theta}_0) = \text{plim} \frac{1}{N} \sum_{i=1}^N \mathbf{w}_i \left. \frac{\partial m_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right|_{\boldsymbol{\theta}_0} \quad (12.37)$$

and

$$\mathbf{B}(\boldsymbol{\theta}_0) = \text{plim} \frac{1}{N} \sum_{i=1}^N \mathbf{w}_i V[\widehat{m}_i(\boldsymbol{\theta}_0)] \mathbf{w}_i', \quad (12.38)$$

with the variance $V[\cdot]$ being with respect to both the conditional distribution of y_i given x_i and the draws \mathbf{u}_{iS} given after (12.35).

Before giving a derivation for this proposition we note the following. First, the MSM estimator has the remarkable property of being consistent even if $S = 1$. Second, there is an efficiency loss for finite S . The variance matrix for $\widehat{\boldsymbol{\theta}}_{\text{MM}}$ is the same as that for $\widehat{\boldsymbol{\theta}}_{\text{MSM}}$, except that for MM estimation $V[\widehat{m}_i]$ in (12.38) is replaced by the smaller $V[m_i]$. Third, the efficiency loss caused by simulation disappears as $S \rightarrow \infty$, since then $V[\widehat{m}_i] = V[m_i]$. Fourth, as for MM estimation, the MSM estimator with $S \rightarrow \infty$ may be inefficient compared to other estimators if the instruments \mathbf{w} are poorly chosen.

Consistency of the MSM estimator requires that condition (12.29) is satisfied for $\widehat{Q}_N(\boldsymbol{\theta})$ and $Q_N(\boldsymbol{\theta})$ given in (12.34) and (12.32). By a law of large numbers

$$\text{plim} \frac{1}{N} \sum_{i=1}^N \mathbf{w}_i \widehat{m}_i = \text{plim} N^{-1} \sum_{i=1}^N \mathbf{w}_i E_{\mathbf{u}_{iS}}[\widehat{m}_i],$$

where the first plim is with respect to all stochastic variables whereas the second plim is with respect to all stochastic variables aside from the simulated draws \mathbf{u} . Here $E_{\mathbf{u}_{iS}}[\widehat{m}_i] = m_i$ since \widehat{m}_i is an unbiased simulator, so

$$\text{plim} \frac{1}{N} \sum_{i=1}^N \mathbf{w}_i \widehat{m}_i = \text{plim} N^{-1} \sum_{i=1}^N \mathbf{w}_i m_i.$$

This in turn implies that $\text{plim } \widehat{Q}_N(\boldsymbol{\theta}) = \text{plim } Q_N(\boldsymbol{\theta})$. So $\widehat{\boldsymbol{\theta}}_{\text{MSM}}$ is consistent, provided $\boldsymbol{\theta}_0$ maximizes $\text{plim } Q_N(\boldsymbol{\theta})$, which is necessary for the original MM estimator to be consistent.

For the limit distribution, differentiating $\widehat{Q}_N(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ yields

$$\left(\frac{1}{N} \sum_{i=1}^N \mathbf{w}_i \frac{\partial \widehat{m}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right)' \frac{1}{N} \sum_{i=1}^N \mathbf{w}_i \widehat{m}_i(\widehat{\boldsymbol{\theta}}) = \mathbf{0}.$$

The first matrix is a full-rank square matrix, so equivalently $\widehat{\boldsymbol{\theta}}_{\text{MSM}}$ satisfies the first-order conditions

$$\frac{1}{N} \sum_{i=1}^N \mathbf{w}_i \widehat{m}_i(\widehat{\boldsymbol{\theta}}) = \mathbf{0},$$

where $\widehat{m}_i(\boldsymbol{\theta}) = \widehat{m}_i(y_i, \mathbf{x}_i, \mathbf{u}_{iS}, \boldsymbol{\theta})$. By the usual exact first-order Taylor series expansion about $\boldsymbol{\theta}_0$

$$\sum_{i=1}^N \mathbf{w}_i \widehat{m}_i(\boldsymbol{\theta}_0) + \sum_{i=1}^N \mathbf{w}_i \left. \frac{\partial \widehat{m}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right|_{\boldsymbol{\theta}^*} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \mathbf{0},$$

and hence

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = - \left(N^{-1} \sum_{i=1}^N \mathbf{w}_i \left. \frac{\partial \widehat{m}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right|_{\boldsymbol{\theta}^*} \right)^{-1} N^{-1/2} \sum_{i=1}^N \mathbf{w}_i \widehat{m}_i(\boldsymbol{\theta}_0).$$

Now $E_{\mathbf{u}} [\partial \widehat{m}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}] = \partial E_{\mathbf{u}} [\widehat{m}(\boldsymbol{\theta})] / \partial \boldsymbol{\theta} = \partial m(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$, so the first matrix on the right-hand side converges to $\mathbf{A}(\boldsymbol{\theta}_0)$ given in Proposition 12.2. The second term on the right-hand side has a limit normal distribution with mean zero and variance matrix

$$\mathbf{B}(\boldsymbol{\theta}_0) = \text{plim} \frac{1}{N} \sum_{i=1}^N \mathbf{w}_i V[\widehat{m}_i(\boldsymbol{\theta}_0)] \mathbf{w}_i'$$

as in Proposition 12.2, where $V[\widehat{m}_i(\boldsymbol{\theta}_0)]$ is a variance with respect to both \mathbf{u}_{iS} and the distribution of y_i given \mathbf{x}_i .

Since \mathbf{u}_{iS} is independent of y_i we have

$$\begin{aligned} V_{y, \mathbf{u}}[\widehat{m}(\boldsymbol{\theta}_0)] &= V_y [E_{\mathbf{u}}[\widehat{m}(\boldsymbol{\theta}_0)]] + E_y [V_{\mathbf{u}}[\widehat{m}(\boldsymbol{\theta}_0)]] \\ &= V_y[m(\boldsymbol{\theta}_0)] + E_y[V_{\mathbf{u}}[\widehat{m}(\boldsymbol{\theta}_0)]]. \end{aligned}$$

Substitution yields a more detailed definition of $\mathbf{B}(\boldsymbol{\theta}_0)$ given in Proposition 12.2.

Simulation inflates the variance of the MSM estimator because of the term $E_y[V_{\mathbf{u}}[\widehat{m}(\boldsymbol{\theta}_0)]]$, which goes to zero as $S \rightarrow \infty$. In the special case that the simulator is the frequency simulator, it can be shown that $V_{y, \mathbf{u}}[\widehat{m}(\boldsymbol{\theta}_0)] = (1 + 1/S)V_y[m(\boldsymbol{\theta}_0)]$, so that the effect of simulation using the frequency simulator is to inflate the variance of the MM estimator by $(1 + (1/S))!$

12.5.5. Choosing between MSM and MSL

The practitioner will weigh the pros and cons of MSL versus MSM. Given that MSM is consistent for small S , and further given the difficulty of ensuring that one has set S at a large enough value to ensure a good approximation to the MLE, why would MSL be ever preferred to MSM?

First, observe that MSL is in principle straightforward and simple to implement. Given the parametric assumptions, the optimal weighting of observations is inherent to the MLE method. The MSM, analogous to the GMM, in contrast requires us to work with products of weight (or instrumental variable) functions and residuals, and these

components may be correlated. The numerical instability of the GMM estimator (without simulation) has been documented by, for example, Altonji and Segal (1996) (see Section 6.3.5). Similarly, Geweke, Keane, and Runkle (1997) and McFadden and Ruud (1994) have provided evidence of the instability of the MSM estimator. Nevertheless, although simplicity favors MSL, some of the problems associated with ensuring that sufficient number of simulations are applied should not be underestimated.

12.5.6. Unobserved Heterogeneity Example

We return to the example of Section 12.4.5. Then $y_i \sim \mathcal{N}[\theta + u_i, 1]$, where u_i has density $g(u_i)$ given in (12.24). Since $E[y_i - \theta - u_i] = 0$, we can estimate θ by the method of moments estimator that solves

$$\frac{1}{N} \sum_{i=1}^N (y_i - \theta - E[u_i]) = 0, \quad (12.39)$$

yielding $\hat{\theta}_{\text{MM}} = \bar{y} - E[\bar{u}]$. Suppose that $E[\bar{u}]$ is unknown. Then we can instead use the MSM estimator $\hat{\theta}_{\text{MSM}}$ that solves

$$\frac{1}{N} \sum_{i=1}^N \left(y_i - \theta - \frac{1}{S} \sum_{s=1}^S u_i^s \right) = 0, \quad (12.40)$$

where u_i^s are iid random draws from the extreme value distribution.

The estimating equation (12.40) can be solved, yielding

$$\hat{\theta}_{\text{MSM}} = \bar{y} - \bar{u}, \quad (12.41)$$

where $\bar{u} = (NS)^{-1} \sum_i \sum_s u_i^s$ is an average over both N and S . More generally, however, an iterative method may be needed to compute the MSM estimator.

The variance of $\hat{\theta}_{\text{MSM}}$ is easily obtained. By construction the simulated draws of u are independent of each other and of the original data y , so that $V[\hat{\theta}_{\text{MSM}}] = V[\bar{y}] + V[\bar{u}]$. Now $V[\bar{y}] = (\sigma_u^2 + 1)/N$. Since \bar{u} is the average of NS draws of u , $V[\bar{u}] = \sigma_u^2/NS$, it follows that

$$\begin{aligned} V[\hat{\theta}_{\text{MSM}}] &= V[\bar{y}] + V[\bar{u}] \\ &= \frac{\sigma_u^2 + 1}{N} + \frac{\sigma_u^2}{NS}. \end{aligned} \quad (12.42)$$

This can be consistently estimated using $\hat{\sigma}_u^2 = (NS)^{-1} \sum_{i=1}^N \sum_{s=1}^S (u_i^s - \bar{u})^2$.

We consider a sample $\{y_1, \dots, y_{100}\}$ of size $N = 100$ generated from the model (12.24) with $\theta = 1$. Table 12.3 gives the MSM estimator as the number of draws $S \rightarrow \infty$. As the number of simulations S increases the MSM estimator approaches the method of moments estimate, and the standard error falls.

Table 12.3. Method of Simulated Moments Estimation: Example

Number of Simulations	$S = 1$	$S = 10$	$S = 100$	$S = 1,000$	$S = \infty (\text{MM})$
MSM estimate $\hat{\theta}$	1.0073	1.1096	1.2012	1.1887	1.1879
Standard error	(.2471)	(.1657)	(.1681)	(.1676)	(.1684)

12.6. Indirect Inference

In this section we outline another simulation-based approach to model estimation that is sometimes used when one wants to use or estimate a model that is relatively simple to estimate, even when the underlying dgp is thought to be more complex and harder to estimate. There are several variants and interpretations of the approach; see Gouriéroux, Monfort, and Renault (1993), Smith (1993), and Gallant and Tauchen (1996). The approach has also been called the **moment matching** approach. Our exposition essentially follows the first of the aforementioned references.

Suppose that the parametrically specified dgp is denoted by the pdf $f(\mathbf{y}; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \mathcal{R}^q$, whose parameters are relatively difficult to estimate. Suppose that we can specify an **auxiliary model** with the dgp $f^a(\mathbf{y}; \boldsymbol{\beta})$, $\boldsymbol{\beta} \in \mathcal{R}^r$, which is easier to estimate by the quasi-(sometimes also called “pseudo-”) maximum likelihood method. For reasons of identification that are further discussed in the following, we assume that the dimension of $\boldsymbol{\beta}$ is not smaller than the dimension of $\boldsymbol{\theta}$, that is, $r \geq q$. For example, the auxiliary model may be an approximation to the exact likelihood, or it may be an exact likelihood of an approximate model. For a given sample, let $\hat{\boldsymbol{\beta}}$ denote the QML estimates. Then, by the results covered in Section 5.7, we know that $\hat{\boldsymbol{\beta}}$ is in general an inconsistent estimator of $\boldsymbol{\theta}$, and under some regularity conditions it converges in probability to a value called the **pseudo-true** value, which is a function of $\boldsymbol{\theta}$. The function that connects the parameters of the auxiliary model to those of the dgp is called the **binding function**, denoted as $\mathbf{h}(\boldsymbol{\theta})$. The analytical form of this function may or may not be known. Therefore, it may not always be possible to obtain $\boldsymbol{\theta} = \mathbf{h}^{-1}(\hat{\boldsymbol{\beta}})$ or $\hat{\boldsymbol{\theta}} = \mathbf{h}^{-1}(\hat{\boldsymbol{\beta}})$.

The method of indirect inference can be used to obtain an improved QML estimator with a smaller asymptotic bias than $\hat{\boldsymbol{\beta}}$. The idea is to use the model under $f(\mathbf{y}; \boldsymbol{\theta})$ to generate by simulation pseudo-observations $\mathbf{y}^{(s)}$ and to use the auxiliary model under $f^a(\mathbf{y}^{(s)}; \boldsymbol{\beta})$ to estimate $\hat{\boldsymbol{\beta}}^{(s)}$, where s refers to the s th simulation. The indirect estimator is defined by the solution of

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} (\hat{\boldsymbol{\beta}}^{(s)} - \hat{\boldsymbol{\beta}})' \boldsymbol{\Omega} (\hat{\boldsymbol{\beta}}^{(s)} - \hat{\boldsymbol{\beta}}), \quad (12.43)$$

where $\boldsymbol{\Omega}$ is a given symmetric positive definite matrix. This estimator is similar to the minimum distance estimator considered in Section 6.7. That is, we sequentially generate pseudo-observations and estimate the parameters of the auxiliary model based on the pseudo-observations. The iterations continue until the quadratic form in (12.43) is minimized. A very important point is that the seed that generates the pseudo-random observations $\mathbf{y}^{(s)}$ is kept unchanged, so that variations in the pseudo-observations across simulations are due to the variation in $\hat{\boldsymbol{\beta}}^{(s)}$.

Before further discussion, we consider a simple but specific example involving a nonlinear dgp and a linear auxiliary model. The motivation is that the auxiliary model should be easy to estimate, and the dgp should be easy to simulate.

Let the dgp be as follows:

$$\begin{aligned} y_i &= \exp(\mathbf{x}'_i \boldsymbol{\gamma}) + u_i, \\ u_i &\sim \mathcal{N}[0, \sigma^2]. \end{aligned} \quad (12.44)$$

Let the auxiliary model be the following:

$$\begin{aligned} y_i &= \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i, \\ \varepsilon_i &\sim \mathcal{N}[0, \sigma_\varepsilon^2]. \end{aligned} \quad (12.45)$$

Note the following interpretations:

$$\begin{aligned} \frac{\partial E[y|\mathbf{x}]}{\partial \mathbf{x}} &= \boldsymbol{\beta} \quad (\text{under the auxiliary model}), \\ \frac{\partial \ln E[y|\mathbf{x}]}{\partial \mathbf{x}} &= \frac{\partial E[y|\mathbf{x}]}{\partial \mathbf{x}} \times \frac{1}{E[y|\mathbf{x}]} = \boldsymbol{\gamma} \quad (\text{under the dgp}). \end{aligned}$$

Therefore, the binding function is $\boldsymbol{\gamma}E[y|\mathbf{x}] = \boldsymbol{\beta}$, or $\boldsymbol{\gamma} = (E[y|\mathbf{x}])^{-1}\boldsymbol{\beta}$. Note that $\dim[\boldsymbol{\beta}]$ equals $\dim[\boldsymbol{\gamma}]$.

Given the data $(\mathbf{x}_i, y_i, i = 1, \dots, N)$ and the least-squares estimator $\widehat{\boldsymbol{\beta}}$, and given a N -dimensional pseudo-random draw, denoted $u^{(0)}$, we generate $y_i^{(1)} (i = 1, \dots, N)$ using

$$y_i^{(1)} = \exp(\mathbf{x}'_i \widehat{\boldsymbol{\beta}}) + u_i^{(0)}$$

and obtain a revised estimator $\widehat{\boldsymbol{\beta}}^{(1)} = (\sum \mathbf{x}_i \mathbf{x}'_i)^{-1} \sum \mathbf{x}_i y_i^{(1)}$, which in turn is used to generate another set of pseudo-observations. The entire simulation cycle is repeated, holding $\mathbf{u}^{(0)}$ fixed, until $(\widehat{\boldsymbol{\beta}}^{(s)} - \widehat{\boldsymbol{\beta}})' \Omega (\widehat{\boldsymbol{\beta}}^{(s)} - \widehat{\boldsymbol{\beta}})$ approaches a constant value to desired accuracy. In the present case it is reasonable to set Ω equal to either the identity matrix or $\mathbf{X}'\mathbf{X}$, the latter choice implying that prediction from the auxiliary model is a modeling objective. The resulting estimate of $\boldsymbol{\gamma}$ is the indirect estimator.

In other applications $\dim(\boldsymbol{\beta})$ will exceed $\dim(\boldsymbol{\theta})$, so a unique value of $\boldsymbol{\theta}$ may not be available. Indeed, in the absence of an analytical binding function, we cannot recover $\boldsymbol{\theta}$, even if the two dimensions are the same. Then one settles for the best indirect estimates of the auxiliary model parameters.

To see the connection between the indirect estimator and **moment matching**, set $\Omega = \mathbf{X}'\mathbf{X}$; then $(\widehat{\boldsymbol{\beta}}^{(s)} - \widehat{\boldsymbol{\beta}})' \mathbf{X}'\mathbf{X} (\widehat{\boldsymbol{\beta}}^{(s)} - \widehat{\boldsymbol{\beta}}) = (\widehat{\boldsymbol{\beta}}^{(s)} \mathbf{X} - \widehat{\boldsymbol{\beta}} \mathbf{X})' (\widehat{\boldsymbol{\beta}}^{(s)} \mathbf{X} - \widehat{\boldsymbol{\beta}} \mathbf{X})$, which indicates that the indirect estimator is “matching” the first moment of distribution. If one also wants to match the second moment, the vector $\boldsymbol{\beta}$ can be augmented by additional parameters, such as the variance parameter. Thus one can match several moments if so desired.

Under regularity conditions the indirect estimator is consistent and asymptotically normal. The reader is referred to the previously cited works for additional detail.

12.7. Simulators

As in Section 12.3.2 we consider computation of

$$I = E[h(x)] = \int h(x)g(x)dx, \quad (12.46)$$

where for simplicity x is often a scalar. As in Section 12.3, x is being used here to denote the variable being integrated out, whereas in application sections \mathbf{u} denotes the variable being integrated out as \mathbf{x} denotes the regressors.

A **simulator** is a method to compute I . There are many ways to do so, aside from direct Monte Carlo integration given in (12.14). Ideally, simulators should be unbiased, because many of the results in Sections 12.4 and 12.5 assume an unbiased simulator, and smooth so that standard iterative gradient methods can be used. Even then the computing time for estimation of empirically interesting models can be a formidable obstacle. We present a few of the many clever procedures that have been developed to speed up simulation by reducing, for any given number of simulation draws, the simulation variance relative to crude methods such as direct Monte Carlo integration. A more complete survey is given in Geweke and Keane (2001).

12.7.1. Frequency Simulator

We begin with an example, the frequency simulator, that can be used for some discrete models. This highlights well some of the complications that can arise in simulation.

Suppose the function $h(x)$ is an indicator function that takes value 1 if $x \in A$ and 0 otherwise. Then we wish to compute

$$I = \int \mathbf{1}(x \in A)g(x)dx.$$

Direct Monte Carlo integration yields the estimate

$$\hat{I}_{\text{FREQ}} = \frac{1}{S} \sum_{s=1}^S \mathbf{1}(x^s \in A),$$

where $x^s, s = 1, \dots, S$, are S draws from $g(x)$. This is called the **frequency simulator** as it estimates I by the relative frequency with which the S draws of x^s fall in A .

A leading potential application – one that has motivated much of the econometrics literature on simulation methods – is the multinomial discrete choice model introduced in Section 12.2.2. For a three-alternative model, the probability p_1 of choosing the first alternative is given by (12.3), an integral over the positive orthant of a bivariate normal distribution. The frequency simulator \hat{p}_1 is then the proportion of draws (u_1^s, u_2^s) from the bivariate normal with $u_1^s \geq 0$ and $u_2^s \geq 0$.

The frequency simulator has several limitations. First, it is neither differentiable nor continuous in parameters θ , which appear in $\mathbf{1}(x \in A)$ and/or $g(x)$. So small changes in θ lead to the same number of draws falling in the positive orthant. For this reason McFadden (1989) and Pakes and Pollard (1989) presented a more general asymptotic theory that covers such nonsmooth simulators. In practice, however, it is best to use

alternative **smooth simulators** that are differentiable in parameters as this permits computation using the usual gradient methods.

Second, this simulator is very inefficient if only a small fraction of $x \in A$. For example, for a discrete choice model with $p_1 = 0.001$, even with 10,000 draws of S the estimate \hat{p}_1 will be very noisy. Similar problems arise more generally in direct Monte Carlo evaluation of (12.46) with continuous $h(x)$ if the probability of drawing x is low in regions where $h(x)$ is relatively large.

Third, this simulator may have problems at the boundary and give an estimate $\hat{I} = 0$ or $\hat{I} = 1$ even if the model imposes $0 < I < 1$ and this condition is necessary for model estimation.

12.7.2. Importance Sampling

The **importance sampling simulator** reexpresses the integral (12.46) as

$$\begin{aligned} I &= \int \left(\frac{h(x)g(x)}{p(x)} \right) p(x)dx \\ &= \int w(x)p(x)dx, \end{aligned} \tag{12.47}$$

where $p(x)$ is a density function chosen so that (a) it is easy to draw from $p(x)$, (b) $p(x)$ has the same support as the original domain of integration, and (c) $w(x) = h(x)g(x)/p(x)$ is easy to evaluate, is bounded, and has finite variance. We then use the direct Monte Carlo integral estimate based on (12.47) rather than (12.46),

$$\hat{I}_{IS} = \frac{1}{S} \sum_{s=1}^S w(x^s), \tag{12.48}$$

where x^s , $s = 1, \dots, S$, are draws from $p(x)$ rather than $g(x)$. The term importance sampling is used because $w(x)$ determines the weight or “importance” of different points in the sample space. Importance sampling has been employed in the Bayesian simulation literature for many years and was introduced into Bayesian econometrics by Kloek and van Dijk (1978) as a way of evaluating posterior distributions. This material is further discussed in Section 13.4.

The importance sampler \hat{I}_{IS} has variance $S^{-1}\text{V}_p[w(x)]$, given independent draws from $p(x)$. This variance is clearly minimized if $w(x)$ is a constant over the entire range of integration, since then $\text{V}_p[w(x)]$ is zero. This is done by setting $w(x) = E_g[h(x)]$, as then $p(x) = h(x)g(x)/E_g[h(x)]$ is a density that integrates to 1. Unfortunately, this theoretically ideal importance sampling estimate is not practicable, as $E_g[h(x)]$ is unknown. However, it does indicate the potential gains to importance sampling, especially if $p(x)$ is chosen so that $w(x)$ is fairly flat.

Even if importance sampling leads to an increased variance, which can occur in practice, it does have other attractions. It produces a smooth sampler if $w(x)$ is smooth in the parameters to be estimated. Moreover, it is useful if draws from $g(x)$ are difficult, as can often be the case if x is a vector of correlated random variables.

For the multinomial probit discrete choice model a popular importance sampler is the **GHK simulator**, due to Geweke (1992), Hajivassiliou and McFadden (1994), and

Keane (1994). This recursively truncates the multivariate normal pdf so that draws are restricted to the positive orthant. Advantages of this simulator compared to the frequency simulator are that it is smooth, requires many fewer draws for alternatives with low probability of being chosen, and is unlikely to have boundary problems.

12.7.3. Variance Reduction by Antithetic Acceleration

The preceding methods assume independent draws, using methods to be detailed in Section 12.8, from an appropriate distribution such as $g(x)$ or, if importance sampling is used, from $p(x)$.

Variance reduction methods instead use dependent draws as these can reduce the variance of a simulator. A leading example is **antithetic sampling** that uses negatively correlated draws. Ripley (1987, pp. 129–132), Geweke (1988), and Hajivassiliou (2000) provide a discussion of this technique and Geweke (1995) surveys this and several other variance reduction techniques.

Suppose we wish to evaluate the integral I in (12.46), where x is assumed to have zero mean and symmetric density $g(x)$. The direct Monte Carlo integral, based on $2S$ simulated iid draws from $g(x)$, is

$$\hat{h}_{2S}(x) = \frac{1}{2S} \sum_{s=1}^{2S} h(x^s)$$

and, given independence of the $2S$ draws, has variance

$$V[\hat{h}_{2S}(x)] = \frac{1}{2S} V[h(x)].$$

Antithetic sampling uses an alternative estimate based on only S iid draws,

$$\hat{h}_{A,S}(x) = \frac{1}{S} \sum_{s=1}^S \frac{1}{2} (h(x^s) + h(-x^s)), \quad (12.49)$$

which is an average of $h(x)$ evaluated at x^s and $-x^s$. The pair $(x^s, -x^s)$ is said to be an **antithetic pair** and yields an unbiased estimate of I since we assume x is symmetrically distributed with zero mean. If the mean is instead μ then $(x^s, 2\mu - x^s)$ is an antithetic pair. Given S independent draws of x^s the variance of $\hat{h}_{A,S}(x)$ is

$$\begin{aligned} V[\hat{h}_{A,S}(x)] &= \frac{1}{S^2} \sum_{s=1}^S \frac{1}{4} (V[h(x^s)] + 2\text{Cov}[h(x^s), h(-x^s)] + V[h(-x^s)]) \\ &= \frac{1}{2S} (V[h(x)] + \text{Cov}[h(x), h(-x)]). \end{aligned}$$

Antithetic sampling will therefore be more efficient than regular iid sampling if the covariance term is negative, since then the variance of $\hat{h}_{A,S}(x)$ is smaller than that of $\hat{h}_{2S}(x)$. By switching the sign of the draw, and then reusing the draw, an attempt is made to induce negative correlation in the simulator. Negative correlation is assured when the function is linear, and also if the nonlinearity is not too severe. However, in general, one cannot be certain that efficiency gains will be realized. For example, if $h(\cdot)$ is symmetric about zero then $\text{Cov}[h(x), h(-x)] = V[h(x)]$.

Antithetic sampling can be extended to asymmetric density $g(x)$. Suppose x can be drawn using the inverse transformation method given later in Section 12.8.2. Then one can draw u , say, from the uniform $[0, 1]$, generate the antithetic transform $(1 - u)$, and then use the inverse transformation method to draw from the distribution of choice, so $x_1 = G^{-1}(u)$ and $x_2 = G^{-1}(1 - u)$, where $G(\cdot)$ is the known cdf of x . Then (x_1, x_2) form an antithetic pair and variance reduction occurs if

$$\text{Cov}[h(G^{-1}(u)), h(G^{-1}(1 - u))] = \text{Cov}[f(u), f(1 - u)] < 0,$$

where $f(u)$ is the composite function $h(G^{-1}(u))$. If $f(\cdot)$ is a monotonic function then the variance is reduced (Robert and Casella, 1999, p. 112). However, this property of the function may be difficult to verify. Further, the argument applies to the inverse transformation approach only, whereas in practice other methods are used in pseudo-random number generation (see Section 12.8). Therefore it is difficult to verify in advance whether the conditions for efficiency gains are attainable in a specific application.

Although the dramatic gains in efficiency possible in some special cases may not materialize in more complex settings, worthwhile efficiency gains are realized in many cases. Antithetic sampling can also be used to accelerate importance sampling (Danielsson and Richard, 1993).

Antithetic sampling extends to multivariate draws. Consider bivariate draws of (x, y) , where the density is symmetric about $(0, 0)$. In this case sign reversal is done first element by element and then for the pair. Thus the antithetic quadruple consists of $((x^s, y^s), (-x^s, y^s), (x^s, -y^s), (-x^s, -y^s))$. For an m -dimensional draw the same idea is repeated for all tuples.

12.7.4. Computation Using Quasi-Random Sequences

A second method of variance reduction involves replacing pseudo-random numbers by **quasi-random numbers**, which are systematic simulation draws designed to provide better coverage of the sample space. A potential limitation of the approach is that randomness is required to apply the laws of large numbers and central limit theorems that justify the simulation-based approach.

Quasi-Monte Carlo methods use nonrandom points within the domain of integration instead of using S pseudo-random points. A leading example is **Halton sequences**, summarized in Press et al. (1993) and introduced into the econometrics literature by Bhat (2001) and Train (2003).

Halton sequences have two desirable properties. First, they are designed to give fairly even coverage over the domain of the sampling distribution. With more evenly spread draws for each observation, the simulated probabilities vary less over observations, relative to those calculated with random draws. This is similar to deterministic evaluation of an integral over a specified grid. Second, with Halton sequences, the draws for one observation tend to fill in the spaces left empty by the previous observations. The simulated probabilities are, therefore, negatively correlated over observations. As in the case of antithetic variates, this negative correlation reduces the variance of the simulated function. Under suitable regularity conditions it can be shown

that the integration error using pseudo-random sequences is of order N^{-1} , compared to pseudo-random sequences where the convergence rate is $N^{-1/2}$ (Bhat, 2001).

Halton sequences are best described by example. Suppose that the function to be simulated depends on one random variable. The starting point is a prime number. The Halton sequence based on the prime number 2 is constructed as follows. Divide the unit interval $(0, 1)$ into two parts. The dividing point $1/2$ becomes the first element of the Halton sequence. Next divide each part into two more parts. The dividing points, $1/4$ and $3/4$, become the next two elements of the sequence. Divide each of the four parts into two parts each, and continue to obtain the sequence $\{1/2, 1/4, 3/4, 1/8, 3/8, \dots\}$. Similarly, the sequence based on the prime number 3 is $\{1/3, 2/3, 1/9, 2/9, 4/9, \dots\}$. Halton sequences on nonprime numbers are not unique because the Halton sequence for a nonprime number divides the unit space in the same way as each of the prime numbers that constitute the nonprime.

The length of each sequence is determined by the number of observations N and the numbers of simulation draws S . One discards the first few (say 20) elements of the sequence as the early elements have a tendency to be correlated over Halton sequences with different primes (see Train, 2003, for an example). Consequently, one could begin by generating Halton sequences of length $N \times S + 20$ and discard the first 20 elements of each sequence. For each element of each sequence, calculate the inverse of the cumulative normal distribution. The resulting values are the **Halton draws** from the sampling distribution.

One major advantage of quasi-random number draws is that the draws are designed to cover the sample space of random numbers in a more uniform fashion than in the case of pseudo-random numbers. This can be seen visually in Figure 12.1. In this figure, Panel 2 shows a draw from a bivariate normal distribution constructed using a Halton sequence. The remaining three panels show pseudo-random number draws from the same distribution. The more even coverage of the sample space is evident in the former case.

For more thorough discussion and examples of simulation-based estimation that use Halton draws and impressive evidence of the relative efficiency of the approach in one or more dimensions, see Train (2003, Chapter 9). The method works very well for multinomial logit model with normally distributed random parameters (Section 15.7).

12.8. Methods of Drawing Random Variates

The preceding simulators require draws of random variates. In this section we summarize methods to take such draws from a density, denoted $g(x)$ or $p(x)$ in Section 12.7 and denoted $f(x)$ in this section. Usually it is sufficient to obtain draws from the uniform or the standard normal (which is possible in most popular software) since these can form the basis for making draws from distributions other than the uniform or normal.

If the draws are to be used for simulation-based estimation then all draws from the uniform or standard normal should be made before any estimation, to prevent “chatter,” whereby iterative methods fail to converge owing to noise created by new draws at

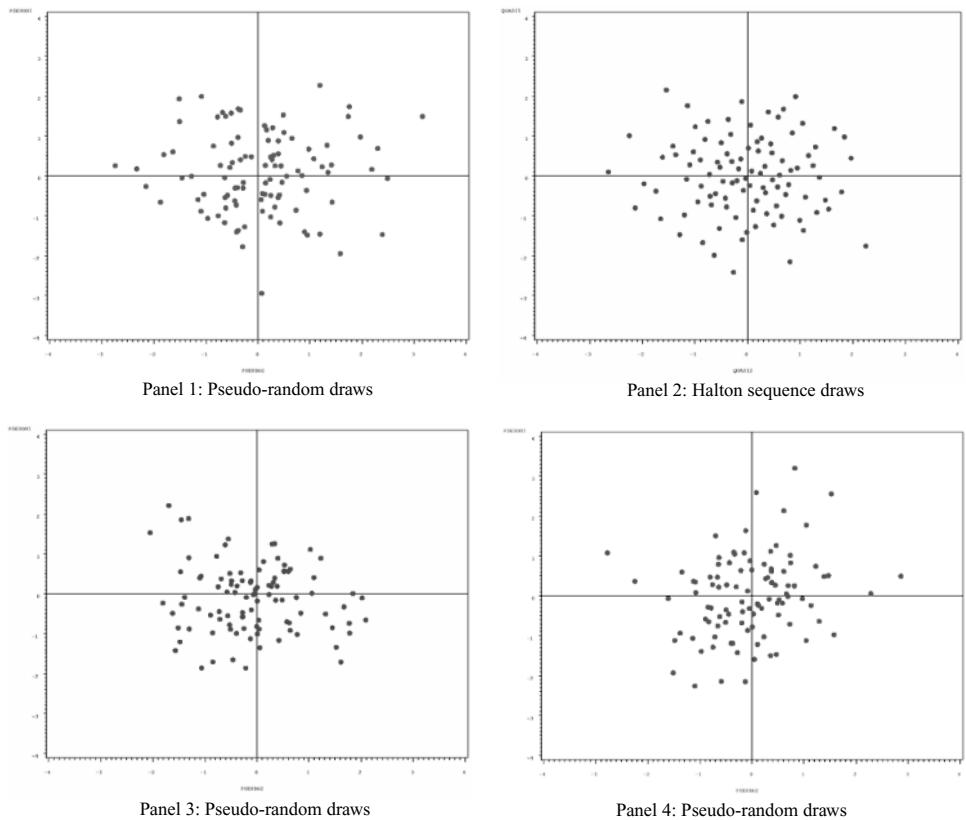


Figure 12.1: Halton sequence draws (panel 2) compared to pseudo-random draws.

each iteration. For example, if $x \sim \mathcal{N}[\mu, \sigma^2]$ and estimates of μ and σ change over iterations, then we make NS initial draws of $z \sim \mathcal{N}[0, 1]$ and then over iterations recompute $x = \mu + \sigma z$ using the original draws of z .

This section provides a basic discussion of some standard methods for generating random variates. For more advanced or extensive treatments, there are many good monographs and surveys, including those by Bradley, Fox, and Schrage (1983), Dagpunar (1988), Devroye (1986), and Ripley (1987).

Before presenting the methods, note that the term random number generation is an oxymoron. A more accurate description is given by the term **pseudo-random numbers**. The essential characteristic of these generators is that they use deterministic devices to produce long chains of numbers that mimic the properties of the realizations from some target distribution. The specific target distribution will depend on the context, but for the purposes of this book uniform, normal, exponential, gamma, logistic, and Poisson distributions are standard. The chain process is started up by supplying a **seed**. After some finite but large number of values have been generated the cycle of numbers repeats itself. That is, the computer algorithms will generate exactly the same numbers beginning with a given seed. Good random number generators are those that generate a long chain of numbers without recycling and without any built-in dependence. The key consideration in choosing generators is whether the generated

distribution closely mimics the properties of the target distribution at a reasonable computational cost.

12.8.1. Pseudo-Random Uniform Number Generators

Pseudo-random uniform numbers are constructed using a deterministic sequence that mimics the statistical properties of a sequence of uniform random numbers. A good generator has a long period, has a distribution close to uniform, and produces independent draws. It is important to have a good generator, as pseudo-random numbers from virtually any distribution can then be obtained by transforming uniform pseudo-random numbers (Bradley et al., 1983, p. 24).

A standard generator begins with the equation

$$X_j = (kX_{j-1} + c) \bmod m,$$

where the modulus operator $a \bmod b$ forms the remainder when a is divided by b . This produces a sequence of integers between 0 and m , and the uniform random variable is then obtained as $R_j = X_j/m$ (Ripley, 1987, p. 20). A value for X_0 , referred to as the **seed**, is needed to initiate the generator. The uniform random sequences generated are deterministic, which permits replication as the same numbers should be drawn if analysis is repeated with the same value of the seed. The periodicity of the cycle depends on X_0 , k , and c . If computation is done using 32-bit integer arithmetic the maximum periodicity is approximately $2^{31} \simeq 2.1 \times 10^9$. However, it is easy to choose poor values of X_0 , k , and c so that the periodicity is much lower than this. Books such as that by Press et al. (1993) should be consulted for potential pitfalls.

12.8.2. Nonuniform Variates

Random variables from many other distributions, including the normal itself, are usually based on an initial draw of a uniform random number. Four commonly used methods are (1) inverse transformation, (2) transformation, (3) accept–reject, and (4) mixing and compounding.

Inverse Transformation

Let $F(x)$ denote the cdf of the continuous random variable x , that is,

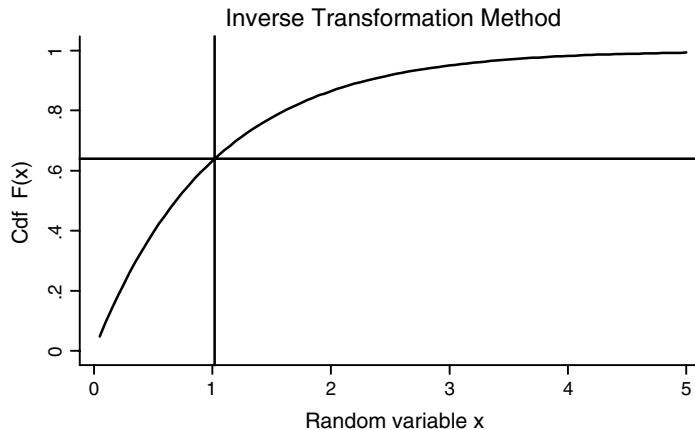
$$F(x) = \Pr[X \leq x].$$

Given a draw of a uniform variate r , $0 \leq r \leq 1$, the **inverse transformation**

$$x = F^{-1}(r)$$

gives a unique value of x because F is continuous and monotonically increasing.

For example, the cdf of the unit exponential is $1 - e^{-x}$. Solving $r = 1 - e^{-x}$ yields $x = -\ln(1 - r)$. If we make a draw from uniform $[0, 1]$ and get 0.64, then $x = -\ln(1 - 0.64) = 1.0217$. Figure 12.2 plots the cdf of X and shows graphically how this method works. An arbitrary point on the vertical axis at height r is selected and the corresponding value on the horizontal axis is obtained by completing a rectangle. This is the inverse transformation.



Draw of 0.64 (vertical axis) yields $x = 1.02$ (horizontal axis).

Figure 12.2: Inverse transformation method for making draws from the unit exponential. A random uniform draw of 0.64 (so $F(x) = 1 - \exp(-x) = 0.64$) yields $x = 1.02$.

This method is particularly easy to use if the analytical form of $F(\cdot)$ is given and x is a continuous random variable. If there is no closed-form expression available, then the method is still often feasible, albeit computationally more costly, as the inverse cdfs of standard distributions are often available as functions in programs.

The method can be extended to discrete random variables with a cdf that is a step function. For example, if x takes integer values then a uniform draw $r = 0.312$ leads to a draw of $x = j$, where the integer j is such that $F(j - 1) < 0.312$ and $F(j) \geq 0.312$.

A standard method for generating normal random variates is the Box–Muller method. This uses the inverse transformation method, applied to the joint distribution of two independent normal variates rather than to a single variate. Specifically, if r_1 and r_2 are iid uniform then $x_1 = \sqrt{-2 \ln r_1} \cos(2\pi r_2)$ and $x_2 = \sqrt{-2 \ln r_1} \sin(2\pi r_2)$ are iid $\mathcal{N}[0, 1]$.

Transformation

In some cases a random variable with the desired density can be obtained by suitable **transformation** of a random variable whose distribution is easy to draw from. Then random variates can be obtained by applying this same transformation.

This transformation method is an obvious way to make draws from distributions based on the normal. Examples include squaring standard normal variates to obtain random variables with central chi-square distribution, adding squared values of r independent standard normal variates to yield chi-squared variates with r degrees of freedom, and computing the mean square of independent chi-squares to yield F -distributed random variables. Transformation methods are not restricted to distributions based on the normal.

Accept–Reject Methods

Suppose we want to draw from the density $f(x)$ but this is difficult, however, there is another density $g(x)$ that covers $f(x)$ in the sense that $f(x) \leq kg(x)$ for all x for some

Accept-reject Method

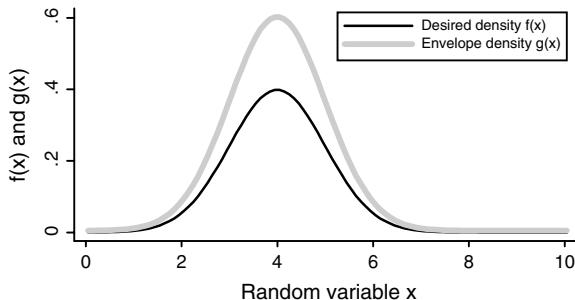


Figure 12.3: Accept-reject method draws from density $g(x)$ where $kg(x)$ envelopes the desired density $f(x)$.

finite constant k . This is depicted in Figure 12.3, where the thick line serves to mimic the envelope $kg(x)$.

The **accept–reject method** draws from $g(x)$, rather than $f(x)$. The draw is accepted, $x = r$, if

$$r \leq \frac{f(x)}{kg(x)},$$

where r is a draw from the uniform distribution. If the condition is not satisfied then the draw is rejected and further draws are made until the condition is satisfied. The appeal of the method depends on the ease of drawing from $g(x)$ rather than $f(x)$. The limitation is that on average a draw will be accepted with probability $1/k$, so that many draws are needed if k is large.

To see how this method works, let Y denote the random variable generated by the accept–reject method, X denote a random variable with density $g(x)$, and U denote a draw from the uniform. Then Y has cdf

$$\begin{aligned} \Pr[Y \leq y] &= \Pr[X \leq y | U \leq f(x)/kg(x)] \\ &= \frac{\Pr[X \leq y, U \leq f(x)/kg(x)]}{\Pr[U \leq f(x)/kg(x)]} \\ &= \frac{\int_{-\infty}^y \int_0^{f(x)/kg(x)} dug(x) dx}{\int_{-\infty}^{\infty} \int_0^{f(x)/kg(x)} dug(x) dx} \\ &= \frac{\int_{-\infty}^y [f(x)/kg(x)]g(x) dx}{\int_{-\infty}^{\infty} [f(x)/kg(x)]g(x) dx} \\ &= \frac{\int_{-\infty}^y [f(x)/k] dx}{\int_{-\infty}^{\infty} [f(x)/k] dx} \\ &= \int_{-\infty}^y f(x) dx, \end{aligned}$$

which is the cdf corresponding to the density $f(x)$ as desired.

Composition

Sometimes the density $f(x)$ can be expressed as being that from a mixture or a compound distribution, with

$$f(x) = \int g(x|\varepsilon)h(\varepsilon) d\varepsilon.$$

Then a draw from $f(x)$ can be obtained by first making a draw of ε from density $h(\varepsilon)$ and then making a draw of x from the conditional density $g(x|\varepsilon)$.

As an example, consider drawing from the negative binomial distribution with mean λ and variance $\lambda(1 + \alpha\lambda)$, where both λ and α are given constants. Here we may use the fact that the negative binomial distribution can be regarded as a Poisson–gamma mixture (see Chapter 20). First, one draws ε from a gamma distribution with mean 1 and variance α , which can be done by a transformation of the exponential. Second, one draws from the Poisson distribution with mean $\lambda\varepsilon$, given ε from the previous step.

If $h(\varepsilon)$ is a discrete distribution with point mass p_j at C points, $j = 1, \dots, C$, then the previous integration step is replaced by summation. Thus,

$$f(x) = \sum_{j=1}^C p_j g(x|\varepsilon = \varepsilon_j).$$

Then, to make S draws from $f(x)$, we draw Sp_j observations each from $g(x|\varepsilon = \varepsilon_j)$, and “compose” the required sample of S values by pooling the draws.

Some Standard Generators

The tables in Appendix B describes pseudo-random number generation for several standard continuous and discrete cases. They are based on the assumption that r, r_1, r_2, \dots are values of independent uniform $[0, 1]$ random variables R, R_1, R_2, \dots . Note that there may exist different methods to generate the corresponding random variable; we list only one or two of these methods.

12.8.3. Multivariate Distributions

Draws from multivariate distributions are generally much more complicated than draws from univariate distributions. For example, methods such as inverse transformation and transformation may no longer be applicable. For many multivariate distributions the method of mixing or composition can be used, as many multivariate distributions are mixture distributions.

Quite general methods are Gibbs sampling and other Markov chain Monte Carlo methods. These are deferred to Section 13.5, as they are extensively applied in Bayesian analysis, which uses complicated multivariate distributions. As will be explained the draws made using the Gibbs sampler may show some tendency to be correlated, a fact that will reduce the efficiency of the simulator.

Here we restrict attention to the multivariate normal. Then draws are easily obtained by transformation of univariate standard normal draws. Specifically, suppose we wish

to make draws from a q -dimensional normal distribution, so $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \Sigma)$. This can be done by transformation based on the result that a positive definite Σ has **Choleski decomposition**

$$\Sigma = \mathbf{L}\mathbf{L}',$$

where \mathbf{L} is a lower triangular matrix. For example, for $q = 2$ the Choleski decomposition is

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} \\ 0 & l_{22} \end{bmatrix},$$

yielding three equations $l_{11}^2 = \sigma_{11}$, $l_{11}l_{21} = \sigma_{12}$, and $l_{21}^2 + l_{22}^2 = \sigma_{22}$ that can be solved for l_{11} , l_{21} , and l_{22} . Given a q -dimensional vector ε whose elements have standard normal distribution, it is easy to verify that if $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, then $\mathbf{x} = \mathbf{L}\varepsilon$, a linear combination of normals, has distribution $\mathcal{N}(\mathbf{0}, \Sigma)$. Specifically, $E[\mathbf{L}\varepsilon] = \mathbf{0}$, and $V[\mathbf{L}\varepsilon] = E[\mathbf{L}\varepsilon\varepsilon'\mathbf{L}'] = \mathbf{L}\mathbf{L}' = \Sigma$. The key to this method is that linear combinations of the normal are also normally distributed, result that does not hold for nonnormal distributions.

12.9. Bibliographic Notes

Press et al. (1993) provide a good starting point for both quadrature and Monte Carlo integration and give further references, including some given elsewhere in this chapter.

The econometrics literature on simulation-based estimation emphasizes the multinomial probit model. The methods have much wider applicability, however, and can be more easily and successfully implemented in other models that are less challenging to fit than the multinomial probit. Lerman and Manski (1981) used simulated frequencies to estimate choice probabilities and found that many draws were needed. McFadden (1989) proposed MSM and demonstrated its consistency and asymptotic normality. Pakes and Pollard (1989) provide a quite general treatment of the asymptotic theory for both MSM and MSL. The relatively accessible survey of Stern (1997) is an excellent place to start. Gouriéroux and Monfort (1996) provide a textbook treatment of the basic methods. Many other references are better read in the specific context of models that are discussed in later chapters. In particular, Hajivassiliou and Ruud (1994) emphasize truncated normal models including the multinomial probit and Train (2003) considers a range of discrete choice models including the random parameters logit.

Exercises

12–1 To estimate the integral $I = \int t(x)g(x)dx$ by Monte Carlo, the sum $\hat{I} = N^{-1} \sum t(x_i)g(x_i)/p(x_i)$ is used, where x_i are draws from the importance sampling distribution $p(x)$. Show that $\text{plim } \hat{I} = I$.

12–2 For $f(\theta) = |\Sigma|^{-1/2} [1 + \frac{1}{\nu}(\theta - \mu)' \Sigma^{-1}(\theta - \mu)]^{-(\nu+d)/2}$, consider the d -dimensional integral $\int_{R^d} f(\theta)d\theta$. The integrand is the kernel of a multivariate- t density, so the correct answer is the inverse of the normalizing constant.

(a) Evaluate this integral as a Monte Carlo average $S^{-1} \sum_{s=1}^S f(\theta^{(s)})/h(\theta^{(s)})$, $\theta^{(s)} \sim h(\theta)$, where the importance density $h(\theta)$ is multivariate- t with the

same location and scale as $f(\theta)$, but with a different degrees-of-freedom parameter.

- (b) Explore the stability of this average as you vary the degrees of freedom of $h(\theta)$. Increase the mismatch between $f(\theta)$ and $h(\theta)$ by changing the location and scale of $h(\theta)$ and explore further.

12–3 For the MSM estimator in Section 12.5.3 suppose that the simulator is the frequency simulator.

- (a) Show that $V_{y,u}[\hat{m}(\theta_0)] = (1+1/S)V_y[m(\theta_0)]$.
 (b) Hence show that the effect of simulation using the frequency simulator is to inflate the variance of the method of moments estimator by $(1 + (1/S))$.
 (c) How large is the efficiency loss for the standard errors if $S = 10$?

12–4 For the example in Section 12.5.6 consider the estimator $\hat{\alpha}$ that solves $\sum_{i=1}^N [y_i - \frac{1}{S} \sum_{s=1}^S (\alpha + u_i^s)] = 0$. Obtain analytical expressions for this estimator and its variance.

12–5 (a) Write an algorithm for drawing a pseudo-random sample from a three-dimensional multivariate normal distribution $\mathcal{N}[\mathbf{0}, \Sigma]$ with $\sigma_{jj} = 1$, $j = 1, 2, 3$, and covariances $\sigma_{12} = \sigma_{13} = \sigma_{23} = 0.5$. Draw a sample of 1,000 realizations and compare the estimated means and variances with those of the dgp.
 (b) Repeat part (a) with the trivariate normal being replaced by a Student's t -distribution with five degrees of freedom.

12–6 Write a computing procedure to make draws from a univariate truncated normal density $\mathcal{T}\mathcal{N}_{[a,b]}[\mu, \sigma^2]$ using the inverse transform method given in Section 12.8.2. Here $[a, b]$ are lower and upper truncation points. Choose $\mu = 1$, $\sigma^2 = 4$, and $a = 3$, $b = 4$.

12–7 Consider the standard binary logit regression model (see Section 14.3).

- (a) Write down the log-likelihood function.
 (b) Introduce a random intercept assumption in which the intercept is drawn from a suitable distribution with finite mean and variance. What justification can you offer for introducing an unobserved heterogeneity term in this way? If the logit model is derived from the random utility model with extreme value errors, how does the random intercept affect that interpretation and/or derivation? [See Revelt and Train, 1998.]
 (c) Suggest a suitable distributional assumption for the random intercept; rewrite the likelihood function conditional on unobserved heterogeneity. Next write down the likelihood function with unobserved heterogeneity integrated out.
 (d) Describe in a step-by-step manner how to use the maximum simulated likelihood estimation procedure to estimate this model. Explain, with details, how to calculate the variance matrix of unknown parameters. How would you decide how many simulations you will use?
 (e) Consider the method of simulated moments as an alternative to the MSL procedure for the random parameter logit. Write down the moment condition(s) conditional on the unobserved heterogeneity term. Then outline an MSM estimation procedure for this model.

12–8 Some computing packages allow you to draw both Poisson and Gamma pseudo-random numbers directly. It is also known that the negative binomial distribution

SIMULATION-BASED METHODS

can be derived as a mixture of Poisson and gamma random variables (see Section 20.4).

- (a) Write down a procedure for drawing negative binomial-distributed variables using the method of mixtures.
- (b) Apply your method by drawing a sample of 10,000 on a Poisson-distributed variable with mean 0.25.
- (c) Draw a corresponding sample from a Gamma distribution with mean 1 and variance α , with α set to produce negative binomial random variables with variance 0.3125.

Bayesian Methods

13.1. Introduction

This chapter serves as an introduction to Bayesian econometrics. Bayesian regression analysis has grown in a spectacular fashion since the publication of books by Zellner (1971) and Leamer (1978). Application to routine data analysis has also expanded enormously, greatly aided by revolutionary advances in computer hardware and software technology. In the light of such major developments, a single chapter can never do adequate justice to the many facets of this subject. This chapter therefore has the very modest goal of providing a rough road map to the major ideas and developments in Bayesian econometrics. Despite this modest objective some parts are still quite technical.

The Bayesian approach, unlike the likelihood or frequentist or classical approach presented in previous chapters, requires the specification of a probabilistic model of prior beliefs about the unknown parameters, given an initial specification of a model. Many researchers are uncomfortable about this step, both philosophically and practically. This has traditionally been the basis of the concern that the Bayesian approach is subjective rather than objective. It will be shown that in large samples the role of the prior may be negligible, that relatively uninformative priors can be specified, and that there are methods available for studying the sensitivity of inferences to priors. Therefore, the charge of subjectivity may not always be as serious as many claim.

Bayesian approaches play a potentially large role in applied microeconomics, especially when dealing with complex models that lack analytically tractable likelihood functions. Chapter 12 introduced simulation-based methods for such situations. These methods, particularly simulated likelihood, are potentially problematic as they generally require maximization of a function using a sufficiently large number of simulation draws that increases at an appropriate rate as the sample size grows. Even with today's powerful computers, analysis of large samples and high-dimensional models can require a formidable amount of computation. Bayesian methods, in contrast, do not require maximization algorithms. Bayesian procedures are flexible enough to be adapted to produce estimates that are excellent (if not perfect) substitutes for maximum

likelihood estimates, which are obtained in many cases more efficiently. Indeed, it is not necessary that one goes through a philosophical conversion to use these procedures; they can be adapted for pragmatic reasons.

The foregoing remarks do not mean that Bayesian procedures do not have a deeper rationale and justification. They do. Three features in particular deserve to be mentioned. First, Bayesian procedures can yield the entire posterior distribution of the parameters of interest, leaving the user to decide which moment or quantile of the distribution to report, potentially on the basis of decision-theoretic criteria. One does not need separate estimators for means, medians, quantiles, and so forth as the posterior distribution has them all! Second, Bayesian analysis, being conditional on the data, yields exact finite-sample results, obviating the need for finite-sample corrections or adjustments. This distribution approaches the normal distribution in large samples where the influence of the priors vanishes. Third, Bayesian methods provide a natural way to select models.

Section 13.2 introduces the basic concepts and components of Bayesian analysis and the key properties of Bayesian estimators. These ideas are illustrated in Section 13.3 for the relatively tractable linear regression model. More generally, no closed-form solution exists for the posterior distribution. Section 13.4 presents Monte Carlo integration methods, notably importance sampling, used to obtain numerical estimates of posterior moments. Section 13.5 details Markov chain Monte Carlo methods, notably Gibbs sampling and the Metropolis–Hastings algorithm, used to obtain draws from the (intractable) posterior distribution. An example of these methods is given in Section 13.6. The additional topics of data augmentation and Bayesian model selection are presented in Sections 13.7 and 13.8.

13.2. Bayesian Approach

In the Bayesian approach uncertainty about the value of the parameters θ is explicitly modeled by introducing a density $\pi(\theta)$ for the **prior distribution**, so named because it is specified without considering the data currently in hand. It expresses subjective beliefs about the true unknown parameter in the language of probability. Specification of the prior is studied in detail in Section 13.2.4. As an example, suppose that θ is an income elasticity and on the basis of an economic model or previous studies it is felt that θ lies between 0.8 and 1.2 with probability 0.95. Then a prior for θ is $\theta \sim \mathcal{N}[1, 0.1^2]$.

The other ingredient of Bayesian inference is the sample joint density or likelihood $f(\mathbf{y}|\theta)$, where in the single-equation case \mathbf{y} is an $N \times 1$ vector. Dependence on regressors is suppressed throughout this section, for notational simplicity. Exogenous regressors are introduced in Section 13.3, in which case $f(\mathbf{y}|\theta)$ becomes $f(\mathbf{y}|\mathbf{X}, \theta)$ and Bayesian analysis is then conditional on regressors. Note also that in this chapter $f(\cdot)$ usually denotes the joint density of all observations, rather than the density of the i th observation.

If no data are available then all we have is the prior. After data are observed, the classical approach is to estimate the unknown parameter θ using the maximum likelihood principle. The Bayesian approach instead combines the likelihood of the sample with

the prior, reflecting the view that any prior information should be exploited, even if it is in the form of a probability distribution. This process can be thought of as a revision of the prior given the data (likelihood). Indeed, we can derive a distribution of θ after combining the likelihood and the prior. The resulting distribution is called a **posterior distribution**, and it reflects the investigator's beliefs about θ a posteriori, that is, after observing the data.

13.2.1. Bayes' Theorem

The basic result that delivers the posterior distribution is **Bayes' Theorem**, also referred to sometimes as Bayes' **inverse law of probability**, that

$$f(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{f(y)}, \quad (13.1)$$

where $f(y)$ denotes the marginal probability distribution of y , formally defined as

$$f(y) = \int_{R(\theta)} f(y|\theta)\pi(\theta)d\theta, \quad (13.2)$$

where $R(\theta)$ denotes the support of $\pi(\theta)$. This result is obtained by noting that, for events A and B , the conditional probability

$$\begin{aligned} \Pr[A|B] &= \frac{\Pr[A \cap B]}{\Pr[B]} \\ &= \frac{\Pr[B|A]\Pr[A]}{\Pr[B]}, \end{aligned}$$

where the second equality follows because $\Pr[B|A] = \Pr[A \cap B]/\Pr[A]$.

Because the denominator $f(y)$ in (13.1) is free of θ , we can more simply write $p(\theta|y)$ as proportional to the product of the pdf and the prior; thus

$$p(\theta|y) \propto L(y|\theta)\pi(\theta). \quad (13.3)$$

This simplifies derivation and representation of the posterior, by omitting inessential constants that can be recovered later, as will be illustrated in Section 13.2.2. When a density function is written without normalizing constants it is referred to as a **density kernel**.

In many cases (13.1) or (13.3) do not yield a closed-form expression for the posterior density. A closed-form expression is not needed, however, and later sections present recent simulation-based techniques for obtaining good numerical approximations to the posterior density. These techniques permit Bayesian analysis for almost any parametric microeconomics application.

It is common to use a special symbol for the posterior density, so we will replace $f(\theta|y)$ by $p(\theta|y)$. Also, the original joint density, $f(y|\theta)$ is the likelihood function $L(y|\theta)$. Henceforth we will write the **posterior density** as

$$p(\theta|y) \propto L(y|\theta)\pi(\theta). \quad (13.4)$$

This representation, the key one for the Bayesian approach, emphasizes an important difference between the frequentist and Bayesian approaches. In the frequentist

approach, the true value of the parameter is constant but parameter estimates are treated as random variables. In contrast, in the Bayesian approach the parameter is treated as if it is random.

13.2.2. Bayes' Theorem Example

Suppose $y \sim \mathcal{N}[\theta, \sigma^2]$, where σ^2 is known but the scalar parameter θ is unknown. Given a random sample (y_1, \dots, y_N) , the joint density of \mathbf{y} is

$$\begin{aligned} L(\mathbf{y}|\theta) &= \prod_{i=1}^N (2\pi\sigma^2)^{-1/2} \exp\left\{-(y_i - \theta)^2 / 2\sigma^2\right\} \\ &= (2\pi\sigma^2)^{-N/2} \exp\left\{-\sum_{i=1}^N (y_i - \theta)^2 / 2\sigma^2\right\} \\ &\propto \exp\left\{-\frac{N}{2\sigma^2} (\bar{y} - \theta)^2\right\}, \end{aligned}$$

where $\bar{y} = N^{-1} \sum_i y_i$, and we use $\sum_i (y_i - \theta)^2 = \sum_i (y_i - \bar{y} + \bar{y} - \theta)^2 = \sum_i (y_i - \bar{y})^2 + \sum_i (\bar{y} - \theta)^2$. Multiplicative terms not involving θ , which are absorbed in the constant of proportionality, are dropped. The frequentist approach maximizes the log-likelihood with respect to θ , leading to the MLE $\hat{\theta} = \bar{y}$.

The Bayesian approach additionally specifies a prior for θ . An analytically convenient choice is the normal prior, with $\theta \sim \mathcal{N}[\mu, \tau^2]$, where we suppose that values of the prior mean μ and prior variance τ^2 are specified. A large value of τ^2 indicates greater prior uncertainty than a small value. Then the prior density is

$$\begin{aligned} \pi(\theta) &= (2\pi\tau^2)^{-1/2} \exp\left\{-(\theta - \mu)^2 / 2\tau^2\right\} \\ &\propto \exp\left\{-(\theta - \mu)^2 / 2\tau^2\right\}, \end{aligned}$$

where $(2\pi\tau^2)^{-1/2}$, which is free of θ , is absorbed into the factor of proportionality. Using (13.4), we obtain the posterior density

$$p(\theta|\mathbf{y}) = \frac{L(\mathbf{y}|\theta)\pi(\theta)}{\int_{-\infty}^{\infty} L(\mathbf{y}|\theta)\pi(\theta)d\theta}, \quad -\infty < \theta < \infty. \quad (13.5)$$

The denominator ensures that the posterior is proper (i.e., it integrates to 1). For some purposes the denominator can be ignored, in which case we work with $p(\theta|\mathbf{y}) \propto L(\mathbf{y}|\theta)\pi(\theta)$. The numerator can be expanded as follows:

$$\begin{aligned} L(\mathbf{y}|\theta)\pi(\theta) &= (2\pi\sigma^2)^{-N/2} \exp\left\{-\sum_{i=1}^N \frac{(y_i - \theta)^2}{2\sigma^2}\right\} (2\pi\tau^2)^{-1/2} \exp\left\{-\frac{(\theta - \mu)^2}{2\tau^2}\right\} \\ &= (2\pi)^{-(N+1)/2} (\sigma^2)^{-N/2} (\tau^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - \theta)^2 - \frac{(\theta - \mu)^2}{2\tau^2}\right\}. \end{aligned}$$

Because

$$\sum_{i=1}^N (y_i - \theta)^2 = \sum_{i=1}^N (y_i - \bar{y})^2 + N(\bar{y} - \theta)^2,$$

and noting that the constant of integration in (13.5) and other multiplicative constants independent of θ can be absorbed into the proportionality constant, we have

$$p(\theta|y) \propto \exp \left\{ -\frac{N}{2\sigma^2} (\theta - \bar{y})^2 \right\} \exp \left\{ -\frac{1}{2} \frac{(\theta - \mu)^2}{\tau^2} \right\} \quad (13.6)$$

$$\propto \exp \left\{ -\frac{1}{2} \left[\frac{(\theta - \mu)^2}{\tau^2} + \frac{(\bar{y} - \theta)^2}{N^{-1}\sigma^2} \right] \right\}$$

$$\propto \exp \left\{ -\frac{1}{2} \left[\frac{(\theta - \mu_1)^2}{\tau_1^2} \right] \right\}. \quad (13.7)$$

The last line is the kernel of $\mathcal{N}[\mu_1, \tau_1^2]$ distribution, where

$$\begin{aligned} \mu_1 &= \tau_1^2 (N\bar{y}/\sigma^2 + \mu/\tau^2), \\ \tau_1^2 &= (N/\sigma^2 + 1/\tau^2)^{-1}. \end{aligned} \quad (13.8)$$

The final line in (13.7) is obtained by completing the square, using the result that for arbitrary scalars z, y, a_1, a_2, c_1 , and c_2 , we have

$$c_1(z - a_1)^2 + c_2(z - a_2)^2 = (c_1 + c_2) \left(z - \left(\frac{c_1 a_1 + c_2 a_2}{(c_1 + c_2)^2} \right) \right)^2 + \frac{c_1 c_2}{(c_1 + c_2)} (a_1 - a_2)^2,$$

where $z = \theta, a_1 = \mu, a_2 = \bar{y}, c_1 = 1/\tau^2$, and $c_2 = 1/(N^{-1}\sigma^2 + \tau^2)$. The terms free of θ are dropped.

In summary, we have the following:

Data: $y|\theta \sim \mathcal{N}[\theta, \sigma^2], \sigma^2$ known.

Prior: $\theta \sim \mathcal{N}[\mu, \tau^2], \mu, \tau^2$ specified.

Posterior: $\theta|y \sim \mathcal{N}[\mu_1, \tau_1^2], \mu_1, \tau_1^2$ given in (13.8).

The **posterior mean** μ_1 is a weighted sum of the prior mean μ and the sample mean \bar{y} with weights that reflect the precision of the likelihood via σ^2/N and the prior via τ^2 . Bayesian practice is to summarize variability using the **precision parameter**, defined as the reciprocal of the variance. Here the **posterior precision** τ_1^{-2} is the sum of the sample precision of $\bar{y}, N/\sigma^2$, and the **prior precision** $1/\tau^2$, so precision is increased by pooling the sample and prior information.

If the prior information is imprecise, so that $1/\tau^2$ is small, then the weight assigned to the prior mean is also small relative to the sample information and the prior plays a minor role in generating the posterior. Similarly, the sample information also dominates as the sample size gets large, since then N/σ^2 gets large relative to $1/\tau^2$. The posterior distribution tends to the familiar asymptotically normal, except the Bayesian result is that $\theta \xrightarrow{a} \mathcal{N}[\bar{y}, \sigma^2/N]$ rather than $\bar{y} \xrightarrow{a} \mathcal{N}[\theta, \sigma^2/N]$.

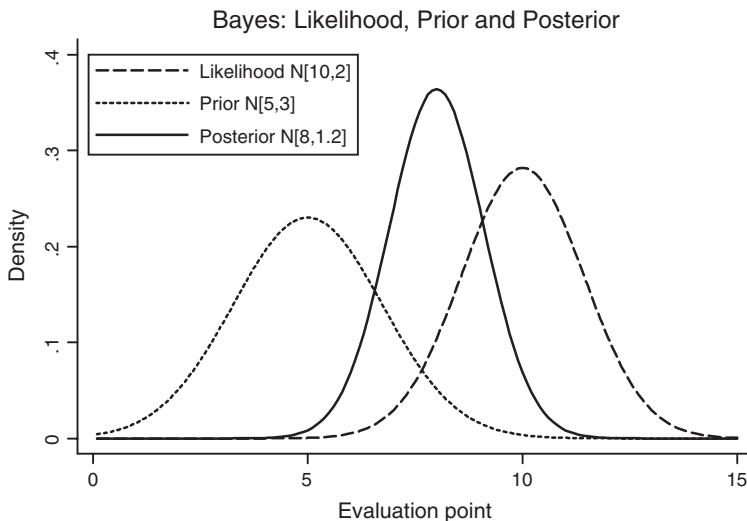


Figure 13.1: Bayesian analysis for mean parameter of normal density: plot of normal likelihood (right), normal prior density (left), and resulting posterior density (center).

As a concrete example, suppose $\sigma^2 = 100$, the prior sets $\mu = 5$ and $\tau^2 = 3$, and a sample of size $N = 50$ has sample mean $\bar{y} = 10$. Then the likelihood is $\mathcal{N}[10, 2]$, the prior is $\mathcal{N}[5, 3]$, and from (13.7) and (13.8) the posterior is $\mathcal{N}[8, 1.2]$. These densities are plotted in Figure 13.1. The posterior mean lies between the prior mean and the sample mean, whereas the posterior has variance that is smaller than the variance of both the prior and the likelihood.

13.2.3. Bayesian and Non-Bayesian Approaches Compared

It is useful to draw parallels and contrasts between the frequentist and Bayesian approaches.

In a parametric frequentist formulation the likelihood function is the main basis of statistical inference. Under suitable regularity conditions the MLE is consistent and asymptotically normal. Sampling theory of estimators provides a basis for probability statements about the estimated magnitudes, or functions thereof, or conditional prediction. Prior information on parameters is incorporated by restricted ML estimation.

In a Bayesian analysis, summarized in Table 13.1, the data-generating process and the data are combined with a prior distribution on the parameters. Specification of this prior distribution is discussed in detail in Section 13.2.4. The prior embodies probabilistically specified information before the current data are analyzed and may be based on “received information.” The prior information and the data are combined using Bayes’ Theorem.

The outcome of this exercise is the posterior distribution of the parameters θ , which we may think of as the translated likelihood function. Alternatively, given the data, the posterior distribution reflects our “revised prior.” If the sample is small, and perhaps

Table 13.1. Bayesian Analysis: Essential Components

Component	Formula
Sampling model	(y_1, \dots, y_N) iid from $f(y \theta)$
Joint density/likelihood	$f(y \theta), L(y \theta); \theta \in \Theta$
Prior distribution	$\pi(\theta), \theta \in \Theta$
Posterior density	$p(\theta y) : \begin{cases} = f(y \theta)\pi(\theta)/\int f(y \theta)\pi(\theta)d\theta \\ \propto f(y \theta)\pi(\theta) \\ \propto L(y \theta)\pi(\theta) \end{cases}$
Posterior pdf \rightarrow posterior inference \rightarrow	$\begin{cases} \text{parameter estimation} \\ \text{probability statements} \\ \text{prediction} \\ \text{model comparison} \end{cases}$

relatively uninformative, the posterior may look like the prior, but if the sample is large, the posterior distribution will reflect the features of the data.

13.2.4. Specification of the Prior

Bayesian analysis requires specification of the dgp $f(y|\theta)$ and of the prior $\pi(\theta)$. The dgp is usually specified to be the same as that used in a fully parametric likelihood-based analysis. For binary outcomes a logit or probit model might be specified, for count data the Poisson or negative binomial model would be specified, and so on.

The principle challenge introduced by Bayesian analysis, compared to classical analysis, is the need to additionally specify a prior distribution. Results can vary with the choice of prior, as different priors lead to different posterior distributions unless the sample is large enough that the sample information dominates.

One approach is to choose a prior such that it has little impact on the posterior, so that results essentially are based on the sampled data. An alternative approach, warranted when strong prior information is available, is to specify a prior that reflects this information. Both approaches, especially the latter, were historically constrained by issues of tractability of the resulting posterior, but this has now become much less of a consideration given recent computational advances. A popular intermediate approach is to use **hierarchical priors**, with uncertainty about parameters expressed in terms of probability functions that themselves involve other parameters about which we are also uncertain.

Noninformative Priors

A **noninformative prior** is one that has little impact on the resulting posterior distribution.

The obvious way to try to obtain a noninformative prior is to use a **uniform prior** with $\pi(\theta) = c$ for all θ , where $c > 0$ is a constant, since this places equal weight on all possible values of θ .

One disadvantage of the uniform prior is that if it is used in settings where the parameters θ are unbounded then the prior is an **improper density** because then necessarily $\int \pi(\theta) d\theta = \infty$. The resulting posterior distribution may then also be improper, though in several leading examples the posterior is nonetheless proper.

Another disadvantage of the uniform prior is that it is not invariant to reparameterization. For example, for a scalar parameter $\theta > 0$ an alternative obvious parameterization of the density of y is in terms of the parameter $\gamma = \ln \theta$, as then $-\infty < \gamma < \infty$. If θ has a uniform prior, $\pi(\theta) = c$, then the corresponding prior $\pi^*(\gamma)$ for γ is not the uniform since $\pi^*(\gamma) = \pi(\theta) |d\theta/d\gamma| = ce^\gamma$. Although seemingly uninformative for one parameterization, the prior is informative in another parameterization.

The uniform prior can be emulated by specifying a proper prior that has very large variances. For example, suppose the scalar θ has $\mathcal{N}[\mu, \tau^2]$ prior, where τ^2 is very large. Then for values of θ likely to be supported by the data the prior $\pi(\theta) \simeq 1/(2\pi\tau^2)$, a constant, because $\exp[-(\theta - \mu)/2\tau^2] \simeq 1$. It is important to note that this obvious approach, called a **vague** or **diffuse** or **flat prior**, has the same weakness as the uniform prior. It is not invariant to reparameterization.

Instead, a widely used noninformative prior is **Jeffreys' prior**,

$$\pi(\theta) \propto |\mathcal{I}(\theta)|^{1/2}, \quad (13.9)$$

where for a vector θ , $|\mathcal{I}(\theta)|$ is the determinant of the information matrix $\mathcal{I}(\theta) = -E[\partial^2 \mathcal{L}/\partial \theta \partial \theta']$ with $\mathcal{L} = \ln L(y|\theta)$. Jeffreys' prior, named after the pioneering Bayesian Harold Jeffreys, has the property of **invariance** to reparameterization or transformation of model parameters, so that same prior information is being given regardless of the particular parameterization chosen.

To verify Jeffrey's rule, for simplicity consider the scalar parameter case. Given transformation $\gamma = h(\theta)$, $\partial \mathcal{L}/\partial \gamma = \partial \mathcal{L}/\partial \theta \times \partial \theta/\partial \gamma$ and

$$\frac{\partial^2 \mathcal{L}}{\partial \gamma^2} = \frac{\partial^2 \mathcal{L}}{\partial \theta^2} \left(\frac{\partial \theta}{\partial \gamma} \right)^2 + \frac{\partial \mathcal{L}}{\partial \theta} \frac{\partial^2 \theta}{\partial \gamma^2}.$$

Taking expectations with respect to the sample density and noting that $E[\partial \mathcal{L}/\partial \theta] = 0$ by the property of likelihood scores yields

$$\mathcal{I}(\gamma) = \mathcal{I}(\theta) \left(\frac{\partial \theta}{\partial \gamma} \right)^2.$$

It follows that

$$|\mathcal{I}(\gamma)|^{1/2} = |\mathcal{I}(\theta)|^{1/2} \left| \frac{\partial \theta}{\partial \gamma} \right|.$$

In general the prior $\pi(\theta)$ for θ implies the prior for γ is $\pi^*(\gamma) = \pi(\theta) \times |d\theta/d\gamma|$. Specializing to prior (13.9), we have $\pi^*(\gamma) \propto |\mathcal{I}(\theta)|^{1/2} \times |d\theta/d\gamma|$, but this is $|\mathcal{I}(\gamma)|^{1/2}$ as desired.

As an example, suppose $y \sim \mathcal{N}[\mu, \sigma^2]$, and consider three cases. First, if μ is the unknown parameter and σ^2 is known, then the information measure for μ is $\mathcal{I}(\mu) = N/\sigma^2$, and Jeffrey's prior $|\mathcal{I}(\mu)|^{1/2} \propto c$, a constant since here σ^2 is known. Note that this prior is an improper prior. Second, if σ^2 is unknown and μ is known, then the

information measure for σ^2 is $\mathcal{I}(\sigma^2) = N/(2\sigma^4)$, and Jeffrey's prior $|\mathcal{I}(\sigma^2)|^{1/2} \propto \sigma^{-2}$. Third, if both μ and σ^2 are unknown then the information matrix $|\mathcal{I}(\mu, \sigma^2)| = (N/\sigma^2)(N/2\sigma^4) = N^2/2\sigma^6$. Therefore, Jeffreys' rule implies that the joint prior $\pi(\mu, \sigma^2) \propto \sigma^{-3}$. Note that this is different from what we get if we apply Jeffreys' rule to the separate priors for μ and σ^2 , as $\pi(\mu) \propto c$ and $\pi(\sigma^2) \propto \sigma^{-2}$ yields $\pi(\mu)\pi(\sigma^2) \propto \sigma^{-2}$.

Jeffreys' rule can serve as a method of generating a prior when there are no obvious candidate priors available. However, the literature does not seem to have resolved the issue of whether the rule produces a noninformative prior and if so in what sense. Further, as is clear from the preceding example Jeffreys' prior can be improper, which may lead to an improper posterior.

Conjugate Priors

When a proper prior is specified, either as an informative prior or as a diffuse prior, it is convenient to choose a functional form for the prior that, given the specified sample density for the data, leads to a “nice” analytically tractable expression for the posterior, such as (13.7).

Such tractable results most often arise if the sample and prior densities form a **natural conjugate pair**, defined as having the property that sample density and prior and posterior distributions all lie in the same class of densities. Then the prior is called a **natural conjugate prior**. Section 13.2.2 gave an example, where for normally distributed data a normal prior for the mean leads to a posterior that was also normal.

The exponential family is essentially the only class of densities to have natural conjugate priors. A one-parameter member of the exponential family has a density that for a single observation can be expressed as

$$\begin{aligned} f(y|\theta) &= \exp\{a(\theta) + b(y) + c(\theta)u(y)\} \\ &\propto \exp\{a(\theta) + c(\theta)u(y)\}, \end{aligned} \tag{13.10}$$

where different functions $a(\cdot)$, $c(\cdot)$, and $u(\cdot)$ lead to different densities in the family, and $b(\cdot)$ is a normalizing constant. For example, setting $c(\theta) = \mu/\sigma^2$, $a(\theta) = -\mu^2/2\sigma^2$, and $u(y) = y$ yields the kernel of the $\mathcal{N}[\mu, \sigma^2]$ distribution (for σ^2 known). Note that setting $u(y) = y$ yields the linear exponential family, presented in some detail in Section 5.7.3. More generally, if θ is a vector then $c(\theta)u(y)$ is replaced by $\mathbf{c}(\theta)' \mathbf{u}(y)$, where usually $\mathbf{u}(\cdot)$ has the same dimension as θ .

For a random sample of size N the exponential family leads to sample density

$$L(\mathbf{y}|\theta) \propto \exp\{Na(\theta) + c(\theta)t(\mathbf{y})\}, \tag{13.11}$$

where $t(\mathbf{y}) = \sum_i u(y_i)$. Consider the following prior on θ :

$$\pi(\theta|\beta, \alpha) \propto \exp\{\beta a(\theta) + \alpha c(\theta)\}, \tag{13.12}$$

where α and β are specified parameters of the prior and the functions $a(\cdot)$ and $c(\cdot)$ are the same as those in (13.10). This density is an exponential family density for θ once

Table 13.2. Conjugate Families: Leading Examples

Distribution	Sample Density	Conjugate Prior Density
Normal	$\mathcal{N}[\theta, \sigma^2]$	$\theta \sim \mathcal{N}[\mu, \tau^2]$
Normal	$\mathcal{N}[\mu, 1/\theta^2]$	$\theta \sim \mathcal{G}[\alpha, \beta]$
Binomial	$\mathcal{B}[N, \theta]$	$\theta \sim \text{Beta}[\alpha, \beta]$
Poisson	$\mathcal{P}[\theta]$	$\theta \sim \mathcal{G}[\alpha, \beta]$
Gamma	$\mathcal{G}[\nu, \theta]$	$\theta \sim \mathcal{G}[\alpha, \beta]$
Multinomial	$\mathcal{MN}[\theta_1, \dots, \theta_k]$	$\theta_1, \dots, \theta_k \sim \text{Dirichlet}[\alpha_1, \dots, \alpha_k]$

α is viewed as fixed. Applying Bayes' Theorem and simplifying, we get

$$\begin{aligned} p(\theta|\mathbf{y}) &\propto L(\mathbf{y}|\theta)\pi(\theta|\beta, \alpha) \\ &\propto \exp\{(\beta + N)a(\theta) + (\alpha + t(\mathbf{y}))c(\theta)\}, \end{aligned} \quad (13.13)$$

which is readily verified to have the same kernel as the original prior in (13.12). Comparison of the posterior with the sample density reveals that the prior is treated as providing an additional β observations \mathbf{y}_p , say, with $t(\mathbf{y}_p) = \alpha$.

Table 13.2 presents some standard conjugate families, where the relevant densities are provided in Appendix B. The gamma includes exponential and chi-square as special cases. Negative binomial, uniform, and Pareto likelihoods also have conjugate prior densities.

An attraction of a conjugate prior is the resulting computational and analytical simplicity. Nevertheless, using a conjugate prior is a restriction and the justification for imposing it is less compelling now than it was in the past when computational resources available to a typical researcher were rather limited.

Another advantage of having a posterior that is in the same class as the prior is that the posterior can easily replace the prior as a new (data-based) prior for a later analysis. If a prior is to be interpreted as “received information,” then one may take the posterior from one investigation as a prior for the next.

Hierarchical Priors

Hierarchical priors are those that arise when the parameters in a prior are themselves modeled as having a distribution. The parameters that appear in such a “prior on a prior” are called **hyperparameters**.

The data have joint density $L(\mathbf{y}|\theta)$, as in Section 13.2.1, but now the prior on θ depends on parameters τ , say, that are random rather than fixed. Thus the prior on θ is $\pi(\theta|\tau)$, where the parameters τ in turn have a prior $\pi(\tau)$. The joint prior is $\pi(\theta, \tau) = \pi(\theta|\tau)\pi(\tau)$, and Bayes' rule yields the joint posterior

$$p(\theta, \tau|\mathbf{y}) \propto L(\mathbf{y}|\theta)\pi(\theta|\tau)\pi(\tau).$$

Interest will usually lie in the marginal posterior for θ , which is obtained by integrating the joint posterior with respect to τ . The specified parameters of the prior $\pi(\tau)$ are called hyperparameters. Alternatively, these parameters in turn can be given a prior, in which case another hierarchical level is introduced leading to joint prior $\pi(\theta|\tau)\pi(\tau|\phi)\pi(\phi)$, and so on. Recent advances in computational methods for Bayesian analysis, particularly the Gibbs sampler, are well suited to hierarchical priors because of their recursive structure.

Hierarchical priors can be viewed as a Bayesian analogue of random coefficient models in a classical setting. For example, for iid count data we might suppose that $y_i \sim \mathcal{P}[\theta_i]$, where the Poisson parameter is now random. A convenient distribution for θ_i is the conjugate gamma distribution, so $\theta_i \sim \mathcal{G}[\alpha, \beta]$. The classical approach estimates α and β by maximum likelihood. A nonhierarchical Bayesian model specifies values for α and β and obtains the posterior for θ_i . A hierarchical Bayesian model specifies priors for α and β , such as the gamma that is conjugate, and first obtains the joint posterior for θ_i , α , and β before finding the marginal posterior for θ_i .

Hierarchical priors arise naturally in the context of **hierarchical models**, also known as **multilevel models**. Such models are widely applied in classical settings using special purpose software (Bryk and Raudenbusch, 1992, 2002). An early contribution by Lindley and Smith (1972) analyzed hierarchical regression models in a Bayesian setting. Hierarchical modeling has a natural appeal when the data to be analyzed naturally fall into strata, groups, or layers, and further one may expect to see groupwise parameter variation in the relationship of interest. For example, observations on test scores could come from students in specific grades and schools. Modeling of test scores could involve individual characteristics that by definition vary across individuals, class characteristics that vary across grades, and school characteristics that only vary across schools. Because such data will involve clustering of observations, this topic is also discussed in Chapter 24. Such models also have a close relationship with random effects formulation for panel data.

As an example, suppose that data naturally fall into J groups, and that the population mean of y differs across the groups. For individual i in group j suppose $y_{ij} \sim \mathcal{N}[\theta_j, \sigma^2]$, where for simplicity we assume σ^2 is known. Then the sample mean in the j th group $\bar{y}_j \sim \mathcal{N}[\theta_j, \sigma^2/N_j]$, where N_j denotes the number of individuals in the group and independence is assumed. A hierarchical model specifies the means θ_j to have prior $\theta_j \sim \mathcal{N}[\mu, \tau^2]$, for example, where additional priors are specified for the parameters μ and τ^2 of the higher level prior.

Sensitivity Analysis

In a frequentist analysis one may entertain a variety of exact prior restrictions in formulating a model for estimation. For example, a model may be estimated under one or more sets of restrictions, and the results can be compared to form an idea of the sensitivity of the estimates to prior assumptions.

The same logic and approach applies in Bayesian analysis. One need not take the prior to be literally true, and one can perform a sensitivity analysis that studies how the

posterior changes with different choice of prior. Similarly, one can vary assumptions about the dgp and see how posterior beliefs change in response.

13.2.5. Densities and Measures Related to the Posterior

Bayesian analysis is based on the posterior distribution. For convenience Bayesian regression results usually report only summary measures, such as posterior moments, quantiles, or marginal distributions of components of θ . However, the posterior distribution is also used for prediction and probability statements, detailed in this Section, and for model comparison, presented in Section 13.8.

Several quantities play an important role in a Bayesian analysis.

Marginal Posterior

In general θ is multidimensional, denoted by $\theta' = (\theta_1, \dots, \theta_q)$ and interest may lie in the posterior distribution of individual components of θ . The **marginal posterior density** of the k th parameter, θ_k , is obtained by integrating out of the joint posterior all the remaining $(q - 1)$ elements of θ . Formally, this is denoted as $p(\theta_k | \mathbf{y})$ and is obtained by calculating the $(q - 1)$ -fold integral

$$\begin{aligned} p(\theta_k | \mathbf{y}) &= \int p(\theta_1, \dots, \theta_p | \mathbf{y}) d\theta_1 \dots d\theta_{k-1} d\theta_{k+1} \dots d\theta_q \\ &= \int p(\theta | \mathbf{y}) d\theta_{-k}, \end{aligned} \tag{13.14}$$

where the more compact notation in the second line contains θ_{-k} , which means all elements of θ other than θ_k . The marginal posterior density is usually asymmetric and need not be unimodal, whereas the asymptotic normal distribution for classical estimators is symmetric and unimodal. It can be useful to graph the posterior, especially if it departs considerably from a symmetric unimodal distribution.

Posterior Moments

Classical regression output reports the parameter estimate and standard error. For Bayesian regression one can similarly report the mean or median and the standard deviation of the marginal posterior density of each parameter.

Point Estimation

In classical analysis there is an unknown true parameter value θ_0 such that the dgp is $f(\mathbf{y} | \theta_0)$, and we seek a point estimate that is a good estimate of θ_0 . In Bayesian analysis, in contrast, interest lies in the entire distribution of θ , which is determined by both θ_0 and prior beliefs about θ_0 .

Point estimation is therefore emphasized much less in Bayesian analysis. For convenience the posterior mean or the posterior median are nonetheless commonly reported as point estimates. By specifying a loss function an optimal point estimate of a parameter can be obtained; see Section 13.2.7.

Posterior Intervals

Once the posterior distribution has been obtained, it can be used to make probability statements analogous to those in the frequentist analysis. In particular, we can consider Bayesian confidence intervals and regions.

For the k th parameter, a $100(1 - \alpha)\%$ **posterior density interval** $\mathcal{R}(\theta_k)$ is any interval that θ_k falls into with posterior probability α , or formally

$$1 - \alpha = \Pr [\theta_k \in \mathcal{R}(\theta_k) | \mathbf{y}] = \int_{\mathcal{R}(\theta_k)} p(\theta_k | \mathbf{y}) d\theta. \quad (13.15)$$

There are many regions that correspond to this probability. The simplest posterior interval is one between the $\alpha/2$ and $(1 - \alpha/2)$ quantiles, such as between the 2.5 and 97.5 percentiles. More complicated is a **highest posterior density (HPD) interval** that satisfies (13.15) and additionally the condition that no point in $\mathcal{R}(\theta)$ has a smaller probability density than any point outside the region. This interval need not be contiguous if the posterior is multimodal, and it differs from the simpler interval unless the posterior is symmetric and unimodal.

These intervals can be extended to regions. A $100(1 - \alpha)\%$ **highest posterior density region** $\mathcal{R}(\theta)$ is a region such that

$$1 - \alpha = \Pr [\theta \in \mathcal{R}(\theta) | \mathbf{y}] = \int_{\mathcal{R}(\theta)} p(\theta | \mathbf{y}) d\theta. \quad (13.16)$$

An attraction of the Bayesian approach is that a posterior interval is much simpler to interpret than a confidence interval in frequentist analysis. If a 95% posterior interval for θ_k is $(1, 4)$, then θ_k lies between 1 and 4 with posterior probability 0.95. In contrast, for a frequentist 95% confidence interval for θ_k equal to $(1, 4)$ we can only say that if it were possible to repeat the analysis with many different samples yielding many different confidence intervals, then 95% of these confidence intervals will include the true value of θ_k .

Hypothesis Testing

Hypothesis testing receives little attention in the Bayesian context. As noted in the discussion of point estimation, interest does not lie in determining the true parameter value θ_0 . Instead, interest lies in the distribution of the range of values that θ might take given the data and a prior. For model comparison see Section 13.8.

Conditional Posterior Density

The **conditional posterior density** of θ_k , given θ_j , can be obtained from the joint and marginal posterior densities as

$$p(\theta_k | \theta_j, \theta_j \in \boldsymbol{\theta}_{-k}, \mathbf{y}) = \frac{p(\theta_k, \theta_j | \mathbf{y})}{p(\theta_j | \mathbf{y})}. \quad (13.17)$$

Of special interest and significance is the set of q conditional distributions $p(\theta_k | \boldsymbol{\theta}_{-k})$, $k = 1, \dots, q$, also known as the set of **full conditional distributions**. These play an

important role in the modern computational techniques for obtaining the joint posterior distribution presented in later sections.

The definitions of marginal and conditional posteriors in (13.15) and (13.17) can be extended from individual parameters to blocks of parameters.

Marginal Likelihood

The marginal probability or **marginal likelihood** is the denominator in Bayes' rule and is defined as

$$f(\mathbf{y}) = \int L(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}. \quad (13.18)$$

It is the expected value of the likelihood, $E[L(\mathbf{y}|\boldsymbol{\theta})]$, where the expectation is with respect to the prior density. The marginal likelihood constitutes a basis for Bayesian inference (see Section 13.8), as it contains information about the support in the data for the prior.

Posterior Predictive Density

Consider out-of-sample prediction of a single observation y^p . This has density $f(y^p|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is unknown. The **posterior predictive density** of y^p weights this density by the posterior probability distribution of $\boldsymbol{\theta}$, yielding

$$f^P(y^p) = \int f(y^p|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}. \quad (13.19)$$

If covariates appear in the likelihood function as in a regression model, then these densities are conditioned on them also.

13.2.6. Large-Sample Behavior of the Posterior

The influence of even informative priors on the posterior diminishes as the sample becomes large, as illustrated in the Section 13.2.2 example. This is the basis of the statement that asymptotically the likelihood dominates the inference or that the weight assigned to the prior essentially goes to zero as the sample size grows.

Because the posterior distribution can be awkward to manipulate, an asymptotic approximation to the posterior is of interest as it can be used in place of the true finite-sample posterior distribution. This approximation is easy to obtain since asymptotically the posterior equals the likelihood. We follow Gelman et al. (1995), to which the reader is referred for additional detail.

For simplicity assume that observations are iid. Then the log-posterior

$$\sum_{i=1}^N \ln p(\boldsymbol{\theta}|y_i) = \ln \pi(\boldsymbol{\theta}) + \sum_{i=1}^N \ln f(y_i|\boldsymbol{\theta}). \quad (13.20)$$

This representation makes it clear that in a large sample the posterior is dominated by the likelihood contribution, since the contribution of the prior to the posterior remains fixed whereas the contribution of the sample to the posterior grows with N .

Assume that the posterior $p(\theta|y)$ is unimodal and approximately symmetric. We consider the asymptotic properties of the posterior mode, denoted by $\widehat{\theta}$, which is then the local and global maximum of the posterior.

To establish consistency of $\widehat{\theta}$, we note that the posterior mode converges to the MLE as $N \rightarrow \infty$, since the second term in (13.20) dominates. The posterior mode is therefore consistent if the MLE is consistent. So $\widehat{\theta} \xrightarrow{P} \theta_0$ if the dgp for y has density $f(y|\theta_0)$ and the usual regularity conditions for ML estimation are satisfied.

To obtain the asymptotic distribution of $\widehat{\theta}$, consider a second-order Taylor series expansion of the log posterior density around the posterior mode $\widehat{\theta}$. Then

$$\ln p(\theta|y) \simeq \ln p(\widehat{\theta}|y) + \frac{1}{2}(\theta - \widehat{\theta})' \left[\frac{\partial^2 \ln p(\theta|y)}{\partial \theta \partial \theta'} \Big|_{\theta=\widehat{\theta}} \right] (\theta - \widehat{\theta}), \quad (13.21)$$

where simplification occurs because $\partial \ln p(\theta|y)/\partial \theta = \mathbf{0}$ when evaluated at the posterior mode, and we assume that third- and higher order derivatives of θ can be ignored asymptotically. Define

$$\mathcal{I}(\widehat{\theta}) = - \left. \frac{\partial^2 \ln p(\theta|y)}{\partial \theta \partial \theta'} \right|_{\theta=\widehat{\theta}}$$

to be the observed information based on the posterior density $\ln p(\theta|y)$, evaluated at the posterior mode. Then exponentiating (13.21) yields

$$p(\theta|y) \propto \exp \left(-\frac{1}{2}(\theta - \widehat{\theta})' \mathcal{I}(\widehat{\theta})(\theta - \widehat{\theta}) \right),$$

which is the kernel of multivariate normal distribution with mean $\widehat{\theta}$ and variance matrix $\mathcal{I}(\widehat{\theta})^{-1}$. It follows that a posteriori

$$\theta|y \xrightarrow{a} \mathcal{N}[\widehat{\theta}, \mathcal{I}(\widehat{\theta})^{-1}]. \quad (13.22)$$

As the sample size N grows large, the likelihood component of the posterior becomes dominant and the influence of the prior becomes negligible. In this case we may replace the mode $\widehat{\theta}$ by the MLE, which is the mode of the likelihood density. This yields a result that is sometimes called a **Bayesian central limit theorem** (Gamerman, 1997). Asymptotically, frequentist and Bayesian inferences will be based on the same limiting multivariate normal distribution, and hence there should be no significant inconsistency between them.

This result has been labeled as the **Bernstein–von Mises Theorem** in the literature; see Train (2003, chapter 12) for an accessible discussion of the three components of this theorem. These components comprise (1) the result that the posterior mean converges in probability to the maximum likelihood estimator, (2) that it has a limiting normal distribution, and (3) that the limiting distribution of the posterior mean is the same as that of the maximum likelihood estimator. These results are all implicit in the Bayesian central limit theorem. That theorem is of great interest and relevance to those who wish to apply the likelihood principles of estimation and inference. The full force of its implications will become apparent after we examine numerical methods for approximating the posterior distribution.

Do the preceding arguments imply that Bayesian and likelihood-based methods will produce essentially similar results? Is the choice between the two approaches largely a matter of computational efficiency? A definitive treatment of these issues is not available. However, there are a number of examples in the literature that show not only that the two approaches may produce similar results, but also that Bayesian methods are frequently computationally more efficient.

13.2.7. Bayesian Decision Analysis

Given the full posterior distribution $p(\theta|y)$, which point estimate of θ should be reported? This question was studied in Section 4.2 for best prediction of y using, for example, squared error loss. Here instead we consider best estimation of θ using, for example, quadratic loss.

Let $L(\theta, \hat{\theta})$ denote the specified loss function, where $\hat{\theta}$ is an estimate of the unknown θ . The loss is unknown, as it depends on θ , which is unknown. We can, however, find the expected value over θ of the loss since Bayesian analysis, unlike classical analysis, provides the distribution of θ . The **optimal estimator** $\hat{\theta}_{\text{OPT}}$ is the estimator $\hat{\theta}$ that **minimizes expected posterior loss**, or

$$\min_{\hat{\theta}} E[L(\theta, \hat{\theta})] = \min_{\hat{\theta}} \int L(\theta, \hat{\theta}) p(\theta|y) d\theta, \quad (13.23)$$

Losses associated with different $(\theta, \hat{\theta})$ are weighted by the posterior probability $p(\theta|y)$.

It can be shown that the posterior mean is the optimal estimator under quadratic loss, $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})'(\theta - \hat{\theta})$. If instead absolute error loss is used, with $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$, then the posterior median is the optimal estimator. Once the posterior distribution has been established these point estimates can be computed either analytically or numerically.

Under some conditions minimizing expected posterior loss can be shown to be equivalent to minimizing **expected posterior risk**. The risk function averages the possible loss over hypothetical samples of y from the population, so

$$R(\theta, \hat{\theta}) = \int L(\theta, \hat{\theta}) f(y|\theta) dy.$$

To avoid the possible confusion between loss function and likelihood function, here and in the next equation block, we have used $f(y|\theta)$ as equivalent to the likelihood $L(y|\theta)$. Expected posterior risk averages this risk over different values of the parameters $\theta \in \Theta$ by weighting with respect to the posterior density, so

$$\begin{aligned} E[R(\theta, \hat{\theta})] &= \int_{\Theta} \left\{ \int L(\theta, \hat{\theta}) f(y|\theta) dy \right\} p(\theta|y) d\theta \\ &= \int \left\{ \int_{\Theta} L(\theta, \hat{\theta}) p(\theta|y) d\theta \right\} f(y|\theta) dy \\ &= \int E[L(\theta, \hat{\theta})] f(y|\theta) dy, \end{aligned} \quad (13.24)$$

where in the first equality the outer integral ranges over the domain of θ , in the second equality the order of integration is interchanged, and in the third line the conclusion follows. These operations presume that appropriate restrictions on $L(\theta, \hat{\theta})$ and $p(\theta|y)$ are satisfied. For example, $p(\theta|y)$ must be a proper density function and the loss function must be integrable. Hence expected risk will remain bounded and minimizing it is a well-defined operation.

The foregoing argument establishes a well-known and important result that the Bayes estimator is **admissible** in the sense that it minimizes expected risk for a specified loss function.

13.3. Bayesian Analysis of Linear Regression

Because the analysis of linear regression is a familiar topic, it provides a useful portal to more general nonlinear models. The data are assumed to be generated by the standard linear regression model

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

where \mathbf{X} denotes the $N \times K$ full column rank matrix of weakly exogenous regressors. The errors are assumed to be independent, homoskedastic, and normally distributed, with $\mathbf{u} \sim \mathcal{N}[\mathbf{0}, \sigma^2 \mathbf{I}_N]$. The sample conditional density is therefore $y|\mathbf{X}, \beta, \sigma^2 \sim \mathcal{N}[\mathbf{X}\beta, \sigma^2 \mathbf{I}_N]$. Our exposition follows Zellner (1971).

We deal in turn with noninformative and informative priors. In both cases a closed-form expression for the posterior can be obtained after some considerable algebra. For noninformative prior it will be seen that the OLS estimator has a Bayesian interpretation as the mean of the posterior distribution. In the informative prior case it will be seen that the posterior moments are weighted functions of the sample and prior means.

Subsequent sections present methods for less tractable models, but even then analysis is simplified if results similar to those given in this section can be applied to some subcomponents of the model.

13.3.1. Noninformative Priors

For noninformative priors we use Jeffreys' priors. From Section 13.2.4, for $y \sim \mathcal{N}[\mu, \sigma^2]$ this prior for μ (given σ^2 known) is a constant, whereas the prior for σ^2 (given μ known) is proportional to σ^2 . For the regression case this extends to constant prior for β_j , $j = 1, \dots, K$, so $\pi(\beta_j) \propto c$, and the prior for σ^2 is $\pi(\sigma^2) \propto 1/\sigma^2$. The prior views all values of β_j as equally likely, whereas smaller values of σ^2 are viewed as being more likely. Assuming independence of β and σ^2 the joint prior is

$$\pi(\beta, \sigma^2) \propto 1/\sigma^2.$$

The likelihood function can be reexpressed as

$$\begin{aligned} L(\beta, \sigma^2 | \mathbf{y}, \mathbf{X}) &= (2\pi\sigma^2)^{-N/2} \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\beta)' (\mathbf{y} - \mathbf{X}\beta) \right\} \\ &\propto (\sigma^2)^{-N/2} \exp \left(-\frac{1}{2\sigma^2} \{ \widehat{\mathbf{u}}' \widehat{\mathbf{u}} + (\beta - \widehat{\beta})' \mathbf{X}' \mathbf{X} (\beta - \widehat{\beta}) \} \right) \\ &\propto (\sigma^2)^{-N/2} \exp \left(-\frac{1}{2\sigma^2} (N - K) s^2 + (\beta - \widehat{\beta})' \mathbf{X}' \mathbf{X} (\beta - \widehat{\beta}) \right), \end{aligned} \quad (13.25)$$

where $\widehat{\beta} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y}$ and $\widehat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widehat{\beta}$; the second line uses $\mathbf{y} - \mathbf{X}\beta = \widehat{\mathbf{u}} - \mathbf{X}(\beta - \widehat{\beta})$ and $\mathbf{X}' \widehat{\mathbf{u}} = \mathbf{0}$; and the third line uses $s^2 = \widehat{\mathbf{u}}' \widehat{\mathbf{u}} / (N - K)$.

Combining the likelihood in (13.25) and the prior, we obtain the posterior density

$$\begin{aligned} p(\beta, \sigma^2 | \mathbf{y}, \mathbf{X}) & \propto \left(\frac{1}{\sigma^2} \right)^{N/2} \exp \left(-\frac{1}{2\sigma^2} \{ (N - K) s^2 + (\beta - \widehat{\beta})' \mathbf{X}' \mathbf{X} (\beta - \widehat{\beta}) \} \right) \frac{1}{\sigma^2} \\ & \propto \left(\frac{1}{\sigma^2} \right)^{N/2+1} \exp \left(-\frac{1}{2\sigma^2} \{ (N - K) s^2 + (\beta - \widehat{\beta})' \mathbf{X}' \mathbf{X} (\beta - \widehat{\beta}) \} \right) \\ & \propto \left\{ \left(\frac{1}{\sigma^2} \right)^{K/2} \exp \left(-\frac{1}{2} (\beta - \widehat{\beta})' (\sigma^2 (\mathbf{X}' \mathbf{X})^{-1})^{-1} (\beta - \widehat{\beta}) \right) \right\} \\ & \quad \times \left\{ \left(\frac{1}{\sigma^2} \right)^{(N-K)/2+1} \exp \left(-\frac{(N - K) s^2}{2\sigma^2} \right) \right\}. \end{aligned} \quad (13.26)$$

The conditional posterior distribution $p(\beta | \sigma^2, \mathbf{y}, \mathbf{X})$ of β , given σ^2 , and the data \mathbf{y}, \mathbf{X} , is clearly the K -dimensional multivariate normal with mean $\widehat{\beta}$ and variance $\sigma^2 (\mathbf{X}' \mathbf{X})^{-1}$, since β appears only in the first line of the final expression. The conditional posterior of σ^2 given β is more difficult to obtain as σ^2 appears in both lines.

The marginal posterior of β , obtained by integrating out σ^2 , is much more useful for posterior inference about β . We integrate the second line of (13.26), change variables to $z = 1/\sigma^2$ and use the result that $\int_0^\infty z^c \exp(-az) dz = \Gamma(c+1)/a^{c+1}$ for given constants $a > 0$, $c > -1$, where here $c = N/2 + 1$ and $a = \{\cdot\}$ is the lengthy term in braces. This yields the kernel of the marginal posterior distribution

$$\begin{aligned} p(\beta | \mathbf{y}, \mathbf{X}) &\propto \{ (N - K) s^2 + (\beta - \widehat{\beta})' \mathbf{X}' \mathbf{X} (\beta - \widehat{\beta}) \}^{-N/2} \\ &\propto \left\{ 1 + (\beta - \widehat{\beta})' (s^2 (N - K) (\mathbf{X}' \mathbf{X})^{-1})^{-1} (\beta - \widehat{\beta}) \right\}^{-(N-K+K)/2}, \end{aligned} \quad (13.27)$$

which from Section 13.3.5 is the kernel of a multivariate Student t -distribution centered at $\widehat{\beta}$ with $N - K$ degrees of freedom and covariance matrix $s^2 (\mathbf{X}' \mathbf{X})^{-1}$ multiplied by $(N - K) / (N - K - 2)$. Thus

$$\beta \sim t_K(\widehat{\beta}, s^2 (\mathbf{X}' \mathbf{X})^{-1}). \quad (13.28)$$

An individual element of β has a univariate Student t -distribution.

The marginal posterior for σ^2 is more easily obtained, by integrating the final expression in (13.26) with respect to β and noting that β appears in only the first line,

which is the kernel of the $\mathcal{N}[\hat{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]$ density and integrates to one. It follows that the marginal posterior for σ^2 is

$$p(\sigma^2|\mathbf{y}, \mathbf{X}) \propto (\sigma^2)^{-(N-K+1)/2} \exp\left(-\frac{(N-K)s^2}{2\sigma^2}\right). \quad (13.29)$$

This expression is known to be the kernel of an inverted square-root gamma density. That is, it is the density of a random variable that is the reciprocal of the square-root of a gamma-distributed random variable with degrees-of-freedom parameter $N - K$. This result is identical to that obtained under the frequentist analysis of the distribution of $\hat{\beta}$.

For normal linear regression, Bayesian analysis with noninformative priors therefore yields qualitatively similar conclusions to those from the standard frequentist analysis in finite samples. Conditional on σ^2 the posterior of β is the $\mathcal{N}[\hat{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]$ distribution, and unconditionally the posterior of β is the multivariate t -distribution.

The interpretation is quite different, however, as these distributions are of the unknown parameter β with mean $\hat{\beta}$, rather than of an estimate $\hat{\beta}$ with unknown mean β . For example, the Bayesian 95% HPD interval for β_j is $\hat{\beta}_j \pm t_{0.025, N-K} \times \text{se}[\hat{\beta}_j]$, where $\text{se}[\hat{\beta}_j] = (s^2(\mathbf{X}'\mathbf{X})^{jj})^{1/2}$. From Section 13.2.5 the interpretation is that β_j lies in this interval with posterior probability 0.95.

13.3.2. Informative Priors

Bayesian analysis of the normal linear regression model under informative priors is especially insightful if we use independent conjugate priors for β and σ . From Section 13.2.4, the conjugate prior for β is the normal, and the conjugate prior for $1/\sigma^2$ is the gamma. This leads to the **normal–gamma prior**

$$\pi(\beta, 1/\sigma^2) = \pi_N(\beta|1/\sigma^2)\pi_\gamma(1/\sigma^2),$$

where $\pi_N(\beta|1/\sigma^2)$ is the $\mathcal{N}[\beta_0, \sigma^2\Omega_0^{-1}]$ density, with β_0 and Ω_0 known, and the kernel is

$$\pi_N(\beta|1/\sigma^2) \propto \sigma^{-K} \exp\left[-\frac{(\beta - \beta_0)' \Omega_0 (\beta - \beta_0)}{2\sigma^2}\right], \quad (13.30)$$

and $\pi_\gamma(1/\sigma^2)$ is the $\mathcal{G}[v_0, s_0^2]$ density where v_0 and s_0^2 are known constants, and

$$\pi_\gamma(1/\sigma^2) = \sigma^{-(v_0+1)} \exp\left[-\frac{v_0 s_0^2}{2\sigma^2}\right]. \quad (13.31)$$

Note that the prior for the (location) parameter β depends on the (scale) parameter σ . This makes sense as σ reflects the scale on which y is measured and hence should affect β . Given this prior and the likelihood in (13.25), the posterior density is of a

normal–gamma type. After some algebra it is as follows:

$$\begin{aligned}
 p(\beta, 1/\sigma^2 | \mathbf{y}, \mathbf{X}) &\propto (\sigma^2)^{-N/2} \exp\left[-\frac{s^2(N-K)}{2\sigma^2}\right] \exp\left[-\frac{(\beta-\widehat{\beta})' \mathbf{X}' \mathbf{X} (\beta-\widehat{\beta})}{2\sigma^2}\right] \\
 &\quad \times (\sigma^2)^{-K/2} \exp\left[-\frac{(\beta-\beta_0)' \Omega_0 (\beta-\beta_0)}{2\sigma^2}\right] \\
 &\quad \times (\sigma^2)^{-(v_0/2)-1} \exp\left[-\frac{v_0 s_0^2}{2\sigma^2}\right] \\
 &\propto (\sigma^2)^{(v_0+N)/2-1} \exp\left[-\frac{s_1^2}{2\sigma^2}\right] (\sigma^2)^{-K/2} \\
 &\quad \times \exp\left[-\frac{1}{2\sigma^2} (\beta-\overline{\beta})' \Omega_1 (\beta-\overline{\beta})\right], \tag{13.32}
 \end{aligned}$$

where $\overline{\beta}$ and Ω_1^{-1} denote the posterior mean and variance of β and s_1^2 denotes the posterior mean of σ^2 defined as

$$\begin{aligned}
 \overline{\beta} &= (\Omega_0 + \mathbf{X}' \mathbf{X})^{-1} (\Omega_0 \beta_0 + \mathbf{X}' \mathbf{X} \widehat{\beta}), \\
 \Omega_1 &= (\Omega_0 + \mathbf{X}' \mathbf{X}), \\
 s_1^2 &= s_0^2 + \widehat{\mathbf{u}}' \widehat{\mathbf{u}} + (\beta - \overline{\beta})' [\Omega_0^{-1} + (\mathbf{X}' \mathbf{X})^{-1}] (\beta - \overline{\beta}). \tag{13.33}
 \end{aligned}$$

The posterior mean $\overline{\beta}$ is obtained by using the matrix version of the “completing the square” operation. Specifically, given the $K \times 1$ vectors β , $\overline{\beta}$, β_0 , and $\widehat{\beta}$, and $K \times K$ symmetric square matrices \mathbf{A} and \mathbf{B} , it can be shown that

$$\begin{aligned}
 &(\beta - \beta_0)' \mathbf{A} (\beta - \beta_0) + (\beta - \widehat{\beta})' \mathbf{B} (\beta - \widehat{\beta}) \\
 &= (\beta - \overline{\beta})' (\mathbf{A} + \mathbf{B}) (\beta - \overline{\beta}) + (\beta_0 - \overline{\beta})' \mathbf{AB} (\mathbf{A} + \mathbf{B})^{-1} (\beta_0 - \overline{\beta}),
 \end{aligned}$$

where $\overline{\beta} = (\mathbf{A} + \mathbf{B})^{-1} (\mathbf{A} \beta_0 + \mathbf{B} \widehat{\beta})$.

The joint marginal posterior of β and σ^2 is of the same normal–gamma form as the prior.

The conditional posterior of β given σ^2 has mean $\overline{\beta}$, a matrix-weighted average of the prior mean β_0 and the sample mean $\widehat{\beta}$.

In general using a conjugate prior is algebraically equivalent to augmenting the data with a sample from the same distribution. In this case the normal–gamma prior is equivalent to an additional sample of the same process with regression parameter estimate of β_0 , $\mathbf{X}' \mathbf{X}$ matrix equal to Ω_0 , degrees-of-freedom parameter equal to v_0 , and error sum of squares equal to $v_0 s_0^2$. Since Ω_0 is a fixed matrix, $\Omega_0/N \rightarrow \mathbf{0}$ as $N \rightarrow \infty$, whereas $\mathbf{X}' \mathbf{X}/N$ converges to a matrix of constants. Hence $\overline{\beta} \rightarrow \widehat{\beta}$, verifying that in large samples the ML estimator and the posterior mean are equivalent. The posterior variance Ω_1^{-1} is proportional to $(\Omega_0 + \mathbf{X}' \mathbf{X})^{-1}$. See Leamer (1978) for a more detailed exposition.

The marginal posterior of β is obtained by integrating σ^2 out of the joint posterior. This yields

$$p(\beta|\mathbf{y}, \mathbf{X}) \propto \left[s_1^2 + (\beta - \bar{\beta})' (\Omega_0 + \mathbf{X}'\mathbf{X}) (\beta - \bar{\beta}) \right]^{-(v_1+K/2)}; \quad (13.34)$$

hence a marginal posterior is a multivariate Student t -distribution, one that is centered around $\bar{\beta}$ rather than around $\hat{\beta}$ as in the case of uninformative prior.

Because the conjugate prior treats the prior information like a previous sample from the same process, the sample and prior information are handled symmetrically even though the information from the two sources may be in conflict. Thus the mathematical convenience of using conjugate priors comes at a price. If the prior information and the sample information are apparently in conflict, the posterior distribution can be expected to be bimodal with the modes corresponding to sample and prior means. A prior distribution that allows one to capture such a feature is a prior that specifies that β has a multivariate Student t -density independent of $1/\sigma^2$ and $1/\sigma^2$ has a gamma prior distribution independent of $\mathbf{X}\beta$. This has been called “**Dickey’s prior**” (Leamer, 1978, p. 79). Under this assumption the marginal posterior is a product of two multivariate Student t -densities; this product can also be expressed as a mixture of two t -distributions. Such a distribution can potentially exhibit bimodality. Leamer (1978) has provided a more extensive analysis of this case.

13.3.3. Mixed Estimation

We seek to place Bayesian analysis of linear regression in a frequentist setting.

Frequentist analysis usually incorporates prior information as equality constraints, which is a limiting case of Bayesian analysis where the variance parameters in the prior go to zero. Prior information that is instead stochastic can also be incorporated into frequentist analysis, by using **mixed estimation**. The algebra is simple, and the approach also provides an intuitive understanding of how Bayesian procedures pool prior and sample information.

We continue with the linear regression model under normality. Assume prior information for the regression parameters that $\beta \sim \mathcal{N}[\mathbf{0}, \sigma_v^2 \mathbf{I}_K]$, where extension to nonzero mean is relatively easy. The prior information can be written as

$$\beta = \mathbf{0} + \mathbf{v},$$

where \mathbf{v} is a $K \times 1$ error with $\mathbf{v} \sim \mathcal{N}[\mathbf{0}, \sigma_v^2 \mathbf{I}_K]$. Now augment the sample information $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$ by this prior, and write the full model as an **augmented regression model**

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{X} \\ \mathbf{I}_K \end{bmatrix} \beta + \begin{bmatrix} \mathbf{u} \\ -\mathbf{v} \end{bmatrix}.$$

This can be reparameterized as

$$\begin{aligned} \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} &= \begin{bmatrix} \mathbf{X} \\ \frac{\sigma}{\sigma_v} \mathbf{I}_K \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{u} \\ -\frac{\sigma}{\sigma_v} \mathbf{v} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{X} \\ \lambda \mathbf{I}_K \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{u} \\ \mathbf{v}^* \end{bmatrix}, \end{aligned} \quad (13.35)$$

where $\lambda = \sigma/\sigma_v$ and the transformation $\mathbf{v}^* = -\lambda \mathbf{v}$ has been used so that all errors have common variance σ^2 .

The estimator based on this augmented data set is a **pooled estimator** or a **mixed estimator**. Conditional on λ , the mixed estimator is

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_\lambda &= [\mathbf{X}'\mathbf{X} + \lambda^2 \mathbf{I}_K]^{-1} \mathbf{X}'\mathbf{y} \\ &= [\mathbf{X}'\mathbf{X}(\mathbf{I}_K + \lambda^2 (\mathbf{X}'\mathbf{X})^{-1})]^{-1} \mathbf{X}'\mathbf{y} \\ &= [\mathbf{I}_K + \lambda^2 (\mathbf{X}'\mathbf{X})^{-1}]^{-1} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} \\ &= \mathbf{A}_\lambda \widehat{\boldsymbol{\beta}}, \end{aligned} \quad (13.36)$$

where $\mathbf{A}_\lambda = [\mathbf{I}_K + \lambda^2 (\mathbf{X}'\mathbf{X})^{-1}]^{-1}$, and $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$ is the unrestricted OLS estimator.

This estimator is the so-called ridge-regression estimator introduced without a Bayesian justification by Hoerl and Kennard (1970) to combat the problem of multicollinearity in small samples. This estimator also belongs to a class of **shrinkage estimators**, in which the estimator is shrunk toward (or pulled toward) a prior mean, in this case the zero vector. This sometimes makes some sense in a finite sample with highly multicollinear data where the “*t*-ratios” tend to zero, making it difficult to distinguish between variables whose coefficients are truly close to zero and those that only appear to be that way. In the limit shrinkage leads to variable exclusion.

Several features of $\widehat{\boldsymbol{\beta}}_\lambda$ are noteworthy: (1) Conditional on λ , $\widehat{\boldsymbol{\beta}}_\lambda$ is the mean of a posterior distribution of $\boldsymbol{\beta}$. (2) The estimator is a **matrix-weighted average of $\mathbf{0}$** vector and $\widehat{\boldsymbol{\beta}}$. (3) The algebra changes very little if we chose to shrink the estimator toward some nonzero $\boldsymbol{\beta}$, say $\boldsymbol{\beta}_0$. Then the resulting estimator is a **matrix-weighted average of vectors $\boldsymbol{\beta}_0$ and $\widehat{\boldsymbol{\beta}}$** .

The symmetric weighting matrix $\mathbf{A}_\lambda = [\mathbf{I}_K + (\lambda^2/N) (N^{-1}\mathbf{X}'\mathbf{X})^{-1}] \rightarrow \mathbf{I}_K$ as $N \rightarrow \infty$, since $\lambda^2/N \rightarrow 0$. Therefore,

$$\widehat{\boldsymbol{\beta}}_\lambda \rightarrow \widehat{\boldsymbol{\beta}} \text{ as } N \rightarrow \infty,$$

so the effect of the prior on the posterior mean vanishes as the sample becomes large. Similarly, the conditional posterior variance of $\widehat{\boldsymbol{\beta}}_\lambda$ is given by

$$\begin{aligned} V[\widehat{\boldsymbol{\beta}}_\lambda] &= \mathbf{A}_\lambda V[\widehat{\boldsymbol{\beta}}] \mathbf{A}_\lambda \\ &= \sigma^2 \mathbf{A}_\lambda (\mathbf{X}'\mathbf{X})^{-1} \mathbf{A}_\lambda, \end{aligned}$$

so $V[\widehat{\boldsymbol{\beta}}_\lambda] \rightarrow \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$ as the sample size $N \rightarrow \infty$.

For finite samples, conditional on λ and σ^2 , the conditional **posterior distribution** of $\widehat{\beta}_\lambda$ is

$$\widehat{\beta}_\lambda | \lambda, \sigma^2 \sim \mathcal{N}[\mathbf{A}_\lambda \widehat{\beta}, \sigma^2 \mathbf{A}_\lambda (\mathbf{X}' \mathbf{X})^{-1} \mathbf{A}'_\lambda]. \quad (13.37)$$

The marginal posterior distribution of $\widehat{\beta}_\lambda$ is obtained by integrating out λ and σ^2 . Treating λ as given, and assuming a vague or uninformative prior on σ^2 , we can integrate out σ^2 as was shown in Section 13.3.1. This integration operation is analytically feasible and yields a marginal posterior of β_λ that is the multivariate Student t -distribution. Finally, we can specify a prior distribution on λ , possibly a gamma prior since $\lambda > 0$, and proceed to integrate it out. However, λ enters the conditional posterior in an awkward fashion and cannot be integrated out analytically. At this stage we would need to resort to a numerical technique. Assuming that this is accomplished then we have a Bayesian treatment of this model.

13.3.4. Hierarchical Priors

We consider a three-stage linear regression model that is hierarchical in regression parameters but not in variance parameters.

The first stage is a linear regression model denoted $\mathbf{y} = \mathbf{X}_1 \beta_1 + \mathbf{u}$, where the subscript 1 is added to distinguish between first- and second-stage parameters and regressors. The parameters β_1 are random and are modeled to depend on both parameters and data, so $\beta_1 = \mathbf{X}_2 \beta_2 + \mathbf{v}$. For example, the first level models individual student test performance and the second level brings in school characteristics. The errors are assumed to be normally distributed. The second-level parameters β_2 are treated as unknown and a prior is specified. A prior is also specified for the variance parameter σ_1^2 in the first-stage model.

Assuming normally distributed errors and using conjugate priors leads to the following model:

$$\mathbf{y} | \mathbf{X}_1, \beta_1, \sigma_1^2 \sim \mathcal{N}[\mathbf{X}_1 \beta_1, \sigma_1^2 \mathbf{I}_N], \quad (13.38)$$

$$\beta_1 | \mathbf{X}_2, \beta_2, \Sigma_2 \sim \mathcal{N}[\mathbf{X}_2 \beta_2, \Sigma_2], \quad (13.39)$$

$$\beta_2 \sim \mathcal{N}[\beta^*, \Sigma^*], \quad (13.40)$$

$$\sigma_1^{-2} | v^*, \sigma^{*2} \sim \mathcal{G}[v^*/2, v^* \sigma^{*2}/2], \quad (13.41)$$

where \mathbf{X}_1 is $N \times K$, \mathbf{X}_2 is $K \times M$, β_1 is $K \times 1$, β_2 is $M \times 1$, Σ_2 is $K \times K$, β^* is $M \times 1$, and Σ^* is $M \times M$. For the regression parameter β_1 the second line gives the prior, and the third line gives the subsequent second-stage prior, or a prior on a prior, for β_2 (while Σ_2 is assumed known). The parameters (β^*, Σ^*) are often referred to as hyperparameters. For variance parameters, the fourth line gives a prior for the variance parameter σ_1^2 with v^* and σ^{*2} specified. The innovation is the addition of (13.40).

Note that we can collapse the stages and convert this into a two-level model. Specifically, we can write a two-stage model with an informative prior in one of two

ways, either

$$\begin{aligned} \mathbf{y} | \mathbf{X}_1, \boldsymbol{\beta}_1, \sigma_1^2 &\sim \mathcal{N}[\mathbf{X}_1 \boldsymbol{\beta}_1, \sigma_1^2 \mathbf{I}_N], \\ \boldsymbol{\beta}_1 | \mathbf{X}_2, \Sigma_2 &\sim \mathcal{N}[\mathbf{X}_2 \boldsymbol{\beta}^*, \Sigma_2 + \mathbf{X}_2 \boldsymbol{\Sigma}^* \mathbf{X}_2'] \end{aligned}$$

or

$$\begin{aligned} \mathbf{y} | \mathbf{X}_1, \mathbf{X}_2, \boldsymbol{\beta}_2, \Sigma_2, \sigma_1^2 &\sim \mathcal{N}[\mathbf{X}_1 \mathbf{X}_2 \boldsymbol{\beta}_2, \sigma_1^2 \mathbf{I}_N + \mathbf{X}_1 \Sigma_2 \mathbf{X}_2'], \\ \boldsymbol{\beta}_2 &\sim \mathcal{N}[\boldsymbol{\beta}^*, \boldsymbol{\Sigma}^*]. \end{aligned}$$

If σ_1^2 were, given this setup corresponds to **conditionally conjugate** normal priors. Using results introduced earlier we can derive expressions for the posterior means of either $\boldsymbol{\beta}_1$ or $\boldsymbol{\beta}_2$ as matrix-weighted averages of either $\boldsymbol{\beta}^*$ and $\widehat{\boldsymbol{\beta}}_1$ or of $\boldsymbol{\beta}^*$ and $\widehat{\boldsymbol{\beta}}_2$.

The use of the normal distribution is only illustrative. Hierarchical models for generalized linear models, members of the linear exponential family, have been widely used (Albert, 1988).

In hierarchical models it may not be possible to obtain the full posterior probability distribution of first-stage parameters such as $\boldsymbol{\beta}_1$ in an analytically tractable form. Fortunately, the advances in computational methods presented in the next section are especially well suited to models with a hierarchical structure.

Another approach, which is an application of the **empirical Bayes** method, involves estimation of parameters in the higher stage priors, similar to that in the likelihood approach. This approach avoids, for example, assuming that Σ_2 and $\boldsymbol{\Sigma}^*$ are known matrices.

13.3.5. Multivariate t - and Wishart Distributions

Bayesian analysis makes use of a wider range of distributions than classical analysis. Here we present details on two multivariate distributions that are used in Bayesian analysis of linear regression under normality.

The multivariate t -distribution is a multivariate extension of the univariate student t . It is similar to the multivariate normal, except that the tails of the distribution can be considerably fatter. In Bayesian analysis it arises as the marginal posterior for $\boldsymbol{\beta}$ given a conjugate normal prior (see Section 13.3.2) or can be used directly as the prior for $\boldsymbol{\beta}$ if tails fatter than the normal are desired. A $q \times 1$ random variable \mathbf{t} that is multivariate Student- t distributed with degrees-of-freedom parameter v , mean parameters μ , and dispersion parameters Σ , has joint density

$$\begin{aligned} f_t(\mathbf{t} | v, \mu, \Sigma) &= \frac{\Gamma((v+1)/2)}{\Gamma(v/2)(\pi v)^{(1/2)} |\Sigma|^{1/2}} \\ &\quad \times \left\{ 1 + \frac{1}{v} (\mathbf{t} - \mu)' \Sigma^{-1} (\mathbf{t} - \mu) \right\}^{-(v+q)/2}, \end{aligned}$$

where $\Gamma(\cdot)$ is the gamma function. This distribution is symmetric with mode μ , mean μ if $v > 1$, and variance $[v/(v-2)]\Sigma$ if $v > 2$. The tails can be much fatter than the normal (e.g., the variance is 3Σ if $v = 3$) and the normal is obtained as $v \rightarrow \infty$. If

$\mathbf{z} \sim \mathcal{N}[\mathbf{0}, \mathbf{I}]$ and $s \sim \chi^2(v)$ then $\mathbf{t} = \boldsymbol{\mu} + \Sigma^{-1/2}\mathbf{z}/\sqrt{s/v}$ has the multivariate t -distribution given here, providing an easy way to obtain draws.

The Wishart distribution is a multivariate extension of the univariate chi-square distribution, or more generally the gamma distribution. In Bayesian analysis it is used as the conjugate prior for the inverse of the covariance matrix of a multivariate normal distribution. A $q \times q$ random positive definite matrix \mathbf{W} that is **Wishart distributed** with degrees of freedom parameter $v \geq q$ and scale matrix \mathbf{S} has joint density

$$f_{\mathbf{W}}(\mathbf{W}|v, \mathbf{S}) = 2^{vq/2} \pi^{q(q-1)/4} \prod_{j=1}^q \Gamma\left(\frac{v+1-j}{2}\right) \times |\mathbf{S}|^{-v/2} |\mathbf{W}|^{(v-q-1)/2} \exp(-\text{tr}(\mathbf{S}^{-1}\mathbf{W})/2),$$

where $\Gamma(\cdot)$ is the gamma function and $\text{tr}(\cdot)$ denotes the trace of a matrix. This distribution has mean $v\mathbf{S}$. The sample covariance matrix for iid multivariate normal data is Wishart distributed. More generally, given $v(q)$, independent $q \times 1$ vectors $\mathbf{x}_j \sim \mathcal{N}[\mathbf{0}, \mathbf{S}]$, $j = 1, \dots, v$, then $\sum_{j=1}^v \mathbf{x}_j \mathbf{x}_j'$ is Wishart distributed. If \mathbf{W}^{-1} is Wishart distributed with density $f_{\mathbf{W}}(\mathbf{W}^{-1}|v, \mathbf{S})$ then \mathbf{W} is **inverse-Wishart distributed** with density

$$f_{\mathbf{W}}(\mathbf{W}|v, \mathbf{S}) = 2^{vq/2} \pi^{q(q-1)/4} \prod_{j=1}^q \Gamma\left(\frac{v+1-j}{2}\right) |\mathbf{S}|^{v/2} |\mathbf{W}|^{-(v+q+1)/2} \exp(-\text{tr}(\mathbf{S}^{-1}\mathbf{W})/2).$$

13.4. Monte Carlo Integration

In many modeling situations the posterior distribution of the parameters of interest is analytically intractable. In such cases numerical methods are needed to estimate either the full posterior distribution or some key moments of this distribution such as the posterior mean.

In this section we consider computation of key posterior moments, without explicitly obtaining the posterior distribution. The methods of Chapter 12 can be applied, with potentially less computational burden since the integral needs to be computed once for the entire sample rather than for every individual at every iteration. In the subsequent section we present methods to simulate the posterior distribution.

13.4.1. Importance Sampling

Suppose the problem is to evaluate the posterior moment function $E[m(\boldsymbol{\theta}|\mathbf{y})]$, where expectation is with respect to the posterior density $p(\boldsymbol{\theta}|\mathbf{y})$. We wish to compute

$$E[m(\boldsymbol{\theta})] = \int_{R(\boldsymbol{\theta})} m(\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}. \quad (13.42)$$

For example, the posterior mean of the k th parameter is $E[\theta_k] = \int \theta_k p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}$. Other examples include posterior standard deviations, marginal posterior densities, posterior intervals, and posterior expectations of a given function of parameters.

From Chapter 12 a direct Monte Carlo integral estimate of $E[m(\boldsymbol{\theta})]$ is $\widehat{E}[m(\boldsymbol{\theta})] = S^{-1} \sum_s m(\boldsymbol{\theta}^s)$, where $\boldsymbol{\theta}^s$, $s = 1, \dots, S$, are S draws of $\boldsymbol{\theta}$ from the posterior density $p(\boldsymbol{\theta}|\mathbf{y})$. However, this estimate is infeasible in the current Bayesian setting if there is no closed-form solution for the posterior density defined formally in (13.1), as then it is not possible to make draws from the posterior $p(\boldsymbol{\theta}|\mathbf{y})$. Instead, we use importance sampling, introduced in Section 12.7.2. The integral considered in (13.42) can be rewritten as

$$E[m(\boldsymbol{\theta})] = \int_{R(\boldsymbol{\theta})} \left(\frac{m(\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y})}{g(\boldsymbol{\theta})} \right) g(\boldsymbol{\theta}) d\boldsymbol{\theta}, \quad (13.43)$$

where $g(\boldsymbol{\theta}) > 0$ is a known density function, with the same support as $p(\boldsymbol{\theta}|\mathbf{y})$, that is easy to make draws from. The corresponding Monte Carlo integral estimate is

$$\widehat{E}[m(\boldsymbol{\theta})] = \frac{1}{S} \sum_{s=1}^S \frac{m(\boldsymbol{\theta}^s) p(\boldsymbol{\theta}^s|\mathbf{y})}{g(\boldsymbol{\theta}^s)},$$

where $\boldsymbol{\theta}^s$, $s = 1, \dots, S$, are S draws from of $\boldsymbol{\theta}$ from the **importance sampling density** $g(\boldsymbol{\theta})$ rather than from the original **target density** $p(\boldsymbol{\theta}|\mathbf{y})$. Note that the requirement that $p(\boldsymbol{\theta}|\mathbf{y})$ and $g(\boldsymbol{\theta})$ should have the same support is potentially problematic if $p(\boldsymbol{\theta}|\mathbf{y})$ depends on additional parameters or if the functional form of the full conditional densities is known but that of the marginal posterior is not.

Application to the posterior density additionally needs to account for the constant of integration in the denominator of (13.1). Let $p^{\text{ker}}(\boldsymbol{\theta}|\mathbf{y})$ denote the **kernel** of the posterior density, where $p^{\text{ker}}(\boldsymbol{\theta}|\mathbf{y}) = L(\mathbf{y}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta})$ or a multiple of this quantity. However, for notational simplicity the dependence on \mathbf{y} is suppressed in what follows. The posterior density is then

$$p(\boldsymbol{\theta}) = \frac{p^{\text{ker}}(\boldsymbol{\theta})}{\int p^{\text{ker}}(\boldsymbol{\theta}) d\boldsymbol{\theta}},$$

with corresponding posterior moment

$$\begin{aligned} E[m(\boldsymbol{\theta})] &= \int m(\boldsymbol{\theta}) \left(\frac{p^{\text{ker}}(\boldsymbol{\theta})}{\int p^{\text{ker}}(\boldsymbol{\theta}) d\boldsymbol{\theta}} \right) d\boldsymbol{\theta} \\ &= \frac{\int m(\boldsymbol{\theta}) p^{\text{ker}}(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int p^{\text{ker}}(\boldsymbol{\theta}) d\boldsymbol{\theta}} \\ &= \frac{\int (m(\boldsymbol{\theta}) p^{\text{ker}}(\boldsymbol{\theta}) / g(\boldsymbol{\theta})) g(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int (p^{\text{ker}}(\boldsymbol{\theta}) / g(\boldsymbol{\theta})) g(\boldsymbol{\theta}) d\boldsymbol{\theta}}. \end{aligned}$$

The **importance sampling-based estimate** of the posterior moment $E[m(\boldsymbol{\theta})]$ is then

$$\widehat{E}[m(\boldsymbol{\theta})] = \frac{\frac{1}{S} \sum_{s=1}^S m(\boldsymbol{\theta}^s) p^{\text{ker}}(\boldsymbol{\theta}^s) / g(\boldsymbol{\theta}^s)}{\frac{1}{S} \sum_{s=1}^S p^{\text{ker}}(\boldsymbol{\theta}^s) / g(\boldsymbol{\theta}^s)}, \quad (13.44)$$

where $\boldsymbol{\theta}^s$, $s = 1, \dots, S$, are S draws of $\boldsymbol{\theta}$ from the importance sampling density $g(\boldsymbol{\theta})$.

This method was proposed by Kloek and van Dijk (1978). Geweke (1989) established consistency and asymptotic normality under some regularity conditions. These conditions include the assumptions that the importance sampling density $g(\boldsymbol{\theta}) > 0$

over the support $R(\theta)$ of $p(\theta)$; that $E[m(\theta)] < \infty$, so the posterior moment exists; and that $\int p(\theta|y)d\theta = 1$, so the posterior density is proper. As previously noted, usually we work with the kernel $p^{\text{ker}}(\theta|y) = L(y|\theta)\pi(\theta)$, which need not integrate to one. The prior $\pi(\theta)$ need not be proper, but to ensure that $\int p(\theta|y)d\theta = 1$ it is necessary that $\int \pi(\theta)d\theta < \infty$.

The importance sampling approach is simple, but implementation entails subtleties well explained in Geweke (1989). A critical requirement is that the $g(\theta)$ should have thicker tails than the $p(\theta|y)$, to ensure that the **importance weight** $w(\theta) = p(\theta|y)/g(\theta)$ remains bounded. In view of the asymptotic normality of the log posterior, a good choice of $g(\theta)$ is a multivariate t -distribution, with the mean set to the posterior mode, and the covariance matrix proportional to the inverse of the Hessian of the log of the posterior, and degrees of freedom set to a value sufficiently small to ensure thick tails. Geweke (1989) also provides a measure, called the relative numerical efficiency, that estimates the number of replications required to achieve a given level of precision of $\hat{E}[m(\theta)]$ computed using draws from $g(\theta)$ relative to the number of replications needed if draws from $p(\theta|y)$ were possible. From Chapter 12, for a higher dimensional integral more simulation draws are required to get a good approximation to the integral and one might additionally use simulation acceleration methods presented in Chapter 12, such as antithetic sampling.

The importance sampling method uses each draw θ^s from the sampling density $g(\theta)$ with equal probability. A more efficient approximation would weight the draws according to how close $g(\theta^s)$ is to the target $p(\theta^s|y)$. This can be done by importance resampling (see Gelman et al., 1995).

The importance sampling method can be used to provide many useful summary measures of the posterior, as presented in Section 13.2.5. This includes estimates of the quantiles and percentiles of the posterior, permitting calculation of 95% posterior intervals and plots of the posterior density of θ_k .

13.5. Markov Chain Monte Carlo Simulation

A modern idea in Bayesian analysis is that rather than concentrating on the estimation of key summary measures of the posterior distribution (see the previous section) it is desirable to obtain a large sample from the posterior distribution. Then the summary statistics of this sample from the posterior will provide desired information about the moment characteristics of the sample of estimates and about other interesting associated measures such as marginal distributions of parameters or functions of parameters. For example, given S draws from the posterior distribution, $E[\theta_k]$ can be estimated by $S^{-1} \sum_s \theta_k^s$.

The challenge is to make draws from the joint posterior distribution when there is no tractable closed-form expression for the posterior density. If a suitable density exists for computation of posterior moments using importance sampling, then it might also be suitable for making draws from the posterior using the accept–reject method presented in Section 12.8. However, this method can be very inefficient as a high percentage of draws may be rejected.

Instead, sequential draws are made yielding simulated values that, if the sequence is run long enough, converge to a stationary distribution that coincides with the target posterior density $p(\theta|y)$. The method is called **Markov chain Monte Carlo (MCMC)**, because it involves simulation (Monte Carlo) and the sequence is that of a Markov chain. After convergence of the chain, S sequential draws can be used to compute summary measures for the posterior, such as estimating $E[\theta_k]$ by $\widehat{E}[\theta_k] = S^{-1} \sum_s \theta_k^s$. The draws are positively correlated, however, so the precision of the estimate will be reduced for given S because its estimated variance will exceed the usual $(S - 1)^{-1} \sum_s (\theta_k^s - \widehat{E}[\theta_k])^2$.

The sequential method entails constructing a Markov chain. Two widely used algorithms are the **Gibbs sampler** and the **Metropolis–Hastings** algorithm, the former being a special case of the latter, see Hastings (1970). Excellent detailed treatments of the subject can be found in Gelman et al. (1995), Gamerman (1997), and Robert and Casella (1999). What follows is a bare-bones sketch.

13.5.1. Markov Chains

Before presenting the Gibbs sampler and the Metropolis–Hastings algorithm we provide some key definitions and concepts used in the MCMC literature. These definitions are given in the context of a model with discrete states. They can be extended to the continuous state model, relevant to applications where the posterior is continuous in the parameters.

A **Markov chain** is defined as a sequence of random variables x_n ($n = 0, 1, 2, \dots$), where x_n takes values in a finite space A , together with a **transition kernel** $K(\cdot)$ that defines the probability that x_n equals a particular value given previous values x_{n-j} . We consider a **Markov chain** with the property that

$$\Pr[x_{n+1} = x | x_n, x_{n-1}, \dots, x_0] = \Pr[x_{n+1} = x | x_n], \quad (13.45)$$

so that the *distribution* of x_{n+1} given the past is completely determined only by the preceding value x_n . The transition kernel is a **transition matrix** \mathbf{T} with element

$$t_{xy} = \Pr[x_{n+1} = y | x_n = x], \quad (13.46)$$

which informally is the probability of transition from x to y . For a **finite-state** Markov chain the set A of values (or states) that x_n may take is finite with, say, m elements. Then

$$\mathbf{T} = \begin{bmatrix} t_{11} & \cdots & t_{1m} \\ \vdots & \ddots & \vdots \\ t_{m1} & \cdots & t_{mm} \end{bmatrix}, \quad (13.47)$$

with $\sum_{j=1}^m t_{ij} = 1$, $i = 1, \dots, m$.

Now consider the transition from x to y in n steps (stages). The transition probability is given by \mathbf{T}^n , the n -times matrix product of \mathbf{T} . The rows of the matrix \mathbf{T}^n give the marginal distribution across the m states at the n th stage, and the j th row vector $\mathbf{t}_j^{(n)} = (t_{j1}^{(n)}, \dots, t_{jm}^{(n)})$ gives the marginal distribution of transition probabilities from state j to

the other states at stage n . If the initial distribution of transition probabilities is denoted $\mathbf{t}_j^{(0)}$, then $\mathbf{t}_j^{(n)} = \mathbf{t}_j^{(0)} \mathbf{T}^n = \mathbf{t}_j^{(n-1)} \mathbf{T}$. So the marginal distribution of transition probabilities at the n th stage is determined solely by the initial distribution and the transition matrix.

In the Markov simulation context, the asymptotic behavior of the chain as $n \rightarrow \infty$ is of interest. The chain is said to yield a **stationary distribution** or **invariant distribution** with transition probabilities t_{xy} if

$$\sum_{x \in A} \mathbf{t}_x \mathbf{T}_{x,y} = \mathbf{t}_y \quad \forall y \in A, \quad (13.48)$$

where transition is from state \mathbf{t}_x to \mathbf{t}_y . Then applying the transition matrix leads to no change in the marginal distribution of transition probabilities. The existence and uniqueness of a stationary distribution is an important issue.

If the stationary distribution exists, and if $\lim_{n \rightarrow \infty} \mathbf{t}_x \mathbf{T}_{x,y}^n = \mathbf{t}_y$, then the chain will asymptotically approach \mathbf{t}_y independently of the initial distribution. In this sense \mathbf{t}_y is a limiting distribution. Although here the stationary distribution is defined for a finite-state Markov chain, MCMC methods can handle Markov chains that are not finite state; see Gilks, Richardson, and Spiegelhalter (1996, pp. 60–61).

A state y may be recurrent or transient. A **recurrent state** is one that will be revisited with probability one, and a **transient state** is one that will not be revisited with some positive probability.

For Bayesian applications the goal is to obtain draws from the posterior $p(\boldsymbol{\theta})$. Applying a Markov chain to obtain these draws, the initial value of a parameter vector, $\boldsymbol{\theta}^{(0)}$ (which is analogous to the distribution of states), is assigned or sampled from the transition kernel. Using a suitable method of drawing pseudo-random numbers, a new vector of values $\boldsymbol{\theta}^{(1)}$ is drawn from the transition kernel evaluated at $\boldsymbol{\theta}^{(0)}$, that is, $K(\boldsymbol{\theta}^{(0)})$. At the n th stage the draws are from a transition kernel $K(\boldsymbol{\theta}^{(n-1)})$ and so forth. The Markov chain used is one such that as $n \rightarrow \infty$ the limiting distribution is the posterior $p(\boldsymbol{\theta})$. Once convergence to the limiting distribution occurs all subsequent draws are also from this distribution, though they will be correlated.

These ideas provide the intuitive basis for a class of MCMC procedures that can be used to recover Bayesian posterior distributions for many different, and possibly high-dimensional, models such as, for example, the linear hierarchical models discussed in Section 13.3.4. Provided that one specifies a transition kernel $K(\boldsymbol{\theta}^{(n-1)}, \cdot)$ from which draws of $\boldsymbol{\theta}$ can be made and within which is embedded the chain's limiting distribution, the target posterior distribution can be recovered in the sense of being approached arbitrarily closely.

The current description is at a very general level. In practice, the choice of the transition kernel is not unique and there are many possible chains one can construct. Some choices may be better than others in terms of speed of convergence to the limiting distribution. If convergence is found to be very slow and computationally expensive, alternative chains may need to be substituted. Clearly, criteria are needed to determine whether convergence has occurred and how close to the target distribution the chain is at the n th stage.

13.5.2. Gibbs Sampler

We begin with the Gibbs sampler, a member of the MCMC class that is easy to describe and implement.

Let $\boldsymbol{\theta} = [\theta_1 \ \theta_2]'$ have posterior density $p(\boldsymbol{\theta}) = p(\theta_1, \theta_2)$, where for notational simplicity we suppress dependence on \mathbf{y} . If the conditional densities are known, which is not guaranteed as knowledge of both $p(\theta_1|\theta_2)$ and $p(\theta_2|\theta_1)$ is necessary, then alternating sequential draws from $p(\theta_1|\theta_2)$ and $p(\theta_2|\theta_1)$ in the limit converge to draws from $p(\theta_1, \theta_2)$.

Example

A simple illustration is to consider bivariate normal data with uniform prior for the mean and known covariance matrix. Let $\mathbf{y} = (y_1, y_2) \sim \mathcal{N}[\boldsymbol{\theta}, \Sigma]$, where $\boldsymbol{\theta} = [\theta_1 \ \theta_2]'$ and Σ has diagonal entries 1 and off-diagonal entries ρ . Then given a uniform prior for $\boldsymbol{\theta}$ the posterior can be shown to be $\boldsymbol{\theta}|\mathbf{y} \sim \mathcal{N}[\bar{\mathbf{y}}, N^{-1}\Sigma]$, a bivariate normal. Since the conditional posterior distributions are

$$\begin{aligned}\theta_1|\theta_2, \mathbf{y} &\sim \mathcal{N}[(\bar{y}_1 + \rho(\theta_2 - \bar{y}_2)), (1 - \rho^2)/N], \\ \theta_2|\theta_1, \mathbf{y} &\sim \mathcal{N}[(\bar{y}_2 + \rho(\theta_1 - \bar{y}_1)), (1 - \rho^2)/N],\end{aligned}$$

we can iteratively sample from each conditional normal distribution using updated values of θ_1 and θ_2 . If the chain is run long enough then it will converge to the bivariate normal. In this example it is easy to make direct draws from the joint posterior of $\boldsymbol{\theta}|\mathbf{y}$, using Choleski's transformation given in Section 12.8, but in other examples it can be possible to draw from the conditionals but not the joint posterior.

Gibbs Sampler

More generally, consider a q -dimensional target distribution $p(\boldsymbol{\theta})$, where the notation suppresses the dependence on data. Suppose that $\boldsymbol{\theta}$ is partitioned into d blocks. For example, $\boldsymbol{\theta}' = [\beta \ \sigma^2]'$ in a linear regression example. Let $\boldsymbol{\theta}_k$ denote the k th block and $\boldsymbol{\theta}_{-k}$ denote all components of $\boldsymbol{\theta}$ aside from $\boldsymbol{\theta}_k$. Assume that the full conditional distributions $p(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{-k})$, $k = 1, \dots, d$, are known. Then sequential sampling from the full conditionals can be set up as follows:

1. Let the initial values of $\boldsymbol{\theta}$ be $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \dots, \theta_d^{(0)})$.
2. The next iteration involves sequentially revising all components of $\boldsymbol{\theta}$ to yield $\boldsymbol{\theta}^{(1)} = (\theta_1^{(1)}, \dots, \theta_d^{(1)})$ generated using d draws from the d conditional distributions as follows:

$$\begin{aligned}p(\theta_1^{(1)}|\theta_2^{(0)}, \dots, \theta_d^{(0)}) \\ p(\theta_2^{(1)}|\theta_1^{(1)}, \theta_3^{(0)}, \dots, \theta_d^{(0)}) \\ \vdots \\ p(\theta_d^{(1)}|\theta_1^{(1)}, \theta_2^{(1)}, \dots, \theta_{d-1}^{(1)}).\end{aligned}$$

3. Return to step 1, reinitialize the vector θ at $\theta^{(1)}$, and cycle through step 2 again to obtain the new draw $\theta^{(2)}$. Repeat the steps until convergence is achieved.

Gilks et al. (1996, p. 7) provide a sketch of the proof of the statement that the stationary distribution is the posterior. After convergence the draws are from the target joint posterior. Geman and Geman (1984) showed that the stochastic sequence $\{\theta^{(n)}\}$ is a Markov chain with the correct stationary distribution. Gelfand and Smith (1990) showed that, under some conditions, as the number of cycles of draws from the full set of conditionals tends to infinity, the chain converges to the stationary posterior distribution. See also Tanner and Wong (1987). Once convergence occurs, numerous draws can be made and used to calculate sample analogues of the posterior moments of marginal or joint distributions.

The results mentioned here do not tell us how many cycles are needed for convergence, which is model dependent. It is very important to ensure that sufficient number of cycles are executed for the chain to converge. A variety of diagnostic tests of **convergence** are available. Because estimates of posterior moments should be based on draws from the posterior distribution it is standard practice to discard the earlier results from the chain, the so-called **burn-in phase**.

Sequential simulation algorithms can be modified so that each draw depends not simply on the immediately preceding draw but also on earlier draws, the key requirement being that probability of improvement on the current approximation to the posterior should be positive and (preferably) high. The attraction of the more restrictive Markovian property is that it facilitates the proof that the transition distributions converge to the target posterior.

For Bayesian analysis the Gibbs sampler is useful when the joint posterior is intractable but the full conditional distributions are available in a convenient form. Many applications use considerable ingenuity and knowledge of conjugate priors and related Bayesian results, many from the earlier presimulation literature, to specify priors that lead to known full conditional distributions.

We consider two examples that apply the MCMC methods.

Linear Regression Example

In Section 13.3.2 we analyzed the posterior distribution of the normal linear homoskedastic regression model, given normal–gamma conjugate priors. The conditional posterior of β given σ^{-2} was shown to be multivariate normal, and the conditional posterior of σ^{-2} given β is the gamma. Even though integration is feasible and we can derive the posterior in an explicit form (see (13.32)) it is actually easier to use the Gibbs sampler to draw a large sample from the joint posterior distribution. The chain consists of recursive draws from the normal conditional on the precision parameter σ^{-2} and from the gamma distribution conditional on the β .

The structure of the algorithm resembles that given later in Section 13.6 for a slightly more complicated case of a two-equation seemingly unrelated regressions model.

In many cases it would be natural to work with **blocks** of parameters. For example, in a multiequation multivariate linear regression model with a nondiagonal contemporaneous covariance matrix, the conditional mean parameters $(\beta_1, \beta_2, \dots)$ form one block of parameters, and Σ forms a second. Then the full conditional distributions will have the form $\beta_1, \beta_2, \dots | data, \Sigma$ and $\Sigma | data, \beta_1, \beta_2, \dots$. Chib and Greenberg (1996, pp. 418–419) outline the Gibbs algorithm for this case.

Hierarchical Prior Example

The Gibbs sampler has been deployed with much success in the analysis of the hierarchical prior model. From the structure of the linear hierarchical model given in (13.39)–(13.41), it can be seen that formulating a Markov chain based on a full set of conditionals is feasible in this case. The same general approach can be extended to a nonlinear hierarchical prior model, although some additional steps are necessary if the nonlinearity occurs in conjunction with a latent variable model (Albert, 1988).

13.5.3. Metropolis Algorithm

The Gibbs sampler is the best-known MCMC algorithm. Its applicability is limited, however, as it requires direct sampling from the full conditional distributions, which may not be known. Two extensions that allow the MCMC to be applied more generally are the Metropolis algorithm and the Metropolis–Hastings algorithm. Chib and Greenberg (1995) provide a tutorial and references. The following summary is simpler but avoids many details that are necessary if the reader seeks a more complete understanding.

The Metropolis algorithm constructs a sequence $\{\theta^{(n)}, n = 1, 2, \dots\}$ whose distributions converge to the target posterior, assumed to be computable up to a normalizing constant.

For notational simplicity we again suppress dependence of $p(\theta|y)$ on y . The algorithm consists of the following steps:

1. Draw a starting point $\theta^{(0)}$ from an initial approximation to the posterior for which $p(\theta^{(0)}) > 0$. For example, the draw may be from a multivariate t -distribution centered on the mode of the marginal posterior distribution.
2. Next set $n = 1$. Draw θ^* from a *symmetric jumping distribution* $J_1(\theta^{(1)}|\theta^{(0)})$, with the property that for any arbitrary pair (θ^a, θ^b) , $J_n(\theta^a|\theta^b) = J_n(\theta^b|\theta^a)$. An example is $\theta^{(1)}|\theta^{(0)} \sim \mathcal{N}[\theta^{(0)}, V]$ for some fixed V . Symmetry of the jumping distribution leads to simplicity but is not otherwise essential.
3. Calculate the ratio of densities $r = p(\theta^*)/p(\theta^{(0)})$.
4. Set

$$\theta^{(1)} = \begin{cases} \theta^* & \text{with probability } \min(r, 1), \\ \theta^{(0)} & \text{with probability } (1 - \min(r, 1)), \end{cases}$$

which means that the draw $\theta^{(1)}$ is a draw from a mixture distribution with components θ^* and $\theta^{(0)}$.

5. Return to step 2, increase the counter, and repeat the following steps.
6. After a suitably large number of iterations apply the necessary checks for the convergence of the distribution. If convergence has occurred the target posterior has been recovered.

This algorithm can be viewed as an iterative method to maximize $p(\theta)$. If θ^* increases $p(\theta)$ then $\theta^{(n)} = \theta^*$ always, whereas if θ^* decreases $p(\theta)$ then $\theta^{(n)} = \theta^*$ with probability $r < 1$.

The algorithm is similar in spirit to accept–reject sampling (see Section 12.8), though there is no requirement here that a fixed multiple of the jumping distribution always covers the posterior.

The Metropolis algorithm generates a Markov chain that has properties of reversibility, irreducibility, and Harris recurrence that ensure convergence to a stationary distribution. Gelman et al. (1995) demonstrate that this stationary distribution is the desired posterior $p(\theta)$ as follows. Let θ_a and θ_b be two points such that $p(\theta_b) \geq p(\theta_a)$. If $\theta^{(n-1)} = \theta_a$ and $\theta^* = \theta_b$ then $\theta^{(n)} = \theta_b$ with certainty and $\Pr[\theta^{(n)} = \theta_b, \theta^{(n-1)} = \theta_a] = J_n(\theta_b|\theta_a)p(\theta_a)$. If the order is reversed and $\theta^{(n-1)} = \theta_b$ and $\theta^* = \theta_a$, then $\theta^{(n)} = \theta_a$ with probability $r = p(\theta_a)/p(\theta_b)$ and $\Pr[\theta^{(n)} = \theta_a, \theta^{(n-1)} = \theta_b] = J_n(\theta_a|\theta_b)p(\theta_b)[p(\theta_a)/p(\theta_b)] = J_n(\theta_a|\theta_b)p(\theta_a) = J_n(\theta_b|\theta_a)p(\theta_a)$ given the assumption of symmetric jumping distribution. The marginal distributions of $\theta^{(n)}$ and $\theta^{(n-1)}$ are therefore equal, since their joint distribution is symmetric, so $p(\theta)$ is the symmetric stationary distribution of the Markov chain.

13.5.4. The Metropolis–Hastings Algorithm

The performance of the Metropolis algorithm varies with the choice of initial approximating distribution and choice of jumping distribution. A potential problem is that the Metropolis algorithm may be slow, as would be the case if the move from the current to a new value is not made sufficiently often, causing the chain to move infrequently. The algorithm can be speeded up by permitting use of jumping distributions that are not symmetric.

The **Metropolis–Hastings (M–H) algorithm** is the same as the Metropolis algorithm, except that in step 2 the jumping distribution need not be symmetric, and in step 3 the acceptance probability r for general n becomes

$$r_n = \frac{p(\theta^*)/J_n(\theta^*|\theta^{(n-1)})}{p(\theta^{(n-1)})/J_n(\theta^{(n-1)}|\theta^*)} = \frac{p(\theta^*)J_n(\theta^{(n-1)}|\theta^*)}{p(\theta^{(n-1)})J_n(\theta^*|\theta^{(n-1)})}.$$

The remaining steps are executed with this revised definition. Note that if any normalizing constants are present in either $p(\cdot)$ or $J_n(\cdot)$, then they cancel in this definition of r_n . So both posterior and jumping probabilities need only be computed up to this constant. See Hastings (1970).

13.5.5. M–H Examples

Different jumping distributions lead to different M–H algorithms with different efficiency in terms of the number of draws needed to obtain the desired draws from the posterior. We give several examples, noting that there are few general guidelines available for choice of jumping distribution, except to use the Gibbs sampler wherever possible.

The Gibbs sampler is a special case of the M–H algorithm. If θ is partitioned into d blocks, then there are d Metropolis steps at the n th step of the algorithm. The jumping distribution is the conditional distribution given in Section 13.5.2 and it can be shown that the acceptance probability is always 1. Gibbs sampling is also called **alternating conditional sampling**.

It is possible to use mixed strategies, whereby different transition kernels are used for different subsets of parameters. For example, an M–H step can be combined with a Gibbs sampler, the latter being used for components for which direct sampling is feasible.

The **independence chain** makes all draws from a fixed density $g(\theta)$, say, in which case the acceptance probability simplifies to the ratio $r_n = w(\theta^*)/w(\theta^{(n-1)})$ of importance weights $w(\theta) = p(\theta)/g(\theta)$. A **random walk chain** sets the draw $\theta^* = \theta^{(n-1)} + \varepsilon$, where ε is a draw from $g(\varepsilon)$.

Gelman et al. (1995, p. 334) consider simulating the q -variate normal with variance Σ . For a Metropolis algorithm with jumping distribution $\theta^*|\theta^{(n-1)} \sim \mathcal{N}[\theta^{(n-1)}, c^2 \Sigma]$, the choice $c \simeq 2.4/\sqrt{q}$ leads to greatest efficiency relative to direct draws from the q -variate normal. The efficiency is about 0.3, compared to $1/q$ for the Gibbs sampler in the case that $\Sigma = \sigma^2 \mathbf{I}_q$.

13.6. MCMC Example: Gibbs Sampler for SUR

We illustrate the application of the Gibbs sampler to the analysis of the seemingly unrelated regression model. This example is slightly more challenging than an application to single-equation regression, because errors correlated across equations are introduced.

We consider a two-equation example with i th observation

$$\begin{aligned} y_{1i} &= \mathbf{x}'_{1i}\beta_1 + \varepsilon_{1i}, \\ y_{2i} &= \mathbf{x}'_{2i}\beta_2 + \varepsilon_{2i}, \end{aligned}$$

where $(\varepsilon_1, \varepsilon_2)$ are bivariate normal with zero mean and covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}.$$

Combining the two equations gives the i th observation

$$\mathbf{y}_i = \mathbf{x}'_i\beta + \varepsilon_i,$$

where $\varepsilon_i \sim \mathcal{N}[\mathbf{0}, \Sigma]$. In summary, the dgp is

$$y_i | \mathbf{x}_i, \boldsymbol{\beta}, \Sigma \sim \mathcal{N}[\mathbf{x}'_i \boldsymbol{\beta}, \Sigma]$$

and interest lies in estimating the posterior means of the regression parameters $\boldsymbol{\beta}$ and variance parameters Σ , given data \mathbf{y} , \mathbf{X} .

We consider independent informative priors, with

$$\begin{aligned}\boldsymbol{\beta} &\sim \mathcal{N}[\boldsymbol{\beta}_0, \mathbf{B}_0^{-1}], \\ \Sigma^{-1} &\sim \text{Wishart}[v_0, \mathbf{D}_0],\end{aligned}$$

where \mathbf{B}_0 is defined as precision, the inverse of the prior variance, and the inverse Wishart, defined in Section 13.3.5, is a generalization of the inverse gamma. An alternative approach, not taken here, uses dependent priors similar to those in Section 13.3.2, in which case $\boldsymbol{\beta} | \Sigma \sim \mathcal{N}[\boldsymbol{\beta}_0, \omega_0 \Sigma]$ for specified ω_0 .

Performing some algebra yields the conditional posteriors

$$\begin{aligned}\boldsymbol{\beta} | \Sigma, \mathbf{y}, \mathbf{X} &\sim \mathcal{N}\left[\mathbf{C}_0 \left(\mathbf{B}_0 \boldsymbol{\beta}_0 + \sum_{i=1}^N \mathbf{x}'_i \Sigma^{-1} \mathbf{y}_i \right)_i, \mathbf{C}_0\right], \\ \Sigma^{-1} | \boldsymbol{\beta}, \mathbf{y}, \mathbf{X} &\sim \text{Wishart}\left[v_0 + N, \left(\mathbf{D}_0^{-1} + \sum_{i=1}^N \mathbf{u}_i' \mathbf{u}_i\right)^{-1}\right],\end{aligned}$$

where $\mathbf{C}_0 = (\mathbf{B}_0 + \sum_{i=1}^N \mathbf{x}'_i \Sigma^{-1} \mathbf{x}_i)^{-1}$ and $\mathbf{u}_i = \mathbf{y}_i - \mathbf{x}'_i \boldsymbol{\beta}$. The Gibbs sampler can be used since the conditional posteriors are known and sampling from both distributions is straightforward.

For a simulation example we let the regressors in each equation be an intercept plus a single scalar regressor, different in the two equations, generated from a standard normal. Then y_1 and y_2 are generated with the four regression parameters $\beta_{11} = \beta_{12} = \beta_{21} = \beta_{22} = 1$, the error variances $\sigma_{11} = \sigma_{22} = 1$, and the error covariance $\sigma_{12} = \sigma_{21} = -0.5$. The sample size is either $N = 1,000$ or $N = 10,000$. Given these data, we present Bayesian estimates of the parameters, where the prior distributions set $\boldsymbol{\beta}_0 = \mathbf{0}$, $\mathbf{B}_0^{-1} = \tau \mathbf{I}$, $\mathbf{D}_0 = \mathbf{I}$, and $v_0 = 5$. To check the impact of different priors three values of τ are considered, $\tau = 10$, $\tau = 1$, and $\tau = 1/10$, with smaller values of τ corresponding to tighter priors.

The Gibbs sampler makes draws recursively from the conditional posterior distributions. We reject the first 5,000 replications that constitute the “burn-in” phase and report results using the subsequent 50,000 and 100,000 replications.

A selection of the results is given in Table 13.3, which reports the mean and variance of the marginal posterior distribution of each coefficient in five different samples that themselves are independent draws. The first three columns present a sensitivity analysis for different values of τ , which shows that the results are not very sensitive. The fourth column, compared to the first, shows that doubling the number of replications has very little effect. The fifth column, compared to the first, shows that increasing the sample size tenfold to 100,000 greatly increases the precision as expected, reducing the standard deviation of the coefficient by a factor of more than 3, but with relatively small impact on the point estimates.

Table 13.3. Gibbs Sampling: Seemingly Unrelated Regressions Example^a

Prior parameter τ	$\tau = 10$	$\tau = 1$	$\tau = 1/10$	$\tau = 10$	$\tau = 10$
Sample size N	1,000	1,000	1,000	1,000	10,000
Gibbs sample replications	50,000	50,000	50,000	100,000	100,000
β_{11} (eq. 1 intercept)	0.971 (0.0310)	1.013 (0.0312)	0.983 (0.0316)	1.020 (0.0324)	1.010 (0.0100)
β_{12} (eq. 1 slope)	1.026 (0.0265)	0.9835 (0.0271)	1.006 (0.0265)	1.006 (0.0268)	1.015 (0.0086)
β_{21} (eq. 2 intercept)	1.016 (0.0309)	0.972 (0.0325)	0.993 (0.0322)	1.017 (0.0326)	0.991 (0.0100)
β_{22} (eq. 2 slope)	0.983 (0.0256)	0.992 (0.0285)	0.979 (0.0272)	1.005 (0.0277)	1.007 (0.0085)
σ_{11} (eq. 1 error variance)	0.960 (0.0429)	0.969 (0.0434)	1.012 (0.0453)	1.043 (0.0466)	1.010 (0.0143)
σ_{12} (error covariance)	-0.499 (0.0340)	-0.507 (0.0358)	-0.519 (0.0368)	-0.576 (0.0379)	-0.515 (0.0113)
σ_{22} (eq. 2 error variance)	0.950 (0.425)	1.066 (0.0476)	1.049 (0.0467)	1.062 (0.0472)	1.002 (0.0141)

^a Model is a two-equation seemingly unrelated regression. Table gives the mean and standard deviation of the posterior distribution for each parameter. Smaller values of τ correspond to tighter priors.

One way to check for convergence is to look at the means and standard deviations of the output and see whether they drift or stay at the same level. If the change is small, say less than 0.1 for 10,000 replications, then convergence is presumed. One also might look at several chains at a time. The draws will always be correlated but the important question is how fast the autocorrelation function decays to zero. Sometimes this problem cannot be fixed and it is simply inherent to the algorithm. One can also take every tenth or hundredth observation to purge serial correlation.

To check whether the Gibbs sampler has converged to the stationary posterior distribution in the present case, we compute the first 20 autocorrelation coefficients of draws from the posterior after convergence for each coefficient. Lack of convergence would be indicated by the presence of serial correlation in the draws from the target distribution. When the number of replications is small, say 1,000, the autocorrelation coefficients are found to be as high as 0.06 in some cases. However, when the number of replications is 50,000 and greater, there is virtually no evidence of serial correlation up to order 20, and correlation disappears with the order. In most cases the estimates are smaller than 0.005. It is easy to verify that for $N = 1,000$, the prior parameters τ has very little impact on the posterior. This computation is very simple and takes little more than a few seconds.

13.7. Data Augmentation

The Gibbs sampler can sometimes be applied to a wider range of models by introduction of auxiliary variables. In particular, this is the case for models involving latent

variables, such as discrete choice models, truncated and censored models, and finite mixture models introduced in later chapters.

In the scalar case the latent dependent variable y^* is not observed; instead, we observe only $y = g(y^*)$ for some specified function y . For example, in a logit or probit model (see Chapter 14) we may observe only whether y^* is positive or negative, in which case $y = \mathbf{1}(y^* > 0)$ and we observe $y = 1$ if $y^* > 0$ and $y = 0$ if $y^* \leq 0$.

Bayesian analysis of latent variable models, and especially the application of the Gibbs sampler, is greatly aided by the replacement of the latent variable by **imputed values**. This step is feasible if we can write down the predictive density of the latent variables in terms of the observed variables. The procedure of adding imputed values as if they were observed data is called **data augmentation**. (An example was given in Section 10.3.7 where the EM algorithm was exposited.) The essential insight, due to Tanner and Wong (1987), is that the posterior based only on the observed data is intractable, but that obtained after data augmentation is often tractable using the Gibbs sampler.

Consider the posterior expressed in terms of both directly observed variables \mathbf{y} and the latent variables \mathbf{y}^* ,

$$p(\boldsymbol{\theta}|\mathbf{y}) \equiv \int_{\mathbf{y}^*} p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{y}^*) f(\mathbf{y}^*|\mathbf{y}) d\mathbf{y}^*, \quad (13.49)$$

where the right-hand-side integral may be interpreted as an averaging operation with respect to \mathbf{y}^* .

Analogous to the EM method, data augmentation involves cycling between an **imputation step**, I-step, and a **posterior step**, P-step.

In the imputation step we make draws from the full conditional density of \mathbf{y}^* . This averages over the parameters ψ that appear in the probability distribution that links \mathbf{y}^* and \mathbf{y} . The predictive distribution is

$$f(\mathbf{y}^*|\mathbf{y}) = \int_{\psi} f(\mathbf{y}^*|\mathbf{y}, \psi) f(\psi|\mathbf{y}) d\psi. \quad (13.50)$$

Given the current draw from $p(\boldsymbol{\theta}|\mathbf{y})$ we can make a draw of \mathbf{y}^* from $f(\mathbf{y}^*|\mathbf{y})$, repeating both parts of the step m times to generate m multiple imputations \mathbf{y}_i^* , $i = 1, \dots, m$. This completes the I-step.

Given the augmented data from the I-step, the P-step is implemented by updating the current approximation to $p(\boldsymbol{\theta}|\mathbf{y})$; thus,

$$\text{updated } p(\boldsymbol{\theta}|\mathbf{y}) = \frac{1}{m} \sum_{i=1}^m p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{y}_i^*). \quad (13.51)$$

Then the algorithm returns to the I-step.

If $m = 1$, the procedure amounts to performing integration in (13.49) by Gibbs sampling. If m is chosen to be sufficiently large, the posterior distribution is approximated better. An extended example of the data augmentation method applied to the missing data problem is given in Chapter 26.

13.8. Bayesian Model Selection

Chapters 7 and 8 dealt with issues of hypothesis testing, specification diagnostics, and model comparison from a frequentist viewpoint. In this section we consider the principal tool, **Bayes factors**, that is used in Bayesian analysis to evaluate the strength of evidence in favor of the null hypothesis (model). It also serves as a criterion for model selection, irrespective of whether nested or nonnested pairs of models are under consideration. In the econometrics literature, Zellner (1971, 1978) provided an early discussion in the context of model selection. Our treatment is based on Kass and Raftery's (1995) review article.

Denote the data by \mathbf{y} and the two hypotheses under consideration, possibly nonnested, by H_1 and H_2 . Prior probabilities of the two hypotheses are $\Pr[H_1]$ and $\Pr[H_2]$. The corresponding dgps are $\Pr[\mathbf{y}|H_1]$ and $\Pr[\mathbf{y}|H_2] = 1 - \Pr[\mathbf{y}|H_1]$. The prior probabilities of the models are transformed to posterior probabilities by the sample evidence as reflected in the likelihood. By Bayes' Theorem

$$\Pr[H_k|\mathbf{y}] = \frac{\Pr[\mathbf{y}|H_k]\Pr[H_k]}{\Pr[\mathbf{y}|H_1]\Pr[H_1] + \Pr[\mathbf{y}|H_2]\Pr[H_2]}, \quad k = 1, 2, \quad (13.52)$$

and the **posterior odds ratio**

$$\frac{\Pr[H_1|\mathbf{y}]}{\Pr[H_2|\mathbf{y}]} = \frac{\Pr[\mathbf{y}|H_1]\Pr[H_1]}{\Pr[\mathbf{y}|H_2]\Pr[H_2]} \equiv B_{12} \frac{\Pr[H_1]}{\Pr[H_2]}, \quad (13.53)$$

where $B_{12} = \Pr[\mathbf{y}|H_1]/\Pr[\mathbf{y}|H_2]$, is called the Bayes factor. Hypothesis 1 is preferred if the posterior odds ratio exceeds one. The right-hand side of (13.53) expresses the posterior odds ratio as the product of the Bayes factor and the prior odds. If a priori the two models are equally probable, so $\Pr[H_1] = \Pr[H_2]$, then the Bayes factor equals the posterior odds in favor of H_1 . If several hypotheses are involved the Bayes factor can be computed for all pairs of hypotheses. The Bayes factor is defined even if the hypotheses are not nested.

The Bayes factor has the form of a likelihood ratio. It depends on unknown parameters, denoted by vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, that are eliminated by averaging or integrating over the parameter space with respect to the prior, so

$$\Pr[\mathbf{y}|H_k] = \int \Pr[\mathbf{y}|\boldsymbol{\theta}_k, H_k] \pi(\boldsymbol{\theta}_k|H_k) d\boldsymbol{\theta}, \quad k = 1, 2. \quad (13.54)$$

From Section 13.2.5, this equation gives the marginal and the predictive probability of the data given the prior distribution.

A complication is that this expression depends on all the constants that appear in the likelihood. These constants can be neglected when evaluating the posterior, but they are required for the computation of the Bayes factor. The integral in (13.54) may need to be numerically evaluated if it does not have an explicit solution using, for example, importance sampling. There is a substantial literature, reviewed in Kass and Raftery (1995), on the computation of the Bayes factor that we will not pursue here. We note that there are some asymptotic approximations to the Bayes factors that are readily computable using output from packages that maximize likelihoods.

Table 13.4. Interpretation of Bayes Factors

Bayes Factor B_{12}	$2 \ln(B_{12})$	Evidence against H_1
1 to 3	0 to 2	weak
3 to 20	2 to 6	positive
20 to 150	6 to 10	strong
>150	>10	very strong

Interpretation of the Bayes factor is in terms of evidence against the H_1 . “The Bayes factor is a summary of the evidence provided by the data in favor of one scientific theory, represented by a statistical model, as opposed to another” (Kass and Raftery, 1995, p. 777). In the frequentist analysis twice the log-likelihood ratio is an often-used quantity. Similarly, twice the log of the Bayes factor is a criterion used in evaluating the evidence. Kass and Raftery present the following categorization of the strength of evidence against the null model that they have found useful in their own work; see Table 13.4.

Suppose that two models under comparison are nested. Denote by H_0 the constrained model and H_1 the model that is unconstrained. A pairwise comparison of the two models using the posterior odds ratio requires computation of the Bayes factor, as shown earlier. The Bayes factor for the null hypothesis model is defined as

$$B_{01} = \frac{m(\mathbf{y}|H_0)}{m(\mathbf{y}|H_1)}$$

where $m(\mathbf{y}|H_j)$ is the marginal likelihood of the model specification H_j . If the models H_0 and H_1 are nested, then the Savage-Dickey density ratio approach (see Verdinelli and Wasserman, 1995) can be taken to calculate the Bayes factors.

An important insight due to Chib (1995) has made the computation of Bayes factors a great deal easier than suggested by the earlier literature, irrespective of whether the models are nested or nonnested. His approach consists of two related ideas. The first rewrites the marginal density, for a given model H_k , $m(\mathbf{y})$ as a ratio

$$m(\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta}|\mathbf{y})}, \quad (13.54)$$

where the numerator is the product of the density (inclusive of constants) and the prior, and the denominator is the posterior density of $\boldsymbol{\theta}$. This result is a rearrangement of the terms in equation (13.1), with the qualification that we have used the notation $m(\mathbf{y})$ in place of $f(\mathbf{y})$ or $\Pr[\mathbf{y}|H_k]$ used earlier; it merely states that the marginal density is the normalizing constant. Second, after a successful application of an MCMC algorithm, we will have available a Monte Carlo estimate of the posterior density estimate $\pi(\tilde{\boldsymbol{\theta}}|\mathbf{y})$ at a given point $\tilde{\boldsymbol{\theta}}$. Then it follows that

$$\ln \hat{m}(\mathbf{y}) = \ln f(\mathbf{y}|\tilde{\boldsymbol{\theta}}) + \ln \pi(\tilde{\boldsymbol{\theta}}) - \ln \pi(\tilde{\boldsymbol{\theta}}|\mathbf{y}). \quad (13.55)$$

Therefore, given estimates of the terms on the right-hand side, the marginal density can be readily computed using the output from a Gibbs sampler. This approach has been

extended in Chib and Jeliazkov (2001) to the case where the output is instead from a Metropolis-Hastings algorithm.

In complex and highly parameterized models, the computation of the Bayes factor is a nontrivial matter. However, it can be shown that the Schwarz criterion, also known as the Bayes information criterion (see Section 8.5), gives a rough approximation to the log of the Bayes factor. Recall that $BIC = -2 \ln L(\hat{\theta}_{ML}) + \ln Nq$. This is easy to compute if the value of the log-likelihood is available.

From (13.52) it is obvious that the ratio of prior probabilities of the model plays a role in evaluating the evidence against the null. In many situations, the investigator may have little to go on in assigning these probabilities. This consideration has received some attention in the literature that deals with the sensitivity of the Bayes factor to the prior model probabilities.

13.9. Practical Considerations

The use of Markov chain methods has now become dominant in the Bayesian literature. Because the methods are computer intensive, good software is essential. At the time of writing, the WinBUGS package, a later version of the BUGS (Bayesian inference Using Gibbs Sampling) package (Gilks et al., 1996), has been widely recommended and found to be especially useful for hierarchical models and missing data problems. It is available at the BUGS Web site. More detailed information about other Bayesian software can be found in Gamerman (1997, Section 5.6).

The issue of how long to run the chain continues to be an active area of research. Diagnostic checks for convergence are available and have been mentioned, but they often do not have universal applicability. Cappè and Robert (2000) provide a review of the issues of implementation including stopping rules. The complexity of the conditional distributions is clearly an important factor. Graphs of output for scalar parameters from the Markov are a visually attractive way of confirming convergence, but more formal approaches are available (Geweke, 1992). Another suggestion, due to Gelman and Rubin (1992), is to use multiple (parallel) Gibbs samplers, each beginning with different starting values to see if different chains converge to the same posterior distribution. Zellner and Min (1995) propose several convergence criteria that can be used if the posterior can be written explicitly.

13.10. Bibliographic Notes

There are several excellent book-length treatments with emphasis on modern computational methods for Bayesian analysis, including those by Gamerman (1997) and Gelman et al. (1995). Relatively accessible treatments are provided by Gill (2002), Koop (2003), and Lancaster (2004). Koop presents Bayesian methods for many standard nonlinear cross-section models and for panel data. The older texts by Zellner (1971) and Leamer (1978) are still valuable sources of results.

13.2 Stigler (1986) provides a good exposition of the work of Bayes (1764). Bayes first presented some properties of probability, notably $\Pr[A|B] = \Pr[A \cap B]/\Pr[B]$. Bayes then

applied this result to obtaining the posterior probability $\Pr[a < \theta < b|y]$, where a and b are specified bounds, y is the number of successes in N binomial trials, and θ is the unknown probability of success in each trial. Bayes chose a uniform prior, in which case the posterior density $f(\theta|y) \propto f(y|\theta)$. Bayes' example was challenging as he could not accurately calculate the posterior probability, which involved the incomplete gamma, not tabulated until the 20th century. Bayes' paper was initially neglected. A more commonly used approach due to Laplace and others was the method of inverse probability that also let $f(\theta|y) \propto f(y|\theta)$. These methods were supplanted by maximum likelihood, introduced by Fisher (1922), whose paper directly critiqued Bayesian and inverse-probability methods.

The regularity conditions for convergence to posterior normality are discussed in Heyde and Johnstone (1979). Train (2003) provides an excellent but less formal treatment of the so-called Bernstein–von Mises Theorem.

13.3 Zellner (1971) and Leamer (1978) are excellent sources for Bayesian analysis of linear regression.

13.4 Geweke (1989) and Geweke and Keane (2001) are valuable references on Monte Carlo integration.

13.5 Casella and George (1992) provide an expository treatment of the Gibbs sampler. Numerous papers by Chib and his collaborators and Geweke and his collaborators cover many topics of interest in microeconomics. Chib and Greenberg (1996, Section 3) provide a number of applications of MCMC, including the seemingly unrelated regression model and the Tobit and probit models. In the latter case they show the computational simplification that results from combining Gibbs sampling with the data augmentation approach. Data augmentation is used to handle latent variables that are introduced to deal with the underlying unobservables that arise naturally in many censored and discrete choice models. Chib (2001) provides a detailed and up-to-date survey that includes MCMC algorithms for many leading linear and nonlinear regression models. Geweke and Keane (2000) concentrate on the methods of integration; they cover both Bayesian and non-Bayesian topics.

Exercises

21–1 Show that if $\beta|\lambda \sim \mathcal{N}[\mu, \lambda^{-1}\Sigma]$, and $\lambda \sim \text{Gamma}[\alpha/2, \alpha/2]$, then the unconditional distribution of β is a multivariate t -distribution with parameters (μ, Σ, α) .

21–2 (Adapted from Chib, 1992). Consider the censored regression or Tobit model (see Section 16.3) where $y^* = \mathbf{x}'\beta + \varepsilon$, $\varepsilon \sim \text{iid } \mathcal{N}[0, \sigma^2]$, and y is observed when $y^* > 0$ but is not observed (censored) when $y^* \leq 0$. There are N_0 censored observations on y , and y_0 refers to them. Introduce a latent variable z that corresponds to the censored observations such that $z_i < 0$ if the i th observation belongs to the censored set. The data augmentation method can be used to draw latent variables z_i , a set of independent random variables distributed as truncated normal, with support $(-\infty, 0)$ and pdf $\phi(z_i|y_i, \beta, \sigma^2)/(1 - \Phi(\mathbf{x}_i'\beta/\sigma))$, $-\infty < z_i < 0$, where ϕ and Φ are, respectively the normal pdf and cdf. Use a normal prior for β and a gamma prior for σ^{-2} .

- (a) Show that it is possible to specify a full set of conditionals for z_i , β , and σ^{-2} .
- (b) Use the results of part (a) to outline the Gibbs algorithm for simulating z_i , β , and σ^{-2} .
- (c) Explain how suitable initial values of β and σ^{-2} may be obtained.

PART FOUR

Models for Cross-Section Data

Part 4, consisting of chapters 14 to 20, covers the core nonlinear limited dependent variable models for cross-section data, defined by the range of values taken by the dependent variable. Topics covered include models for binary, multinomial, duration and count data. The complications of censoring, truncation and sample selection are also studied. The essential base for Part 4 is least squares and maximum likelihood estimation.

Chapters 14–15 cover models for binary and multinomial data that are standard in the analysis of discrete outcomes and discrete choice. Maximum likelihood methods are dominant. Different parameterizations for the conditional probabilities in these models lead to different models, notably logit and probit models, which are well-established. Recent literature has focused on less restrictive modeling with more flexible functional forms for conditional probabilities and on accommodating individual unobserved heterogeneity. These objectives motivate the use of semiparametric methods and simulation-based estimation methods.

Censoring, truncation, or sample selection generate several important classes of models that are analyzed in Chapter 16. The long-established Tobit model is central to this literature, but its estimation and inference rely on strong distributional assumptions to permit consistent estimation. We also examine the newer semiparametric methods that rely on weaker assumptions.

Chapters 17–19 consider duration models in which the focus is on either the determinants of spell lengths, such as length of an unemployment spell, or on modeling the hazard rate of transitions from one initial state to another. The analysis covers both discrete and continuous time models, and both parametric and semiparametric formulations, including the standard models like the exponential, the Weibull, and the proportional hazards model. Chapter 18 covers formulation and interpretation of richer models that incorporate unobserved heterogeneity. The relative importance of state dependence and unobserved heterogeneity as determinants of the average length of spell is a central issue, whose resolution raises fundamental questions about alternative modeling approaches. Chapter 19 deals with models with several types of events using the competing risks formulation and models of multiple spells.

Chapter 20 covers the analysis of event count of the kind very common in health economics. There are many strong connections and parallels between count data models and duration models because of their common foundation in stochastic processes. We analyze the widely-used Poisson and negative binomial regression models, together with important variants such as the two-part or hurdle model, zero-inflated models, latent class models, and endogenous regressor models, all of which accommodate different facets of the event processes.

Binary Outcome Models

14.1. Introduction

Discrete outcome or qualitative response models are models for a dependent variable that indicates in which one of m mutually exclusive categories the outcome of interest falls. Often there is no natural ordering of the categories. For example, categorization may be on the occupation of a worker.

This chapter considers the simplest case of **binary outcomes**, where there are **two possible outcomes**. Examples include whether or not an individual is employed and whether or not a consumer makes a purchase. Binary outcomes are simple to model and **estimation is usually by maximum likelihood** because the distribution of the data is necessarily defined by the Bernoulli model. If the probability of one outcome equals p , then the probability of the other outcome must be $(1 - p)$. For regression applications the probability p will vary across individuals as a function of regressors. The two standard binary outcome models, the logit and the probit models, specify different functional forms for this probability as a function of regressors. The difference between these estimators is qualitatively similar to use of different functional forms for the conditional mean in least-squares regression.

Section 14.2 provides a data example. Section 14.3 presents a summary of statistical results for standard models including logit and probit models. In Section 14.4 binary outcome models are presented as arising from an underlying latent variable. This formulation is useful as it extends readily to multinomial models (see Chapter 15) and models for censored and selected samples (see Chapter 16). Section 14.5 details necessary modifications to standard estimation methods when one of the outcomes is deliberately oversampled. Aggregation issues are considered in Section 14.6. Semiparametric methods for binary outcome models that place less structure on the model for the probability p are given in Section 14.7.

14.2. Binary Outcome Example: Fishing Mode Choice

This section models choice between fishing from a charter boat and fishing from a pier. The dependent variable is binary with

$$y_i = \begin{cases} 1 & \text{if fishing from a charter boat,} \\ 0 & \text{if fishing from a pier,} \end{cases}$$

where the values 1 and 0 are chosen for simplicity. The single explanatory variable is $x_i = \ln \text{relp}_i = \ln(\text{relp}_i)$ where relp denotes the price of charter fishing relative to the price of fishing from the pier, so

$$x_i = \ln \text{relp}_i = \ln(\text{price}_{\text{charter},i} / \text{price}_{\text{pier},i}).$$

The prices of charter boat and pier fishing vary across individuals owing to various factors, for example, to differences in access. It is expected that the probability of charter boat fishing decreases as its relative price increases.

The data are summarized in Table 14.1. The sample of 630 individuals is a subset of the data described in greater detail in Section 15.2, where four different modes of fishing and additional regressors are considered. Charter boat fishing was selected by 71.7% of the sample. For people choosing to fish from the charter boat, the charter boat was on average less expensive than pier fishing, as $\$75 < \121 . For people choosing to fish from the pier the reverse was true. So it appears that price has the expected effect.

An **OLS regression** of y_i on x_i ignores the discreteness of the dependent variable and does not constrain predicted probabilities to be between zero and one.

A more appropriate model is the **logit model** (see Section 14.3.4), which specifies

$$p_i = \Pr[y_i = 1 | x_i] = \frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)}$$

and clearly ensures that $0 < p_i < 1$. Maximum likelihood estimation (see Section 14.3.3) leads to parameter estimates given in the first column of Table 14.2. The implied marginal effect for the logit model equals

$$\frac{dp_i}{dx_i} = \frac{\exp(\beta_1 + \beta_2 x_i)}{(1 + \exp(\beta_1 + \beta_2 x_i))^2} \beta_2.$$

Table 14.1. *Fishing Mode Choice: Data Summary*

Variable	Subsample Averages		
	$y = 1$ Charter	$y = 0$ Pier	All y Overall
Price charter (\$)	75	110	85
Price pier (\$)	121	31	95
In relp	-0.264	1.643	0.275
Sample probability	0.717	0.283	1.000
Observations	452	178	630

Table 14.2. Fishing Mode Choice: Logit and Probit Estimates^a

Regressor	Logit	Model Probit	OLS
Constant	2.053 (12.15)	1.194 (13.34)	0.784 (65.58)
ln relp	-1.823 (-12.61)	-1.056 (-13.87)	-0.243 (-28.15)
-ln L	-206.83	-204.41	-
Pseudo R^2	0.449	0.455	0.463

^a Dependent variable $y = 1$ if charter boat fishing and $y = 0$ if pier fishing. Regressor $x = \ln relp$, the natural logarithm of the price of charter boat fishing relative to pier fishing. Intercept and slope parameter estimates with t -statistics in parentheses are from ML estimation of logit and probit models and from OLS estimation.

Since $\hat{\beta}_{2,\text{LOGIT}} < 0$ it follows that $dp_i/dx_i < 0$, as expected. The actual magnitude of the marginal effect varies with the point of evaluation x_i (see Section 14.3.2). An approximation for the logit model, though not other models, is that $dp_i/dx_i \simeq \bar{y}(1 - \bar{y})\hat{\beta}_2 = -0.370$. An OLS regression instead provides a direct estimate of -0.243 .

An alternative model is the **probit model** (see Section 14.3.5), which specifies

$$p_i = \Pr[y_i = 1|x_i] = \Phi(\beta_1 + \beta_2 x_i),$$

where $\Phi(\cdot)$ is the cumulative distribution function for the standard normal, so $p_i = \int_{-\infty}^{\beta_1 + \beta_2 x_i} (2\pi)^{-1/2} e^{-z^2/2} dz$. The ML coefficients are given in the second column of Table 14.2 and differ appreciably from the logit coefficients. Since different specifications are being estimated the coefficients are not comparable. This is similar to our inability to compare coefficients in models with conditional mean $\mathbf{x}'\boldsymbol{\beta}$ and $\exp(\mathbf{x}'\boldsymbol{\beta})$. For the probit model $dp_i/dx_i = \phi(\beta_1 + \beta_2 x_i)\beta_2$, where $\phi(\cdot)$ is the density for the standard normal. So again $dp_i/dx_i < 0$ since $\hat{\beta}_{2,\text{PROBIT}} < 0$.

Although the slope coefficients necessarily differ across the models, from Table 14.2 the t -statistics are similar and are all very high. The log-likelihood for the probit model is 2.42 higher than that for the logit, favoring the probit model since both models use the same number of parameters. In many other examples there is little difference in $\ln L$ across the models. The predicted probabilities from the three models are plotted as a function of x in Figure 14.1. In OLS we assume that $\Pr[y_i = 1|x_i] = \beta_1 + \beta_2 x_i$ is linear in x_i , whereas the nonlinear functions for logit and probit are essentially equivalent.

14.3. Logit and Probit Models

We now provide more formal theory for these models. We present binary outcomes as a direct extension of the coin-toss example of introductory statistics to situations where the probability of success is modeled to depend on regressors. Two commonly used parameterizations lead to the logit and probit models. Motivation for these parameterizations, using latent variables, is deferred to Section 14.4.

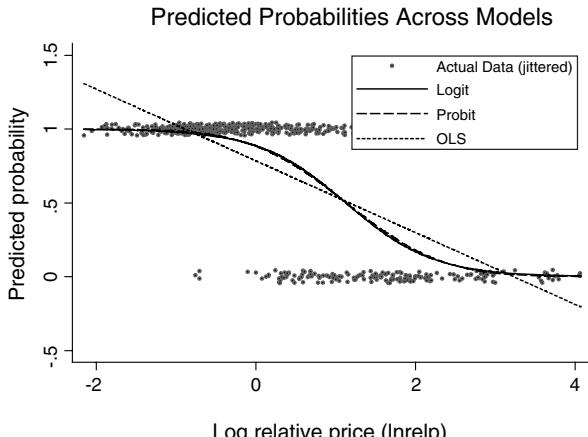


Figure 14.1: Charter boat fishing: predicted probability from logit and probit models and OLS prediction when the single regressor is the natural logarithm of relative price. Actual outcomes of 1 or 0 are also plotted after jittering for readability. Data for 620 individuals.

14.3.1. General Binary Outcome Model

For binary outcome data the dependent variable y takes one of two values. We let

$$y = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1 - p. \end{cases}$$

There is no loss of generality in setting the values to 1 and 0 if all that is being modeled is p , which determines the probability of the outcome. In introductory statistics this model describes the outcome of a coin toss where heads leads to $y = 1$ and occurs with probability p .

A regression model is formed by parameterizing the probability p to depend on a regressor vector \mathbf{x} and a $K \times 1$ parameter vector β . The commonly used models are of single-index form with **conditional probability** given by

$$p_i \equiv \Pr[y_i = 1 | \mathbf{x}] = F(\mathbf{x}_i' \boldsymbol{\beta}), \quad (14.1)$$

where $F(\cdot)$ is a specified function. To ensure that $0 \leq p \leq 1$ it is natural to specify $F(\cdot)$ to be a cumulative distribution function.

Table 14.3 presents the most commonly used binary outcome models. **The logit model** arises if $F(\cdot)$ is the cdf of the logistic distribution and the **probit model** arises if $F(\cdot)$ is the standard normal cdf. Note that if $F(\cdot)$ is a cdf, then this cdf is only being used to model the parameter p and does not denote the cdf of y itself. The less-used **complementary log-log model** arises if $F(\cdot)$ is the cdf of the extreme value distribution. It differs from the other models in being asymmetric around zero and is used when one of the outcomes is rare. The **linear probability model** does not use a cdf and instead lets $p_i = \mathbf{x}_i' \boldsymbol{\beta}$.

Table 14.3. Binary Outcome Data: Commonly Used Models

Model	Probability ($p = \Pr[y = 1 \mathbf{x}]$)	Marginal Effect ($\partial p / \partial x_j$)
Logit	$\Lambda(\mathbf{x}'\beta) = \frac{e^{\mathbf{x}'\beta}}{1 + e^{\mathbf{x}'\beta}}$	$\Lambda(\mathbf{x}'\beta)[1 - \Lambda(\mathbf{x}'\beta)]\beta_j$
Probit	$\Phi(\mathbf{x}'\beta) = \int_{-\infty}^{\mathbf{x}'\beta} \phi(z)dz$	$\phi(\mathbf{x}'\beta)\beta_j$
Complementary log-log	$C(\mathbf{x}'\beta) = 1 - \exp(-\exp(\mathbf{x}'\beta))$	$\exp(-\exp(\mathbf{x}'\beta))\exp(\mathbf{x}'\beta)\beta_j$
Linear probability	$\mathbf{x}'\beta$	β_j

14.3.2. Marginal Effects

Interest lies in determining the **marginal effect** of change in a regressor on the conditional probability that $y = 1$. For general probability model (14.1) and change in the j th regressor, assumed to be continuous, this is

$$\frac{\partial \Pr[y_i = 1|\mathbf{x}_i]}{\partial x_{ij}} = F'(\mathbf{x}'_i\beta)\beta_j, \quad (14.2)$$

where $F'(z) = \partial F(z) / \partial z$. The marginal effects differ with the point of evaluation \mathbf{x}_i , as for any nonlinear model, and differ with different choices of $F(\cdot)$. The last column of Table 14.3 gives the marginal effects for the common binary outcome models.

Marginal effects for nonlinear models are discussed in Section 5.2.4. Given a specific model there are several ways to compute an average marginal effect. It is best to use $N^{-1} \sum_i F'(\mathbf{x}'_i\hat{\beta})\hat{\beta}_j$, the sample average of the marginal effects. Some programs instead evaluate at the sample average of the regressors, $F'(\bar{\mathbf{x}}'\hat{\beta})\hat{\beta}_j$. An easily constructed measure evaluates at \bar{y} , the sample average of y , so that $F(\mathbf{x}'\beta) = \bar{y}$ and $F'(\mathbf{x}'\beta) = F'(F^{-1}(\bar{y}))$. This is especially simple for the logit model as then this yields estimated marginal effect $\bar{y}(1 - \bar{y})\hat{\beta}_j$. Further discussion for specific models is given in Sections 14.3.4–14.3.7.

Many studies instead report only the regression coefficients. The standard binary outcome models are single-index models, so the ratio of coefficients for two different regressors equals the ratio of the marginal effects. The sign of the coefficient gives the sign of the marginal effect, since $F'(\cdot) > 0$. The coefficients can be used to obtain an upper bound on the marginal effects. For the logit model $\partial p / \partial x_j \leq 0.25\hat{\beta}_j$, since $\Lambda(\mathbf{x}'\beta)(1 - \Lambda(\mathbf{x}'\beta)) \leq 0.25$, with maximum when $\Lambda(\mathbf{x}'\beta) = 0.5$ and $\mathbf{x}'\beta = 0$. For the probit model $\partial p / \partial x_j \leq 0.4\hat{\beta}_j$, since $\phi(\mathbf{x}'\beta) \leq 1/\sqrt{2\pi} \approx 0.4$, with maximum when $\Phi(\mathbf{x}'\beta) = 0.5$ and $\mathbf{x}'\beta = 0$.

14.3.3. ML Estimation

We consider estimation given a sample $(y_i, \mathbf{x}_i), i = 1, \dots, N$, where we assume independence over i . Results are given for p_i defined in (14.1), with specialization to logit and probit specifications given later.

MLE for General Binary Outcome Models

The outcome is Bernoulli distributed, the binomial distribution with just one trial. A very convenient compact notation for the density of y_i , or more formally its **probability mass function**, is

$$f(y_i | \mathbf{x}_i) = p_i^{y_i} (1 - p_i)^{1-y_i}, \quad y_i = 0, 1, \quad (14.3)$$

where $p_i = F(\mathbf{x}'_i \boldsymbol{\beta})$. This yields probabilities p_i and $(1 - p_i)$ since $f(1) = p^1(1 - p)^0 = p$ and $f(0) = p^0(1 - p)^1 = 1 - p$.

The density (14.3) implies log density $\ln f(y_i) = y_i \ln p_i + (1 - y_i) \ln(1 - p_i)$. Given independence over i and model (14.1) for p_i , the log-likelihood function is

$$\mathcal{L}_N(\boldsymbol{\beta}) = \sum_{i=1}^N \left\{ y_i \ln F(\mathbf{x}'_i \boldsymbol{\beta}) + (1 - y_i) \ln(1 - F(\mathbf{x}'_i \boldsymbol{\beta})) \right\}. \quad (14.4)$$

Differentiating with respect to $\boldsymbol{\beta}$, we have that the MLE $\hat{\boldsymbol{\beta}}_{\text{ML}}$ solves

$$\sum_{i=1}^N \left\{ \frac{y_i}{F_i} F'_i \mathbf{x}_i - \frac{1 - y_i}{1 - F_i} F'_i \mathbf{x}_i \right\} = \mathbf{0},$$

where $F_i = F(\mathbf{x}'_i \boldsymbol{\beta})$, $F'_i = F'(\mathbf{x}'_i \boldsymbol{\beta})$, and $F'(z) = \partial F(z)/\partial z$. Converting to fractions with common denominator $F_i(1 - F_i)$ and simplifying yields the ML first-order conditions

$$\sum_{i=1}^N \frac{y_i - F(\mathbf{x}'_i \boldsymbol{\beta})}{F(\mathbf{x}'_i \boldsymbol{\beta})(1 - F(\mathbf{x}'_i \boldsymbol{\beta}))} F'(\mathbf{x}'_i \boldsymbol{\beta}) \mathbf{x}_i = \mathbf{0}. \quad (14.5)$$

There is no explicit solution for $\hat{\boldsymbol{\beta}}_{\text{MLE}}$, but the Newton–Raphson iterative procedure usually converges very quickly since for the probit and logit models, at least, the log-likelihood is globally concave.

Consistency of the MLE

The MLE is **consistent** if the conditional density of y given \mathbf{x} is correctly specified. Since the density here must be the Bernoulli, the only possible misspecification is that the Bernoulli probability is misspecified. So the MLE is consistent if $p_i \equiv F(\mathbf{x}'_i \boldsymbol{\beta})$ and is inconsistent otherwise.

More formally, note that for binary data, $E[y] = 1 \times p + 0 \times (1 - p) = p$. Given (14.1) this implies

$$E[y_i | \mathbf{x}_i] = F(\mathbf{x}'_i \boldsymbol{\beta}), \quad (14.6)$$

which in turn implies that the left-hand side of the first-order equations (14.5) has expected value zero, the essential condition for consistency. This special result of consistency provided the conditional mean is correctly specified holds for LEF densities (see Section 5.7.3) and the Bernoulli is an LEF density.

Distribution of the MLE

Given correct specification of the density, $\widehat{\beta}_{\text{ML}} \stackrel{a}{\sim} \mathcal{N}[\beta, (-E[\partial^2 \mathcal{L}_N / \partial \beta \partial \beta'])^{-1}]$ (see Section 5.6.4). Differentiating (14.4) with respect to β' , and taking minus the expected value yields the estimated **asymptotic variance** matrix

$$\widehat{\text{V}}[\widehat{\beta}_{\text{ML}}] = \left(\sum_{i=1}^N \frac{1}{F(\mathbf{x}'_i \widehat{\beta})(1 - F(\mathbf{x}'_i \widehat{\beta}))} F'(\mathbf{x}'_i \widehat{\beta})^2 \mathbf{x}_i \mathbf{x}'_i \right)^{-1}, \quad (14.7)$$

where simplification occurs because $E[y_i - F(\mathbf{x}'_i \beta)] = 0$. This variance matrix is of the simple form $(\sum_i \widehat{w}_i \mathbf{x}_i \mathbf{x}'_i)^{-1}$, where the weights \widehat{w}_i are given in (14.7).

Since consistency requires only correct specification of the conditional mean or probability, it is natural to consider the quasi-MLE (see Section 5.7) and base inference on the sandwich form of the variance matrix $\mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1}$ rather than $-\mathbf{A}^{-1}$ used in (14.7). Here

$$\text{V}[y_i | \mathbf{x}_i] = F(\mathbf{x}'_i \beta)(1 - F(\mathbf{x}'_i \beta)), \quad (14.8)$$

since $\text{V}[y] = (1 - p)^2 \times p + (0 - p)^2 \times (1 - p) = p(1 - p)$. Some algebra shows that this implies that $\mathbf{A} = -\mathbf{B}$ and hence $\mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1} = -\mathbf{A}^{-1}$, assuming independence over i . The only way that (14.8) does not hold is if $p \neq F(\mathbf{x}' \beta)$ in which case the MLE would suffer from the more fundamental problem of inconsistency.

Binary outcome models are unusual in that there is no advantage in using the sandwich form if data are independent over i . The only reason for moving to a robust variance matrix estimate is if observations are correlated over i as the result of clustering. Then the robust estimate needs to be one that is robust to clustering (see Section 24.5) rather than to misspecification of the conditional variance.

14.3.4. Logit Model

The **logit model** or **logistic regression model** specifies

$$p = \Lambda(\mathbf{x}' \beta) = \frac{e^{\mathbf{x}' \beta}}{1 + e^{\mathbf{x}' \beta}}, \quad (14.9)$$

where $\Lambda(\cdot)$ is the logistic cdf (see Section 14.4.1 for further details), with $\Lambda(z) = e^z / (1 + e^z) = 1 / (1 + e^{-z})$.

The logit MLE first-order conditions (14.5) simplify to

$$\sum_{i=1}^N (y_i - \Lambda(\mathbf{x}'_i \beta)) \mathbf{x}_i = \mathbf{0}, \quad (14.10)$$

since $\Lambda'(z) = \Lambda(z)[1 - \Lambda(z)]$. So the raw residual $y_i - \Lambda(\mathbf{x}'_i \beta)$ is orthogonal to the regressors, similar to OLS regression. This simple form arises because $\Lambda(\cdot)$ is the canonical link function (see Section 5.7.4) for the Bernoulli density.

If the regressors \mathbf{x}_i include an intercept, then (14.10) implies that $\sum_i (y_i - \Lambda(\mathbf{x}'_i \widehat{\beta})) = 0$, so the logit residuals sum to zero. This implies that the average in-sample predicted probability $N^{-1} \sum_i \Lambda(\mathbf{x}'_i \widehat{\beta})$ necessarily equals the sample frequency \bar{y} .

The **marginal effects** for the logit model can be fairly easily obtained from the coefficients, since $\partial p_i / \partial x_{ij} = p_i(1 - p_i)\beta_j$, where $p_i = \Lambda_i = \Lambda(\mathbf{x}'_i \boldsymbol{\beta})$. Evaluating at $p_i = \bar{y}$ yields a crude estimated marginal effect of $\bar{y}(1 - \bar{y})\hat{\beta}_j$. For $0.3 < p_i < 0.7$, for example, $\partial p_i / \partial x_{ij}$ lies between $0.21\beta_j$ and $0.25\beta_j$. For data where $p_i \simeq 0.0$, in which case most outcomes are zero, $\partial p_i / \partial x_{ij} = p_i\beta_j$ so β_j gives the proportionate effect on the probability that $y_i = 1$ as x_{ij} changes.

In the statistics literature a very common interpretation of the coefficients is in terms of **marginal effects on the odds ratio rather than on the probability**. For the logit model

$$\begin{aligned} p &= \exp(\mathbf{x}' \boldsymbol{\beta}) / (1 + \exp(\mathbf{x}' \boldsymbol{\beta})) \\ \Rightarrow \frac{p}{1-p} &= \exp(\mathbf{x}' \boldsymbol{\beta}) \\ \Rightarrow \ln \frac{p}{1-p} &= \mathbf{x}' \boldsymbol{\beta}. \end{aligned} \tag{14.11}$$

Here $p/(1 - p)$ measures the probability that $y = 1$ relative to the probability that $y = 0$ and is called the **odds ratio or relative risk**. For example, consider a pharmaceutical drug study where $y = 1$ denotes survival and $y = 0$ denotes death and regressors include a measure of drug intake. An odds ratio of 2 means that the odds of survival are twice those of death. **For the logit model the log-odds ratio is linear in the regressors.**

Statistical analyses and packages use the second equality in (14.11). Suppose the j th regressor increases by one unit. Then $\exp(\mathbf{x}' \boldsymbol{\beta})$ increases to $\exp(\mathbf{x}' \boldsymbol{\beta} + \beta_j) = \exp(\mathbf{x}' \boldsymbol{\beta}) \times \exp(\beta_j)$. It follows from (14.11) that the odds ratio has increased by a multiple $\exp(\beta_j)$. Thus a logit model slope parameter of 0.1, for example, means that a **one unit increase in the regressor multiplies the initial odds ratio by $\exp(0.1) \simeq 1.105$.** This is a proportionate increase of 0.105 times the initial odds ratio, so the **relative probability of survival increases by 10.5%.** This interpretation of the logit model is **widely used in biostatistics applications.**

For economists it is more natural to interpret either the second or third equality in (14.11) as implying that β_j is a **semi-elasticity**. Then, taking a calculus approach, we interpret a logit model slope parameter of 0.1 as meaning that a one-unit increase in the regressor increases the odds ratio by a multiple 0.1. This coincides exactly with the interpretation used in statistics for very small β_j , since then $\exp(\beta_j) - 1 \simeq \beta_j$.

14.3.5. Probit Model

The **probit model** specifies the conditional probability

$$p = \Phi(\mathbf{x}' \boldsymbol{\beta}) = \int_{-\infty}^{\mathbf{x}' \boldsymbol{\beta}} \phi(z) dz, \tag{14.12}$$

where $\Phi(\cdot)$ is the standard normal cdf, with derivative $\phi(z) = (1/\sqrt{2\pi}) \exp(-z^2/2)$, which is the standard normal density function.

The probit MLE first-order conditions are that

$$\sum_{i=1}^N w_i (y_i - \Phi(\mathbf{x}'_i \boldsymbol{\beta})) \mathbf{x}_i = \mathbf{0},$$

where, unlike the logit model, the weight $w_i = \phi(\mathbf{x}'_i \boldsymbol{\beta}) / [\Phi(\mathbf{x}'_i \boldsymbol{\beta})(1 - \Phi(\mathbf{x}'_i \boldsymbol{\beta}))]$ varies across observations.

The probit model marginal effects are $\partial p_i / \partial x_{ij} = \phi(\mathbf{x}'_i \boldsymbol{\beta})\beta_j = \phi(\Phi^{-1}(p_i))\beta_j$, where $p_i = \Phi(\mathbf{x}'_i \boldsymbol{\beta})$. There are no further simplifications similar to those for the logit model, though $\partial p_i / \partial x_{ij} \leq 0.40\beta_j$ since $\phi(z) \leq \phi(0.5) = 1/\sqrt{2\pi}$.

The probit model is not as simple as the logit model. It is nevertheless widely used as it is the natural model if the starting point is a latent normal regression model (see Section 14.4).

14.3.6. OLS Estimation

A simple alternative to either logit or probit is **OLS regression of y on \mathbf{x}** . This has the **obvious deficiency** that it is possible to obtain predicted probabilities $\mathbf{x}'_i \hat{\boldsymbol{\beta}}$ that are **negative or that exceed one**.

The OLS estimator is nonetheless useful as an exploratory tool. In practice it provides a **reasonable direct estimate of the sample-average marginal effect on the probability that $y = 1$ as \mathbf{x} changes, even though it provides a poor model for individual probabilities**. In practice it provides a good guide to which variables are statistically significant. In many applications it turns out that $0 < \mathbf{x}'_i \hat{\boldsymbol{\beta}} < 1$ for all sample observations, in which case OLS is more reasonable.

If the OLS estimator is used then **standard errors should correct for heteroskedasticity**. Linear regression is justified if the probability $p_i = \mathbf{x}'_i \boldsymbol{\beta}$. Then $y_i | \mathbf{x}_i$ has mean $\mathbf{x}'_i \boldsymbol{\beta}$ and heteroskedastic variance $\mathbf{x}'_i \boldsymbol{\beta}(1 - \mathbf{x}'_i \boldsymbol{\beta})$ that varies with \mathbf{x}_i .

In theory more efficient ML estimation is possible if $p_i = \mathbf{x}'_i \boldsymbol{\beta}$. From (14.5) the ML first-order conditions are $\sum_i \mathbf{x}_i(y_i - \mathbf{x}'_i \boldsymbol{\beta}) / [\mathbf{x}'_i \boldsymbol{\beta}(1 - \mathbf{x}'_i \boldsymbol{\beta})] = \mathbf{0}$. The estimator can be numerically unstable as it places very high weight on to observations with $\mathbf{x}'_i \boldsymbol{\beta}$ close to 0 or 1. Moreover, the efficiency gains compared to OLS are often small.

Although OLS estimation with heteroskedastic standard errors can be a useful exploratory data analysis tool, it is best to use the logit or probit MLE for final data analysis.

14.3.7. Choosing a Binary Model

Which model should be used – logit or probit? This question is explored in this section.

Theoretical Considerations

Theoretically the answer depends on the dgp, which is unknown. Unlike other applications of ML there is no problem in specifying the distribution – the only possible distribution for a (0, 1) variable is the Bernoulli. The problem lies in specifying a functional form for the parameter of this distribution. If the dgp has $p = \Lambda(\mathbf{x}' \boldsymbol{\beta})$ then a logit model should be used, and estimators based on other models such as probit are potentially inconsistent. Similar qualitative conclusions hold if instead the dgp has

$p = \Phi(\mathbf{x}'\boldsymbol{\beta})$, in which case the probit model should be used. It is very unlikely that $p = \mathbf{x}'\boldsymbol{\beta}$ since then p is not restricted to be between 0 and 1.

The theoretical consequences of model misspecification, however, are not as great as this. If the regressors are distributed such that the mean of each regressor, conditional on the linear combination $\mathbf{x}'\boldsymbol{\beta}$, is linear in $\mathbf{x}'\boldsymbol{\beta}$, then choosing the wrong function F can be shown to affect all slope parameters equally so that the ratio of slope parameters is constant across models; see Ruud (1983). This condition is satisfied by the family of spherical distributions, including the multivariate normal.

The logit model has a relatively simple form for the first-order conditions and asymptotic distribution. Berkson (1951), who popularized the logit model, gave this as one of several reasons for preferring the logit model to the original probit model. Within the framework of generalized linear models, which are widely used in biostatistics, the logit model is the natural model as it corresponds to use of the canonical link for the binomial distribution. The interpretation of coefficients in terms of the log-odds ratio is also an attraction of the logit model.

Yet another motivation for the logit model is discriminant analysis. In **discriminant analysis** both y and \mathbf{x} are random variables; \mathbf{x} is observed but y is not observed. Given \mathbf{x} we need to determine whether y equals zero or one. A classic example is classifying what type of humanoid ($y = 0$ or 1) a skull belongs to given various dimensions (\mathbf{x}) of the skull. If the conditional distributions of the characteristics \mathbf{x} given y are multivariate normal distributed, the posterior probability of y given \mathbf{x} is similar to the probability in the logit model. For more details, see Amemiya (1981, pp. 1507–1510) and Maddala (1983, pp. 17–21).

The probit model, in contrast, has the attraction of being motivated by a latent normal random variable (see Section 14.4) and extends naturally to Tobit models (see Chapter 16). For these reasons many economists use the probit model.

Empirical Considerations

Empirically, either logit and probit can be used. There is often little difference between the predicted probabilities from probit and logit models. The difference is greatest in the tails where probabilities are close to 0 or 1. The difference is much less if interest lies only in marginal effects averaged over the sample rather than for each individual.

The natural metric to use to compare models is the fitted log-likelihood, since there is agreement that the log-likelihood is correct, given the model for p_i , and the logit and probit models have the same number of parameters. Thus for each model compute

$$\mathcal{L}_N(\widehat{\boldsymbol{\beta}}) = \sum_i \{y_i \ln \widehat{p}_i + (1 - y_i) \ln(1 - \widehat{p}_i)\},$$

where $\widehat{p}_i = \Lambda(\mathbf{x}'_i \widehat{\boldsymbol{\beta}}_{\text{Logit}})$ or $\widehat{p}_i = \Phi(\mathbf{x}'_i \widehat{\boldsymbol{\beta}}_{\text{Probit}})$. Often the fitted log-likelihoods are very similar for the two models, again suggesting little additional gain to using one rather than the other model. For more formal nonnested model tests see Pesaran and Pesaran (1995) and Section 8.5.

The different models do yield quite different estimates $\hat{\beta}$ of regression parameters. However, this is just an artifact of using different formulas for the probabilities. It is more meaningful to compare the marginal effect across models, as this measure is scaled similarly across the three models. From Section 14.2.3, $\partial p / \partial x_j \leq 0.25\hat{\beta}_j$ for logit, $\partial p / \partial x_j \leq 0.4\hat{\beta}_j$ for probit, and $\partial p / \partial x_j = \hat{\beta}_j$ for OLS. This suggests the **rule of thumb**

$$\begin{aligned}\hat{\beta}_{\text{Logit}} &\simeq 4\hat{\beta}_{\text{OLS}}, \\ \hat{\beta}_{\text{Probit}} &\simeq 2.5\hat{\beta}_{\text{OLS}}, \\ \hat{\beta}_{\text{Logit}} &\simeq 1.6\hat{\beta}_{\text{Probit}}.\end{aligned}\tag{14.13}$$

Amemiya (1981, p. 1488) demonstrates that these comparisons work quite well for slope parameters if $0.1 \leq p \leq 0.9$. Greater departures across the models occur in the tails. For logit an alternative method, based on (14.18) given later; uses $\hat{\beta}_{\text{Logit}} \simeq (\pi/\sqrt{3})\hat{\beta}_{\text{Probit}}$.

Endogenous Regressors

Logit and probit models can be extended to handle many of the complications that commonly arise in microeconomic analysis. In particular, **endogenous regressors are accommodated using methods similar to those for censored data given in Section 16.8.2, and panel data methods are presented in Chapter 23.**

For such complications it is easier to work with the linear probability model, since then standard linear model methods can be applied provided standard errors adjust for heteroskedasticity. Even if logit and probit models are ultimately used, a linear model can be useful for exploratory analysis.

14.3.8. Determining Model Adequacy

Model diagnostics and selection for nonlinear models were presented in Section 8.7. Here we consider specialization to binary outcome models. There is no single best measure, and statistical packages accordingly report several measures detailed in Amemiya (1981) and Maddala (1983).

Pseudo- R^2

A standard measure of goodness of fit in the linear regression model is R^2 . Generalizations to nonlinear models are called **pseudo- R^2** , with several generalizations possible.

A preferred measure is the relative gain measure denoted R_{RG}^2 in Section 8.7.1. This measure is not always computable, but it is for the binary outcome model since Q_{\max} , the maximum possible value of the log-likelihood, is zero. To obtain this result note that the best possible fit is clearly a y^* that predicts $y = 1$ with probability $p = 1$ and $y = 0$ with probability $1 - p = 0$, in which case $f(y^*) = 1$ and $\ln f(y^*) = 0$. Then $R_{\text{RG}}^2 = 1 - (0 - Q_{\text{fit}})/(0 - Q_0) = 1 - Q_{\text{fit}}/Q_0$. This yields the R^2 measure for binary

outcome models proposed by McFadden (1974):

$$\begin{aligned} R_{\text{Binary}}^2 &= 1 - \frac{\mathcal{L}_N(\hat{\beta})}{\mathcal{L}_N(\bar{y})} \\ &= 1 - \frac{\sum_i \{y_i \ln \hat{p}_i + (1 - y_i) \ln(1 - \hat{p}_i)\}}{N[\bar{y} \ln \bar{y} + (1 - \bar{y}) \ln(1 - \bar{y})]}, \end{aligned} \quad (14.14)$$

where $\hat{p}_i = F(\mathbf{x}'_i \hat{\beta})$ and $\bar{y} = N^{-1} \sum_i y_i$.

Additional R^2 measures, many specific to binary data, are given in Amemiya (1981) and Maddala (1983). An obvious one is the squared sample correlation between y_i and \hat{p}_i . One of these additional measures is also attributed to McFadden, and many references give this measure rather than the R^2 in (14.14).

Predicted Outcomes

In the linear regression model goodness of fit is often evaluated by comparison of fitted and actual values. For binary data the fitted value \hat{y} should be binary since y is binary. The criterion $\sum_i (y_i - \hat{y}_i)^2$ gives the number of wrong predictions, which arise if (y, \hat{y}) equals $(1, 0)$ or $(0, 1)$. An obvious prediction rule is to set $\hat{y} = 1$ when $\hat{p} = F(\mathbf{x}' \hat{\beta}) > 0.5$. However, this has the weakness that if most of the sample has $y = 1$ then often $\sum_i (y_i - \hat{y}_i)^2 = n(1 - \bar{y})$ since it is likely that $\hat{p} > 0.5$ and hence $\hat{y} = 1$ for all the observations. Similar problems arise if most of the sample has $y = 0$.

More generally, a range of cutoff values may be considered. Letting $\hat{y} = 1$ when $\hat{p} > c$, we obtain the **receiver operating characteristics** (ROC) curve which plots the fraction of $y = 1$ values correctly classified against the fraction of $y = 0$ values incorrectly specified as the cutoff c varies. For $c = 1$ all values are predicted to be 1, so all $y = 1$ values are correctly specified and all $y = 0$ values are incorrectly specified and the ROC curve takes value $(0, 0)$. Similarly, for $c = 0$ the ROC curve takes value $(1, 1)$.

If the model has no predictive ability the ROC curve is a straight line between these points. The more bowed the curve, and the more area under it, the better the predictive power of the model.

Predicted Probabilities

Since binary data have a simple discrete distribution, an obvious approach is to compare the sample average predicted probability that $y = 1$, $N^{-1} \sum_i \hat{p}_i$, where $\hat{p}_i = F(\mathbf{x}'_i \hat{\beta})$, with the sample frequency \bar{y} . However, this is not useful for the logit model with an intercept, since $N^{-1} \sum_i \hat{p}_i = \bar{y}$ always holds as the ML first-order conditions imply $\sum_i [y_i - \Lambda(\mathbf{x}'_i \hat{\beta})] = 0$. A similar result holds for estimation by OLS; for the probit model the result is not exact but in practice is quite close.

This approach can be used for predictions over subsamples, however, and can then form the basis for the chi-square goodness-of-fit test given in Section 8.2.6.

14.4. Latent Variable Models

A **latent variable** is a variable that is incompletely observed. Latent variables can be introduced into binary outcome models in two different ways. In the first the latent variable is an index of an unobserved propensity for the event of interest to occur. In the second the latent variable is the difference in utility that occurs if the event of interest occurs, which presumes that the binary outcome is a result of individual choice. The latter method makes clear the need to distinguish between regressors that vary across alternatives for a given individual and regressors such as socioeconomic characteristics that for a given individual are invariant across alternatives.

It should be stressed that the binary outcome is Bernoulli distributed, as in Section 14.3. Latent variable models merely provide a rationale for a particular functional form for the Bernoulli parameter.

Latent variable models do provide extensions to multinomial outcomes and censored outcomes (detailed in Chapters 15 and 16). They also provide a framework that permits Bayesian analysis using data augmentation (see Section 13.7). Brief discussion of Bayesian analysis of binary and multinomial data is given in Sections 15.7.2 and 15.8.2.

14.4.1. Index Function Models

In the **index function** formulation interest lies in explaining an underlying unobserved continuous random variable y^* , but all we observe is the binary variable y , which takes value 1 or 0 according to whether or not y^* crosses a threshold. Different distributions for y^* lead to different binary outcome models.

Let y^* be a latent (or unobserved) variable, such as the desire to work if labor supply is being modeled. The natural regression model for y^* is the **index function model**

$$y^* = \mathbf{x}'\boldsymbol{\beta} + u. \quad (14.15)$$

However, this model cannot be estimated as y^* is not observed. Instead, we observe

$$y = \begin{cases} 1 & \text{if } y^* > 0, \\ 0 & \text{if } y^* \leq 0, \end{cases} \quad (14.16)$$

where the threshold of zero is a normalization explained in the following.

Given (14.16),

$$\begin{aligned} \Pr[y = 1|\mathbf{x}] &= \Pr[y^* > 0] \\ &= \Pr[\mathbf{x}'\boldsymbol{\beta} + u > 0] \\ &= \Pr[-u < \mathbf{x}'\boldsymbol{\beta}] \\ &= F(\mathbf{x}'\boldsymbol{\beta}), \end{aligned} \quad (14.17)$$

where F is the cdf of $-u$, which equals the cdf of u in the usual case of density symmetric about 0.

The index function model therefore provides motivation for the functional form of $F(\cdot)$ in (14.1).

Probit and Logit Models

The probit model arises if the error u is standard normal distributed, since then (14.17) yields $\Pr[-u < \mathbf{x}'\beta] = \Phi(\mathbf{x}'\beta)$, where $\Phi(\cdot)$ is the cdf of the standard normal.

Now introduce the **logistic distribution**. In its standard form the logistic has cdf

$$\Lambda(u) = e^u / (1 + e^u), \quad -\infty < u < \infty. \quad (14.18)$$

The density function $\Lambda'(u) = e^u / (1 + e^u)^2$ is symmetric about 0, and a logistic random variable has mean 0 and variance $\pi^2/3 \simeq 1.814^2$.

The logit model arises if the error u is logistic distributed, since then (14.17) yields $\Pr[-u < \mathbf{x}'\beta] = \Lambda(\mathbf{x}'\beta)$. Note that β is scaled differently in the two models due to different $V[u]$.

Identification Considerations

Identification of the single-index model requires a restriction on the variance of u , as the single-index model can only identify β up to scale. All that is observed is whether or not $y^* > 0$, or equivalently whether or not $\mathbf{x}'\beta + u > 0$. However, this is equivalent to whether or not $\mathbf{x}'\beta^+ + u^+ > 0$, where $\beta^+ = a\beta$ and $u^+ = au$ for any $a > 0$. Placing a restriction on the variance of the error (u or u^+) secures uniqueness of β . The error variance is set to one in the probit model and $\pi^2/3$ in the logit model.

The threshold for the index model need not be zero. If more generally $y = 1$ when $y^* > \mathbf{z}'\delta$ then (14.17) becomes $\Pr[y = 1] = F(\mathbf{x}'\beta - \mathbf{z}'\delta)$. Then δ can be separately identified if and only if all components of \mathbf{z} and \mathbf{x} differ. In particular, if both \mathbf{x} and \mathbf{z} include intercepts these cannot be separately identified, so we normalize the threshold intercept to be zero. Note also that the mean of the error distribution needs to be normalized. For the logit and probit models it is set to zero.

Discussion

The index function model implies a direct interpretation of β as the change in the latent variable y^* when \mathbf{x} changes by one unit. Even though y^* is unobserved, this interpretation is meaningful if one uses knowledge of the specified variance of u . For example, a slope parameter of 0.5 in the probit model means a one-unit change in the regressor leads to a 0.5 standard deviation change in y^* , since in this model the variance of y^* equals 1.

Commonly used **extensions** of the index function approach are to ordered discrete choice models (see Section 15.9) and to models for censored and selected samples (see Chapter 16).

14.4.2. Random Utility Models

In the random utility formulation a consumer chooses between alternatives 0 and 1 according to which has the higher satisfaction or utility. The discrete variable y then takes value 1 if alternative 1 has higher utility, and it takes value 0 if alternative 0 has higher utility.

The **additive random utility model** (ARUM) specifies the utilities of alternatives 0 and 1 to be

$$\begin{aligned} U_0 &= V_0 + \varepsilon_0, \\ U_1 &= V_1 + \varepsilon_1, \end{aligned} \tag{14.19}$$

where V_0 and V_1 are **deterministic components of utility** and ε_0 and ε_1 are **random components of utility**. A simple example is $V_0 = \mathbf{x}'\beta_0$ and $V_1 = \mathbf{x}'\beta_1$, though from Section 14.4.3 only $(\beta_1 - \beta_0)$ is then identified.

The alternative with higher utility is chosen. We observe $y = 1$, say, if $U_1 > U_0$. Owing to the presence of the random components of utility this is a random event with

$$\begin{aligned} \Pr[y = 1] &= \Pr[U_1 > U_0] \\ &= \Pr[V_1 + \varepsilon_1 > V_0 + \varepsilon_0] \\ &= \Pr[\varepsilon_0 - \varepsilon_1 < V_1 - V_0] \\ &= F(V_1 - V_0), \end{aligned} \tag{14.20}$$

where F is the cdf of $(\varepsilon_0 - \varepsilon_1)$. This yields $\Pr[y = 1] = F(\mathbf{x}'\beta)$ if $V_1 - V_0 = \mathbf{x}'\beta$.

The ARUM requires a scale normalization since if $U_1 > U_0$ then $aU_1 > aU_0$. This is usually done by specifying the variance of $\varepsilon_0 - \varepsilon_1$ or the variances of ε_0 and ε_1 .

Different specifications for the distributions of ε_0 and ε_1 give different $F(\cdot)$ and hence different discrete choice models. The random utility formulation is especially useful for specifying unordered multinomial choice models (see Section 15.5).

Probit and Logit Models

An obvious choice for error distribution in (14.19) is that ε_0 and ε_1 are normal. Then $(\varepsilon_0 - \varepsilon_1)$ is normally distributed. Normalization of the variance of $(\varepsilon_0 - \varepsilon_1)$ to unity gives the probit model since then $F(\cdot)$ in (14.20) is the standard normal cdf.

Now introduce the **type 1 extreme value distribution** or **log Weibull distribution**. Then the random variable ε has density

$$f(\varepsilon) = e^{-\varepsilon} \exp(-e^{-\varepsilon}), \quad -\infty < \varepsilon < \infty, \tag{14.21}$$

and cdf $F(\varepsilon) = \exp(-e^{-\varepsilon})$. The extreme value distributions, rarely used in econometrics, are obtained as limiting distributions as $N \rightarrow \infty$ of the maximum of N random variables drawn from the same distribution. The type 1 extreme value distribution is a special case that is right-skewed over $(-\infty, \infty)$ with most of the mass between -2 and 5 . It has median $-\ln(-\ln(0.5)) \simeq 0.36651$, mean $\Gamma'(1) \simeq 0.57722$, where $\Gamma'(x)$ denotes the derivative of the gamma function, and variance $\pi^2/6 \simeq 1.28255^2$. The distribution is well approximated by a log-normal.

The logit model arises if ε_0 and ε_1 are assumed to be independent type 1 extreme value distributed. Then the difference $(\varepsilon_0 - \varepsilon_1)$ can be shown to be logistic distributed (see Johnson and Kotz, 1970), so $F(\cdot)$ in (14.20) is the logistic cdf.

An alternative derivation of this result, working directly with the extreme value distribution, is given later in Section 14.8. The derivation indicates the difficulty in

obtaining closed-form solutions for probabilities when the ARUM is **extended** to choice among three or more alternatives in Section 15.5. Recent computational advances permit estimation even in the absence of a closed-form solution.

14.4.3. Alternative-Varying Regressors

In most applications of binary choice models, some regressors vary across individuals, but regressors do not necessarily vary across alternatives.

At the one extreme regressors do not vary across alternatives. For example, in labor supply models of the decision to work, socioeconomic characteristics such as income and gender do not vary across alternatives. A potential regressor, the wage rate, does vary across the alternatives of work or not work, but this regressor is usually not included as it is only observed for those who choose to work.

At the other extreme all regressors may vary across alternatives. For example, in transportation mode choice models the regressors may be the time cost and money cost of the two modes of transportation.

A general hybrid ARUM defines the deterministic components of utility in (14.19) to be

$$V_{ij} = \mathbf{z}'_{ij}\boldsymbol{\alpha}_j + \mathbf{w}'_i\gamma_j, \quad j = 0, 1, \quad (14.22)$$

where \mathbf{z}_{ij} are regressors that take different values across the two alternatives, whereas \mathbf{w}_i are individual characteristics that do not vary with the choice. Then (14.20) yields

$$\Pr[y_i = 1] = F(\mathbf{z}'_{i1}\boldsymbol{\alpha}_1 - \mathbf{z}'_{i0}\boldsymbol{\alpha}_0 + \mathbf{w}'_i(\gamma_1 - \gamma_0)).$$

For **alternative-invariant regressors** only the parameter difference $(\gamma_1 - \gamma_0)$ can be identified. For **alternative-varying regressors** that do vary across alternatives and across individuals the coefficients can vary over alternatives, but it is customary to set $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_0 = \boldsymbol{\alpha}$. For example, the loss of utility resulting from a one-dollar increase in travel costs is expected to be the same across different transportation modes. Thus the ARUM leads to

$$\Pr[y_i = 1] = F((\mathbf{z}'_{i1} - \mathbf{z}'_{i0})'\boldsymbol{\alpha} + \mathbf{w}'_i(\gamma_1 - \gamma_0)), \quad (14.23)$$

which is the original binary choice model (14.1) where the regressors are alternative-invariant regressors \mathbf{w} and the difference across alternatives of alternative-varying regressors \mathbf{z} .

14.5. Choice-Based Samples

Choice-based sampling arises whenever selection of the sample is determined in part by values taken by the dependent variable y , rather than being completely random or being based in part by values taken by \mathbf{x} .

Discrete data models are a leading example since surveys often deliberately oversample choices that are made infrequently. For example, if few people choose to

commute by bus, an oversampling of bus riders may be undertaken. In the medical literature the same problem arises with **case-control analysis** where, for example, a binary data analysis may be based on a full sample of those who had a heart attack and a subsample of people with similar characteristics who did not have a heart attack. The standard term choice-based sampling is a little misleading since it does not arise from individual choice.

To see the inconsistency of standard binary choice methods, consider estimation of the logit model when the only regressor is the intercept. Then $\Lambda(\mathbf{x}'_i \boldsymbol{\beta}) = \Lambda(\boldsymbol{\beta})$ and the logit MLE first-order conditions become $N^{-1} \sum_i (y_i - \Lambda(\boldsymbol{\beta})) = 0$, so $\hat{\boldsymbol{\beta}} = \ln(\bar{y}/(1 - \bar{y}))$. Consistency of $\hat{\boldsymbol{\beta}}$ clearly requires a random sample because, for example, oversampling $y = 1$ leads to overestimation of \bar{y} and hence $\hat{\boldsymbol{\beta}}$.

Methods to obtain consistent estimates given endogenous sampling such as choice-based sampling are covered in detail in Section 24.4. Analysis is straight-forward if the degree of oversampling is known. Let Q_1 denote the fraction of the population with $y = 1$ and $H_1 = \bar{y}$ denote the fraction of the sample with $y = 1$. Similarly define $Q_0 = 1 - Q_1$ and $H_0 = 1 - H_1$. Then consistent estimation is possible using the **weighted MLE** proposed by Manski and Lerman (1977). For binary outcome models this maximizes the weighted log-likelihood

$$\mathcal{L}_N^W(\boldsymbol{\beta}) = \sum_{i=1}^N \left\{ \left(\frac{Q_1}{H_1} \right) y_i \ln F(\mathbf{x}'_i \boldsymbol{\beta}) + \left(\frac{Q_0}{H_0} \right) (1 - y_i) \ln(1 - F(\mathbf{x}'_i \boldsymbol{\beta})) \right\}.$$

For example, if outcomes $y = 1$ are oversampled, then $Q_1/H_1 < 1$ and the oversampled observations with $y = 1$ are downweighted. This estimator is easily implemented using any program for binary outcome models that permits weighting of observations. Then observations with $y = 1$ are given weight Q_1/H_1 and observations with $y = 0$ are given weight Q_0/H_0 .

A detailed summary of ML methods for choice-based sampling of binary and multinomial data, including methods when Q_1 and Q_0 are unknown, is given in Amemiya (1985, Section 9.5). The weighted MLE is inefficient but simple to implement and the efficiency loss may not be great. Manski and McFadden (1981a) proposed a variation that is more efficient (see Amemiya and Vuong, 1987). Cosslett (1981a,b) proposed further refinements that are fully efficient but impractical to implement. Imbens (1992) and Lancaster and Imbens (1996) proposed GMM estimation as an alternative method that is feasible to implement and is fully efficient. King and Zeng (2001) give a summary for the binary logit model; additionally, they consider small-sample corrections that, even with oversampling, make a difference when the population probability of interest occurs with low probability. For further details see Section 24.4.

The epidemiological literature has focused on the logit model for case-control studies. The method is attributed to Prentice and Pyke (1979). See Breslow (1996), especially his Section 4.3, which discusses links between the econometrics and epidemiological literature.

14.6. Grouped and Aggregate Data

In some applications only grouped or aggregate data may be available, yet individual behavior is felt to be best modeled by a binary choice model. Grouping poses no problem when the grouping is based on unique values of the regressors and there are many observations per unique value of the regressors. We begin with this simple example before moving to more realistic ones.

14.6.1. Berkson's Minimum Chi-Square Estimator

Suppose the regressor vector \mathbf{x}_i , $i = 1, \dots, N$, takes only T distinct values, where T is much smaller than N . Then for each value of the regressors we have multiple observations on y . This type of grouped data is called **many observations per cell**. It can arise particularly in experimental data where \mathbf{x} is of low dimension and is set by experimental design to just a few values. Let \mathbf{x}_t , $t = 1, \dots, T$, be the T distinct values, N_t be the number of observations on y_t for the t th distinct value of \mathbf{x} , so $\sum_{t=1}^T N_t = N$, and \bar{p}_t be the proportion of times $y_i = 1$ when $\mathbf{x}_i = \mathbf{x}_t$. Note that the subscript t is being used to denote grouping and does not necessarily denote time.

For individual i with $\mathbf{x}_i = \mathbf{x}_t$, the Bernoulli probability is

$$p_t = \Pr[y_i = 1 | \mathbf{x}_i = \mathbf{x}_t] = F(\mathbf{x}'_t \boldsymbol{\beta}), \quad (14.24)$$

as before. Inverting (14.24) implies that

$$F^{-1}(p_t) = \mathbf{x}'_t \boldsymbol{\beta}.$$

Now p_t is unknown but can be estimated by \bar{p}_t , so Berkson proposed regressing $F^{-1}(\bar{p}_t)$ on \mathbf{x}_t . Thus we estimate by LS the transformation model

$$F^{-1}(\bar{p}_t) = \mathbf{x}'_t \boldsymbol{\beta} + v_t, \quad t = 1, \dots, T. \quad (14.25)$$

The error term $v_t = F^{-1}(\bar{p}_t) - F^{-1}(p_t)$ is heteroskedastic with variance that will decrease as N_t increases, since then \bar{p}_t is a better estimate of p_t , and will also depend on the shape of $F(\cdot)$. By Taylor series expansion (see Amemiya (1981, p. 1498) or Maddala (1983, p. 31)), v_t has variance that can be consistently estimated by

$$\tilde{\sigma}_t^2 = \frac{\bar{p}_t(1 - \bar{p}_t)}{N_t[F'(F^{-1}(\bar{p}_t))]^2}. \quad (14.26)$$

Berkson's minimum chi-square estimator $\hat{\boldsymbol{\beta}}_{MC}$ minimizes the weighted sum of residuals $\sum_{t=1}^T (F^{-1}(\bar{p}_t) - \mathbf{x}'_t \boldsymbol{\beta})/\tilde{\sigma}_t^2$ with respect to $\boldsymbol{\beta}$. This is easily computed by OLS regression of $F^{-1}(\bar{p}_t)/\tilde{\sigma}_t$ on $\mathbf{x}_t/\tilde{\sigma}_t$.

This estimator is simple to implement, as it only requires an OLS package. Yet it is fully efficient, as it can be shown to have the same asymptotic distribution as the MLE that treats each observation separately, rather than grouping them into cells with common regressor value \mathbf{x}_t . For the logit model this estimator is especially simple, as $F^{-1}(\bar{p}_t) = \ln(\bar{p}_t/(1 - \bar{p}_t))$ and $\tilde{\sigma}_t^2 = 1/[N_t \bar{p}_t(1 - \bar{p}_t)]$.

The advantage of the minimum chi-square estimator is its computational simplicity, although advances in computer power now make this point moot. Grouped economics

data are rarely such that there are many observations within group per unique value of the regressors, unless the regressors are just a few indicator variables. The method does provide insights to aggregation, however, a topic we now consider.

14.6.2. Estimation with Aggregate Data

Econometrics examples of **data aggregation** include data on the proportion of people working and data on the proportion of those commuting by bus in different regions, explained by data on the average characteristics of people in the region.

As a concrete example, suppose \bar{p}_t equals the unemployment rate in region t and $\bar{\mathbf{x}}_t$ equals the average level of schooling in region t . One possible model is LS regression of \bar{p}_t on $\bar{\mathbf{x}}_t$. Because $0 < \bar{p}_t < 1$, many studies instead transform to a dependent variable that is unbounded, estimating the model

$$\ln\left(\frac{\bar{p}_t}{1 - \bar{p}_t}\right) = \bar{\mathbf{x}}_t'\beta + u_t, \quad (14.27)$$

where u_t is an error.

This model looks similar to the minimum chi-square estimator for the logit model, when $F^{-1}(\bar{p}_t) = \ln(\bar{p}_t/(1 - \bar{p}_t))$. However, it is not because Berkson's estimator is only appropriate if all regressors in the t th cell take the same value. Here instead the regressors can take different values, as different people in region t will have different levels of schooling.

To see the consequences of aggregation when there is **within-cell heterogeneity** in the regressors, suppose the individual-level model is an index model (see Section 14.4.1) with

$$\begin{aligned} y_i^* &= \mathbf{x}'_i\beta + u_i, \\ u_i &\sim \mathcal{N}[0, 1]. \end{aligned}$$

We choose to work with normal errors, corresponding to a probit rather than logit model, because it is then possible to obtain analytical results. Model the heterogeneity as

$$\mathbf{x}_i \sim \mathcal{N}[\mu_t, \Sigma_t],$$

for individuals in cell t . This realistically permits variation across cells, and the complication is that $\Sigma_t \neq \mathbf{0}$, so there is within-cell heterogeneity. Then in region t , conditional on β , μ_t , and Σ_t ,

$$\begin{aligned} \Pr[y_i = 1] &= \Pr[\mathbf{x}'_i\beta + u_i > 0] \\ &= \Pr\left[\frac{\mathbf{x}'_i\beta + u_i - \mu'_t\beta}{\sqrt{1 + \beta'\Sigma_t\beta}} > \frac{-\mu'_t\beta}{\sqrt{1 + \beta'\Sigma_t\beta}}\right] \\ &= \Phi\left(\frac{-\mu'_t\beta}{\sqrt{1 + \beta'\Sigma_t\beta}}\right), \end{aligned}$$

where we use $\mathbf{x}'_i\beta + u_i \sim \mathcal{N}[\mu'_t\beta, (1 + \beta'\Sigma_t\beta)]$ given the preceding assumptions and then subtract the mean and divide by the standard deviation to transform to a standard normal variate.

By similar argument to that leading to (14.25) given (14.24), the underlying individual-level binary choice parameters β can be consistently estimated by nonlinear LS estimation of β in the regression

$$\Phi^{-1}(\bar{y}_t) = \frac{\bar{x}'_t \beta}{\sqrt{1 + \beta' S_t \beta}} + w_t, \quad (14.28)$$

where \bar{y}_t and \bar{x}_t are cell averages and S_t is the sample variance of x_i in cell t . The Berkson minimum chi-square estimate instead regresses $\Phi^{-1}(\bar{y}_t)$ on \bar{x}_t and is inconsistent for β unless $\Sigma_t = \mathbf{0}$.

14.6.3. Discussion

Aggregation issues are much more complicated in nonlinear models. If the original individual-level model was the linear model $y_i = x'_i \beta + u_i$ with $x_i \sim \mathcal{N}[\mu_t, \Sigma_t]$ in the t th cell, then the corresponding linear regression of \bar{y}_t on \bar{x}_t would yield a consistent estimate of β . With nonlinear models similar aggregation leads to inconsistent estimation of individual-level parameters, unless adjustment such as that in (14.28) is undertaken. Furthermore, the example in Section 14.6.2, due to McFadden and Reid (1975), is unusual in that aggregation of a nonlinear model leads to tractable results. This example is discussed in considerable detail by Cameron (1990), who considers it in the wider context of **aggregation in nonlinear models**.

An active area of aggregation in discrete choice, usually multinomial choice, is the marketing literature on **market shares of branded goods**. Allenby and Ross (1991) present examples where the bias of fitting aggregate logit models may not be great. More importantly, recent computational advances permit estimation of individual-level parameters with aggregate data even if aggregation yields no closed-form solution. See, for example, Berry (1994) and Nevo (2001), who estimate models qualitatively similar to the random parameters logit model in Section 15.7.

Finally, note that in many applications with aggregate proportions data, such as unemployment rate by region, there is no desire to estimate individual-level parameters. The only goal is a reasonable model for dependent variable \bar{p}_t that lies between zero and one. Then the linear regression (14.27) may be fine. The error u_t in (14.27) will no longer have the variance given in (14.26). It will still be heteroskedastic, however, so statistical inference should be based on White heteroskedastic-robust standard errors.

14.7. Semiparametric Estimation

The binary outcome model is perhaps the leading example of semiparametric regression. Most econometrics studies presume a single-index form $F(x'_i \beta)$, where the functional form for F is not specified. The goal is to obtain an estimate of β that is consistent for β , ideally \sqrt{N} -consistent and asymptotically normal, while $F(\cdot)$ is viewed as a nuisance function. The single-index model semiparametric estimators of Section 9.7.4 can be applied. Additional estimators exploit the index function model interpretation for binary outcomes. In addition, semiparametric ML estimation that

attains the semiparametric efficiency bounds is possible with little need for additional assumptions, since it is clear that the distribution is Bernoulli and only $F(\mathbf{x}_i'\beta)$ is not known.

14.7.1. Semiparametric Conditional Mean Estimation

The estimation problem in general is one where the dependent variable y_i takes value 0 or 1 with conditional mean

$$E[y_i|\mathbf{x}_i] = m(\mathbf{x}_i),$$

where $m(\cdot)$ is unknown. Note that $m(\mathbf{x}_i)$ also equals the conditional probability that $y_i = 1$.

The nonparametric regression methods of Sections 9.4–9.6 can be applied, despite the binary nature of the dependent variable. This is easily seen from Figure 14.1, a scatterplot of binary variable y on scalar regressor x , a natural candidate for kernel regression of y on x . The fitted values will lie between 0 and 1, aside from unusual cases such as when higher order kernels are used, in which case the fitted variable can take negative values.

In many microeconomics applications \mathbf{x} is of too high a dimension for nonparametric methods to work well (the curse of dimensionality). Semiparametric regression models that partially specify $m(\cdot)$ are given in Section 9.7. Additive models are fairly popular in statistical applications. In econometrics single-index models are instead used, since a popular starting point is the index function model of Section 14.4.1. This yields a **single-index model** if the latent variable $y^* = \mathbf{x}'\beta + u$. Thus we suppose

$$E[y_i|\mathbf{x}_i] = F(\mathbf{x}_i'\beta),$$

where we follow the notation of this chapter and use $F(\cdot)$ rather than $g(\cdot)$ to denote the unknown function.

From Section 9.7.4, β is only identified up to location and scale. This is also clear from Section 14.4.1, where the error u in the index model was normalized to have mean 0 (location) and the variance needed to be specified (scale). Here restrictions are not placed on u , so β is not completely identified but the ratios of slope coefficients are identified. See Manski (1988b) for a detailed analysis of identification in binary choice models.

\sqrt{N} -consistent asymptotically normal estimates of β can be obtained by average derivative estimation or by semiparametric least squares (see Section 9.7.4). However, alternative estimators, specific to binary outcomes, are more often used.

14.7.2. Maximum Score Estimation

Semiparametric estimators for binary outcomes are often based on the index function model $y^* = \mathbf{x}'\beta + u$ for binary outcomes. In such cases it is convenient to write the model as

$$y_i = \mathbf{1}(\mathbf{x}_i'\beta + u_i > 0),$$

where $\mathbf{1}(A) = 1$ if event A occurs.

Manski (1975) noted that the predicted value of y_i is $\mathbf{1}(\mathbf{x}'_i \boldsymbol{\beta} > 0)$, setting $u_i = 0$ since u_i is unknown, in which case a score of the number of correct predictions is

$$S_N(\boldsymbol{\beta}) = \sum_{i=1}^N \{y_i \mathbf{1}(\mathbf{x}'_i \boldsymbol{\beta} > 0) + (1 - y_i) \mathbf{1}(\mathbf{x}'_i \boldsymbol{\beta} \leq 0)\}, \quad (14.29)$$

since correct predictions occur if $y_i = 1$ and $\mathbf{1}(\mathbf{x}'_i \boldsymbol{\beta} > 0)$, or if $y_i = 0$ and $\mathbf{1}(\mathbf{x}'_i \boldsymbol{\beta} \leq 0)$. Manski's **maximum score estimator** $\widehat{\boldsymbol{\beta}}_{MS}$ maximizes $S_N(\boldsymbol{\beta})$. This is a nonstandard problem because $\mathbf{1}(\mathbf{x}'_i \boldsymbol{\beta} > 0)$ is not differentiable in $\boldsymbol{\beta}$. Manski (1975, 1985) established consistency assuming $F(0) = 0.5$, or equivalently that $\text{Median}[u_i | \mathbf{x}_i] = 0$. It has subsequently been shown that $N^{1/3}(\widehat{\boldsymbol{\beta}}_{MS} - \boldsymbol{\beta})$ has a nonnormal limit distribution, though inference can be performed using the bootstrap (Manski and Thompson (1986)).

Manski's estimator can be viewed as a least absolute deviations estimator. From Section 4.6.2, the LAD estimator minimizes the sum of absolute differences between y_i and $\text{Median}[y_i | \mathbf{x}_i]$. This less familiar estimator is qualitatively similar to the LS estimator, which minimizes the sum of absolute differences between y_i and $E[y_i | \mathbf{x}_i]$. To implement LAD here requires obtaining $\text{Median}[y_i | \mathbf{x}_i]$. If $\text{Median}[u_i | \mathbf{x}_i] = 0$ then $\text{Median}[y_i^* | \mathbf{x}_i] = \mathbf{x}'_i \boldsymbol{\beta}$, so $\text{Median}[y_i | \mathbf{x}_i] = \mathbf{1}(\mathbf{x}'_i \boldsymbol{\beta} > 0)$. Thus the **binary outcome model LAD estimator** minimizes

$$Q_N(\boldsymbol{\beta}) = \sum_{i=1}^N |y_i - \mathbf{1}(\mathbf{x}'_i \boldsymbol{\beta} > 0)|. \quad (14.30)$$

From Exercise 14.4 $Q_N(\boldsymbol{\beta}) = N - S_N(\boldsymbol{\beta})$, so the maximum score estimator equals the LAD estimator. See Manski (1985, p. 320) for other interpretations of the maximum score estimator as a LAD estimator.

The objective function $S_N(\boldsymbol{\beta})$ for the maximum score estimator given in (14.29) is not differentiable. It can be rewritten as

$$S_N(\boldsymbol{\beta}) = \sum_{i=1}^N (2y_i - 1) \mathbf{1}(\mathbf{x}'_i \boldsymbol{\beta} > 0) + N - \sum_{i=1}^N y_i,$$

see Exercise 14.4. The second sum can be ignored as it does not involve $\boldsymbol{\beta}$.

An estimator with differentiable objective function is the **smooth maximum score estimator** of Horowitz (1992) that maximizes

$$Q_N^S(\boldsymbol{\beta}) = \sum_{i=1}^N (2y_i - 1) K(\mathbf{x}'_i \boldsymbol{\beta} / h_N),$$

where $K(\mathbf{x}' \boldsymbol{\beta} / h_N)$ is a smoothed version of $\mathbf{1}(\mathbf{x}' \boldsymbol{\beta} > 0)$. Since $\mathbf{1}(\mathbf{x}' \boldsymbol{\beta} > 0)$ equals zero for negative values of $\mathbf{x}' \boldsymbol{\beta}$ and equals one for positive values of $\mathbf{x}' \boldsymbol{\beta}$ it is natural to choose $K(\cdot)$ to be a cdf with $K(0) = 0.5$ and choose h_N to be small. Smoothing simplifies computation of the estimator, but analysis is complicated by the need to have $h_N \rightarrow 0$ at appropriate rate as $N \rightarrow \infty$. The estimator converges at rate close to \sqrt{N} and is asymptotically normal. For details see Horowitz (2002), who presents a bootstrap that permits tests with better size properties in finite samples.

LAD estimation can be extended to the censored regression model (see Section 16.9.2).

14.7.3. Maximum Rank Correlation Estimator

Begin with a single-index model with $E[y_i | \mathbf{x}_i] = F(\mathbf{x}'_i \boldsymbol{\beta})$. If $F(\mathbf{x}'_i \boldsymbol{\beta})$ is monotonically increasing in $\mathbf{x}'_i \boldsymbol{\beta}$, then $E[y_i | \mathbf{x}_i] > E[y_j | \mathbf{x}_j]$ if $\mathbf{x}'_i \boldsymbol{\beta} > \mathbf{x}'_j \boldsymbol{\beta}$. Thus it is likely, though not guaranteed, that the observed values $y_i > y_j$ when $\mathbf{x}'_i \boldsymbol{\beta} > \mathbf{x}'_j \boldsymbol{\beta}$. This suggests choosing $\boldsymbol{\beta}$ to ensure that with high frequency $y_i > y_j$ when $\mathbf{x}'_i \boldsymbol{\beta} > \mathbf{x}'_j \boldsymbol{\beta}$.

The **maximum rank correlation (MRC)** estimator of Han (1987) chooses $\boldsymbol{\beta}$ to maximize

$$Q_N^{\text{MRC}}(\boldsymbol{\beta}) = \sum_{i=1}^N \sum_{\substack{j=1 \\ j < i}}^N \mathbf{1}(y_i > y_j) \mathbf{1}(\mathbf{x}'_i \boldsymbol{\beta} > \mathbf{x}'_j \boldsymbol{\beta}) + \mathbf{1}(y_i < y_j) \mathbf{1}(\mathbf{x}'_i \boldsymbol{\beta} < \mathbf{x}'_j \boldsymbol{\beta}).$$

The ij th term in this sum equals one if $y_i > y_j$ when $\mathbf{x}'_i \boldsymbol{\beta} > \mathbf{x}'_j \boldsymbol{\beta}$ or if $y_i < y_j$ when $\mathbf{x}'_i \boldsymbol{\beta} < \mathbf{x}'_j \boldsymbol{\beta}$, and equals zero if instead there is a sign reversal so that $y_i < y_j$ when $\mathbf{x}'_i \boldsymbol{\beta} > \mathbf{x}'_j \boldsymbol{\beta}$ or $y_i > y_j$ when $\mathbf{x}'_i \boldsymbol{\beta} < \mathbf{x}'_j \boldsymbol{\beta}$. The estimator is called the maximum rank correlation estimator because $Q_N^{\text{MRC}}(\boldsymbol{\beta})$ is a multiple of Kendall's rank correlation coefficient between y_i and $\mathbf{x}'_i \boldsymbol{\beta}$.

This estimator is \sqrt{N} -consistent and asymptotically normal (see Sherman, 1993).

14.7.4. Semiparametric ML Estimation

For binary choice data the likelihood function given independent observations is clearly that given in (14.4). The only complication is that $F(\cdot)$ is unknown. Klein and Spady (1993) proposed the **semiparametric MLE** that maximizes

$$\mathcal{L}_N(\boldsymbol{\beta}) = \sum_{i=1}^N \left\{ y_i \ln \widehat{F}(\mathbf{x}'_i \boldsymbol{\beta}) + (1 - y_i) \ln(1 - \widehat{F}(\mathbf{x}'_i \boldsymbol{\beta})) \right\},$$

where $\widehat{F}(\mathbf{x}'_i \boldsymbol{\beta})$ is a nonparametric estimate of $F(\mathbf{x}'_i \boldsymbol{\beta})$.

This estimator is similar in spirit to the WSL estimator of Ichimura (1993) detailed in Section 9.7.4, and similar issues in computation arise with iteration between computation of $\widehat{\boldsymbol{\beta}}$ given \widehat{F} and computation of \widehat{F} given $\widehat{\boldsymbol{\beta}}$. Given the ML first-order conditions (14.5), the semiparametric MLE can also be computed as the solution to the equations

$$\sum_{i=1}^N \frac{\widehat{F}'(\mathbf{x}'_i \boldsymbol{\beta})}{\widehat{F}(\mathbf{x}'_i \boldsymbol{\beta})(1 - \widehat{F}(\mathbf{x}'_i \boldsymbol{\beta}))} (y_i - \widehat{F}(\mathbf{x}'_i \boldsymbol{\beta})) \mathbf{x}_i = \mathbf{0},$$

which are the same as those for the **WSLS estimator** with weights $w_i = \widehat{F}'_i / [\widehat{F}_i(1 - \widehat{F}_i)]$.

The attraction of Klein and Spady's estimator is that it is fully efficient in the sense that it attains the semiparametric efficiency bound. Computation is difficult, however. For details see Section 9.7.4, where similar computational issues are discussed for

Ichimura's WSLS estimator, and see Klein and Spady (1993) and Pagan and Ullah (1999, pp. 283–285).

14.7.5. Comparison of Semiparametric Estimators

Econometricians focus on single-index models, and even then there are a multitude of semiparametric estimators available for the binary outcome model. None of these estimators are particularly simple to implement. The objective functions can have multiple optima and may not be smooth. For example, Horowitz (1992) uses simulated annealing for the smooth maximum score estimator and Dorsey and Mayer (1995) use genetic algorithms to obtain the maximum score estimator.

Interpretation of coefficients is also difficult. For example the maximum score estimator applied to the fishing mode data yielded intercept estimate of 0.776 and slope of −0.631 (with bootstrap-estimated standard error of 0.103), but these coefficients are not directly comparable to those given in Table 14.2. Indeed, since parameter slope estimates are only identified up to scale, the semiparametric estimates are most useful if several coefficients are included in the regression and coefficient estimates are compared to those of a reference variable.

The maximum score and maximum rank correlation estimators are unusual among semiparametric estimators in not requiring use of smoothing parameters, such as choice of a bandwidth, an attractive property. The latter of these estimators is \sqrt{N} -consistent.

In recent work Blundell and Powell (2004) propose semiparametric estimation with **endogenous regressors**.

14.8. Derivation of Logit from Type I Extreme Value

The derivation in Section 14.4.2 of the logit model from the ARUM used knowledge of the statistical result that the difference $(\varepsilon_0 - \varepsilon_1)$ of independent type 1 extreme value random variables is logistic distributed. For completeness we provide a direct derivation based on the distributions of ε_0 and ε_1 .

Rewriting the second line of (14.20) yields

$$\begin{aligned} \Pr[y = 1] &= \Pr[\varepsilon_0 < \varepsilon_1 + V_1 - V_0] \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\varepsilon_1 + V_1 - V_0} f(\varepsilon_0, \varepsilon_1) d\varepsilon_0 d\varepsilon_1 \\ &= \int_{-\infty}^{\infty} f(\varepsilon_1) \left\{ \int_{-\infty}^{\varepsilon_1 + V_1 - V_0} f(\varepsilon_0) d\varepsilon_0 \right\} d\varepsilon_1, \end{aligned} \tag{14.31}$$

where in the last line ε_0 and ε_1 are assumed to be independent. By specializing $f(\varepsilon_0)$ to the type 1 extreme value density, (14.31) becomes

$$\begin{aligned} \Pr[y = 1] &= \int_{-\infty}^{\infty} f(\varepsilon_1) \left\{ \int_{-\infty}^{\varepsilon_1 + V_1 - V_0} e^{-\varepsilon_0} \exp(-e^{-\varepsilon_0}) d\varepsilon_0 \right\} d\varepsilon_1 \\ &= \int_{-\infty}^{\infty} f(\varepsilon_1) [\exp(-e^{-\varepsilon_0})]_{-\infty}^{\varepsilon_1 + V_1 - V_0} d\varepsilon_1 \\ &= \int_{-\infty}^{\infty} f(\varepsilon_1) \exp(-e^{-(\varepsilon_1 + V_1 - V_0)}) d\varepsilon_1. \end{aligned} \tag{14.32}$$

Using the extreme value density for ε_1 in (14.32) yields

$$\begin{aligned}\Pr[y = 1] &= \int_{-\infty}^{\infty} e^{-\varepsilon_1} \exp(-e^{-\varepsilon_1}) \exp(-e^{-(\varepsilon_1 + V_1 - V_0)}) d\varepsilon_1 \\ &= \int_{-\infty}^{\infty} e^{-\varepsilon_1} \{ \exp(-e^{-\varepsilon_1} - e^{-(\varepsilon_1 + V_1 - V_0)}) \} d\varepsilon_1 \\ &= \int_{-\infty}^{\infty} e^{-\varepsilon_1} \{ \exp(-e^{-\varepsilon_1} - e^{-\varepsilon_1} e^{-(V_1 - V_0)}) \} d\varepsilon_1 \\ &= \int_{-\infty}^{\infty} e^{-\varepsilon_1} \exp \{ -e^{-\varepsilon_1}(1 + e^{-(V_1 - V_0)}) \} d\varepsilon_1\end{aligned}\quad (14.33)$$

Since $\int_{-\infty}^{\infty} ae^{-\varepsilon} \exp(-ae^{-\varepsilon}) d\varepsilon = 1$ it follows that $\int_{-\infty}^{\infty} e^{-\varepsilon} \exp(-ae^{-\varepsilon}) d\varepsilon = 1/a$. Using this result with $a = 1 + e^{-(V_1 - V_0)}$ in (14.33) yields

$$\begin{aligned}\Pr[y = 1] &= (1 + e^{-(V_1 - V_0)})^{-1} \\ &= e^{V_1} / (e^{V_0} + e^{V_1}) \\ &= e^{V_1 - V_0} / (1 + e^{V_1 - V_0}).\end{aligned}\quad (14.34)$$

Letting $V_1 - V_0 = \mathbf{x}'\beta$ yields the logit model.

14.9. Practical Considerations

Most packages provide probit and logit model estimators. The main choice for the practitioner is which model to use. In practice there is little difference in the predicted marginal effects obtained from the two models, unless most of the outcomes are zero or most of the outcomes are one.

Semiparametric estimation generally requires special coding in languages such as GAUSS, though Lindep implements the estimators of Manski and Klein and Spady.

14.10. Bibliographic Notes

Logit and probit models are commonly used and relatively simple nonlinear regression models that appear in many standard texts such as Greene's (2003). The surveys by Amemiya (1981) and McFadden (1984) include all the basic results. Maddala (1983) and Amemiya (1985) provide further details. The books by Train (1986) and Ben-Akiva and Lerman (1985) are particularly good for applications. These references cover both binary and multinomial outcomes.

- 14.3** Bliss (1934) proposed the probit transformation to plot dosage–mortality curves. Berkson (1951) popularized use of the simpler logit model.
- 14.4** Latent variable models are especially popular in the psychometrics literature.
- 14.5** Amemiya (1985, Section 9.5) provides an excellent survey of choice-based sampling for binary outcome models. See also Section 24.4.
- 14.6** Cameron (1990) considers aggregation in binary outcome models and summarizes general results of Keljian (1980) and Stoker (1984) on estimability of individual-level parameters in nonlinear models using aggregate data.
- 14.7** The maximum score estimator of Manski (1975) is a leading early example of semiparametric regression. Semiparametric methods for binary outcome models are covered in the books by M-J. Lee (1996), Horowitz (1997), and Pagan and Ullah (1999). The last reference covers many methods.

Exercises

14–1 Consider a latent variable modeled by $y_i^* = \mathbf{x}_i'\beta + \varepsilon_i$, with $\varepsilon_i \sim \mathcal{N}[0, 1]$. Suppose we observe only $y_i = 1$ if $y_i^* < U_i$ and $y_i = 0$ if $y_i^* \geq U_i$, where the upper limit U_i is a known constant for each individual (i.e., data) and may differ over individuals.

- (a) Find $\Pr[y_i = 1 | \mathbf{x}_i]$. [Hint: Note that this differs from the standard case both due to presence of U_i and because the equalities are reversed with $y_i = 1$ if $y_i^* < U_i$.]
- (b) Provide details on an estimation method to consistently estimate β .
- (c) Suppose you estimate this model and find that the third regressor x_{3i} has estimated coefficient $\hat{\beta}_3 = 0.2$. Provide a meaningful interpretation of $\hat{\beta}_3$.

14–2 Consider the logit model with $\Pr[y = 1 | x_1, x_2] = \Lambda(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})$, where $\Lambda(z) = e^z / (1 + e^z)$.

- (a) Write down the likelihood scores and information matrix in an expanded form.
- (a) Use these to derive Wald and LM score tests of $H_0 : \beta_2 = 0$.
- (c) Explain how you would computationally implement the tests.
- (d) In what sense is the logit model *intrinsically* heteroskedastic?

14–3 Suppose we use an index formulation for a discrete choice model but it is felt that the latent variable is strictly positive. This is accommodated by supposing that the latent variable y^* has exponential density with parameter γ , so the density $f(y^*)$ is $f(y^*) = \gamma^{-1} \exp(-y^*/\gamma)$, with $\gamma = \exp(\mathbf{x}'\beta)$. We observe $y = 1$ if $y^* > \mathbf{z}'\alpha$ and $y = 0$ if $y^* \leq \mathbf{z}'\alpha$.

- (a) Give the log-likelihood function for the observed data.
- (b) What is the effect of a one-unit change in x_{ji} on $\Pr[y_i = 1]$?
- (c) Suppose that $y = 1$ if $y^* > \exp(\mathbf{z}'\alpha)$ and $\mathbf{x} = \mathbf{z}$. Do you see any problems in identifying α and/or β ? Explain your answer.

14–4 Consider the maximum score estimator with objective functions $S_N(\beta)$ given in (14.29) and $Q_N(\beta)$ given in (14.30).

- (a) Show that $S_N(\beta) = \sum_i [\mathbf{1}(y_i = 1) \times \mathbf{1}(\mathbf{x}_i'\beta > 0) + \mathbf{1}(y_i = 0) \times \mathbf{1}(\mathbf{x}_i'\beta \leq 0)]$.
- (b) Show that $Q_N(\beta) = \sum_i [\mathbf{1}(y_i = 1) \times \mathbf{1}(\mathbf{x}_i'\beta \leq 0) + \mathbf{1}(y_i = 0) \times \mathbf{1}(\mathbf{x}_i'\beta > 0)]$.
- (c) Using $\mathbf{1}(y_i = 1) = 1 - \mathbf{1}(y_i = 0)$, show that $Q_N(\beta) = N - S_N(\beta)$.
- (d) Using $\mathbf{1}(\mathbf{x}_i'\beta \leq 0) = 1 - \mathbf{1}(\mathbf{x}_i'\beta > 0)$ show that (14.29) can be rewritten as $S_N(\beta) = \sum_i (2y_i - 1)\mathbf{1}(\mathbf{x}_i'\beta > 0) + N - \sum_i y_i$.

14–5 Use the health expenditure data of Section 16.6. The model is a probit regression of DMED, an indicator variable for positive health expenditures, against just one regressor for simplicity, NDISEASE, the number of chronic diseases.

- (a) Obtain the OLS estimate of the slope parameter.
- (b) Obtain the probit estimate of the slope parameter.
- (c) Given part (b), obtain the marginal effect of chronic diseases in two ways: averaged over the sample and evaluated at the sample average of NDISEASE.
- (d) Obtain the logit estimate of the slope parameter.
- (e) Given part (d), obtain the marginal effect of chronic diseases in three ways: averaged over the sample, evaluated at the sample average of NDISEASE, and evaluated at $\Lambda(\mathbf{x}'\beta) = \bar{y}$.

14.10. BIBLIOGRAPHIC NOTES

- (f) For the logit model calculate the proportionate change in the odds ratio when NDISEASE changes.
- 14–6** Continue the analysis of Exercise 14.5.
- (a) Compare the three binary models on the basis of statistical significance of NDISEASE.
 - (b) Compare the three binary models on the basis of the estimated marginal effect.
 - (c) Compare the three binary models on the basis of the predicted probabilities.
 - (d) Compare the logit and probit binary models on the basis of log-likelihood.

Multinomial Models

15.1. Introduction

The preceding chapter considered models for discrete outcome variables that can take one of two possible values. Here we consider several possible outcomes, usually mutually exclusive. Examples include different ways to commute to work (by bus, car, or walking), various types of health insurance (fee-for-service, managed care, or none), different employment status (full-time, part-time, or none), choice of recreational site, occupational choice, and product choice.

Statistical inference is relatively straight forward in principle, as the data must be multinomial distributed, just as binary data must be Bernoulli or binomial distributed. Estimation is most often by maximum likelihood because the data are clearly multinomial distributed. For some complications, however, moment-based estimation is used instead.

Different multinomial models arise owing to different functional forms for the probabilities of the multinomial distribution, similar to the differences between probit and logit in the binary case. A distinction is also made between models where regressors vary across alternatives for a given individual and models where regressors are constant across alternatives. For example, in transportation mode choice some regressors, such as travel time or cost, will vary with choices whereas others, such as age, are choice invariant.

The simplest multinomial model, the conditional or multinomial logit model, is quite straightforward to use but is viewed as too restrictive in practice, especially if the multinomial outcome data arise from individual choice. For unordered outcomes less restrictive models can be obtained using the random utility model. In this model the alternative with the highest utility is chosen, where utility from each alternative is the sum of deterministic and random components. Different specifications of the random components lead to different functional forms for choice probabilities and hence to different multinomial models. Additional models arise in applications where some structure can be placed on the decision-making process, such as a natural ordering of

alternatives or sequencing of decisions. In practice many different multinomial models are used.

Section 15.2 presents an application to illustrate the issues discussed in this chapter. General results for multinomial models are given in Section 15.3. The conditional and multinomial logit models are presented in Section 15.4. The additive random utility model is presented in Section 15.5. The nested logit, random parameters logit, and multinomial probit models are the subject of Sections 15.6–15.8. Ordered and sequential models are detailed in Section 15.9. Multivariate models with more than one discrete outcome variable are presented in Section 15.10. Semiparametric estimators are briefly reviewed in Section 15.11.

15.2. Example: Choice of Fishing Mode

This section illustrates multinomial logit, the simplest unordered multinomial model, and variations detailed in Section 15.4 that permit regressors to vary across alternatives. The emphasis is on interpretation of estimated models. The marginal effect of a change in a regressor is more complicated than the usual impact on a single conditional mean. For multinomial data there is instead a separate marginal effect on the probability of each outcome, and these marginal effects sum to zero since probabilities sum to one.

The application is to choice of fishing mode. The dependent variable y takes value 1, 2, 3, or 4 depending on which of the four mutually exclusive alternative modes of fishing – respectively, beach, pier, private boat, and charter boat – is chosen. An unordered multinomial model such as multinomial logit is appropriate, since there is no clear ordering of the outcome variable. Regressors are individual income, which does not vary with fishing mode, and price and catch rate, which do vary by fishing mode and across individuals.

The sample of 1,182 people comes from a survey conducted by Thomson and Crooke (1991) and analyzed by Herriges and Kling (1999). The data are summarized in Table 15.1, which gives averages for the subsamples of people who chose each of the modes as well as the overall sample average of regressors.

15.2.1. Conditional Logit: Alternative-Varying Regressors

First consider the role of price and catch rate, regressors that vary across alternatives except that for these data the price of beach and pier fishing are the same.

Looking down the columns of Table 15.1, we see that people tend to fish where it is cheapest for them to do so. For example, for people choosing to fish from the beach the average price was \$36 compared to average prices of \$36, \$98, and \$125 for the other modes. More generally, for people choosing the beach and pier these modes were on average much cheaper than the boat modes, and for people fishing from a boat this was on average much cheaper than beach or pier fishing. The relationship between mode choice and catch rate is less clear-cut, though it is clear that the charter boat has the highest catch rate.

Table 15.1. Fishing Mode Multinomial Choice: Data Summary

Explanatory Variable	Sub sample Averages				
	$y = 1$ Beach	$y = 2$ Pier	$y = 3$ Private	$y = 4$ Charter	All y Overall
Income (\$1,000s per month)	4.052	3.387	4.654	3.881	4.099
Price beach (\$)	36	31	138	121	103
Price pier (\$)	36	31	138	121	103
Price private (\$)	98	82	42	45	55
Price charter (\$)	125	110	71	75	84
Catch rate beach	0.28	0.26	0.21	0.25	0.24
Catch rate pier	0.22	0.20	0.13	0.16	0.16
Catch rate private	0.16	0.15	0.18	0.18	0.17
Catch rate charter	0.52	0.50	0.65	0.69	0.63
Sample probability	0.113	0.151	0.354	0.382	1.000
Observations	134	178	418	452	1182

For alternative-specific regressors that vary across alternatives, such as price and catch rate, the multinomial logit model is called a conditional logit model (see Section 15.4.1). The probability of the i th individual choosing fishing mode j is given by

$$p_{ij} = \Pr[y_i = j] = \frac{\exp(\beta_P P_{ij} + \beta_C C_{ij})}{\sum_{k=1}^4 \exp(\beta_P P_{ik} + \beta_C C_{ik})}, \quad j = 1, \dots, 4,$$

where P denotes price, C denotes catch rate, the subscript i denotes the i th individual, and subscript j or k denotes the alternative. This model is an obvious extension of binary logit and gives probabilities that lie between 0 and 1 and sum to one. Other multinomial models use a different functional form for p_{ij} .

The coefficient estimates are given in the CL column of Table 15.2. For the CL model, though not for all multinomial models, the sign of the coefficient is directly interpretable. Anticipating results from Section 15.4.3, since $\beta_P < 0$ we have that an increase in the price of one alternative decreases the probability of choosing that alternative and increases the probability of choosing other alternatives. Similarly, since $\beta_C > 0$, an increase in the catch rate for one alternative increases choice probability for that alternative and decreases the choice probability for other alternatives.

A standard measure of the impact of changes in regressors is $N^{-1} \sum_{i=1}^N \partial p_{ij} / \partial x_{ikr}$, the average marginal response of the probability of choosing alternative j when the r th regressor increases by one unit for alternative k and is unchanged for the other alternatives. For the CL model this is estimated by $N^{-1} \sum_{i=1}^n \hat{p}_{ij} (\delta_{ijk} - \hat{p}_{ik}) \hat{\beta}_r$ (see (15.18)), where $\hat{\beta}$ is the estimate of β and \hat{p}_{ij} , $j = 1, \dots, m$, are the predicted probabilities.

The average responses across the four modes for the two regressors price and catch rate are given in Table 15.3. The table gives the effect on choice probability of a 100-unit (or \$100) change in price and the effect of a one-unit change in the catch rate. For example, an increase of \$100 in the price of beach fishing leads to a decrease of 0.272

15.2. EXAMPLE: CHOICE OF FISHING MODE

Table 15.2. *Fishing Mode Multinomial Choice: Logit Estimates^a*

Regressor	Type	Coefficient	Model type		
			CL	MNL	Mixed
Price (P)	Specific	β_P	-0.021	-	-0.025
Catch rate (C)	Specific	β_{CR}	0.953	-	0.358
Intercept	Invariant	$\alpha_1 : Beach$	-	0.0	0.0
		$\alpha_2 : Pier$	-	0.814	0.778
		$\alpha_3 : Private$	-	0.739	0.527
		$\alpha_4 : Charter$	-	1.341	1.694
Income (I)	Invariant	$\beta_{I1} : Beach$	-	0.0	0.0
		$\beta_{I2} : Pier$	-	-0.143	-0.128
		$\beta_{I3} : Private$	-	0.092	0.089
		$\beta_{I4} : Charter$	-	-0.032	-0.033
- ln L			-1311	-1477	-1215
Pseudo- R^2			0.162	0.099	0.258

^a Type of regressor is alternative-specific (price and catch rate) or alternative-invariant (income). Outcomes are (1) beach, (2) pier, (3) private, and (4) charter. MLE estimates are for conditional logit (CL), multinomial logit (MNL), and mixed logit (Mixed) models. MNL and Mixed models are normalized to base category beach. All estimates except that for β_{I4} are statistically significant at 5%.

in the probability of fishing and an increase of 0.119, 0.080, and 0.068, respectively, in the probability of fishing from a beach, a pier, a private boat, and a charter boat. Note that the changes in probabilities sum to zero, as expected.

Calculation of these marginal effects and probabilities requires postestimation computation. A back-of-the-envelope calculation uses $\bar{p}_j(\delta_{jk} - \bar{p}_k)\widehat{\beta}_r$ for the CL model, where \bar{p}_j is the sample average probability. For the effect of a \$100 change in the price of beach fishing on the probability of beach fishing this yields $100 \times 0.113(1 - 0.113) \times (-0.021) = -0.21$, compared to the sample average value of -0.272 in the table. This approximation becomes less reasonable as probabilities get closer to 0 or 1.

The results in Table 15.3 are consistent with the view that the greatest substitution is between pier and beach fishing and between private boat and charter boat

Table 15.3. *Fishing Mode Choice: Marginal Effects for Conditional Logit Model^a*

	\$100 Change in Price of				One-Unit Change in Catch Rate for				
	Beach	Pier	Private	Charter	Beach	Pier	Private	Charter	
	Change in Pr[beach]	-.272	.119	.085	.068	.126	-.055	-.040	-.032
Change in Pr[pier]		.119	-.263	.080	.064	-.055	.122	-.037	-.030
Change in Pr[private]		.080	.080	-.391	.225	-.040	-.037	.182	-.105
Change in Pr[charter]		.068	.064	.225	-.357	-.032	-.030	-.105	.166

^a Average marginal response of the probability of choosing each alternative when a regressor changes for one of the alternatives and is unchanged for the other alternatives.

fishing. Specifically, price increases, or catch rate decreases, for pier lead to substitution to beach, and vice versa. A similar result holds for charter versus private boat.

These choice probability changes are for large changes in the regressors, given that average price is \$86 and average catch rate is 0.30. One can instead calculate elasticities. Elasticities for choice probabilities need to be used with care, however, because probabilities are bounded between 0 and 1. A change in predicted probability from 0.01 to 0.02 will lead to an elasticity roughly 50 times larger than that for a change in predicted probability from 0.50 to 0.51.

15.2.2. Multinomial Logit: Alternative-Invariant Regressors

Now consider the role of income, measured as monthly income in thousands of dollars. From Table 15.1 it appears that as income rises the fishing mode moves progressively from pier, where average monthly income of people fishing at a pier is \$3,387, to charter boat to beach and finally to private boat, where the average income is \$4,654.

Because income is invariant across alternatives the appropriate model is the multinomial logit model (presented in Section 15.4.1). This lets regressor coefficients vary across alternatives, with

$$p_{ij} = \Pr[y_i = j] = \frac{\exp(\alpha_j + \beta_{Ij} I_i)}{\sum_{k=1}^4 \exp(\alpha_k + \beta_{Ik} I_i)}, \quad j = 1, \dots, 4,$$

where I denotes income. A normalization of parameters is needed as a consequence of the restriction that probabilities sum to one. The empirical results set $\alpha_1 = 0$ and $\beta_{I1} = 0$.

The parameter estimates are given in the MNL column of Table 15.2. Coefficient interpretation is more difficult than for the CL logit model. In particular, for MNL models a positive regression parameter does not mean that an increase in the regressor leads to an increase in the probability of that alternative. Instead, interpretation for the MNL model is relative to the reference or base category group, here beach as the beach coefficients were normalized to zero. Compared to beach fishing a higher income leads to reduced likelihood of fishing from a pier (since $\beta_{I2} = -0.143 < 0$) or a charter boat (since $\beta_{I4} = -0.032$) and greater likelihood of use of a private boat (since $\beta_{I3} = 0.092$).

The magnitude of the response to income changes can be measured using $N^{-1} \sum_{i=1}^N \partial p_{ij} / \partial I_i$, the marginal effect averaged over individuals. For the MNL models this is estimated by $N^{-1} \sum_{i=1}^N \widehat{p}_{ij} (\widehat{\beta}_j - \widehat{\beta}_i)$ (see (15.19)), where $\widehat{\beta}_j$ is the estimate of β_j , $\widehat{\beta}_i = \sum_{l=1}^m p_{il} \beta_l$ is a probability weighted average of the β_l , and \widehat{p}_{ij} , $j = 1, \dots, m$, are the predicted probabilities. For the four choices a \$1,000 increase in monthly income is associated with changes of 0.000, -0.021, 0.033, and -0.012 in, respectively, the probabilities of fishing from beach, pier, private boat, and charter boat. This indicates little change in beach fishing, movement out of pier and charter boat fishing, and movement to private boat fishing. Since average monthly income is \$4,100 the changes in probability are of reasonable size.

However, income alone is not a great discriminator for fishing mode choice. From the bottom of Table 15.2, we see that the MNL model has much lower log-likelihood and pseudo- R^2 than does the CL model. From output not given, across all individuals in the sample the predicted probabilities from the MNL model range from 0.095 to 0.115 for beach, 0.036 to 0.234 for pier, 0.240 to 0.626 for private boat, and 0.244 to 0.416 for charter boat. Since an intercept is included in the MNL model the averages of these predicted probabilities for each choice equal the sample average probabilities. This result for the MNL model is a consequence of (15.16) given later.

15.2.3. Mixed Logit

A richer model combines the two preceding models. This is done using a so-called mixed logit model (see Section 15.4.1) with

$$\Pr[y_i = j] = \frac{\exp(\beta_P P_{ij} + \beta_C C_{ij} + \alpha_j + \beta_{Ij} I_i)}{\sum_{k=1}^4 \exp(\beta_P P_{ik} + \beta_C C_{ik} + \alpha_k + \beta_{Ik} I_i)}.$$

This model, not to be confused with the model of Section 15.7 which is also referred to as a mixed model, can be implemented as a conditional logit model

$$\Pr[y_i = j] = \frac{\exp(\beta_P P_{ij} + \beta_C C_{ij} + \sum_{l=1}^4 (\alpha_l d_{ijl} + \beta_{Il} d I_{ijl}))}{\sum_{k=1}^4 \exp(\beta_P P_{ik} + \beta_C C_{ik} + \sum_{l=1}^4 (\alpha_l d_{ijl} + \beta_{Il} d I_{ijl}))},$$

where d_{ijl} is a dummy variable equal to one if $j = l$ and equal to zero otherwise, and $d I_{ijl} = d_{ijl} I_i$ is equal to income if $j = l$ and equals zero otherwise. In this case we regress y_i on eight regressors: P_{ij} , C_{ij} , d_{ij2} , d_{ij3} , d_{ij4} , $d I_{ij2}$, $d I_{ij3}$, and $d I_{ij4}$. Since $\alpha_1 = 0$ and $\beta_{I1} = 0$ the regressors d_{ij1} and $d I_{ij1}$ are omitted. Note that if we estimate this CL model with just the d_{ijl} and $d I_{ijl}$ as regressors then the CL estimates equal the MNL estimates given earlier. An MNL model can always be estimated as a CL model (see Section 15.3.4).

While the mixed logit model is richer than the CL model, the CL model has the advantage that if an additional alternative was added to the choice set then one can predict its probability of selection, since the parameters of the CL model do not vary across alternatives.

The results are reported in the last column of Table 15.2. Compared to the first two models the coefficients are little changed, except for considerable change in the catch rate coefficient. This change is due to inclusion of the alternative-specific dummies, rather than inclusion of income. The mixed model is strongly preferred to the other models on the basis of much higher log-likelihood value or formal statistical tests.

15.3. General Results

The results in this section pertain to all multinomial models. The remainder of the chapter specializes to the many different specifications of the multinomial model used in practice.

15.3.1. Multinomial Models

There are m alternatives and the dependent variable y is defined to take value j if the j th alternative is taken, $j = 1, \dots, m$. (Some authors instead consider $m + 1$ alternatives with $j = 0, 1, \dots, m$.) Define the probability that alternative j is chosen as

$$p_j = \Pr[y = j], \quad j = 1, \dots, m. \quad (15.1)$$

Introduce m binary variables for each observation y ,

$$y_j = \begin{cases} 1 & \text{if } y = j, \\ 0 & \text{if } y \neq j. \end{cases} \quad (15.2)$$

Thus y_j equals one if alternative j is the observed outcome and the remaining y_k equal zero, so for each observation on y exactly one of y_1, y_2, \dots, y_m will be nonzero. The **multinomial density** for one observation can then be conveniently written as

$$f(y) = p_1^{y_1} \times \cdots \times p_m^{y_m} = \prod_{j=1}^m p_j^{y_j}. \quad (15.3)$$

For regression models introduce a subscript i for the i th individual and regressors \mathbf{x}_i . Specify a model for the probability that individual i chooses the j th alternative,

$$p_{ij} = \Pr[y_i = j] = F_j(\mathbf{x}_i, \boldsymbol{\beta}), \quad j = 1, \dots, m, \quad i = 1, \dots, N. \quad (15.4)$$

The functional form for F_j should be such that probabilities lie between 0 and 1 and sum over j to one. Different functional specifications for F_j correspond to specific models, notably multinomial logit, nested logit, multinomial probit, ordered, sequential, and multivariate models. These models are presented in subsequent sections.

15.3.2. ML Estimation

The multinomial density for one observation is given in (15.3). The likelihood function for a sample of N independent observations is then $L_N = \prod_{i=1}^N \prod_{j=1}^m p_{ij}^{y_{ij}}$, where the subscript i denotes the i th of N individuals and the subscript j denotes the j th of m alternatives. The **log-likelihood function** is

$$\mathcal{L} = \ln L_N = \sum_{i=1}^N \sum_{j=1}^m y_{ij} \ln p_{ij}, \quad (15.5)$$

where $p_{ij} = F_j(\mathbf{x}_i, \boldsymbol{\beta})$ is a function of parameters $\boldsymbol{\beta}$ and regressors, defined in (15.4). More generally, the number of alternatives may vary across different individuals, so that m choices become m_i choices.

The first-order conditions for the MLE $\widehat{\boldsymbol{\beta}}$ are that it solves

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} = \sum_{i=1}^N \sum_{j=1}^{m_i} \frac{y_{ij}}{p_{ij}} \frac{\partial p_{ij}}{\partial \boldsymbol{\beta}} = \mathbf{0}, \quad (15.6)$$

which is usually nonlinear in $\boldsymbol{\beta}$. The distribution of y_i is necessarily multinomial, so correct specification of the dgp means correct specification of the functional forms $F_j(\mathbf{x}_i, \boldsymbol{\beta})$ for the probabilities p_{ij} . This ensures consistency as then $E[y_{ij}] = p_{ij}$,

so taking the expectation of (15.6) yields $E[\partial\mathcal{L}/\partial\beta] = \sum_{i=1}^N \sum_{j=1}^m \partial p_{ij}/\partial\beta$, which equals zero since $\sum_{j=0}^m p_{ij} = 1$.

The usual asymptotic theory applies and the variance matrix is minus the inverse of the information matrix. Differentiating the double sum in (15.6) with respect to β' and using $E[y_{ij}] = p_{ij}$ yields upon simplification

$$\widehat{\beta} \stackrel{a}{\sim} \mathcal{N} \left[\beta_0, \left(\sum_{i=1}^N \sum_{j=1}^m \frac{1}{p_{ij}} \frac{\partial p_{ij}}{\partial \beta} \frac{\partial p_{ij}}{\partial \beta'} - \frac{\partial^2 p_{ij}}{\partial \beta \partial \beta'} \Big|_{\beta_0} \right)^{-1} \right]. \quad (15.7)$$

Provided observations are independent over i there is no need to use more general sandwich forms of the variance matrix since the data are definitely multinomial distributed and the information matrix equality will hold.

As already mentioned, different models correspond to different choices of $F_j(\mathbf{x}_i, \beta)$ for p_{ij} and hence different expressions in (15.6) and (15.7).

Maximum likelihood estimation for choice-based samples, such as those that oversample infrequently observed outcomes, is presented in Sections 14.5 and 24.4.

15.3.3. Moment-Based Estimation

For simple cross-section applications the standard estimation procedure is the MLE.

However, when complications such as endogeneity or correlation across observational unit i arise, it can be more convenient to instead use **moment-based estimators**. Assuming the probabilities are correctly specified, we can consider any estimator with estimating equations

$$\sum_{i=1}^N \sum_{j=1}^m (y_{ij} - p_{ij}) \mathbf{z}_i = \mathbf{0}, \quad (15.8)$$

where \mathbf{z}_i , a vector of the same dimension as β , does not depend on y_{ij} , for example, $\mathbf{z}_i = \partial p_{ij}/\partial\beta$. This estimator will be consistent if the functional form for p_{ij} is correctly specified, as then $E[y_{ij}] = p_{ij}$ and the double sum on the left-hand side of (15.8) has expected value zero. The efficiency of the estimator will vary with the choice of \mathbf{z}_i and in the most general case GMM estimation procedures can be used. The estimating equations (15.8) are the basis for the method of simulated moments estimator for the multinomial probit model (see Section 15.8.2).

15.3.4. Alternative-Varying Regressors

Multinomial regression models differ not only in the choice of function $F_j(\cdot)$ in (15.4) but also in how regressors and parameters vary across the alternatives.

At one extreme all regressors may be **alternative-varying**, meaning that they take different values for different alternatives. Let \mathbf{x}_{ij} denote the value of the regressors for individual i and alternative j , and let $\mathbf{x}_i = [\mathbf{x}'_{i1} \mathbf{x}'_{i2} \dots \mathbf{x}'_{im}]'$. Then (15.4) is usually of the form

$$F_j(\mathbf{x}_i, \beta) = F_j(\mathbf{x}'_{i1}\beta, \dots, \mathbf{x}'_{im}\beta),$$

where the parameters β are constant across alternatives. An example is the conditional logit model defined later in (15.10).

At the other extreme all regressors may be **alternative-invariant**, meaning that \mathbf{x}_i does not vary across alternatives. An example is individual socioeconomic characteristics in a model of transportation mode choice. Then (15.4) is usually of the form

$$F_j(\mathbf{x}_i, \beta) = F_j(\mathbf{x}'_i \beta_1, \dots, \mathbf{x}'_i \beta_m),$$

where the parameters β_j differ across alternatives and $\beta = [\beta'_1 \ \beta'_2 \ \dots \ \beta'_m]'$. Parameter identification requires a normalization such as $\beta_1 = \mathbf{0}$. An example is the multinomial logit model defined later in (15.11).

The distinction between alternative-varying and alternative-invariant regressors is of practical importance, as standard notation and computer programs for multinomial models work exclusively with one or the other. In practice, of course, some regressors may be alternative-varying and others alternative-invariant. In such cases it is best to use a program written for alternative-varying regressors, as it is possible to go from alternative-invariant regressors to the alternative-varying format. Let \mathbf{x}_i be a $K \times 1$ vector. Then define \mathbf{x}_{ij} to be a $Km \times 1$ vector with zeros everywhere except that the j th block is \mathbf{x}_i , that is,

$$\mathbf{x}_{ij} = [\mathbf{0}' \cdots \mathbf{0}' \quad \mathbf{x}'_i \quad \mathbf{0}' \cdots \mathbf{0}']',$$

and define $\beta = [\mathbf{0}' \ \beta'_2 \ \cdots \ \beta'_m]'$, where $\beta_1 = \mathbf{0}$ is a normalization. Then $\mathbf{x}'_i \beta_j = \mathbf{x}'_{ij} \beta$. The regressors are essentially included as interactions with alternative-specific dummies. An example was given in Section 15.2.3. It is also possible to go from the alternative-specific to the alternative-invariant format, but then $(m - 1)$ parameter equality constraints need to be imposed for each of the alternative-specific regressors.

15.3.5. Revealed Preference and Stated Preference Data

The multinomial data used in microeconometric studies often arise from individual consumer choice. Consumer choice data may be either **revealed preference data**, which are data on actual decisions and outcomes, or **stated preference data**, which are survey responses to hypothetical questions. An example of revealed preference data would be actual occupational choice. An example of stated preference data would be a marketing study for fuel-efficient vehicles that asks a respondent to choose among various hypothetical vehicles that differ in characteristics such as fuel consumption, range, and price.

Revealed preference data often provide little or no data on alternatives other than that chosen. For example, we may know the price to an individual consumer of the chosen product but not the prices of alternative products. The attraction of stated preference data for multinomial modeling is that data are available on key variables such as price for all possible alternatives. This is particularly advantageous if one wishes to predict the probability of choice or market share of a new alternative on the basis of characteristics of the new alternative, as all parameters can be alternative-invariant if all regressors vary across alternatives.

There is some controversy in using stated preference data, because responses can vary with the wording of questions. Moreover, people may overstate or understate their willingness to pay to support particular policies. For example, some might overstate their willingness to support an environmentally friendly policy.

Shopping **scanner data** are especially attractive because they provide data on revealed choice while at the same time data on prices across all alternatives are also provided.

15.3.6. Model Evaluation and Selection

Regression parameters in multinomial models can be difficult to directly interpret. Instead, it is useful to consider the marginal effect (or elasticities) of changes in regressors on outcome probabilities. Formulas for conditional and multinomial logit models are given in Section 15.4.3 and have been used in the Section 15.2 application.

Several model evaluation methods are presented in Amemiya (1981) and Maddala (1983). Using R^2 measures based on the analogue of squared residuals does not work well. Comparisons of predicted probabilities with actual outcomes are of limited value as MNL models estimated with intercept impose in estimation the restriction that the average of the predicted probabilities equals the sample average probabilities for each alternative. It can be useful to look at the range of the in-sample fitted probabilities for each alternative. The wider the range the more discriminating is the model. For more detail see the discussion in Section 14.3.7 for binary outcomes.

Multinomial models are usually estimated by maximum likelihood. Thus to the extent that models are nested one can use standard likelihood ratio tests. When models are nonnested one can use variants of Akaike's information criteria based on the fitted log-likelihood with a degrees-of-freedom adjustment for the number of parameters in the model (see Section 8.5.1).

A useful pseudo- R^2 measure, due to McFadden (1973), is

$$R^2 = 1 - \ln L_{\text{fit}} / \ln L_0, \quad (15.9)$$

where $\ln L_{\text{fit}}$ denotes the fitted model and L_0 denotes an intercept-only model that estimates the probability of each alternative to be the sample average. For any multinomial model the theoretical maximum value of the log-likelihood is zero. This arises if $p_{ij} = 1$ when $y_{ij} = 1$ and $p_{ij} = 0$ otherwise, for i and j . Thus the R^2 measure can be rewritten as

$$R^2 = \frac{\ln L_{\text{fit}} - \ln L_0}{\ln L_{\max} - \ln L_0}.$$

This can be interpreted as the fraction of the maximum potential gain in log-likelihood that is achieved by the fitted model (see Section 8.7.1).

15.4. Multinomial Logit

The simplest multinomial model is the multinomial logit model, proposed by Luce (1959). The commonly used variants of this model differ according to whether or not regressors vary across alternatives. Many of the issues presented in this section pertain to other models presented more briefly in subsequent sections.

15.4.1. Conditional, Multinomial, and Mixed Logit Models

For alternative-varying regressors (see Section 15.3.4) the **conditional logit model** is used. The CL model specifies

$$p_{ij} = \frac{e^{\mathbf{x}'_{il}\beta}}{\sum_{l=1}^m e^{\mathbf{x}'_{il}\beta}}, \quad j = 1, \dots, m. \quad (15.10)$$

Since $\exp(\mathbf{x}'_{il}\beta) > 0$ these probabilities lie between 0 and 1 and sum over j to one. Indeed, once one has seen the formula (15.10) it appears to be the most simple specification that ensures well-behaved probabilities. Because $\sum_{j=1}^m p_{ij} = 1$ an equivalent model is obtained by defining \mathbf{x}_{ij} to be deviations of regressors from values of alternative 1, say, and setting $\mathbf{x}_{i1} = \mathbf{0}$.

When instead the regressors do not vary over alternatives, the **multinomial logit model** is used. The MNL model specifies

$$p_{ij} = \frac{e^{\mathbf{x}'_i\beta_j}}{\sum_{l=1}^m e^{\mathbf{x}'_i\beta_l}}, \quad j = 1, \dots, m. \quad (15.11)$$

Because $\sum_{j=1}^m p_{ij} = 1$, a restriction is needed to ensure model identification and the usual restriction is that $\beta_1 = \mathbf{0}$.

The two models can be combined into what some authors call a **mixed logit model**, with

$$p_{ij} = \frac{e^{\mathbf{x}'_{il}\beta + \mathbf{w}'_i\gamma_j}}{\sum_{l=1}^m e^{\mathbf{x}'_{il}\beta + \mathbf{w}'_i\gamma_l}}, \quad j = 1, \dots, m, \quad (15.12)$$

where \mathbf{x}_{ij} vary over alternatives and \mathbf{w}_i do not vary over alternatives. As discussed in Sections 15.2.3 and 15.3.4, the mixed and MNL models can be reexpressed as a CL model. Note that the term mixed logit model is also sometimes used for a quite different model detailed in Section 15.7.

All these models can be given the general label of multinomial logit, but we follow the standard convention in distinguishing between the MNL and CL models.

An obvious generalization of the multinomial logit model is

$$p_{ij} = \frac{V_{ij}}{\sum_{l=1}^m V_{il}}, \quad j = 1, \dots, m, \quad (15.13)$$

where $V_{ij} > 0$ can be quite general functions of regressors \mathbf{x}_i and parameters β . This is called the **universal logit model**. Although this can generate a potentially rich class of models it is seldom used in econometrics as it does not arise naturally from choice theory.

15.4.2. ML Estimation of CL and MNL Models

We present key formulas for the conditional logit and multinomial logit models. Complete derivations are given in Section 15.12.

For the CL model, where p_{ij} is defined in (15.10), $\partial p_{ij} / \partial \beta = p_{ij}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)$, where $\bar{\mathbf{x}}_i = \sum_{l=1}^m p_{il}\mathbf{x}_{il}$ is a probability weighted average of the regressors (see Section 15.12.1). The CL first-order conditions, given in (15.6) for general p_{ij} , simplify immediately to

$$\sum_{i=1}^N \sum_{j=1}^m y_{ij}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i) = \mathbf{0}. \quad (15.14)$$

Differentiating with respect to β' , using $E[y_{ij}] = p_{ij}$, and performing some further algebra yields

$$\hat{\beta}_{\text{CL}} \stackrel{a}{\sim} \mathcal{N} \left[\beta, \left(\sum_{i=1}^N \sum_{j=1}^m p_{ij}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)' \right)^{-1} \right]. \quad (15.15)$$

For the MNL model, p_{ij} is defined in (15.11) and it is shown in Section 15.12.2 that $\partial p_{ij} / \partial \beta_k = p_{ij}(\delta_{ijk} - p_{ik})\mathbf{x}_i$, where δ_{ijk} is an indicator variable equal to 1 if $j = k$ and equal to 0 if $j \neq k$, and that the resultant MNL first-order conditions simplify after some algebra to

$$\frac{\partial \mathcal{L}}{\partial \beta_k} = \sum_{i=1}^N (y_{ik} - p_{ik})\mathbf{x}_i = 0, \quad k = 1, \dots, m. \quad (15.16)$$

As usual $\hat{\beta}_{\text{MNL}} \stackrel{a}{\sim} \mathcal{N}[\beta, (E[\partial^2 \mathcal{L} / \partial \beta \partial \beta'])^{-1}]$, where further algebra shows that the information matrix has jk th block

$$E \left[\frac{\partial^2 \mathcal{L}}{\partial \beta_j \partial \beta_k'} \right] = \sum_{i=1}^N p_{ij}(\delta_{ijk} - p_{ik})\mathbf{x}_i \mathbf{x}_i', \quad j = 1, \dots, m, \quad k = 1, \dots, m. \quad (15.17)$$

15.4.3. Regression Parameter Interpretation

Care is needed in the interpretation of parameters in any nonlinear model. This is particularly so for multinomial models where, for example, there is not necessarily a one-to-one correspondence between coefficient sign and coefficient probability. Here we present results used in the Section 15.2 application.

Marginal Effects and Elasticities

We focus on **marginal effects** on the choice probabilities of a change in the regressor for a given individual. **Elasticities** can then be computed by multiplying the marginal effect by the current regressor value and dividing by the probability. Typically these are then averaged over individuals to give an average marginal effect or average elasticity.

For the CL model consider the effect on the j th probability of changing by one unit the value of the regressor for the k th alternative. For example, what is the effect

on the probabilities of choosing various modes of transportation if travel time by bus increases by a minute whereas the travel time by other modes is unchanged? From Section 15.12.1

$$\frac{\partial p_{ij}}{\partial \mathbf{x}_{ik}} = p_{ij}(\delta_{ijk} - p_{ik})\beta_i, \quad (15.18)$$

where δ_{ijk} was defined after (15.15). It follows that if the regression coefficient is positive then an increase in the corresponding component of the regressor value for the k th alternative increases the probability of the k th alternative and decreases the probability of the other alternatives.

For the MNL model consider instead the effect on the j th probability of changing by one unit a regressor that takes the same value across all alternatives. For example, what is the effect on the probabilities of choosing to work if age increases by one year? From Section 15.12.2

$$\frac{\partial p_{ij}}{\partial \mathbf{x}_i} = p_{ij}(\beta_j - \bar{\beta}_i), \quad (15.19)$$

where $\bar{\beta}_i = \sum_l p_{il}\beta_l$ is a probability weighted average of the β_l . It follows that the sign of the response is not necessarily given by the sign of β_j , unless $\beta_j > \beta_k$ for all $k \neq j$, and it does not necessarily make any sense to test whether a particular coefficient is zero. As in other nonlinear models we may compute the average response $N^{-1} \sum_i \partial p_{ij} / \partial \mathbf{x}_i = N^{-1} \sum_i p_{ij}(\beta_j - \bar{\beta}_i)$, or we can use noncalculus methods and compare the change in the average predicted probability as regressors change.

Comparison to Base Category

The coefficients in the CL and MNL models can also be given a more direct logit-like interpretation in terms of relative risk (detailed in Section 14.3.4). This is because the models can be reexpressed as binary logit models.

For the MNL model, comparison is to a base category, which is the alternative normalized to have coefficients equal to zero. To see this note that the multinomial logit probabilities (15.11) imply that the conditional probability of observing alternative j given that either alternative j or alternative k is observed is

$$\begin{aligned} \Pr[y = j | y = j \text{ or } k] &= \frac{p_j}{p_j + p_k} \\ &= \frac{e^{\mathbf{x}'\beta_j}}{e^{\mathbf{x}'\beta_j} + e^{\mathbf{x}'\beta_k}} \\ &= \frac{e^{\mathbf{x}'(\beta_j - \beta_k)}}{1 + e^{\mathbf{x}'(\beta_j - \beta_k)}}, \end{aligned} \quad (15.20)$$

which is a logit model with coefficient $(\beta_j - \beta_k)$. The second equality comes after some simplification. Suppose normalization is on alternative 1, so that $\beta_1 = \mathbf{0}$. Then

$$\Pr[y_i = j | y_i = j \text{ or } 1] = \frac{e^{\mathbf{x}'_i\beta_j}}{1 + e^{\mathbf{x}'_i\beta_j}},$$

and β_j can be interpreted in the same way as the logit model coefficient for binary choice between alternatives j and 1. Similarly to the binary logit model the **relative risk** of choosing alternative j rather than alternative 1 is

$$\frac{\Pr[y_i = j]}{\Pr[y_i = 1]} = e^{\mathbf{x}_i' \boldsymbol{\beta}_j},$$

and hence $e^{\beta_{jr}}$ gives the proportionate change in this relative risk when x_{ir} changes by one unit. Such interpretations will vary according to which alternative is normalized to have zero coefficient, and for this interpretation to be really useful one needs to have a natural **base category**. For example, if interest lies in various alternative commute modes to traveling by car then normalize the coefficients for the car alternative to equal zero.

A similar approach can also be applied to the CL model, with

$$\Pr[y_i = j | y_i = j \text{ or } k] = \frac{e^{(\mathbf{x}_{ij} - \mathbf{x}_{ik})' \boldsymbol{\beta}}}{1 + e^{(\mathbf{x}_{ij} - \mathbf{x}_{ik})' \boldsymbol{\beta}}}, \quad (15.21)$$

and normalization now is with respect to regressor values for a base category.

15.4.4. Independence of Irrelevant Alternatives

A limitation of the CL and MNL models is that discrimination among the m alternatives reduces to a series of pairwise comparisons that are unaffected by the characteristics of alternatives other than the pair under consideration. This is clear from (15.20) and (15.21), which show that the MNL model reduces to a binary choice logit model between any pair of choices. The conditional probability does not depend on other alternatives.

As an extreme example, the conditional probability of commute by car given commute by car or red bus is assumed in an MNL or CL model to be independent of whether commuting by blue bus is an option. However, in practice we would expect introduction of a blue bus, which is the same as a red bus in every aspect except color, to have little impact on car use and to halve use of the red bus, leading to an increase in the conditional probability of car use given commute by car or red bus.

This weakness of MNL is known in the literature as the red bus–blue bus problem, or more formally as the assumption of **independence of irrelevant alternatives**. It can be tested by a Hausman test (see Hausman and McFadden, 1984). For example, we could compare the coefficient estimates of red bus in a three-choice model of car, red bus, and blue bus, again with car the base category, with the coefficient estimates of red bus in a binary choice model of car and red bus, again with car the base category.

Much of the econometrics literature has focused on alternative unordered models that do not have this weakness. These models are presented in Sections 15.6–15.8.

15.5. Additive Random Utility Models

Unordered multinomial models more general than multinomial and conditional logit can be obtained using the general framework of additive random utility models, presented in this section. Subsequent sections describe the leading examples.

15.5.1. ARUM

The **additive random utility model** was introduced in Section 14.4.2 for binary outcomes. In the general m -choice multinomial model the utility of the j th choice is specified to be given by

$$U_j = V_j + \varepsilon_j, \quad j = 1, 2, \dots, m, \quad (15.22)$$

where V_j denotes the deterministic component of utility and ε_j denotes the random component of utility. For the i th individual usually $V_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta}$ or $V_{ij} = \mathbf{x}'_i\boldsymbol{\beta}_j$, though more structural analysis may specify direct or indirect utility functions used in consumer demand theory. For notational simplicity we suppress the individual subscript i in the following.

The chosen alternative is that with the highest utility, so that

$$\begin{aligned} \Pr[y = j] &= \Pr[U_j \geq U_k, \text{ all } k \neq j] \\ &= \Pr[U_k - U_j \leq 0, \text{ all } k \neq j] \\ &= \Pr[\varepsilon_k - \varepsilon_j \leq V_j - V_k, \text{ all } k \neq j] \\ &= \Pr[\tilde{\varepsilon}_{kj} \leq -\tilde{V}_{kj}, \text{ all } k \neq j], \end{aligned} \quad (15.23)$$

where the tilda and second subscript j denotes differencing with respect to reference alternative j .

Different multinomial models can be generated by different assumptions about the joint distribution of the error terms. These models are valid statistically, with probabilities summing to one. Additionally, they are consistent with the standard economic theory of decision making.

For example, consider the expression for $\Pr[y = 1]$ in a three-choice model. Using the last equality in (15.23) and defining $\tilde{\varepsilon}_{31} = \varepsilon_3 - \varepsilon_1$ and $\tilde{\varepsilon}_{21} = \varepsilon_2 - \varepsilon_1$ we have

$$\begin{aligned} \Pr[y = 1] &= \Pr[\tilde{\varepsilon}_{21} \leq -\tilde{V}_{21}, \quad \tilde{\varepsilon}_{31} \leq -\tilde{V}_{31}] \\ &= \int_{-\infty}^{-\tilde{V}_{31}} \int_{-\infty}^{-\tilde{V}_{21}} f(\tilde{\varepsilon}_{21}, \tilde{\varepsilon}_{31}) d\tilde{\varepsilon}_{21} d\tilde{\varepsilon}_{31}, \end{aligned} \quad (15.24)$$

which is a bivariate integral that generally does not have an analytical solution. More generally, an m -choice model involves an $(m - 1)$ -variate integral that may or may not yield a closed-form solution for $\Pr[y = j]$.

In general all the errors $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_m$ may be correlated across choices. Some **covariance restrictions** are necessary, however, as the model is **identified** only up to the $(m - 1)$ error-difference pairs (see the last equality in (15.23)), and additionally one variance needs to be specified since the U_j are only determined up to scale.

15.5.2. Different Unordered Multinomial Models

Different unordered multinomial models arise from different assumptions on the joint distribution of $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_m$. Analysis is simplest if the error assumptions lead to a closed-form solution for the choice probabilities. However, in many applications these assumptions are felt to be too restrictive.

The computationally-intensive methods summarized in Chapter 12 permit estimation even if there is no closed-form solution for the choice probabilities. Sections 15.7.2 and 15.8.2 present multinomial examples of these methods.

Type 1 Extreme Value Errors

We first assume that the errors ε_j are iid type 1 extreme value, with density

$$f(\varepsilon_j) = e^{-\varepsilon_j} \exp(-e^{-\varepsilon_j}), \quad j = 1, 2, \dots, m. \quad (15.25)$$

The properties of this density were given in Section 14.4.2, where it was shown to lead to a logit model in the binary outcome case.

For multinomial outcomes modelled using the ARUM with type I extreme value errors it can be shown that (15.23) yields

$$\Pr[y = j] = \frac{e^{V_j}}{e^{V_1} + e^{V_2} + \dots + e^{V_m}}. \quad (15.26)$$

This is the CL model when $V_j = \mathbf{x}'_j \boldsymbol{\beta}$ and the MNL model when $V_j = \mathbf{x}' \boldsymbol{\beta}_j$. The result can be obtained either by integration and simplification similar to the binary case (see Section 14.8), or as a special case of the nested logit result derived in Section 15.6. Thus conditional and multinomial logit models can be obtained from an ARUM.

The assumption that the errors ε_j are independent across alternatives j is too restrictive as it is likely to be violated if two alternatives are similar. For example, suppose alternatives 1 and 2 are similar. A low value of ε_1 (i.e., large and negative) leads to overprediction of the utility of alternative 1. We then also expect to overpredict the utility of alternative 2, so that ε_2 also takes a low value. Since low values of ε_1 and ε_2 tend to go together, and similarly for high values, the errors must be correlated. This is another way of viewing the “red bus–blue bus” problem, and it is a manifestation of a failure of the logit assumption of independence of irrelevant alternatives.

The generalized extreme value model and the nested logit model (see Section 15.6) relax the assumption that the extreme value errors are independent across choices. The errors are grouped with independence across groups but correlation permitted within groups. Closed-form solutions are then available for the choice probabilities. Although these models are richer than the MNL model, the special case of no correlation within groups, in many applications the grouping of errors can be somewhat arbitrary.

The random parameters logit model (see Section 15.7) introduces additional randomness into the MNL model that induces correlation of utilities across alternatives. This is an example of a generalized random utility model (see Section 15.7.3).

Normally Distributed Errors

The multinomial probit model (see Section 15.8) arises if the errors $\varepsilon_1, \dots, \varepsilon_m$ are assumed to be joint normal distributed. This error assumption is a more natural starting point than one of type 1 extreme value. It permits a very rich correlation structure, at the expense of the need to use numerical or simulation methods that accommodate an $(m - 1)$ -variate normal integral.

15.5.3. Consistency with Random Utility Models

It is always possible to present an analytical expression for choice probabilities that lie between zero and one and that sum over alternatives to one. A quite general example is the universal logit model in (15.13). The econometrics literature has placed great emphasis in restricting attention to multinomial models that are consistent with maximization of a random utility function. This is similar to restricting analysis to demand functions that are consistent with consumer choice theory.

Let $V = (V_1, \dots, V_m)$. From Borsch-Supan (1987, p. 19), a set of choice probabilities $p_j(V)$, $j = 1, \dots, m$, is compatible with maximization of an ARUM if

1. $p_j(V) \geq 0$, $\sum_{j=1}^m p_j(V) = 1$, and $p_j(V) = p_j(V + \alpha)$ for all $\alpha \in R$;
2. $\partial p_j(V)/\partial V_k = \partial p_k(V)/\partial V_j$; and
3. $\partial^{(m-1)} p_j(V)/\partial V_1 \dots [\partial V_i] \dots \partial V_m \geq 0$, where the square bracket denotes a term to be dropped out.

These conditions, due to Williams (1977), Daly and Zachary (1979), and McFadden (1981), ensure in turn (1) well-behaved probabilities and translation invariance; (2) integrability of p_j similar to the Slutsky condition; and (3) that the distribution function of the errors in the corresponding ARUM has a proper (nonnegative) density function.

15.5.4. Welfare Analysis

A major advantage of using a multinomial model that is a random utility model is that it permits welfare analysis. Then one can place a dollar value on the effect of changing one or more of the determinants of choice, such as price or time cost of travel in transportation mode choice.

Standard **welfare analysis** uses compensating variation or equivalent variation. The deterministic component of utility in (15.22) is specified as the indirect utility function

$$V_j = V(I - p_j, \mathbf{x}_j), \quad (15.27)$$

where I denotes income, p_j is the price of the j th alternative, and \mathbf{x}_j are characteristics associated with the j th alternative. For notational simplicity the unknown regression parameters β are suppressed. Then utility of alternative j is

$$U_j = U(I - p_j, \mathbf{x}_j, \varepsilon_j) = V(I - p_j, \mathbf{x}_j) + \varepsilon_j. \quad (15.28)$$

Suppose we change the characteristics from \mathbf{x}'_j to \mathbf{x}''_j . Then **compensating variation** CV is the change in income needed to hold utility at its initial level, so that the highest utility level attainable with income I and characteristics \mathbf{x}'_j must equal the highest level attainable with income $(I - CV)$ and characteristics \mathbf{x}''_j . Thus compensating variation CV is implicitly defined as the solution to

$$\max_{j=1,\dots,m} U(I - p_j, \mathbf{x}'_j, \varepsilon_j) = \max_{j=1,\dots,m} U(I - CV - p_j, \mathbf{x}''_j, \varepsilon_j). \quad (15.29)$$

As an example, consider a two-choice model where $U_j = I + x_j + \varepsilon_j$, $j = 1, 2$, and the scalar x_j changes from x'_j to x''_j . Then there are four possibilities. If alternative 1 is chosen before and after then $CV = (x''_1 - x'_1)$, since then $U''_1 = I - CV + x''_1 + \varepsilon_1 = I + x'_1 + \varepsilon_1 = U'_1$. Similarly, if alternative 2 is chosen before and after then $CV = (x''_2 - x'_2)$. If switching occurs from alternative 1 to alternative 2 then $U''_2 = U'_1$ implies $I - CV + x''_2 + \varepsilon_2 = I + x'_1 + \varepsilon_1$, which implies $CV = x''_2 - x'_1 + \varepsilon_2 - \varepsilon_1$. Similarly, if switching occurs from alternative 2 to alternative 1 then $CV = x''_1 - x'_2 + \varepsilon_1 - \varepsilon_2$. More generally, for m choices the compensating variation in this simple example is $CV_{jk} = V''_k - V'_j + \varepsilon_k - \varepsilon_j$ if the change in \mathbf{x} leads to a change from alternative j to alternative k .

The compensating variation depends on observables (I , p_j , and \mathbf{x}_j), parameters that can be estimated, and on unobservable errors ε_j . The unobservables are eliminated by computing the expected compensating variation $E[CV]$, which involves integrating over ε_j . From the preceding example it should be clear that this integration can be quite difficult. Dagsvik and Karlström (2004) provide quite general results, discussed further in Section 15.6.5.

For some models there is no analytical solution for $E[CV]$. Then one instead needs to numerically integrate over ε_j the function for CV defined in (15.29). From Section 12.3.2 this integral can be simulated in the following way:

1. At iteration s draw ε^s from the distribution of $\varepsilon = (\varepsilon_1, \dots, \varepsilon_m)$.
2. Calculate CV^s from $\max_{j=1,\dots,m} U(I - p_j, \mathbf{x}'_j, \varepsilon_j^t) = \max_{j=1,\dots,m} U(I - CV^s - p_j, \mathbf{x}''_j, \varepsilon_j^t)$.
3. Repeat steps 1 and 2 S times.
4. Estimate $E[CV]$ by $S^{-1} \sum_{s=1}^S CV^s$.

This method yields $E[CV]$ for each individual in the sample. Averaging, possibly with weighting, provides a population estimate. Application to the GEV model is discussed in Section 15.6.5.

15.6. Nested Logit

The nested logit is the most analytically tractable generalization of the multinomial models. It is the ideal model to use when there is a clear nesting structure, but not all multinomial choice applications have an obvious nesting structure.

15.6.1. Generalized Extreme Value Model

McFadden (1978) proposed a quite general class of model based on the assumption that the joint distribution of the errors is the **generalized extreme value (GEV)** distribution with joint distribution function

$$F(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_m) = \exp[-G(e^{-\varepsilon_1}, e^{-\varepsilon_2}, \dots, e^{-\varepsilon_m})], \quad (15.30)$$

where the function $G(Y_1, Y_2, \dots, Y_m)$ is specified to satisfy a number of assumptions including nonnegativity, homogeneity of degree one, mixed partial derivatives that are continuous and nonpositive for even order and nonnegative for odd order, and $\lim_{Y_j \rightarrow \infty} G(Y_1, Y_2, \dots, Y_m) = \infty$. These assumptions ensure that the joint distribution and resulting marginal distributions are well defined and that probabilities sum to one.

If the errors are GEV distributed then an explicit solution for the probabilities in the random utility model (15.22) can be obtained, with

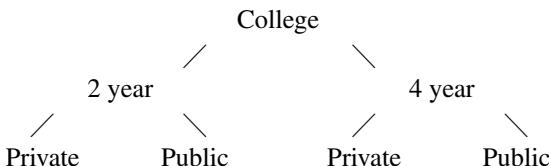
$$p_j = \Pr[y = j] = e^{V_j} \frac{G_j(e^{-V_1}, e^{-V_2}, \dots, e^{-V_m})}{G(e^{-V_1}, e^{-V_2}, \dots, e^{-V_m})}, \quad (15.31)$$

where $G_j(Y_1, Y_2, \dots, Y_m) = \partial G(Y_1, Y_2, \dots, Y_m)/\partial Y_j$ (see McFadden, 1978, p. 81).

A wide range of models can be obtained by different choices of $G(Y_1, Y_2, \dots, Y_m)$. The MNL model is obtained if $G(Y_1, Y_2, \dots, Y_m) = \sum_{k=1}^m Y_k$; hence the MNL model is a GEV model. The other widely used GEV model is the nested logit model.

15.6.2. Nested Logit Model

The nested logit model breaks decision making into groups. A simple example is to consider choice of college, where people first decide whether to go to a two-year or four-year college, and then within each of these paths whether to go to a public or private college. The situation is depicted as follows:

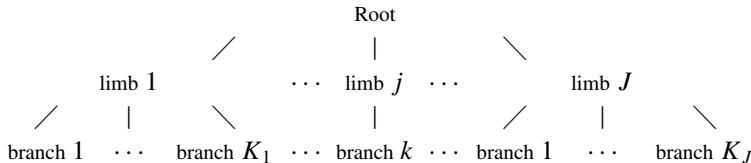


The errors in a random utility model are permitted to be correlated for each option within the two-year and four-year groups, but they are uncorrelated across these two groups.

More generally, we suppose that at the top level there are J limbs to choose from. The j th limb has K_j branches numbered $j1, \dots, jk, \dots, jK_j$. The utility for the alternative in the j th of J limbs and k th of K_j branches is then

$$U_{jk} = V_{jk} + \varepsilon_{jk}, \quad k = 1, 2, \dots, K_j, \quad j = 1, 2, \dots, J, \quad (15.32)$$

where for an m -choice model $K_1 + \dots + K_J = m$. This is illustrated as follows:



$$V_{11} + \varepsilon_{11} \quad \dots \quad V_{1K_1} + \varepsilon_{1K_1} \quad \dots \quad V_{jk} + \varepsilon_{jk} \quad \dots \quad V_{J1} + \varepsilon_{J1} \quad \dots \quad V_{JK_J} + \varepsilon_{JK_J}$$

There can be additional levels, with the third level being a twig, etc. For notational simplicity we present results for a two-level model.

For any model with this nesting p_{jk} , the joint probability of being on limb j and branch k , can be factored as p_j , the probability of choosing limb j , times $p_{k|j}$, the probability of choosing branch k conditional on being on limb j . Thus

$$p_{jk} = p_j \times p_{k|j}.$$

The nested logit model of McFadden (1978) arises when the error terms ε_{jk} have the GEV joint cumulative distribution function

$$F(\varepsilon) = \exp[-G(e^{-\varepsilon_{11}}, \dots, e^{-\varepsilon_{1K_1}}; \dots; e^{-\varepsilon_{J1}}, \dots, e^{-\varepsilon_{JK_J}})] \quad (15.33)$$

for the following particular specification of the function $G(\cdot)$:

$$G(\mathbf{Y}) = G(Y_{11}, \dots, Y_{1K_1}, \dots, Y_{J1}, \dots, Y_{JK_J}) = \sum_{j=1}^J \left(\sum_{k=1}^{K_j} Y_{jk}^{1/\rho_j} \right)^{\rho_j}. \quad (15.34)$$

The parameter ρ_j is a function of the correlation between ε_{jk} and ε_{jl} but does not exactly equal the correlation parameter. In fact ρ_j can be shown to equal $\sqrt{1 - \text{Cor}[\varepsilon_{jk}, \varepsilon_{jl}]}$, so ρ_j is inversely related to the correlation and we expect $0 \leq \rho_j \leq 1$. The choice $\rho_j = 1$ corresponds to independence of ε_{jk} and ε_{jl} and leads to the MNL model. We call the parameters ρ_j the **scale parameters**, as they scale regression parameters in the models considered in the following.

Notation varies considerably across authors. McFadden (1978) and Maddala (1983) instead define this cdf in terms of $\sigma_j = 1 - \rho_j$, called the **dissimilarity parameter**. Others use $\mu_j = 1/\rho_j$. Many authors model alternative ij for the n th individual whereas we model alternative jk and reserve i for the i th individual.

The outcome indicator variables y_{jk} equal one if alternative jk is chosen and zero otherwise. Then from (15.32), $p_{jk} = \Pr[y_{jk} = 1] = \Pr[U_{jk} \geq U_{lm}, \text{ for all } l, m]$. Closed-form solutions for the probabilities p_{jk} , as a function of the V_{jk} and ρ_j , are derived in Section 15.12.3. These are then evaluated for the particular deterministic utility function

$$V_{jk} = \mathbf{z}'_j \boldsymbol{\alpha} + \mathbf{x}'_{jk} \boldsymbol{\beta}_j, \quad k = 1, \dots, K_j, \quad j = 1, \dots, J, \quad (15.35)$$

where \mathbf{z}_j varies over limbs only and \mathbf{x}_{jk} varies over both limbs and branches. The parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_j$ are called **regression parameters**.

The GEV model (15.32)–(15.35) yields the **nested logit model**

$$p_{jk} = p_j \times p_{k|j} = \frac{\exp(\mathbf{z}'_j \boldsymbol{\alpha} + \rho_j I_j)}{\sum_{m=1}^J \exp(\mathbf{z}'_m \boldsymbol{\alpha} + \rho_m I_m)} \times \frac{\exp(\mathbf{x}'_{jk} \boldsymbol{\beta}_j / \rho_j)}{\sum_{l=1}^{K_j} \exp(\mathbf{x}'_{jl} \boldsymbol{\beta}_j / \rho_j)}, \quad (15.36)$$

see Section 15.12.3, where

$$I_j = \ln \left(\sum_{l=1}^{K_j} \exp(\mathbf{x}'_{jl} \boldsymbol{\beta}_j / \rho_j) \right) \quad (15.37)$$

is called the **inclusive value** or the **log-sum**. One attraction of the nested logit model is that the probabilities p_i and $p_{j|i}$ are essentially of conditional logit form.

The preceding results are for regressors that vary across alternatives. The algebra can be adapted to alternative-invariant regressors $V_{jk} = \mathbf{z}' \boldsymbol{\alpha}_j + \mathbf{x}' \boldsymbol{\beta}_{jk}$, with a normalization of one of the $\boldsymbol{\beta}_{jk}$. Algebraically all that is needed is a partition $V_{jk} = A_j + B_{jk}$, where A_j pertains to the limb and B_{jk} pertains to both limb and branch.

15.6.3. Estimation of Nested Logit

For the i th observation we observe $K_1 + \dots + K_J$ outcomes y_{ijk} , where $y_{ijk} = 1$ if alternative jk is chosen and is zero otherwise. Then $p_{ijk} = p_{ik|j} \times p_{ij}$ and the density for one observation $\mathbf{y}_i = (y_{i11}, \dots, y_{iJK_J})$ can be compactly expressed as

$$f(\mathbf{y}_i) = \prod_{j=1}^J \prod_{k=1}^{K_j} [p_{ik|j} \times p_{ij}]^{y_{ijk}} = \prod_{j=1}^J \left(p_{ij}^{y_{ij}} \prod_{k=1}^{K_j} p_{ik|j}^{y_{ijk}} \right),$$

where $y_{ij} = \sum_{l=1}^{K_j} y_{ijl}$ equals one if limb j is chosen and equals zero otherwise.

The density for the sample is $\prod_{i=1}^N f(\mathbf{y}_i)$. The **FIML estimator** maximizes

$$\ln L = \sum_{i=1}^N \sum_{j=1}^J y_{ij} \ln p_{ij} + \sum_{i=1}^N \sum_{j=1}^J \sum_{k=1}^{K_j} y_{ijk} \ln p_{ik|j}, \quad (15.38)$$

with respect to parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}_j$, and ρ_j .

An alternative, less-efficient estimation is the **sequential estimator** or LIML estimator that exploits the partitioning of p_{jk} into the product of $p_{k|j}$ and p_j . The first stage bases estimation on the second term of the right-hand side of (15.38), which from (15.36) is a conditional logit model with estimated parameter $\boldsymbol{\beta}_j / \rho_j$. The second stage bases estimation on the first term of the right-hand side, which from (15.36) is a conditional logit model with added regressor \widehat{I}_{ij} , an estimate of the inclusive value in (15.37) that can be computed using the first-stage parameter estimates. The $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\rho}_j$ are obtained directly from the second stage, whereas $\widehat{\boldsymbol{\beta}}_j$ equals $\widehat{\rho}_j$ times the first-stage estimate $\widehat{\boldsymbol{\beta}}_j / \widehat{\rho}_j$.

This sequential estimator is less efficient than the FIML estimator, and at the second stage the usual CL standard errors underestimate the true standard errors of the sequential estimator since they do not allow for the estimation error in computing the inclusive value. McFadden (1981) gives the formula for correct standard errors, or the bootstrap

can be used. The sequential alternative estimator was originally proposed at a time when even conditional logit model estimation was challenging. Now it is relatively simple to code the likelihood function, so it is best to use FIML. Sequential estimation is potentially useful to provide starting values as the FIML log-likelihood is not globally concave.

As an example we applied the nested logit model to the data of Section 15.2. The nesting structure was shore or boat fishing at the higher level, with lower levels beach or pier (for shore fishing) and private or charter (for boat fishing). The regressors \mathbf{x}_{jk} in (15.36) that vary at the lower level were price (P) and catch rate (C). The regressors \mathbf{z}_j at the higher level that vary across shore or boat were an indicator variable d equal to one if shore fishing and $d \times I$, income interacted with the shore fishing indicator. Estimation by conditional logit (corresponding to $\rho_1 = \rho_2 = 1$) yielded a fitted model with $\ln L = -1252$, as expected smaller than the log-likelihood for the similar but less restricted model given in the last column of Table 15.2. FIML estimation of the corresponding nested logit model, with ρ_1 and ρ_2 now free to vary, led to a much higher log-likelihood model and rejection of the more restricted conditional logit model using the $\chi^2(2)$ likelihood ratio test statistic.

15.6.4. Discussion

The main limitation of the nested logit model is that not all choice problems lend themselves to an obvious nesting structure. One can still select the optimal nesting scheme using likelihood ratio tests, where appropriate, or Akaike's information criteria. However, the resulting scheme does not always accord with a priori expectations.

Another practical issue is that consistency of the nested logit model with choice from an ARUM requires that the three conditions in Section 15.5.2 are satisfied. The third of these conditions is satisfied globally if $0 \leq \rho_j \leq 1$, and with more than two levels of nesting it is additionally required that ρ at higher levels of the nest structure does not exceed ρ at lower levels of nesting. In practice it is possible to obtain estimates of ρ_j outside the unit interval. One can still use the model, as the choice probabilities are proper, but the model may no longer come from an ARUM. Borsch-Supan and others have considered local identification conditions under which the nested logit model may be consistent with ARUM even if ρ_j lie outside the unit interval. It can also be useful to do a grid search over ρ_j to constrain ρ_j to the unit interval and to enumerate the reduction in log-likelihood, if any, caused by doing so.

The nested logit model defined in (15.36) and (15.37) was proposed by McFadden (1978), who derived it as a GEV model. An **earlier variant** of the nested logit model was similar to (15.36) and (15.37), except that $\exp(\mathbf{x}'_{jl}\beta_j/\rho_j)$ was replaced by $\exp(\mathbf{x}'_{jl}\beta_j)$. This had an alternative derivation as a natural extension of the CL model, since CL is the special case of (15.36) and (15.37) with $\rho_j = 1$. See McFadden (1978, p. 79), Maddala (1983, p. 70), and Greene (2003, p. 726).

It is very important to note that the two variants differ if ρ_j differs across alternatives; see Koppelman and Wen (1998) and Train (2003, p. 88). Some early studies obtained sequential estimates that differed substantially from FIML estimates, casting doubt on the robustness of the nested logit model. However, in some of these studies

the different estimators were being applied to different variants of the nested logit model. Furthermore, even today different packages estimate different variants of the model.

The nested logit model can be extended to higher levels of alternatives (or nesting). For example, Goldberg (1995) has five levels: (1) buy any car; (2) buy a new car given yes to 1; (3) which of nine classes of car was purchased given yes to 2; (4) foreign or domestic; (5) model. An added attraction if some nests have numerous choices is that it is sufficient to base estimation on a fixed or randomly selected subset of the alternatives (see McFadden, 1978).

15.6.5. Welfare Analysis

Welfare analysis for the ARUM was presented in Section 15.5.4. In general there is no solution for $E[CV]$, the expected compensating variation.

Remarkably, for GEV models that are linear in income, $V(I - p_j, \mathbf{x}_j) = \alpha(I - p_j) + f(\mathbf{x}_j)$, McFadden (1995) and earlier workers show that there is an explicit solution

$$E[CV] = \frac{1}{\alpha} \left(\ln G \left(e^{V''_1}, \dots, e^{V''_m} \right) - \ln G \left(e^{V'_1}, \dots, e^{V'_m} \right) \right),$$

where the function $G(\cdot)$ for the GEV distribution is defined in (15.34), and V'_j and V''_j are the before and after values of the deterministic component of utility.

For GEV models with income appearing nonlinearly, however, there is no explicit solution. Then one approach is the simulation method given in Section 15.5.4. For a multinomial logit model this is simple as it is easy to draw extreme value errors using the transformation method of Section 12.8.2 – draw u from the uniform on $(0, 1)$ and then set $\varepsilon = -\ln(-\ln(u))$. For a more general nested logit model, however, it is difficult to randomly draw from a GEV distribution even as simple as the bivariate extreme value. McFadden (1995) proposed using the MCMC with the Metropolis–Hastings algorithm (see Section 13.5). Herriges and Kling (1999) give an excellent summary of this simulation method with application to nested logit models for the fishing data of Section 15.2, using various indirect utility functions including the translog.

More recently, Dagsvik and Karlström (2004) show that although there is no explicit solution for $E[CV]$ in the GEV model if income enters nonlinearity, it is analytically possible to reduce $E[CV]$ to a one-dimensional integral. Computing this integral using Gaussian quadrature will be much simpler than employing the afore-mentioned simulation method.

15.7. Random Parameters Logit

The random parameters logit model provides a simple way to generalize the MNL or CL model to permit the utilities of each alternative to be correlated. The model is perhaps the leading microeconomics example of a random parameters model for cross-section data.

15.7.1. Random Parameters Logit Model

The **random parameters logit (RPL) model** specifies the utility to the i th individual for the j th alternative to be

$$U_{ij} = \mathbf{x}'_{ij}\beta_i + \varepsilon_{ij}, \quad j = 1, 2, \dots, m, \quad (15.39)$$

where ε_{ij} are iid extreme value, as for the CL model, but additionally permits the parameters β_i to be random. The most common assumption is that

$$\beta_i \sim \mathcal{N}[\beta, \Sigma_\beta]. \quad (15.40)$$

One variation is to use the log-normal rather than normal distribution for parameters whose sign is known a priori. This model is also called a **mixed logit model**, using terminology borrowed from the panel setting for models with random parameters. By reexpressing the MNL model as a CL model, the results that follow also cover a random parameters MNL model.

The model can be rewritten as

$$\begin{aligned} U_{ij} &= \mathbf{x}'_{ij}\beta + v_{ij}, \\ v_{ij} &= \mathbf{x}'_{ij}\mathbf{u}_i + \varepsilon_{ij}, \end{aligned}$$

where $\mathbf{u}_i \sim \mathcal{N}[\mathbf{0}, \Sigma_\beta]$. Then $\text{Cov}[v_{ij}, v_{ik}] = \mathbf{x}'_{ij}\Sigma_\beta\mathbf{x}_{ik}$, $j \neq k$, so the introduction of random parameters has the attractive property of inducing correlation across alternatives.

In most applications the covariance matrix Σ_β is specified to be diagonal, and additionally some of the diagonal entries may be set to zero. Then the number of covariance parameters to estimate equals the number of components of β_i that are specified to be random.

As an example, consider a mixed CL model with scalar regressor and parameters β and σ_β^2 . Suppose the parameter estimates are $\hat{\beta} = 2.0$ with standard error 0.5 and $\hat{\sigma}_\beta^2 = 1.0$ with standard error 0.2. Then the null hypothesis of constant parameter, that is, $\sigma_\beta^2 = 0$, is strongly rejected since $t = 1.0/0.2 = 5.0$. The effect on $\Pr[y_i = j]$ of an increase in x_{ij} differs across individuals and is positive for about 97.5% of the sample, since it is estimated that $\beta_i \sim \mathcal{N}[2.0, 1.0]$. For an application that emphasizes interpretation of estimated coefficients, see Revelt and Train (1998).

The industrial organization literature considers **aggregation** over consumers of models similar to the RPL model to estimate demand parameters using **market-level data**. See, for example, Berry (1994) and Nevo (2001), and also Allenby and Rossi (1991).

15.7.2. Estimation of Random Parameters Logit

In the linear regression model with random parameters, OLS estimation yields estimates of the means β that are consistent though inefficient. In a nonlinear model, however, estimators that fail to control for the randomness of the parameters will be inconsistent. Thus the usual conditional logit MLE will be inconsistent if the dgp is

given by (15.39) and (15.40). Instead, ML estimation must explicitly account for the stochastic process for β_i .

If β_i were known, so that the only source of randomness is ε_{ij} , a CL model is obtained with probability $p_{ij} = e^{\mathbf{x}'_{ij}\beta_i} / \sum_{l=1}^m e^{\mathbf{x}'_{il}\beta_i}$. Since β_i is in fact random we need to integrate out this randomness. This yields

$$p_{ij} = \Pr[y_i = j] = \int \frac{e^{\mathbf{x}'_{ij}\beta_i}}{\sum_{l=1}^m e^{\mathbf{x}'_{il}\beta_i}} \phi(\beta_i | \boldsymbol{\beta}, \Sigma_{\boldsymbol{\beta}}) d\beta_i, \quad (15.41)$$

where the integral is multidimensional and $\phi(\beta_i | \boldsymbol{\beta}, \Sigma_{\boldsymbol{\beta}})$ denotes the multivariate normal density for β_i with mean $\boldsymbol{\beta}$ and variance $\Sigma_{\boldsymbol{\beta}}$.

The MLE maximizes $\ln L_N = \sum_{i=1}^N \sum_{j=1}^m y_{ij} \ln p_{ij}$ with respect to $\boldsymbol{\beta}$ and $\Sigma_{\boldsymbol{\beta}}$. The challenge is that there is no closed-form solution for the integral, whose dimension is given by the number of components of β_i that are random, with non-zero variance. Estimation is therefore by simulation methods.

One approach is to approximate p_{ij} using the direct simulator (see Section 12.4.1). This replaces the integral (15.41) by the average of S evaluations of the integrand at random draws of β_i from the $\mathcal{N}[\boldsymbol{\beta}, \Sigma_{\boldsymbol{\beta}}]$ distribution. The **MSL estimator** then maximizes

$$\ln \widehat{L}_N(\boldsymbol{\beta}, \Sigma_{\boldsymbol{\beta}}) = \sum_{i=1}^N \sum_{j=1}^m y_{ij} \ln \left[\frac{1}{S} \sum_{s=1}^S \frac{e^{\mathbf{x}'_{ij}\beta_i^{(s)}}}{\sum_{l=1}^m e^{\mathbf{x}'_{il}\beta_i^{(s)}}} \right], \quad (15.42)$$

where $\beta_i^{(s)}$, $s = 1, \dots, S$, are random draws from the density $\phi(\beta_i; \boldsymbol{\beta}, \Sigma_{\boldsymbol{\beta}})$. Since $\boldsymbol{\beta}$ and $\Sigma_{\boldsymbol{\beta}}$ are unknown, this summation is embedded in an iterative procedure with evaluation at $\boldsymbol{\beta}^{(r)}$ and $\Sigma_{\boldsymbol{\beta}}^{(r)}$ at the r th round. Consistency requires that $S \rightarrow \infty$ as well as $N \rightarrow \infty$ and that $\sqrt{N}/S \rightarrow \infty$ (see Section 12.4.3). Methods for speeding up computation include use of Halton sequences (see Section 12.7.4) and alternative simulators.

An alternative estimator uses Bayesian methods with relatively flat priors. Train (2001, 2003) specifies hierarchical priors with $\boldsymbol{\beta} \sim \mathcal{N}[\boldsymbol{\beta}^*, \Omega^*]$, where Ω^* is assumed to be large, and with $\Sigma_{\boldsymbol{\beta}}$ assumed to be inverse-Wishart distributed with degrees of freedom $K = \dim[\boldsymbol{\beta}]$ and scale parameter \mathbf{I}_K . Rather than working with the posterior for just $\boldsymbol{\beta}$ and $\Sigma_{\boldsymbol{\beta}}$ it is computationally quicker to additionally include β_i , $i = 1, \dots, N$. Then (1) the conditional posterior for $\boldsymbol{\beta} | \Sigma_{\boldsymbol{\beta}}, \beta_i$ is normal, (2) the conditional posterior for $\Sigma_{\boldsymbol{\beta}} | \boldsymbol{\beta}, \beta_i$ is inverse Wishart, and (3) the conditional posterior for $\beta_i | \Sigma_{\boldsymbol{\beta}}, \boldsymbol{\beta}$, which is proportional to the integrand in (15.41). Given these conditional posteriors estimation can be done using a variation of the Gibbs sampler (see Section 13.5.2), with the complication that draws for the third posterior need to use one iteration of the Metropolis–Hastings algorithm (see Section 13.5.4) because the full set of conditionals is not available. In an application this took similar computation time to the MSL estimator and, given the relatively flat prior, yielded parameter estimates and standard errors that were generally within 10% of those from MSL estimation.

15.7.3. Generalized Random Utility Models

Models more flexible than multinomial logit are desirable. In this regard there is currently great enthusiasm regarding the random parameters logit model. McFadden and Train (2000) show that any random utility model can be approximated arbitrarily well by a mixed model, though this result requires appropriate choice of regressors and mixing distribution.

There is no reason to restrict the random parameters approach to multinomial logit models. For example, it may be extended to nested logit models. Moreover, additional sources of randomness may be incorporated, notably latent classes and latent variables.

To present these extensions we begin with the ARUM (15.22). This specifies the utility to individual i of the j th alternative to be $U_{ij} = V_{ij}(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_{ij}$, where \mathbf{x}_i denotes observed data, $\boldsymbol{\beta}$ denotes unknown parameters, and ε_{ij} denotes an error independent over i but possibly correlated over j . Assume that the distribution of ε_{ij} is such that (15.23) yields a closed-form solution for the choice probabilities denoted

$$p_{ij} = F_j(\mathbf{V}_i(\mathbf{x}_i, \boldsymbol{\beta}), \boldsymbol{\theta}_\varepsilon),$$

where $\mathbf{V}_i(\mathbf{x}_i, \boldsymbol{\beta}) = [V_{i1}(\mathbf{x}_i, \boldsymbol{\beta}), \dots, V_{im}(\mathbf{x}_i, \boldsymbol{\beta})]$ and $\boldsymbol{\theta}_\varepsilon$ denotes any unknown parameters of the distribution of $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{im})$. Such a closed-form solution is possible if ε_i has a GEV distribution with special cases leading to multinomial logit and nested logit models.

A more general model introduces additional randomness into this model. First, the previously deterministic part of utility becomes $V_{ij} = V_{ij}(\mathbf{x}_i, \boldsymbol{\xi}_i, \boldsymbol{\beta})$. Then assuming that ε_i is such that a closed-form solution for the probabilities exist conditional on $\boldsymbol{\xi}_i$, unconditionally

$$p_{ij} = \int F_j(\mathbf{V}_i(\mathbf{x}_i, \boldsymbol{\xi}_i, \boldsymbol{\beta}), \boldsymbol{\theta}_\varepsilon) f(\boldsymbol{\xi}_i | \boldsymbol{\theta}_\xi) d\boldsymbol{\xi}_i, \quad (15.43)$$

where $f(\boldsymbol{\xi} | \boldsymbol{\theta}_\xi)$ denotes the density of $\boldsymbol{\xi}$. The RPL model is an example with $V_{ij} = \mathbf{x}'_{ij} \boldsymbol{\beta} + \mathbf{x}'_{ij} \boldsymbol{\xi}_i$, where $\boldsymbol{\xi}_i$ is $\mathcal{N}[\mathbf{0}, \Sigma]$ and is motivated via a random parameters argument. However, $\boldsymbol{\xi}_i$ can also be introduced as an additional disturbance term or as a relevant latent variable. Second, individuals may be assumed to come from one of C latent classes; see Section 18.5 for a duration model example and Swait (2003) for a GEV example of latent class or finite mixtures models. If $\boldsymbol{\beta}$ and $\boldsymbol{\theta}_\varepsilon$ vary by class then (15.43) becomes unconditionally

$$p_{ij} = \sum_{c=1}^C \left[\int F_j(\mathbf{V}_i(\mathbf{x}_i, \boldsymbol{\xi}_i, \boldsymbol{\beta}^c), \boldsymbol{\theta}_\varepsilon^c) f(\boldsymbol{\xi}_i | \boldsymbol{\theta}_\xi) d\boldsymbol{\xi}_i \right] \pi_c, \quad (15.44)$$

where π_c denotes probability of membership in the c th class and typically $c = 2$ or $c = 3$. The MSL estimator then maximizes

$$\ln \widehat{L}_N(\boldsymbol{\beta}, \Sigma_{\boldsymbol{\beta}}) = \sum_{i=1}^N \sum_{j=1}^m y_{ij} \ln \left[\frac{1}{S} \sum_{s=1}^S \sum_{c=1}^C F_j(\mathbf{V}_i(\mathbf{x}_i, \boldsymbol{\xi}_i^s, \boldsymbol{\beta}^c), \boldsymbol{\theta}_\varepsilon^c) \pi_c \right],$$

where $\boldsymbol{\xi}_i^s$ denotes the s th draw from $f(\boldsymbol{\xi}_i^s | \boldsymbol{\theta}_\xi)$. Kamakura and Wedel (2004) estimate a finite mixtures MNL model using Bayesian methods.

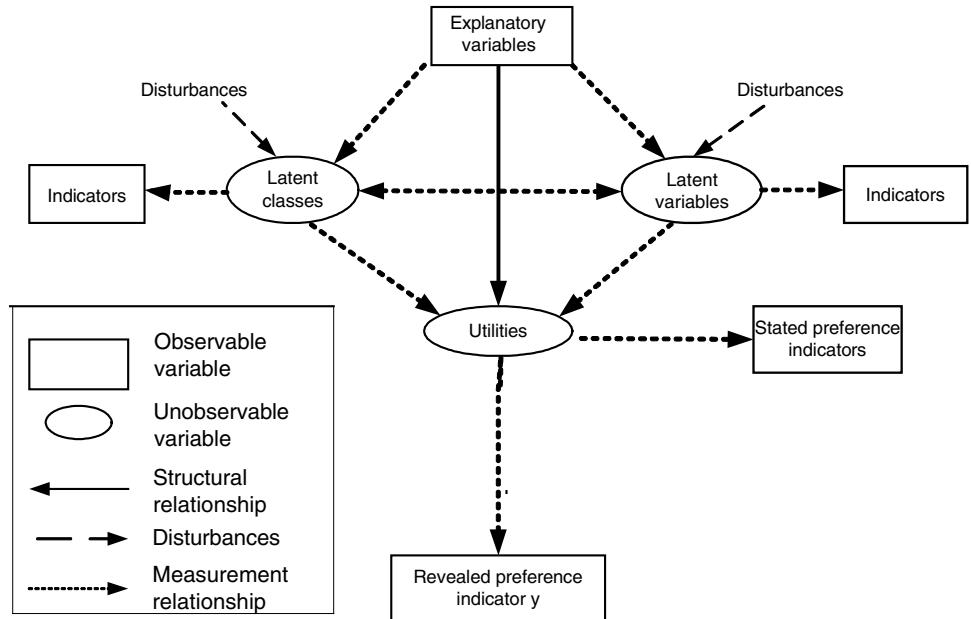


Figure 15.1: Generalized random utility model.

Walker and Ben-Akiva (2002) call such a model a **generalized random utility model**. They cite many articles with such extensions, consider the use of **stated preference data** to supplement revealed preference data, and provide a substantial empirical illustration. Figure 15.1, derived from Walker and Ben-Akiva (2002), summarizes the various extensions.

The multinomial modeling literature has been at the forefront of developing and estimating highly structured parametric models that incorporate random parameters, latent variables, and latent parameters and combine data from more than one source. These methods are applicable to any type of cross-section data, not just discrete outcomes.

15.8. Multinomial Probit

An alternative and obvious way to introduce correlation across choices in the unobserved component is to work with normally distributed errors. However, ML estimation is difficult as in the most general case an $(m - 1)$ -fold integral needs to be calculated.

15.8.1. Multinomial Probit Model

The **multinomial probit** (MNP) model is an m -choice multinomial model, with utility of the j th choice given by

$$U_j = V_j + \varepsilon_j, \quad j = 1, 2, \dots, m, \quad (15.45)$$

where the errors are joint normally distributed, with

$$\boldsymbol{\varepsilon} \sim \mathcal{N}[\mathbf{0}, \boldsymbol{\Sigma}], \quad (15.46)$$

where the $m \times 1$ vector $\boldsymbol{\varepsilon} = [\varepsilon_1 \dots \varepsilon_m]'$. Usually $V_j = \mathbf{x}'_j \boldsymbol{\beta}$ or $V_j = \mathbf{x}' \boldsymbol{\beta}_j$.

Different MNP models arise from different specifications of the covariance matrix $\boldsymbol{\Sigma}$. Some of the off-diagonal entries are specified to be nonzero, to permit correlation across the errors, though some restrictions need to be placed on $\boldsymbol{\Sigma}$. Note that if the errors are uncorrelated the MNP still yields no closed-form solution for the probabilities and it is easier to assume instead that the errors are extreme value and use the CL or MNL models.

Restrictions on $\boldsymbol{\Sigma}$ are needed to ensure **identification**. It is clear from (15.23) that, for any ARUM, choice is determined by the differences in utility or errors. Thus we consider the difference $U_j - U_1$ between utility of alternative j and that of alternative 1, chosen to be the benchmark alternative. Bunch (1991) demonstrated that all but one of the parameters of the covariance matrix of the errors $\varepsilon_j - \varepsilon_1$ is identified; see the discussion at the end of Section 15.5.1. One way to achieve this identification is to normalize $\varepsilon_1 = 0$, say, and then restrict one covariance element. For example, if $m = 2$, set $\varepsilon_1 = 0$ so $\sigma_{11} = 0$ and $\sigma_{12} = 0$ and additionally restrict $\sigma_{22} = 1$. Then $\varepsilon_2 - \varepsilon_1 = \varepsilon_2 \sim \mathcal{N}[0, 1]$, which is the binary probit model.

Additional restrictions on $\boldsymbol{\Sigma}$ or $\boldsymbol{\beta}$ may be needed for successful application. Keane (1992) demonstrated that even if assumptions on the error covariance are made to ensure just-identification, in practice the parameters of the MNP model may be highly imprecisely estimated in models with regressors that do not vary with the alternative. Further restrictions on the MNP model are then needed. This estimation imprecision is qualitatively similar to high multicollinearity among regressors in a linear regression. Keane found that exclusion restrictions on the regressors (with one exclusion for each utility index) work well. Alternatively, and more commonly, further restrictions may be placed on the covariance parameters.

A popular parsimonious model for the errors is the **factor model**

$$\varepsilon_j = v_j + \sum_{l=1}^L c_{jl} \xi_l, \quad j = 1, 2, \dots, m,$$

where v_j and ξ_1, \dots, ξ_L are iid standard normal and c_{jl} are weights called **factor loadings** to be estimated. This model can greatly reduce the number of covariance parameters, from $m(m+1)/2$ to L , and requires an $(L+1)$ -dimensional integral. Numerical methods, usually Gaussian quadrature, can be used for low values of L , whereas simulation methods need to be used for larger L . For panel data the random effects model (see Section 21.2.1) can be viewed as a factor model with error $u_{it} = \alpha_i + \epsilon_{it}$, and the factor model may be especially appropriate in a panel probit setting.

15.8.2. Estimation of Multinomial Probit

The regression and error variance parameters are preferably estimated by ML with log-likelihood given in Section 15.3.2. The challenge is that there is no closed-form expression for the choice probabilities.

For a three-choice MNP model

$$p_1 = \Pr[y = 1] = \int_{-\infty}^{-\tilde{V}_{31}} \int_{-\infty}^{-\tilde{V}_{21}} f(\tilde{\varepsilon}_{21}, \tilde{\varepsilon}_{31}) d\tilde{\varepsilon}_{21} d\tilde{\varepsilon}_{31}$$

(see (15.24)), where $f(\tilde{\varepsilon}_{21}, \tilde{\varepsilon}_{31})$ is a bivariate normal with as many as two free covariance parameters and \tilde{V}_{21} and \tilde{V}_{31} depend on regressors and parameters β . This bivariate normal integral can be quickly evaluated numerically. More generally, however, an m -choice model requires numerical evaluation of an $(m - 1)$ -variate integral. A trivariate normal integral is the limit for numerical methods, limiting standard numerical integration methods to a four-choice MNP model.

For larger models an alternative is to use simulation methods. For simplicity we refer to the three-choice MNP model. One possibility is to use the frequency simulator that approximates p_1 by the fraction of draws of $(\tilde{\varepsilon}_{21}, \tilde{\varepsilon}_{31})$ that are less than $(-\tilde{V}_{21}, -\tilde{V}_{31})$. From Section 12.7.1 this simulator is not smooth and it can be very inefficient (see Section 12.7.2). Furthermore, in the current setting it is possible that it yields boundary values of $\hat{p}_1 = 0$ or 1 . In general it is better to use importance sampling, detailed in Section 12.7.2. For Monte Carlo integration over a region of the multivariate normal a very popular importance sampler is the GHK simulator, due to Geweke (1992), Hajivassiliou and McFadden (1994), and Keane (1994). This recursively truncates the multivariate normal pdf. Compared to the frequency simulator it is smooth, requires many fewer draws for alternatives with low probability of being chosen, and is unlikely to have boundary problems. Train (2003) provides a detailed account of this method.

The preceding discussion considers evaluation of MNP probabilities assuming knowledge of β and Σ . In fact we need to estimate β and Σ . The **maximum simulated likelihood estimator** estimator maximizes

$$\ln \hat{L}_N(\beta, \Sigma) = \sum_{i=1}^N \sum_{j=1}^m y_{ij} \ln \hat{p}_{ij},$$

where the \hat{p}_{ij} are obtained using the GHK or other simulator. Consistency requires the number of draws in the simulator $S \rightarrow \infty$ as well as $N \rightarrow \infty$. The method is very burdensome. At the r th round of an iterative procedure (see Chapter 10) the estimates are $\hat{\beta}^{(r)}$ and $\hat{\Sigma}^{(r)}$ and the update requires recalculating \hat{p}_{ij} , which requires S draws for each of N individuals.

An alternative estimation procedure is the **method of simulated moments** (see Section 12.5). From (15.8) a consistent method of moments estimator solves $\sum_{i=1}^N \sum_{j=1}^m (y_{ij} - p_{ij}) \mathbf{z}_i = \mathbf{0}$, where, for example, $\mathbf{z}_i = \mathbf{x}_i$. The corresponding MSM estimator of β and Σ solves the estimating equations

$$\sum_{i=1}^N \sum_{j=1}^m (y_{ij} - \hat{p}_{ij}) \mathbf{z}_i = \mathbf{0},$$

where the \hat{p}_{ij} are obtained using an unbiased simulator. Then $(y_{ij} - \hat{p}_{ij}) \mathbf{z}_i$ is unbiased for $(y_{ij} - p_{ij}) \mathbf{z}_i$, so consistent estimation is possible even if $S = 1$. This can greatly reduce computation. However, there is an efficiency loss for low S , and even for large

S MSM is less efficient than MSL since in this example the method of moments is less efficient than ML. A less-used related method that is as efficient as MSL is the **method of simulated scores** (see Hajivassiliou and McFadden, 1998).

An alternative estimator uses Bayesian methods. Unlike RPL there is no closed-form solution for the probabilities, which need to be derived from the utilities. The latent utilities $\mathbf{U}_i = (U_{i1}, \dots, U_{ij})$ are introduced as auxiliary variables and the data augmentation approach (see Section 13.7) is used. Letting $\mathbf{U} = (\mathbf{U}_1, \dots, \mathbf{U}_N)$ and $\mathbf{y} = (y_1, \dots, y_N)$ we have that the Gibbs sampler cycles among (1) the conditional posterior for $\beta|\mathbf{y}, \mathbf{U}, \Sigma$, (2) the conditional posterior for $\Sigma|\mathbf{y}, \beta, \mathbf{U}$, and (3) the posterior for $\mathbf{U}_i|\mathbf{y}, \beta, \Sigma$. Albert and Chib (1993) provide a quite general treatment for both unordered and ordered multinomial models. McCulloch and Rossi (1994) provide a substantive MNP application. Chib (2001) discusses the complication of imposing the restrictions on Σ needed for identification (see Section 15.8.1).

15.8.3. Discussion

Both MNP and RPL models lack a closed-form solution for p_{ij} . However, for RPL there is at least a closed-form solution conditional on β_i and the only problem is to integrate out β_i . For the MNP model, which predates the RPL model, there is no such conditional result and approximating p_{ij} becomes more challenging, especially if p_{ij} is close to zero or one. It appears to be easier to get model flexibility through nested logit, RPL or mixture models than by use of MNP.

15.9. Ordered, Sequential, and Ranked Outcomes

In this section we present models with more structure than unordered models, such as those with a natural ordering of alternatives or sequencing of decisions. Analysis is straightforward as appropriate models are well established and estimation is again by MLE based on (15.4), with different models leading to different specifications of the probabilities p_{ij} .

15.9.1. Ordered Multinomial Models

Suppose that there is a natural ordering of alternatives. For example, self-rated health status may be one of excellent, good, fair, or poor. Such data can be estimated by an unordered multinomial model, but a much more parsimonious model and sensible model is one that takes account of this ordering.

The starting point is an index model, with single latent variable

$$y_i^* = \mathbf{x}'\beta + u_i, \quad (15.47)$$

where \mathbf{x} here does not include an intercept, a departure from Section 14.4.1. As y^* crosses a series of increasing unknown thresholds we move up the ordering of alternatives. For example, for very low y^* health status is poor, for $y^* > \alpha_1$ health status improves to fair, for $y^* > \alpha_2$ it improves further to good, and so on.

In general for an m -alternative **ordered model** we define

$$y_i = j \quad \text{if } \alpha_{j-1} < y_i^* \leq \alpha_j, \quad (15.48)$$

where $\alpha_0 = -\infty$ and $\alpha_m = \infty$. Then

$$\begin{aligned} \Pr[y_i = j] &= \Pr[\alpha_{j-1} < y_i^* \leq \alpha_j] \\ &= \Pr[\alpha_{j-1} < \mathbf{x}'_i \boldsymbol{\beta} + u_i \leq \alpha_j] \\ &= \Pr[\alpha_{j-1} - \mathbf{x}'_i \boldsymbol{\beta} < u_i \leq \alpha_j - \mathbf{x}'_i \boldsymbol{\beta}] \\ &= F(\alpha_j - \mathbf{x}'_i \boldsymbol{\beta}) - F(\alpha_{j-1} - \mathbf{x}'_i \boldsymbol{\beta}), \end{aligned} \quad (15.49)$$

where F is the cdf of u_i . The regression parameters $\boldsymbol{\beta}$ and the $(m - 1)$ threshold parameters $\alpha_1, \dots, \alpha_{m-1}$ are obtained by maximizing the log-likelihood (15.5) with p_{ij} defined in (15.49). For the **ordered logit model** u is logistic distributed with $F(z) = e^z/(1 + e^z)$. For the **ordered probit model** u is standard normal distributed and $F(\cdot)$ is the standard normal cdf. Letting K denote the number of regressors excluding the intercept, an m -choice ordered model has $K + m - 1$ parameters whereas an MNL model has $(m - 1)(K + 1)$ parameters.

The sign of the regression parameters $\boldsymbol{\beta}$ can be immediately interpreted as determining whether or not the latent variable y^* increases with the regressor. For marginal effects in the probabilities

$$\frac{\partial \Pr[y_i = j]}{\partial \mathbf{x}'_i} = \{F'(\alpha_{j-1} - \mathbf{x}'_i \boldsymbol{\beta}) - F'(\alpha_j - \mathbf{x}'_i \boldsymbol{\beta})\}\boldsymbol{\beta},$$

where F' denotes the derivative of F . The term in braces can be positive or negative.

This model can also be applied to count data that take just a few values. Cameron and Trivedi (1986) applied the ordered probit model to number of doctor consultations. Hausman, Lo, and MacKinley (1992) applied the ordered probit to data on changes in a count, which can be negative, and additionally modeled the error term u_i to be heteroskedastic.

15.9.2. Sequential Multinomial Models

In some situations decisions are made sequentially. For example, one might first decide whether or not to go to college. If no college is chosen then $y = 1$. If $y \neq 1$ then decide whether to go to a two-year college ($y = 2$) or four-year college ($y = 3$). Given specification of this sequence the probabilities are easily obtained. For example, model the first decision by a probit model and the second decision, if relevant, by a probit model. Then $\Pr[y = 1] = \Phi(\mathbf{x}'_1 \boldsymbol{\beta}_1)$ and $\Pr[y = 2|y \neq 1] = \Phi(\mathbf{x}'_2 \boldsymbol{\beta}_2)(1 - \Phi(\mathbf{x}'_1 \boldsymbol{\beta}_1))$. The unconditional probability is

$$\Pr[y = 2] = \Pr[y = 2|y \neq 1] \times \Pr[y \neq 1] = \Phi(\mathbf{x}'_2 \boldsymbol{\beta}_2)(1 - \Phi(\mathbf{x}'_1 \boldsymbol{\beta}_1)).$$

The parameters $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ can be estimated by maximizing the log-likelihood function (15.5), where $p_{1i} = \Phi(\mathbf{x}'_1 \boldsymbol{\beta}_1)$, p_{2i} is given in the preceding equation, and $p_{3i} = 1 - p_{1i} - p_{2i}$.

This approach relies on correct specification of the sequence of decision making. A better model for this choice example may be a three-choice nested logit model where the errors in the utilities for two-year college and four-year college are correlated with each other and independent of the error in the utility for no college. These models can be compared using the likelihood-based methods given in Section 8.5.

15.9.3. Ranked Data Models

The models discussed thus far have assumed that alternatives are mutually exclusive and only one alternative is chosen. More generally, alternatives may be ranked, especially with stated preference data. For example, both first and second choices may be known.

The **rank-ordered logit model** is simple to estimate (see Beggs, Cardell, and Hausman, 1981). Consider a four-alternative conditional logit model with alternative 2 the first choice and alternative 3 the second choice. Alternative 2 is chosen from all four alternatives and then alternative 3 is chosen from the remaining alternatives 1, 3, and 4. The joint probability of these first and second choices is

$$\frac{e^{\mathbf{x}'_{i2}\beta}}{e^{\mathbf{x}'_{i1}\beta} + e^{\mathbf{x}'_{i2}\beta} + e^{\mathbf{x}'_{i3}\beta} + e^{\mathbf{x}'_{i4}\beta}} \times \frac{e^{\mathbf{x}'_{i3}\beta}}{e^{\mathbf{x}'_{i1}\beta} + e^{\mathbf{x}'_{i3}\beta} + e^{\mathbf{x}'_{i4}\beta}}.$$

Estimation is by ML given similar expressions for the other 11 joint probabilities.

For the multinomial probit model there is no similar simplification. Hajivassiliou and Ruud (1994) present a method to simulate the joint probabilities; they use the **rank-ordered probit model** to illustrate a variety of simulation-based estimators.

15.10. Multivariate Discrete Outcomes

The preceding models, aside from rank-ordered models, are models for one discrete dependent variable that takes one of m mutually exclusive values. Now we consider models when there is more than one discrete outcome. The log-likelihood function is similar to (15.5) for the multinomial model, with different models corresponding to different functional forms for the probabilities. These probabilities may need to account for correlation of the two outcomes and possibly simultaneity.

15.10.1. Bivariate Discrete Outcomes

For simplicity consider **bivariate discrete data** (y_{1i}, y_{2i}) . For example, in a joint model of labor supply and fertility the dependent variables (y_{1i}, y_{2i}) for individual i may be $y_{1i} = 2$ if work and $y_{1i} = 1$ do not work, and $y_{2i} = 2$ if have children and $y_{2i} = 1$ if have no children.

More generally, y_1 may take values $1, \dots, m_1$ and y_2 may take values $1, \dots, m_2$. For individual i define

$$p_{ijk} = \Pr[y_{1i} = j, y_{2i} = k], \quad j = 1, \dots, m_1, \quad k = 1, \dots, m_2. \quad (15.50)$$

Note that p_{ijk} define probabilities of mutually exclusive events and $\sum_j \sum_k p_{ijk} = 1$. Define $m_1 \times m_2$ corresponding binary indicator variables $y_{jk} = 1$ if $(y_1 = j, y_2 = k)$ and $y_{jk} = 0$ otherwise. Then the joint density for the i th observation is

$$f(y_{1i}, y_{2i}) = \prod_{k=1}^{m_1} \prod_{j=1}^{m_2} p_{ijk}^{y_{ijk}}.$$

The log-likelihood is then $\sum_{i=1}^N \sum_{k=1}^{m_1} \sum_{j=1}^{m_2} y_{ijk} \ln p_{ijk}$ and estimation is by ML as in Section 15.4.2.

The essential difference between the multivariate and multinomial models is in the specification of the functional form for the probabilities.

In the simplest case the two discrete dependent variables are independent and $p_{ijk} = \Pr[y_{1i} = j] \times \Pr[y_{2i} = k]$. Then y_1 and y_2 can be modeled using separate multinomial models.

If the two variables are instead viewed as interrelated, a simple approach is to use a multinomial logit model for the probabilities p_{ijk} . Then the bivariate outcomes (y_1, y_2) are essentially treated as $m_1 \times m_2$ univariate outcomes. For example, in the labor supply and fertility example one of the four outcomes is then work and have children.

In the next section we consider models between these two extremes.

15.10.2. Bivariate Probit

The bivariate probit model is a joint model for two binary outcomes that generalizes the index function model (see Section 14.4.1) from one latent variable to two latent variables that may be correlated.

Define the unobserved latent variables

$$\begin{aligned} y_1^* &= \mathbf{x}'_1 \boldsymbol{\beta}_1 + \varepsilon_1, \\ y_2^* &= \mathbf{x}'_2 \boldsymbol{\beta}_2 + \varepsilon_2, \end{aligned} \tag{15.51}$$

where the ε_1 and ε_2 are joint normal with means zero, variances one, and correlation ρ . Then the **bivariate probit model** specifies the observed outcomes to be

$$\begin{aligned} y_1 &= \begin{cases} 2 & \text{if } y_1^* > 0, \\ 1 & \text{if } y_1^* \leq 0, \end{cases} \\ y_2 &= \begin{cases} 2 & \text{if } y_2^* > 0, \\ 1 & \text{if } y_2^* \leq 0, \end{cases} \end{aligned}$$

where we use values (2, 1) rather than (1, 0) to be consistent with the notation of this chapter. This model collapses to two separate probit models for y_1 and y_2 when the error correlation $\rho = 0$.

When $\rho \neq 0$ there is no closed-form solution for the probabilities. For example,

$$\begin{aligned} p_{22} &= \Pr [y_1 = 2, y_2 = 2] \\ &= \Pr [y_1^* > 0, y_2^* > 0] \\ &= \Pr [-\varepsilon_1 < \mathbf{x}'_1 \boldsymbol{\beta}_1, -\varepsilon_2 < \mathbf{x}'_2 \boldsymbol{\beta}_2] \\ &= \Pr [\varepsilon_1 < \mathbf{x}'_1 \boldsymbol{\beta}_1, \varepsilon_2 < \mathbf{x}'_2 \boldsymbol{\beta}_2] \\ &= \int_{-\infty}^{\mathbf{x}'_1 \boldsymbol{\beta}_1} \int_{-\infty}^{\mathbf{x}'_2 \boldsymbol{\beta}_2} \phi(z_1, z_2, \rho) dz_1 dz_2 \\ &= \Phi(\mathbf{x}'_1 \boldsymbol{\beta}_1, \mathbf{x}'_2 \boldsymbol{\beta}_2, \rho), \end{aligned}$$

where $\phi(z_1, z_2, \rho)$ and $\Phi(z_1, z_2, \rho)$ are, respectively, the standardized bivariate normal density and cdf for (z_1, z_2) with zero means, unit variances, and correlation ρ , and the fourth equality holds for the bivariate normal with mean zero.

Performing similar algebra for the other possible outcomes yields

$$\begin{aligned} p_{jk} &= \Pr [y_1 = j, y_2 = k] \\ &= \Phi(q_1 \mathbf{x}'_1 \boldsymbol{\beta}_1, q_2 \mathbf{x}'_2 \boldsymbol{\beta}_2, \rho), \end{aligned}$$

where $q_l = 1$ if $y_l = 2$ and $q_l = -1$ if $y_l = 1$ for $l = 1, 2$. This is the basis for ML estimation, detailed in Greene (2003), who also considers computation of marginal effects.

Implementation requires evaluation of a bivariate normal integral, which is numerically feasible. Generalizations to multivariate probit are obvious though will experience numerical challenges because of higher order integrals. If each outcome is ordered then the model can be generalized to a **bivariate ordered probit model**.

One can also consider a simultaneous equations probit model that generalizes (15.51) to allow the right-hand side variables to be endogenous. For example, the first equation for y_1^* may include y_2^* and/or y_2 as regressors and similarly for y_2^* , with some restrictions required to ensure the model is identified. This model is similar to the simultaneous equations Tobit model discussed in Section 16.8.2.

15.11. Semiparametric Estimation

Some studies have extended semiparametric estimation methods to models for unordered multinomial data. Abe (1999) estimated the conditional logit model with $\mathbf{x}'_{ij} \boldsymbol{\beta}$ in (15.10) replaced by the additive model form $\sum_p \beta_p f_p(\mathbf{x}_{ijp})$, where p denotes the p th component of \mathbf{x}_{ij} and the function $f_p(\cdot)$ is estimated by the data. L-F. Lee (1995) extended the Klein and Spady (1993) estimator (see Section 14.7) from binary outcomes to multinomial outcomes. Semiparametric methods for multiple-index models can also be applied to the multinomial unordered model. The challenge is to ensure that predicted probabilities lie between zero and one and sum to one.

Ordered models lend themselves well to semiparametric analysis since they involve an index $\mathbf{x}' \boldsymbol{\beta}$ that crosses a number of thresholds. See, for example, Klein and Sherman (2002), who present an estimator that is \sqrt{N} -consistent and asymptotically normal for

both regression and threshold points up to location and scale, under the assumption that errors are independent of regressors.

15.12. Derivations for MNL, CL, and NL Models

We consider the conditional and multinomial logit models, deriving first and second derivatives of the log-likelihood function and expressions for the effect of changes in regressors on the probabilities. Then the nested logit (NL) model is derived from the GEV model.

15.12.1. Conditional Logit

The conditional logit probability is $p_{ij} = e^{\mathbf{x}'_{ij}\beta} / \sum_l e^{\mathbf{x}'_{il}\beta}$. Differentiation by parts yields

$$\begin{aligned}\frac{\partial p_{ij}}{\partial \beta} &= \frac{e^{\mathbf{x}'_{ij}\beta}}{\sum_l e^{\mathbf{x}'_{il}\beta}} \mathbf{x}_{ij} - \frac{e^{\mathbf{x}'_{ij}\beta}}{(\sum_l e^{\mathbf{x}'_{il}\beta})^2} \sum_l e^{\mathbf{x}'_{il}\beta} \mathbf{x}_{il} \\ &= p_{ij} \mathbf{x}_{ij} - p_{ij} \sum_l p_{il} \mathbf{x}_{il} = p_{ij} \mathbf{x}_{ij} - p_{ij} \bar{\mathbf{x}}_i = p_{ij} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i),\end{aligned}$$

where $\bar{\mathbf{x}}_i = \sum_l p_{il} \mathbf{x}_{il}$. Then

$$\frac{\partial \mathcal{L}}{\partial \beta} = \sum_i \sum_j \frac{y_{ij}}{p_{ij}} \frac{\partial p_{ij}}{\partial \beta} = \sum_i \sum_j \frac{y_{ij}}{p_{ij}} p_{ij} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i) = \sum_i \sum_j y_{ij} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i).$$

It follows that

$$\begin{aligned}\frac{\partial^2 \mathcal{L}}{\partial \beta \partial \beta'} &= - \sum_i \sum_j y_{ij} \frac{\partial \bar{\mathbf{x}}_i}{\partial \beta'} \\ &= - \sum_i \sum_j y_{ij} \frac{\partial \sum_l p_{il} \mathbf{x}_{il}}{\partial \beta'} \\ &= - \sum_i \sum_j y_{ij} \sum_l p_{il} (\mathbf{x}_{il} - \bar{\mathbf{x}}_i) \mathbf{x}'_{il} \\ &= \sum_i \sum_j p_{ij} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i) \mathbf{x}'_{ij} \\ &= \sum_i \sum_j p_{ij} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i) (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)',\end{aligned}$$

which is (15.15). The second to last equality uses the fact that y_{ij} equals one for exactly one of the choices and zero otherwise, so that $\sum_j y_{ij} \sum_l a_{il} = \sum_j \sum_l y_{ij} a_{il} = \sum_j a_{ij}$, and the last equality uses $\sum_j p_{ij} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i) \bar{\mathbf{x}}'_i = \sum_j (p_{ij} \mathbf{x}_{ij} - p_{ij} \bar{\mathbf{x}}_i) \bar{\mathbf{x}}'_i = \sum_j (\bar{\mathbf{x}}_i - p_{ij} \bar{\mathbf{x}}_i) \bar{\mathbf{x}}'_i = \mathbf{0}$ as $\sum_j p_{ij} = 1$.

Now consider the effect of changing regressors. For the conditional logit model

$$\frac{\partial p_{ij}}{\partial \mathbf{x}_{ij}} = \frac{e^{\mathbf{x}'_{ij}\beta}}{\sum_l e^{\mathbf{x}'_{il}\beta}} \beta - \frac{e^{\mathbf{x}'_{ij}\beta}}{(\sum_l e^{\mathbf{x}'_{il}\beta})^2} e^{\mathbf{x}'_{ij}\beta} \beta = p_{ij} (1 - p_{ij}) \beta,$$

whereas for $j \neq k$

$$\frac{\partial p_{ij}}{\partial \mathbf{x}_{ik}} = -\frac{e^{\mathbf{x}'_i \beta_j}}{\left(\sum_l e^{\mathbf{x}'_i \beta_l}\right)^2} e^{\mathbf{x}'_i \beta_j} \beta_j = -p_{ij} p_{ik} \beta_j.$$

Combining these two results yields (15.18).

15.12.2. Multinomial Logit

The multinomial logit probability is $p_{ij} = e^{\mathbf{x}'_i \beta_j} / \sum_l e^{\mathbf{x}'_i \beta_l}$. Differentiation by parts yields

$$\frac{\partial p_{ij}}{\partial \beta_j} = \frac{e^{\mathbf{x}'_i \beta_j}}{\sum_l e^{\mathbf{x}'_i \beta_l}} \mathbf{x}_i - \frac{e^{\mathbf{x}'_i \beta_j}}{\left(\sum_l e^{\mathbf{x}'_i \beta_l}\right)^2} e^{\mathbf{x}'_i \beta_j} \mathbf{x}_i = p_{ij} \mathbf{x}_i - p_{ij} p_{ij} \mathbf{x}_i,$$

whereas for $k \neq j$

$$\frac{\partial p_{ij}}{\partial \beta_k} = -\frac{e^{\mathbf{x}'_i \beta_j}}{\left(\sum_l e^{\mathbf{x}'_i \beta_l}\right)^2} e^{\mathbf{x}'_i \beta_k} \mathbf{x}_i = -p_{ij} p_{ik} \mathbf{x}_i.$$

Combining we have

$$\frac{\partial p_{ij}}{\partial \beta_k} = \delta_{ijk} p_{ij} \mathbf{x}_i - p_{ij} p_{ik} \mathbf{x}_i = p_{ij} (\delta_{ijk} - p_{ik}) \mathbf{x}_i,$$

where the indicator variable $\delta_{ijk} = 1$ if $j = k$, and

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \beta_k} &= \sum_i \sum_j \frac{y_{ij}}{p_{ij}} \frac{\partial p_{ij}}{\partial \beta_k} \\ &= \sum_i \sum_j \frac{y_{ij}}{p_{ij}} (\delta_{ijk} p_{ij} - p_{ij} p_{ik} \mathbf{x}_i) \\ &= \sum_i \left[\sum_j y_{ij} \delta_{ijk} - y_{ij} p_{ik} \right] \mathbf{x}_i \\ &= \sum_i [y_{ik} - p_{ik}] \mathbf{x}_i, \end{aligned}$$

as stated in (15.16), where the last line uses the definition of δ_{ijk} and $\sum_j y_{ij} = 1$. For the second derivative we have

$$\frac{\partial^2 \mathcal{L}}{\partial \beta_j \partial \beta_k} = -\sum_i \sum_j \frac{\partial p_{ij}}{\partial \beta_k} \mathbf{x}_i = -\sum_i \sum_j p_{ij} (\delta_{ijk} - p_{ik}) \mathbf{x}_i \mathbf{x}'_i,$$

which yields (15.17).

When regressors change

$$\begin{aligned} \frac{\partial p_{ij}}{\partial \mathbf{x}_i} &= \frac{e^{\mathbf{x}'_i \beta_j}}{\sum_l e^{\mathbf{x}'_i \beta_l}} \beta_j - \frac{e^{\mathbf{x}'_i \beta_j}}{\left(\sum_l e^{\mathbf{x}'_i \beta_l}\right)^2} \sum_l e^{\mathbf{x}'_i \beta_l} \beta_l \\ &= p_{ij} \beta_j - p_{ij} \sum_l p_{il} \beta_l = p_{ij} (\beta_j - \bar{\beta}_i), \end{aligned}$$

where $\bar{\beta}_i = \sum_l p_{il} \beta_l$, as stated in (15.19).

15.12.3. Nested Logit

We consider the two-level GEV model given by (15.32) and (15.33) with

$$G(\mathbf{Y}) = G(Y_{11}, \dots, Y_{1K_1}, \dots, Y_{J1}, \dots, Y_{JK_J}) = \sum_{j=1}^J a_j \left(\sum_{k=1}^{K_j} Y_{jk}^{1/\rho_j} \right)^{\rho_j},$$

which is a generalization of (15.34) owing to the coefficients a_j . The general GEV result (15.31) becomes $\Pr[y_{jk} = 1] = Y_{jk} G_{jk}/G(\mathbf{Y})$, where G_{jk} is the derivative of $G(\mathbf{Y})$ with respect to Y_{jk} and evaluation is at $Y_{jk} = e^{V_{jk}}$.

Now

$$G_{jk} = \frac{\partial G(\mathbf{Y})}{\partial Y_{jk}} = a_j \left(\sum_{l=1}^{K_j} Y_{jl}^{1/\rho_j} \right)^{\rho_j-1} \times Y_{jk}^{(1/\rho_j)-1},$$

which gives

$$Y_{jk} G_{jk} = a_j \left(\sum_{l=1}^{K_j} Y_{jl}^{1/\rho_j} \right)^{\rho_j} Y_{jk}^{1/\rho_j}.$$

Then

$$p_{jk} \equiv \frac{Y_{jk} G_{jk}}{G(\mathbf{Y})} = \frac{a_j \left(\sum_{l=1}^{K_j} Y_{jl}^{1/\rho_j} \right)^{\rho_j-1} Y_{jk}^{1/\rho_j}}{\sum_{m=1}^J a_m \left(\sum_{l=1}^{K_m} Y_{ml}^{1/\rho_m} \right)^{\rho_m}}.$$

The probability of choosing limb j is

$$p_j \equiv \sum_{k=1}^{K_j} p_{jk} = \frac{a_j \left(\sum_{l=1}^{K_j} Y_{jl}^{1/\rho_j} \right)^{\rho_j}}{\sum_{m=1}^J a_m \left(\sum_{l=1}^{K_m} Y_{ml}^{1/\rho_m} \right)^{\rho_m}},$$

after some simplification, and the conditional probability of choosing branch k given limb j is

$$p_{k|j} \equiv \frac{p_{jk}}{p_j} = \frac{Y_{jk}^{1/\rho_j}}{\sum_{l=1}^{K_j} Y_{jl}^{1/\rho_j}}.$$

This result is also given in Maddala (1983, p. 72).

We need to evaluate these expression at $Y_{jk} = \exp(V_{jk})$. Suppose

$$V_{jk} = \mathbf{z}'_j \boldsymbol{\alpha} + \mathbf{x}'_{jk} \boldsymbol{\beta}_j.$$

Then performing some algebra yields

$$\begin{aligned} (e^{V_{jk}})^{1/\rho_j} &= \exp(\mathbf{z}'_j \boldsymbol{\alpha}/\rho_j) \exp(\mathbf{x}'_{jk} \boldsymbol{\beta}_j/\rho_j), \\ \sum_{l=1}^{K_j} (e^{V_{jl}})^{1/\rho_j} &= \exp(\mathbf{z}'_j \boldsymbol{\alpha}/\rho_j) \exp(I_j), \\ \left(\sum_{l=1}^{K_j} (e^{V_{jl}})^{1/\rho_j} \right)^{\rho_j} &= \exp(\mathbf{z}'_j \boldsymbol{\alpha} + \rho_j I_j), \end{aligned}$$

where

$$I_j = \ln \left(\sum_{l=1}^{K_l} \exp (\mathbf{x}'_{jk} \boldsymbol{\beta}_j / \rho_j) \right).$$

It follows that the probability of choosing limb j becomes

$$\begin{aligned} p_j &= \frac{a_j \left(\sum_{l=1}^{K_l} (e^{V_{jl}})^{1/\rho_j} \right)^{\rho_j}}{\sum_{m=1}^J a_m \left(\sum_{l=1}^{K_l} (e^{V_{ml}})^{1/\rho_m} \right)^{\rho_m}} \\ &= \frac{a_j \exp(\mathbf{z}'_j \boldsymbol{\alpha} + \rho_j I_j)}{\sum_{m=1}^J a_m (\exp(\mathbf{z}'_m \boldsymbol{\alpha} + \rho_m I_m))}, \end{aligned}$$

as stated for the first term in (15.36). Note that the scalar a_j can be absorbed into \mathbf{z}_j as a limb-specific dummy, as $a_j \exp(\mathbf{z}'_j \boldsymbol{\alpha} + \rho_j I_j) = \exp(\ln a_j + \mathbf{z}'_j \boldsymbol{\alpha} + \rho_j I_j)$. Without loss of generality we therefore set $a_j = 1$.

The probability of branch k within limb j is

$$\begin{aligned} p_{k|j} &= \frac{(e^{V_{jk}})^{1/\rho_j}}{\sum_{l=1}^{K_l} (e^{V_{jl}})^{1/\rho_j}} \\ &= \frac{\exp(\mathbf{z}'_j \boldsymbol{\alpha} / \rho_j) \exp(\mathbf{x}'_{jk} \boldsymbol{\beta} / \rho_j)}{\sum_{l=1}^{K_l} \exp(\mathbf{z}'_j \boldsymbol{\alpha} / \rho_j) \exp(\mathbf{x}'_{jl} \boldsymbol{\beta} / \rho_j)} \\ &= \frac{\exp(\mathbf{x}'_{jk} \boldsymbol{\beta}_j / \rho_j)}{\sum_{l=1}^{K_l} \exp(\mathbf{x}'_{jl} \boldsymbol{\beta} / \rho_j)}, \end{aligned}$$

as stated for the second term in (15.36).

15.13. Practical Considerations

The multinomial logit model is adequate for describing data or estimating the marginal probabilities but is viewed as a poor model if a more structural interpretation of the parameters is required, owing to the independence of irrelevant alternatives assumption. Many packages estimate the multinomial logit model.

The nested logit model can be estimated in STATA and by using the NLOGIT add-on to LIMDEP, and it is easy to code in a language such as GAUSS. It is the obvious model to use if there is an obvious nesting structure, but usually there is no obvious structure.

The random parameters logit model requires special code in a language such as GAUSS and requires use of the simulation-based estimation methods given in Chapter 12. Ken Train provides code at his Web site elsa.berkeley.edu/~train.

The multinomial probit model is even more challenging to estimate, for more than four choices, and has met with relatively little empirical success. For these reasons the random parameters logit model is currently preferred.

15.14. Bibliographic Notes

- 15.3** Good basic references for multinomial models include Amemiya (1981, 1985), Maddala (1983), and Greene (2003). The books by Ben-Akiva and Lerman (1985), Train (1986), and Borsch-Supan (1987) provide extensive applications as well as a review of theory. Train (2003) presents an outstanding treatment of unordered multinomial models and on estimation using simulation methods.
- 15.5** The seminal article by McFadden (1981) provides an advanced treatment of discrete choice modeling, emphasizing the random utility model approach. For welfare analysis see Small and Rosen (1981), Train (2003, pp. 59–61) and Dagsvik and Karström (2004).
- 15.6** Borsch-Supan (1987) gives an excellent exposition and application of the nested logit model.
- 15.7** The random parameters logit model and other recent advances are well covered in Train (2003). Revelt and Train (1998) provide an early application.
- 15.8** Bolduc (1999) presents MSL estimation of a nine-choice multinomial probit model.

Exercises

- 15–1** Consider a latent variable modeled by $y^* = \mathbf{x}'\beta + \varepsilon$, with $\varepsilon \sim \mathcal{N}[0, 1]$. Suppose we observe only $y = 2$ if $y^* < \alpha$, $y = 1$ if $\alpha \leq y^* < U$, and $y = 0$ if $y^* \geq U$, where the upper limit U is a known constant for each individual (i.e., data) and may differ over individuals, but α is unknown.
- Obtain the conditional probabilities that $y = 0$, $y = 1$, and $y = 2$.
 - Provide details on a method to consistently estimate β and α .
- 15–2** Use a 50% subsample of the fishing mode choice data of Section 15.2.
- Estimate the conditional logit model of Section 15.2.1.
 - Comment on the statistical significance of parameter estimates.
 - What is the effect of an increase in price on the various modes of fishing?
- 15–3** Use a 50% subsample of the fishing mode choice data of Section 15.2.
- Estimate the multinomial logit model of Section 15.2.2.
 - Comment on the statistical significance of parameter estimates.
 - What is the effect of an increase in income on the various modes of fishing?
- 15–4** Use a 50% subsample of the fishing mode choice data of Section 15.2. Suppose we collapse the model to three alternatives and order the alternatives, with $y = 0$ if fishing from a pier or beach, $y = 1$ if fishing from a private boat and $y = 2$ if fishing from a charter boat.
- Estimate an ordered logit model with income as the only regressor.
 - Provide an interpretation of the estimated coefficient.
 - Compare the fit of this model with that from a three-choice multinomial model with income as the regressor.

Tobit and Selection Models

16.1. Introduction

In this chapter we consider two closely related topics: regression when the dependent variable of interest is **incompletely observed** and regression when the dependent variable is completely observed but is observed in a **selected sample** that is not representative of the population. This includes limited dependent variable models, latent variable models, generalized Tobit models, and selection models.

All these models share the common feature that even in the simplest case of population conditional mean linear in regressors, OLS regression leads to inconsistent parameter estimates because the sample is not representative of the population. Alternative estimation procedures, most relying on strong distributional assumptions, are necessary to ensure consistent parameter estimation.

Leading causes of incompletely observed data are truncation and censoring. For **truncated data** some observations on both the dependent variable and regressors are lost. For example, income may be the dependent variable and only low-income people are included in the sample. For **censored data** information on the dependent variable is lost, but not data on the regressors. For example, people of all income levels may be included in the sample, but for confidentiality reasons the income of high-income people may be top-coded and reported only as exceeding, say, \$100,000 per year. Truncation entails greater information loss than does censoring. A leading example of truncation and censoring is the **Tobit model**, named after Tobin (1958), who considered linear regression under normality. Similar issues arise for truncation and censoring in other models introduced in later chapters, most notably for censored duration data presented in Chapter 17. More generally, truncation and censoring are examples of missing data problems that are studied in Chapter 27.

The first-generation estimation methods require strong distributional assumptions. Even seemingly minor departures from assumptions, such as heteroskedastic errors when homoskedastic errors are assumed, can lead to inconsistent parameter estimates. For this reason the models presented in this chapter provide a leading econometrics application of semiparametric regression methods. Semiparametric methods for simple

forms of censoring and truncation such as top-coding have been successfully applied. However, for more general models with selection on unobservables there is to date no widely accepted procedure.

Section 16.2 presents general theory for censored and truncated nonlinear regression models, with specialization to the Tobit model given in Section 16.3. An alternative model for censored data, the two-part model, is introduced in Section 16.4. The sample selection model is presented in Section 16.5. An application to health expenditures in Section 16.6 contrasts the two-part and sample selection models. The Roy model for unobserved counterfactuals is presented in Section 16.7. Section 16.8 considers fully structural models obtained by utility maximization with corner solutions or by extension of simultaneous equation models to selected samples. Semiparametric estimation is presented in Section 16.9.

16.2. Censored and Truncated Models

We present general methods for estimation of fully parametric models when data are censored or truncated. These methods can be applied to models presented in later chapters such as count and duration models. The leading example, the Tobit model for censoring or truncation in linear models, is introduced in Section 16.2.1 and given separate treatment in Section 16.3.

16.2.1. Censoring and Truncation Example

Let y^* denote a variable that is incompletely observed. For truncation from below, y^* is only observed if $y^* > 0$. Since negative values do not appear in the sample, the truncated mean exceeds the mean of y^* . For censoring from below at zero, y^* is not completely observed when $y^* \leq 0$, but it is known that $y^* < 0$ and for simplicity y is then set to 0. Since negative values are scaled up to zero, the censored mean also exceeds the mean of y^* . Clearly, sample means in truncated or censored samples cannot be used without adjustment to estimate the original population mean.

This chapter studies similar issues for regression models. With luck, truncation and censoring might lead only to a shift up or down in the intercept, leaving slope coefficients unchanged; however, this is not the case. For example, if $E[y^*|\mathbf{x}] = \mathbf{x}'\beta$ in the original model then truncation or censoring leads to $E[y|\mathbf{x}]$ being nonlinear in \mathbf{x} and β so that OLS gives inconsistent estimates of β and hence inconsistent estimates of marginal effects.

As an illustration we consider the following labor supply example with simulated data. The relationship between desired annual hours worked, y^* , and hourly wage, w , is specified to be of linear-log form with data-generation process

$$\begin{aligned} y^* &= -2500 + 1000 \ln w + \varepsilon, \\ \varepsilon &\sim \mathcal{N}[0, 1000^2], \\ \ln w &\sim \mathcal{N}[2.75, 0.60^2]. \end{aligned} \tag{16.1}$$

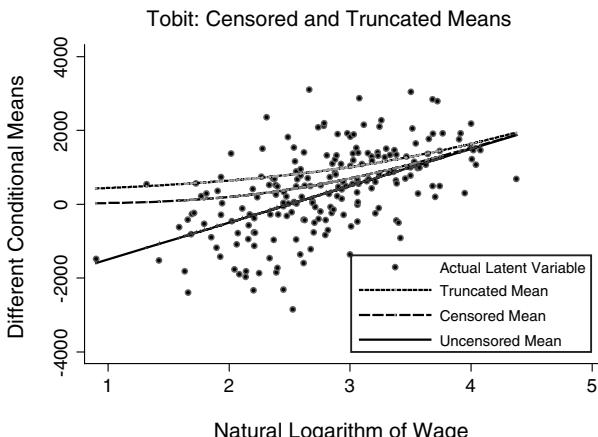


Figure 16.1: Tobit regression of hours on log wage: uncensored conditional mean (bottom), censored conditional mean (middle), and truncated conditional mean (top) for censoring/truncation from below at zero hours. Data are generated from a classical linear regression model.

This is a Tobit model, studied in detail in Section 16.3. The model implies that the wage elasticity is $1000/y^*$, which equals, for example, 0.5 for full-time work (2,000 hours). For each 1% increase in wage, annual hours increase by 10 hours.

Figure 16.1 presents a scatter plot of y^* and $\ln w$ for a generated sample of 200 observations. The unconditional mean for y^* , which is $-2500 + 1000 \ln w$, is given by the lowest curve, which is a straight line.

With censoring at zero, negative values of y^* are set to zero because people with negative desired hours of work choose not to work. For this particular sample this is the case for about 35% of the observations. This pushes up the mean for low wages, since the many negative values of the y^* are shifted up to zero. It has little impact for high wages, since then few observations on y^* are zero. The middle curve in Figure 16.1 gives the resulting censored mean, using the formula given later in (16.23).

With truncation at zero the 35% of the population with negative values of y^* are dropped altogether. This increases the mean above the censored mean, since zero values are no longer included in the data used to form the mean. The upper curve in Figure 16.1 gives the resulting truncated mean, using the formula given later in (16.23).

It is clear that censored and truncated conditional means are nonlinear in x even if the underlying population mean is linear. OLS estimation using truncated or censored data will lead to inconsistent estimation of the slope parameter, since by visual inspection of Figure 16.1 a linear approximation to the nonlinear truncated and censored means will have flatter slope than that for the original untruncated mean. Analysis should instead be based on the formulas for the censored or truncated conditional mean. Unfortunately these are based on strong distributional assumptions, as we will see.

16.2.2. Censoring and Truncation Mechanisms

As is customary for regression analysis, we let y denote the observed value of the dependent variable. The departure from usual analysis is that y is the incompletely observed value of a **latent dependent variable** y^* , where the observation rule is

$$y = g(y^*),$$

for some specified function $g(\cdot)$. Leading examples of $g(\cdot)$ immediately follow.

Censoring

With **censoring** we always observe the regressors \mathbf{x} , completely observe y^* for a subset of the possible values of y^* , and incompletely observe y for the remaining possible values of y^* . If censoring is **from below** (or from the left), we observe

$$y = \begin{cases} y^* & \text{if } y^* > L \\ L & \text{if } y^* \leq L. \end{cases} \quad (16.2)$$

For example, all consumers may be sampled with some having positive durable goods expenditures ($y^* > 0$) and others having zero expenditures ($y^* \leq 0$). If censoring is **from above** (or from the right) we observe

$$y = \begin{cases} y^* & \text{if } y^* < U \\ U & \text{if } y^* \geq U. \end{cases} \quad (16.3)$$

For example, annual income data may be top-coded at $U = \$100,000$. This form of censoring is called type 1 censoring in the duration literature (see Section 17.4.1).

The incompletely observed observations on y^* are set to L or U for simplicity. More generally, we require that for incompletely observed observations y^* is known to be missing (i.e., we observe that y^* lies outside the relevant bound) and regressors \mathbf{x} continue to be completely observed.

Truncation

Truncation entails additional information loss as all data on observations at the bound are lost. With truncation from below we observe only

$$y = y^* \quad \text{if } y^* > L. \quad (16.4)$$

For example, only consumers who purchased durable goods may be sampled ($L = 0$). With truncation from above we observe only

$$y = y^* \quad \text{if } y^* < U. \quad (16.5)$$

For example, only low-income individuals may be sampled.

Interval Data

Interval data are data recorded in intervals. Survey data are often collected in this way to aid recall and to provide some greater anonymity in responses to more personal

questions. For example, income may be reported in intervals of \$10,000 and then top-coded at \$100,000. Such data are censored at multiple points, with the observed data y being the particular interval in which the unobserved y^* lies.

16.2.3. Censored and Truncated MLE

Censoring and truncation are easily dealt with if the researcher applies a fully parametric approach. This may be the case with interval data or top-coded data where, for example, it may be reasonable to assume a log-normal distribution for earnings or a negative binomial model for number of doctor visits.

If the conditional distribution of y^* given regressors \mathbf{x} is specified, then the parameters of this distribution can be consistently and efficiently estimated by ML estimation based on the conditional distribution of the censored or truncated y . Specifically, let $f^*(y^*|\mathbf{x})$ and $F^*(y^*|\mathbf{x})$ denote the conditional probability density function (or probability mass function) and cumulative distribution function of the latent variable y^* . Then one can always obtain $f(y|\mathbf{x})$ and $F(y|\mathbf{x})$, the corresponding conditional pdf and cdf of the observed dependent variable y , since $y = g(y^*)$ is a transformation of y^* .

The limitation of the parametric approach is its reliance on strong distributional assumptions. For example, for the linear regression model under normality the MLE remains consistent even if the errors are nonnormal, but the censored MLE becomes inconsistent if the errors are nonnormal (see Section 16.3.2). More flexible models and semiparametric methods are presented in later sections.

Censored MLE

Censoring and truncation change both the conditional mean and the conditional density. We begin with the density.

Consider ML estimation given censoring from below. For $y > L$ the density of y is the same as that for y^* , so $f(y|\mathbf{x}) = f^*(y|\mathbf{x})$. For $y = L$, the lower bound, the density is discrete with mass equal to the probability of observing $y^* \leq L$, or $F^*(L|\mathbf{x})$. Thus for censoring from below

$$f(y|\mathbf{x}) = \begin{cases} f^*(y|\mathbf{x}) & \text{if } y > L, \\ F^*(L|\mathbf{x}) & \text{if } y = L. \end{cases}$$

As mentioned after (16.3), setting $y = L$ when $y^* \leq L$ is not necessary. Even if no value of y is observed when $y^* \leq L$ the density is still $F^*(L|\mathbf{x})$.

The density is a hybrid of the pdf and cdf of y^* . Similar to analysis for binary outcome models, it is notationally convenient to introduce an indicator variable

$$d = \begin{cases} 1 & \text{if } y > L, \\ 0 & \text{if } y = L. \end{cases} \quad (16.6)$$

Then the conditional density given censoring from below can be written as

$$f(y|\mathbf{x}) = f^*(y|\mathbf{x})^d F^*(L|\mathbf{x})^{1-d}. \quad (16.7)$$

For a sample of N independent observations, the censored MLE maximizes

$$\ln L_N(\boldsymbol{\theta}) = \sum_{i=1}^N \{d_i \ln f^*(y_i | \mathbf{x}_i, \boldsymbol{\theta}) + (1 - d_i) \ln F^*(L_i | \mathbf{x}_i, \boldsymbol{\theta})\}, \quad (16.8)$$

where $\boldsymbol{\theta}$ are the parameters of the distribution of y^* . For generality the censoring lower bound L_i is permitted to vary across individuals, though usually $L_i = L$. The censored MLE is consistent and asymptotically normal, provided the original density of the uncensored variable $f^*(y^* | \mathbf{x}, \boldsymbol{\theta})$ is correctly specified.

When censoring is instead from above, the log-likelihood is similar to (16.8), except now $d = 1$ if $y < U$ and $d = 0$ otherwise, and $F^*(L | \mathbf{x}, \boldsymbol{\theta})$ is replaced by $1 - F^*(U | \mathbf{x}, \boldsymbol{\theta})$. A leading example is right-censored duration data (see Section 17.4).

Truncated MLE

For truncation from below at L , and suppressing dependence on \mathbf{x} , the conditional density of the observed y is

$$\begin{aligned} f(y) &= f^*(y | y > L) \\ &= f^*(y) / \Pr[y | y > L] \\ &= f^*(y) / [1 - F^*(L)]. \end{aligned}$$

The truncated MLE therefore maximizes

$$\ln L_N(\boldsymbol{\theta}) = \sum_{i=1}^N \{\ln f^*(y_i | \mathbf{x}_i, \boldsymbol{\theta}) - \ln[1 - F^*(L_i | \mathbf{x}_i, \boldsymbol{\theta})]\}. \quad (16.9)$$

If instead truncation is from above, the log-likelihood is (16.9), except that $1 - F^*(L | \mathbf{x}, \boldsymbol{\theta})$ is replaced by $F^*(U | \mathbf{x}, \boldsymbol{\theta})$.

Ignoring censoring or truncation leads to inconsistency. For example, if truncation is ignored the MLE maximizes $\sum_i \ln f^*(y_i | \mathbf{x}_i, \boldsymbol{\theta})$, which is the wrong likelihood function as it drops the second term in (16.9). Consistency of the censored and truncated MLE requires correct specification of $f(\cdot)$, which in turn requires correct specification of the latent variable density $f^*(\cdot)$. Even if $f^*(\cdot)$ is an LEF density (see Section 5.7.3), the density, and not just the mean, must be correctly specified if censoring or truncation are present.

Interval Data MLE

Suppose the latent variable y^* is only observed to lie in the $(J + 1)$ mutually exclusive intervals $(-\infty, a_1], (a_1, a_2], \dots, (a_J, \infty)$, where a_1, a_2, \dots, a_J are known. Then since

$$\begin{aligned} \Pr[a_j < y^* \leq a_{j+1}] &= \Pr[y^* \leq a_{j+1}] - \Pr[y^* \leq a_j] \\ &= F^*(a_{j+1}) - F^*(a_j), \end{aligned}$$

the interval data MLE maximizes

$$\ln L_N(\boldsymbol{\theta}) = \sum_{i=1}^N \sum_{j=0}^J d_{ij} \ln [F^*(a_{j+1} | \mathbf{x}_i, \boldsymbol{\theta}) - F^*(a_j | \mathbf{x}_i, \boldsymbol{\theta})], \quad (16.10)$$

where the d_{ij} , $j = 0, \dots, J$, are binary indicators equal to one if $y_{ij} \in (a_j, a_{j+1}]$ and zero otherwise. This is similar to an ordered probit or logit model (see Section 15.9.1), except here the interval boundaries a_1, \dots, a_J are known.

16.2.4. Poisson Censored and Truncated MLE Example

Assume that y^* is Poisson distributed, so that $f^*(y) = e^{-\mu} \mu^y / y!$ and $\ln f^*(y) = -\mu + y \ln \mu - \ln y!$, with mean $\mu = \exp(\mathbf{x}'\beta)$.

Suppose the number of visits to a health clinic is modeled, but data are only available for people who visited the health clinic. Then the data are truncated from below at zero and we only observe $y = y^*$ if $y^* > 0$. Then $F^*(0) = \Pr[y^* \leq 0] = \Pr[y^* = 0] = e^{-\mu}$, and from (16.9) the truncated MLE for β maximizes

$$\ln L_N(\beta) = \sum_{i=1}^N \left\{ -\exp(\mathbf{x}'_i \beta) + y_i \mathbf{x}'_i \beta - \ln y_i! - \ln[1 - \exp(-\exp(\mathbf{x}'_i \beta))] \right\}.$$

Suppose instead that data are censored from above at 10 because of top-coding, so that we observe $y = y^*$ if $y^* < 10$ and that $y = 10$ if $y^* \geq 10$. Then $\Pr[y^* \geq 10] = 1 - \Pr[y^* < 10] = 1 - \sum_{k=0}^9 f^*(k)$. From (16.8) the censored MLE for β maximizes

$$\begin{aligned} \ln L_N(\beta) = \sum_{i=1}^N & \left\{ d_i \left[-\exp(\mathbf{x}'_i \beta) + y_i \mathbf{x}'_i \beta - \ln y_i! \right] \right. \\ & \left. + (1 - d_i) \ln \left[\sum_{k=0}^9 e^{-\exp(\mathbf{x}'_i \beta)} (\exp(\mathbf{x}'_i \beta))^k / k! \right] \right\}. \end{aligned}$$

In both cases the resulting first-order conditions are considerably more complicated than those for the Poisson MLE without truncation or censoring. Also, in both cases ignoring the truncation or censoring and maximizing the original density leads to inconsistent parameter estimates.

16.2.5. Censored and Truncated Conditional Means

Censoring and truncation change the conditional mean.

For example, consider the Poisson truncated from below at zero. The truncated density is $f^*(y)/[1 - F^*(0)]$, $y = 1, 2, \dots$, so the truncated mean is $\sum_{k=1}^{\infty} kf^*(k)/[1 - F^*(0)] = \sum_{k=0}^{\infty} kf^*(k)/[1 - F^*(0)] = \mu/(1 - e^{-\mu})$. Thus

$$E[y|\mathbf{x}] = \exp(\mathbf{x}'\beta)/[1 - \exp(-\exp(\mathbf{x}'\beta))],$$

rather than $\exp(\mathbf{x}'\beta)$ if there were no truncation.

This expression for $E[y|\mathbf{x}]$ can be used for NLS estimation. There is little advantage to NLS rather than ML estimation, however, as given truncation the NLS estimator relies on distributional assumptions that are essentially as strong as those needed for consistency of the more efficient ML estimator.

16.3. Tobit Model

Truncation and censoring arise most often in econometrics in the linear regression model with normally distributed error, when only positive outcomes are completely observed. This model is called the Tobit model after Tobin (1958), who applied it to individual expenditures on consumer durable goods. The model in practice is usually too restrictive. It is nonetheless presented in some detail, as it provides the basis for more general models presented in subsequent sections of this chapter.

16.3.1. Tobit Model

The censored normal regression model, or **Tobit model**, is one with censoring from below at zero where the latent variable is linear in regressors with additive error that is normally distributed and homoskedastic. Thus

$$y^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon, \quad (16.11)$$

where the error term

$$\varepsilon \sim \mathcal{N}[0, \sigma^2] \quad (16.12)$$

has variance σ^2 constant across observations. This implies that the latent variable $y^* \sim \mathcal{N}[\mathbf{x}'\boldsymbol{\beta}, \sigma^2]$. The observed y is defined by (16.2) with $L = 0$, so

$$y = \begin{cases} y^* & \text{if } y^* > 0, \\ - & \text{if } y^* \leq 0, \end{cases} \quad (16.13)$$

where $-$ means that y is observed to be missing. No particular value of y is necessarily observed when $y^* \leq 0$, though in some settings such as durable goods expenditures we observe $y = 0$.

Equations (16.11) – (16.13) define the prototypical Tobit model analyzed by Tobin (1958). More generally, Tobit models begin with (16.11) and (16.12) for the latent variable but can have other censoring mechanisms including censoring from above, censoring from both below and above (the **two-limit Tobit model**), and interval-censored data. The results in this section are restricted to the censoring mechanism given in (16.13). The models of later sections are sometimes called generalized Tobit models.

The normalization $L = 0$ is not only natural in many settings, but some such normalization is necessary for a linear model with intercept and constant threshold parameter L . Then we observe y if $y^* > L$, or equivalently if $\beta_1 + \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon > L$ or $(\beta_1 - L) + \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon > 0$. Thus only the difference $(\beta_1 - L)$ is identified. More generally, the latent model $y^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon$ with variable censoring threshold $L = \mathbf{x}'\boldsymbol{\gamma}$ is observationally equivalent to the latent model $y^* = \mathbf{x}'(\boldsymbol{\beta} - \boldsymbol{\gamma}) + \varepsilon$ with fixed threshold $L = 0$. These results are a consequence of censoring arising in a linear model with additive error and do not carry over to nonlinear models, such as the preceding Poisson example.

16.3. TOBIT MODEL

Applying the general expression (16.7) for the censored density, here $f^*(y)$ is the $\mathcal{N}[\mathbf{x}'\beta, \sigma^2]$ density and

$$\begin{aligned} F^*(0) &= \Pr[y^* \leq 0] \\ &= \Pr[\mathbf{x}'\beta + \varepsilon \leq 0] \\ &= \Phi(-\mathbf{x}'\beta/\sigma) \\ &= 1 - \Phi(\mathbf{x}'\beta/\sigma), \end{aligned}$$

where $\Phi(\cdot)$ is the standard normal cdf and the last equality uses symmetry of the standard normal distribution. Thus the censored density can be expressed as

$$f(y) = \left[\frac{1}{\sqrt{2\pi}\sigma^2} \exp \left\{ -\frac{1}{2\sigma^2}(y - \mathbf{x}'\beta)^2 \right\} \right]^d \left[1 - \Phi \left(\frac{\mathbf{x}'\beta}{\sigma} \right) \right]^{1-d}, \quad (16.14)$$

where the binary indicator d is defined in (16.6) with $L = 0$.

The Tobit MLE $\hat{\theta} = (\hat{\beta}', \hat{\sigma}^2)'$ maximizes the censored log-likelihood function (16.8). Given (16.14) this becomes

$$\begin{aligned} \ln L_N(\beta, \sigma^2) &= \sum_{i=1}^N \left\{ d_i \left(-\frac{1}{2} \ln 2\pi - \frac{1}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (y_i - \mathbf{x}'_i\beta)^2 \right) \right. \\ &\quad \left. + (1 - d_i) \ln \left(1 - \Phi \left(\frac{\mathbf{x}'_i\beta}{\sigma} \right) \right) \right\}, \end{aligned} \quad (16.15)$$

a mixture of discrete and continuous densities. The first-order conditions are

$$\frac{\partial \ln L_N}{\partial \beta} = \sum_{i=1}^N \frac{1}{\sigma^2} \left(d_i(y_i - \mathbf{x}'_i\beta) - (1 - d_i) \frac{\sigma\phi_i}{(1 - \Phi_i)} \right) \mathbf{x}_i = \mathbf{0} \quad (16.16)$$

$$\frac{\partial \ln L_N}{\partial \sigma^2} = \sum_{i=1}^N \left\{ d_i \left(-\frac{1}{2\sigma^2} + \frac{(y_i - \mathbf{x}'_i\beta)^2}{2\sigma^4} \right) + (1 - d_i) \frac{\phi_i \mathbf{x}'_i\beta}{(1 - \Phi_i)} \frac{1}{2\sigma^3} \right\} = 0,$$

using $\partial\Phi(z)/\partial z = \phi(z)$ where $\phi(\cdot)$ is the standard normal pdf, and with the definitions $\phi_i = \phi(\mathbf{x}'_i\beta/\sigma)$ and $\Phi_i = \Phi(\mathbf{x}'_i\beta/\sigma)$. As usual $\hat{\theta}$ is consistent if the density is correctly specified, that is, if the dgp is (16.11) and (16.12) and the censoring mechanism is (16.13). The MLE is asymptotic normal distributed with variance matrix given in, for example, Maddala (1983, p. 155) and Amemiya (1985, p. 373).

Tobin (1958) proposed ML estimation of the Tobit model and asserted that the usual ML theory applied. Amemiya (1973) provided a formal proof that the usual theory did apply, despite the mixed discrete–continuous nature of the censored density. The appendix of this classic paper of Amemiya details the asymptotic theory for extremum estimators presented in Section 5.3.

If data are truncated, rather than censored, from below at zero then the Tobit MLE $\hat{\theta} = (\hat{\beta}', \hat{\sigma}^2)'$ maximizes the truncated normal log-likelihood function

$$\ln L_N(\beta, \sigma^2) = \sum_{i=1}^N \left\{ -\frac{1}{2} \ln \sigma^2 - \frac{1}{2} \ln 2\pi - \frac{1}{2\sigma^2} (y_i - \mathbf{x}'_i \beta)^2 - \ln \Phi(\mathbf{x}'_i \beta / \sigma) \right\}, \quad (16.17)$$

obtained using (16.9) for y^* distributed as in (16.11) and (16.12).

16.3.2. Inconsistency of the Tobit MLE

A very major weakness of the Tobit MLE is its heavy reliance on distributional assumptions. If the error ε is either heteroskedastic or nonnormal the MLE is inconsistent.

This can be seen from the ML first-order conditions (16.16), which are a quite complicated function of variables including d_i , y_i , ϕ_i , and Φ_i . The first equation in (16.16) satisfies $E[\partial \ln L_N / \partial \beta] = \mathbf{0}$, a necessary condition for consistency (see Section 5.3.7), if

$$\begin{aligned} E[d_i] &= \Phi_i, \\ E[d_i y_i] &= \Phi_i \mathbf{x}'_i \beta + \sigma \phi_i. \end{aligned}$$

These moment conditions can be shown to hold if the dgp is (16.11) and (16.12) and the censoring mechanism is (16.13). However, they are unlikely to hold under any other specification of the dgp, as they rely heavily on both normality and homoskedasticity. For example, with *heteroskedastic errors* the estimator is inconsistent, since then $E[d_i] = \Phi(\mathbf{x}'_i \beta / \sigma_i) \neq \Phi_i$ unless $\sigma_i^2 = \sigma^2$.

Consistent estimation with heteroskedastic normal errors is possible by specifying a model for heteroskedasticity, say $\sigma_i^2 = \exp(\mathbf{z}'_i \gamma)$. For censoring from below at zero the log-likelihood $\ln L_N(\beta, \gamma)$ is that given in (16.15) with σ^2 replaced by $\exp(\mathbf{z}'_i \gamma)$. Consistency then requires normal errors and correct specification of the functional form of the heteroskedasticity.

Clearly, with censoring or truncation, distributional assumptions become important even for distributions somewhat robust to misspecification in the uncensored or untruncated case. Specification tests for the Tobit model are discussed in Section 16.3.7. In many censored data applications the Tobit model is not appropriate. More general models presented in subsequent sections of this chapter are instead used.

16.3.3. Censored and Truncated Means in Linear Regression

Censoring and truncation in the linear regression model (16.11) lead to observed dependent variable y that has distribution with conditional mean other than $\mathbf{x}' \beta$, conditional variance other than σ^2 even if ε is homoskedastic, and distribution that is nonnormal even if ε is normally distributed. We present general results for linear regression in this section before specializing to normally distributed errors in Sections 16.3.4–

16.3.7. The results provide additional insights regarding the consequences of truncation and censoring and form the basis for non-ML estimation methods presented in later sections.

We begin with the truncated mean. The effects of truncation are intuitively predictable. Left-truncation excludes small values, so the mean should increase, whereas with right-truncation the mean should decrease. Since truncation reduces the range of variation, the variance should decrease.

For *left-truncation* at zero we only observe y if $y^* > 0$. If we suppress dependence of expectations on \mathbf{x} for notational simplicity, the left-truncated mean becomes

$$\begin{aligned} E[y] &= E[y^*|y^* > 0] \\ &= E[\mathbf{x}'\beta + \varepsilon|\mathbf{x}'\beta + \varepsilon > 0] \\ &= E[\mathbf{x}'\beta|\mathbf{x}'\beta + \varepsilon > 0] + E[\varepsilon|\mathbf{x}'\beta + \varepsilon > 0] \\ &= \mathbf{x}'\beta + E[\varepsilon|\varepsilon > -\mathbf{x}'\beta], \end{aligned} \tag{16.18}$$

where the second equality uses (16.11), and the last equality assumes ε is independent of \mathbf{x} . As expected the truncated mean exceeds $\mathbf{x}'\beta$, since $E[\varepsilon|\varepsilon > c]$ for any constant c will exceed $E[\varepsilon]$.

For data *left-censored* at zero suppose we observe $y = 0$, rather than merely that $y^* \leq 0$. The censored mean is obtained by first conditioning the observable y on the binary indicator d defined in (16.6) with $L = 0$ and then unconditioning. Suppressing dependence on \mathbf{x} for notational simplicity again, we have the left-censored mean

$$\begin{aligned} E[y] &= E_d[E_{y|d}[y|d]] \\ &= \Pr[d = 0] \times E[y|d = 0] + \Pr[d = 1] \times E[y|d = 1] \\ &= 0 \times \Pr[y^* \leq 0] + \Pr[y^* > 0] \times E[y^*|y^* > 0] \\ &= \Pr[y^* > 0] \times E[y^*|y^* > 0], \end{aligned} \tag{16.19}$$

where $\Pr[y^* > 0] = 1 - \Pr[y^* \leq 0] = \Pr[\varepsilon > -\mathbf{x}'\beta]$ is one minus the censoring probability and $E[y^*|y^* > 0]$ is the truncated mean already derived in (16.18).

In summary, for the linear regression model with censoring or truncation from below at zero, the conditional means are given by

$$\begin{aligned} \text{latent variable: } \quad E[y^*|\mathbf{x}] &= \mathbf{x}'\beta \\ \text{left-truncated (at 0): } E[y|\mathbf{x}, y > 0] &= \mathbf{x}'\beta + E[\varepsilon|\varepsilon > -\mathbf{x}'\beta], \\ \text{left-censored (at 0): } E[y|\mathbf{x}] &= \Pr[\varepsilon > -\mathbf{x}'\beta] \{ \mathbf{x}'\beta + E[\varepsilon|\varepsilon > -\mathbf{x}'\beta] \}. \end{aligned} \tag{16.20}$$

It is clear that even though the original conditional mean is linear, censoring or truncation leads to conditional means that are nonlinear so that OLS estimates will be inconsistent.

One possible approach to take is a parametric one of assuming a distribution for ε . This leads to expressions for $E[\varepsilon|\varepsilon > -\mathbf{x}'\beta]$ and $\Pr[\varepsilon > -\mathbf{x}'\beta]$ and hence the truncated or censored conditional mean. We do this in the next section for normally distributed errors.

Inverse Mills Ratio as Cutoff Varies

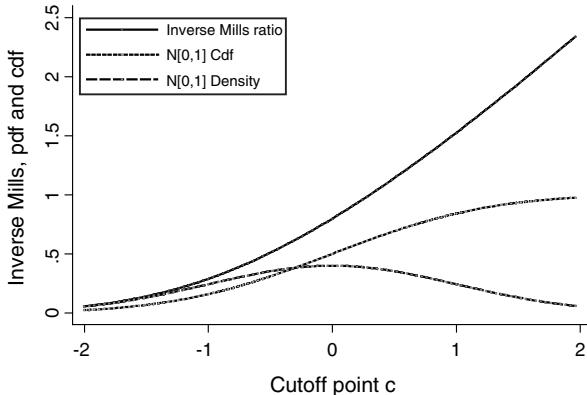


Figure 16.2: Inverse Mills ratio for the standard normal distribution as the censoring or cutoff point c increases. Standard normal cdf and density also plotted.

A second approach seeks to avoid or minimize such parametric assumptions. We consider this in a later section, but note here that regardless of the distribution for ε the truncated mean is a single-index model with correction term decreasing in $\mathbf{x}'\boldsymbol{\beta}$ since $E[\varepsilon|\varepsilon > -\mathbf{x}'\boldsymbol{\beta}]$ is a monotonically decreasing function in $\mathbf{x}'\boldsymbol{\beta}$.

16.3.4. Censored and Truncated Means in the Tobit Model

For the Tobit model the regression error ε is normal and we use the following result, derived in Section 16.10.1.

Proposition 16.1 (Truncated Moments of the Standard Normal): Suppose $z \sim \mathcal{N}[0, 1]$. Then the left-truncated moments of z are

- (i) $E[z|z > c] = \phi(c)/[1 - \Phi(c)]$, and $E[z|z > -c] = \phi(c)/\Phi(c)$,
- (ii) $E[z^2|z > c] = 1 + c\phi(c)/[1 - \Phi(c)]$, and
- (iii) $V[z|z > c] = 1 + c\phi(c)/[1 - \Phi(c)] - \phi(c)^2/[1 - \Phi(c)]^2$

Result (i) of Proposition 16.1 is shown in Figure 16.2. We consider truncation of $z \sim \mathcal{N}[0, 1]$ from below at c , where c ranges from -2 to 2 . The lowest curve is the standard normal density $\phi(c)$ evaluated at c . The middle curve is the standard normal cdf $\Phi(c)$ evaluated at c and gives the probability of truncation when truncation is at c . This probability is approximately 0.023 at $c = -2$ and 0.977 at $c = 2$. The upper curve gives the truncated mean $E[z|z > c] = \phi(c)/[1 - \Phi(c)]$. As expected this is close to $E[z] = 0$ for $c = -2$, since then there is little truncation, and $E[z|z > c] > c$. What is not expected a priori is that $\phi(c)/[1 - \Phi(c)]$ is approximately linear, especially for $c > 0$. Moments when truncation is from above can be obtained using, for example, $E[z|z < c] = -E[-z|z > -c] = -\phi(c)/\Phi(c)$.

16.3. TOBIT MODEL

Appling this result to (16.18), the error term has truncated mean

$$\begin{aligned} E[\varepsilon|\varepsilon > -\mathbf{x}'\beta] &= \sigma E\left[\frac{\varepsilon}{\sigma} \mid \frac{\varepsilon}{\sigma} > \frac{-\mathbf{x}'\beta}{\sigma}\right] \\ &= \sigma\phi(-\frac{\mathbf{x}'\beta}{\sigma})/[1 - \Phi(-\frac{\mathbf{x}'\beta}{\sigma})] \\ &= \sigma\phi(\frac{\mathbf{x}'\beta}{\sigma})/[\Phi(\frac{\mathbf{x}'\beta}{\sigma})] \\ &= \sigma\lambda(\frac{\mathbf{x}'\beta}{\sigma}), \end{aligned} \quad (16.21)$$

where the second line uses Proposition 16.1, the third line uses symmetry about zero of $\phi(z)$, and we define

$$\lambda(z) = \frac{\phi(z)}{\Phi(z)}. \quad (16.22)$$

We follow the definition and terminology of Amemiya (1985) and many others in defining $\lambda(\cdot)$ as in (16.22) and calling it the **inverse Mills ratio**. From Johnson and Kotz (1970, p. 278), Mills actually tabulated the ratio $(1 - \Phi(z))/\phi(z)$ whose inverse $\phi(z)/[1 - \Phi(z)] = \phi(z)/\Phi(-z)$ is the hazard function of the normal distribution. Some authors therefore instead write (16.21) as $E[\varepsilon|\varepsilon > -\mathbf{x}'\beta] = \sigma\lambda^*(-\mathbf{x}'\beta/\sigma)$, where $\lambda^*(z) = \phi(z)/\Phi(-z)$ is referred to as the inverse Mills ratio.

Also, $\Pr[\varepsilon > -\mathbf{x}'\beta] = \Pr[-\varepsilon < \mathbf{x}'\beta] = \Pr[-\varepsilon/\sigma < \mathbf{x}'\beta/\sigma] = \Phi(\mathbf{x}'\beta/\sigma)$. Then the conditional means in (16.20) specialize to

$$\begin{aligned} \text{latent variable: } E[y^*|\mathbf{x}] &= \mathbf{x}'\beta, \\ \text{left-truncated (at 0): } E[y|\mathbf{x}, y > 0] &= \mathbf{x}'\beta + \sigma\lambda(\mathbf{x}'\beta/\sigma), \\ \text{left-censored (at 0): } E[y|\mathbf{x}] &= \Phi(\mathbf{x}'\beta/\sigma)\mathbf{x}'\beta + \sigma\phi(\mathbf{x}'\beta/\sigma). \end{aligned} \quad (16.23)$$

The variance is similarly obtained (see Exercise 16.1). Defining $w = \mathbf{x}'\beta/\sigma$, we have

$$\begin{aligned} \text{latent variable: } V[y^*|\mathbf{x}] &= \sigma^2, \\ \text{left-truncated (at 0): } V[y|\mathbf{x}, y > 0] &= \sigma^2 [1 - w\lambda(w) - \lambda(w)^2], \\ \text{left-censored (at 0): } V[y|\mathbf{x}] &= \sigma^2 \Phi(w) \{w^2 + w\lambda(w) + 1 - \Phi(w)[w + \lambda(w)]\}^2. \end{aligned} \quad (16.24)$$

Clearly truncation and censoring induce heteroskedasticity, and for truncation $V[y|\mathbf{x}] < \sigma^2$ so that truncation reduces variability, as expected.

These results assume normal errors. Maddala (1983, p. 369) gives results similar to Proposition 16.1 for the log-normal, logistic, uniform, Laplace, exponential, and gamma distributions.

16.3.5. Marginal Effects in the Tobit Model

The marginal effect is the effect on the conditional mean of the dependent variable of changes in the regressors. This effect varies according to whether interest lies in the latent variable mean $\mathbf{x}'\beta$ or the truncated or censored means given in (16.23).

Differentiating each with respect to \mathbf{x} yields

$$\begin{aligned} \text{latent variable: } & \partial E[y^*|\mathbf{x}]/\partial \mathbf{x} = \boldsymbol{\beta}, \\ \text{left-truncated (at 0): } & \partial E[y, y > 0|\mathbf{x}]/\partial \mathbf{x} = \{1 - w\lambda(w) - \lambda(w)^2\}\boldsymbol{\beta}, \\ \text{left-censored (at 0): } & \partial E[y|\mathbf{x}]/\partial \mathbf{x} = \Phi(w)\boldsymbol{\beta}, \end{aligned} \tag{16.25}$$

where $w = \mathbf{x}'\boldsymbol{\beta}/\sigma$ and we use $\partial\Phi(z)/\partial z = \phi(z)$ and $\partial\phi(z)/\partial z = -z\phi(z)$. The simple expression for the censored mean is obtained after some manipulation. It can be decomposed into two effects, one for $y = 0$ and one for $y > 0$ (see McDonald and Moffitt, 1980).

In some cases truncation or censoring is just an artifact of data collection, so the truncated and censored means are of no intrinsic interest and we are interested in $\partial E[y^*|\mathbf{x}]/\partial \mathbf{x} = \boldsymbol{\beta}$. For example, with top-coded earnings data we are clearly interested in measuring the effect of schooling on mean earnings rather than earnings of those not top-coded.

In other cases truncation or censoring has behavioral implications. In a model for hours worked, for example, the three marginal effects in (16.25) correspond to the effect of a change in a regressor on, respectively, (1) desired hours of work, (2) actual hours of work for workers, and (3) actual hours of work for workers and nonworkers. For (1) we clearly need an estimate of $\boldsymbol{\beta}$, but for (2) and (3) OLS slope coefficients, although inconsistent for $\boldsymbol{\beta}$, may actually provide a reasonable crude estimate of the marginal effect since the truncated and censored means are still fairly linear in \mathbf{x} .

16.3.6. Alternative Estimators for the Tobit Model

In addition to the MLE, consistent estimation is possible by NLS based on the correct expression for the truncated or censored mean. We consider the NLS estimator and other least-squares estimators.

NLS Estimator

The results in (16.23) can be used to permit consistent estimation of the Tobit model parameters by NLS. For example, with truncated data we minimize

$$S_N(\boldsymbol{\beta}, \sigma^2) = \sum_{i=1}^N (y_i - \mathbf{x}'_i \boldsymbol{\beta} - \sigma \lambda(\mathbf{x}'_i \boldsymbol{\beta}/\sigma))^2$$

with respect to both $\boldsymbol{\beta}$ and σ^2 , but then perform inference controlling for the heteroskedasticity given in (16.24). A similar estimator can be obtained for censored data.

This estimator is not used in practice. Consistency requires correct specification of the truncated mean, which from (16.21) requires both normality and homoskedasticity of the errors. One might as well estimate by ML since this relies on assumptions just as strong and is fully efficient. Moreover, in practice the NLS estimator can be imprecise. From Figure 16.2 it is clear that $\lambda(\mathbf{x}'\boldsymbol{\beta}/\sigma)$ is approximately linear in $\mathbf{x}'\boldsymbol{\beta}/\sigma$, leading to near collinearity because \mathbf{x} is also a regressor. In Section 16.5 we consider models that permit correction terms similar to $\sigma \lambda(\mathbf{x}'\boldsymbol{\beta}/\sigma)$ in (16.23) that have the advantage of depending in part on regressors other than those in \mathbf{x} .

16.3. TOBIT MODEL

Heckman Two-Step Estimator

From (16.23) the truncated (at zero) mean is

$$E[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta} + \sigma\lambda(\mathbf{x}'\boldsymbol{\beta}/\sigma). \quad (16.26)$$

Rather than use NLS, this can be estimated in the following two-step procedure if censored data are available. First, for the full sample do probit regression of d on \mathbf{x} , where the binary variable d equals one if $y > 0$ is observed, to give consistent estimate $\hat{\boldsymbol{\alpha}}$, where $\boldsymbol{\alpha} = \boldsymbol{\beta}/\sigma$. Second, for the truncated sample do OLS regression of y on \mathbf{x} and $\lambda(\mathbf{x}'\hat{\boldsymbol{\alpha}})$ to give consistent estimates of $\boldsymbol{\beta}$ and σ .

This estimation procedure, due to Heckman (1976, 1979), is presented in Section 16.5.4 where it is applied to the more general sample selection model. Section 16.10.2 derives the standard error of $\hat{\boldsymbol{\beta}}$ that accounts for the regressor $\lambda(\mathbf{x}'\hat{\boldsymbol{\alpha}})$ depending on estimated parameters and for heteroskedasticity induced by truncation.

OLS Estimation of the Tobit Model

The OLS estimates using censored or truncated data are inconsistent for $\boldsymbol{\beta}$. This is because the censored and truncated means given in (16.23) are not equal to $\mathbf{x}'\boldsymbol{\beta}$, violating the essential condition for consistency of OLS.

For censored data, OLS provides a linear approximation to the nonlinear censored regression curve. It is clear from Figure 16.1 and (16.25) that this line is flatter than the regression line for uncensored data, which has slope equal to the true slope parameter. Goldberger (1981) showed analytically that if y and \mathbf{x} are joint normally distributed and there is censoring from below at zero, then the OLS slope parameters converge to p times the true slope parameter, where p is the fraction of the sample with positive values of y . These conditions are restrictive but were relaxed somewhat by Ruud (1986). In practice this proportionality result provides a good empirical approximation to the inconsistency of OLS if a Tobit model is instead appropriate.

Similarly, with truncation the regression line is flatter than the untruncated regression line. Goldberger (1981) obtained an analytical result similar to that for the censored case. If y and \mathbf{x} are joint normally distributed and there is censoring from below at zero, then the OLS slope parameters converge to a multiple of the true slope parameter. The multiple, the expression for which is quite lengthy, lies between zero and one, and the shrinkage is the same for all slope coefficients. Truncated OLS therefore understates the absolute magnitude of the true slope parameters.

16.3.7. Specification Tests for the Tobit Model

Given the fragility of the Tobit model it is good practice to test for distributional misspecification. There are four broad strategies.

The first approach is to nest the Tobit model within a richer parametric model and apply a Wald, LR, or LM test. Since the null hypothesis model, the Tobit model, is most easily estimated it is natural to use LM tests. This is particularly straightforward for testing against heteroskedasticity of the form $\sigma_i^2 = \exp(\mathbf{x}_i'\boldsymbol{\alpha})$ in the censored

regression model. Using the OPG form of the LM test (see Section 7.3.5) we compute N times the uncentered R^2 from auxiliary regression of 1 on \tilde{s}_{1i} and \tilde{s}_{2i} , where $f_i = f(y_i | \mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\alpha})$ is the density given in (16.14) with σ replaced by $\exp(\mathbf{x}'\boldsymbol{\alpha})$, the expressions for $s_{1i} = \partial \ln f_i / \partial \boldsymbol{\beta}$ and $s_{2i} = \partial \ln f_i / \partial \boldsymbol{\alpha}$ are obtained by minor adaptation of the expressions in (16.16), and tilde denotes evaluation at the censored Tobit MLE with all components of $\boldsymbol{\alpha}$ except that for the intercept equal to zero. A similar approach for testing the assumption of normally distributed errors is more difficult as there is no standard generalization of the normal.

A second approach is to use conditional moment tests (see Section 8.2) that do not require specification of an alternative hypothesis model. In particular, the first-order conditions (16.16) for the censored Tobit MLE suggest conditional moment tests based on the generalized residual

$$e_i = d_i \frac{y_i - \mathbf{x}'_i \boldsymbol{\beta}}{\sigma^2} - (1 - d_i) \frac{\phi_i}{\sigma(1 - \Phi_i)}.$$

If the Tobit model is correctly specified then $E[e_i | \mathbf{x}_i] = 0$ since the regularity conditions imply that $E[\partial \ln f(y_i) / \partial \boldsymbol{\beta}] = 0$. Then we can implement an m-test of $H_0 : E[ez] = \mathbf{0}$ against $H_a : E[ez] \neq \mathbf{0}$ using $N^{-1} \sum_{i=1}^N \widehat{e}_i \mathbf{z}_i$, where $\widehat{e}_i = e_i$ evaluated at the Tobit MLE $(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)$. From Section 8.2.2 this test can be implemented by computing N times the uncentered R^2 from auxiliary regression of 1 on $\widehat{e}_i \mathbf{z}_i$, \widehat{s}_{1i} , and \widehat{s}_{2i} , where $f_i = f(y_i | \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2)$ is the density given in (16.14) and $s_{1i} = \partial \ln f_i / \partial \boldsymbol{\beta}$ and $s_{2i} = \partial \ln f_i / \partial \sigma^2$ given in (16.16) are evaluated at $(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)$. The variables \mathbf{z}_i may be variables other than \mathbf{x}_i , in which case the test can be interpreted as a test of omitted regressors, or powers of the components of \mathbf{x}_i . Conditional moment tests based on higher order moments have also been developed. For details see Chesher and Irish (1987) and Pagan and Vella (1989).

A third approach is to adapt some of the diagnostic and testing methods developed for right-censored duration data (see Chapter 19) to left-censored normally distributed data.

A final approach contrasts the Tobit MLE $\widehat{\boldsymbol{\beta}}$ with alternative estimates of $\boldsymbol{\beta}$, notably the semiparametric estimates presented in Section 16.9, that are consistent under weaker distributional assumptions.

For further details see Pagan and Vella (1989), who present theory with some application, and Melenberg and Van Soest (1996), who provide a more complete application. Both papers consider specification tests for the richer sample selection model (see Section 16.5) in addition to those for the Tobit model.

16.4. Two-Part Model

The preceding models for censored data restrict the censoring mechanism to be from the same model as that generating the outcome variable. More generally, the censoring mechanism and outcome may be modeled using separate processes. For example, in explaining individual annual hospital expenses one process may determine hospitalization and a second process may explain consequent hospital expenses. The case for

postulating two separate mechanisms is strong if there is compelling reason to believe that certain realized values occur with too large or too small a frequency than is consistent with a simpler model. For example, one might observe many more zeros than is consistent with, for example, the Poisson distribution. A two-part model that permits the zeros and non-zeros to be generated by different densities adds flexibility. Indeed it is a specific type of mixture model.

There are two approaches to such generalization. The two-part model, given in this section, specifies a model for the censoring mechanism and a model for the outcome *conditional* on the outcome being observed. The sample selection model, presented in the subsequent section, instead specifies a joint distribution for the censoring mechanism and outcome, and then finds the implied distribution conditional on the outcome observed. These approaches are contrasted in Section 16.5.7.

16.4.1. Two-Part Model

Let an individual with fully observed outcome be called a **participant** in the activity being studied. Define a binary indicator variable $d = 1$ for participants and $d = 0$ for nonparticipants. Suppose that $y > 0$ is observed for participants and $y = 0$ is observed for nonparticipants. For nonparticipants we observe only $\Pr[d = 0]$. For participants the *conditional density* of y given $y > 0$ is specified to be $f(y|d = 1)$, for some choice of density $f(\cdot)$. The **two-part model** for y is then given by

$$f(y|\mathbf{x}) = \begin{cases} \Pr[d = 0|\mathbf{x}] & \text{if } y = 0, \\ \Pr[d = 1|\mathbf{x}]f(y|d = 1, \mathbf{x}) & \text{if } y > 0. \end{cases} \quad (16.27)$$

This model was presented in detail by Cragg (1971) as a generalization of the Tobit model, which can be presented as a special case of (16.27). An obvious model for the participation decision d is a probit or logit model. A latent variable formulation is that $d = 1$ if $I = \mathbf{x}'\beta + \varepsilon$ exceeds zero, and the model is then viewed as a **hurdle model** since crossing a hurdle or threshold leads to participation. To ensure positive values for the participants, the density $f(y|d = 1, \mathbf{x})$ should be that for a positive-valued random variable, such as the log-normal, or an appropriate density such as the normal truncated from below at zero.

For simplicity the same regressors usually appear in both parts of the model, but this can be relaxed and should be if there are obvious exclusion restrictions. Maximum likelihood estimation is straightforward as it separates into estimation of a discrete choice model using all observations and estimation of the parameters of the density $f(y|d = 1, \mathbf{x})$ using only observations with $y > 0$.

16.4.2. Two-Part Model Examples

Duan et al. (1983) present a leading application of this model to forecasting medical expenses using data from the Rand Health Insurance Experiment. They specified a probit model for whether or not any medical expenses were incurred during the year, so $\Pr[d = 1|\mathbf{x}] = \Phi(\mathbf{x}'_1\beta_1)$, and a log-normal model for medical expenses given that some expenses were incurred, so $\ln y|d = 1, \mathbf{x} \sim \mathcal{N}[\mathbf{x}'_2\beta_2, \sigma_2^2]$. Then expected

medical expenses over the entire population are given by

$$E[y|\mathbf{x}] = \Phi(\mathbf{x}'_1\boldsymbol{\beta}_1) \exp[\sigma_2^2/2 + \mathbf{x}'_2\boldsymbol{\beta}_2], \quad (16.28)$$

where the second term uses the result that if $\ln y \sim \mathcal{N}[\mu, \sigma^2]$ then $E[y] = \exp(\mu + \sigma^2/2)$. Mullahy (1998) considers such retransformation in further detail.

Two-part models are especially popular for modeling count data. For example, in modeling the number of doctor visits there is one model to determine whether or not a patient visits a physician at all and a second model to determine the consequent number of visits for those with at least one visit. Then $\Pr[d = 1]$ is specified to be the probability that a Poisson or negative binomial variable exceeds zero, whereas the density $f(y|d = 1)$ is specified to be a Poisson or negative binomial density truncated from below at zero. This model, due to Mullahy (1986), is called a hurdle model in the count literature and is detailed in Section 20.4.5.

For continuous data two-part models are used for expenditure models with excess zeros (Cragg's original motivation). An alternative, a sample selection model, is presented next.

16.5. Sample Selection Models

Sample selection can arise in many settings and so there are many sample selection models. This section begins with a general discussion of sample selection before focusing on a leading example, the **bivariate sample selection model** studied by Heckman (1979). Another leading example, the **Roy model**, is treated separately in Section 16.7.

16.5.1. Sample Selection Models

Observational studies are rarely based on pure random samples. Most often exogenous sampling is used (see Section 3.2.4) and the usual estimators can be applied. If instead a sample, intentionally or unintentionally, is based in part on values taken by a dependent variable, parameter estimates may be inconsistent unless corrective measures are taken. Such samples can be broadly defined as **selected samples**.

There are many **selection models**, since there are many ways that a selected sample may be generated. Indeed it is very easy to be unaware that a selected sample is being used. For example, consider interpretation of average scores over time on an achievement test such as the Scholastic Aptitude Test, when test taking is voluntary. A decline over time may be due to real deterioration in student knowledge. However, it may just reflect the selection effect that relatively more students have been taking the test over time and the new test takers are the relatively weaker students.

Selection may be due to **self-selection**, with the outcome of interest determined in part by individual choice of whether or not to participate in the activity of interest. It can also result from **sample selection**, with those who participate in the activity of interest deliberately oversampled – an extreme case being sampling only participants. In either case, similar issues arise and selection models are usually called sample selection models.

This chapter presents only three of the many selection models in the literature. The simplest model is the Tobit model already presented in Section 16.3. A prototypical commonly used model that we call the bivariate sample selection model is presented in the remainder of this section. This model generalizes the Tobit model by introducing a censoring latent variable that differs from the latent variable generating the outcome of interest. Another popular model called the Roy model is presented in Section 16.7. This model considers an outcome that takes one of two values depending on the value taken by a censoring random variable. These models correspond to, respectively, the Tobit model types 1, 2, and 5 in the terminology of Amemiya (1985, p. 384).

Consistent estimation in the presence of sample selection on unobservables relies on relatively strong distributional assumptions, even in the case of semiparametric estimation. Experimental data studies provide an attractive alternative as selection problems can then be avoided by random assignment. However, experiments can be difficult to implement in economics applications for cost and ethical reasons. The treatment effects approach, detailed in Chapter 25, seeks to apply the experimental approach to observational data.

16.5.2. A Bivariate Sample Selection Model (Type 2 Tobit)

Let y_2^* denote the outcome of interest. In the standard truncated Tobit model this outcome is observed if $y_2^* > 0$. A more general model introduces a different latent variable, y_1^* , and the outcome y_2^* is observed if $y_1^* > 0$. For example, y_1^* determines whether or not to work and y_2^* determines how much to work, and $y_1^* \neq y_2^*$ since there are fixed costs to work such as commuting costs that are more important in determining participation than hours of work once working.

The **bivariate sample selection model** comprises a **participation equation** that

$$y_1 = \begin{cases} 1 & \text{if } y_1^* > 0, \\ 0 & \text{if } y_1^* \leq 0 \end{cases} \quad (16.29)$$

and a resultant **outcome equation** that

$$y_2 = \begin{cases} y_2^* & \text{if } y_1^* > 0 \\ - & \text{if } y_1^* \leq 0. \end{cases} \quad (16.30)$$

This model specifies that y_2 is observed when $y_1^* > 0$, whereas y_2 need not take on any meaningful value when $y_1^* \leq 0$. The standard model specifies a linear model with additive errors for the latent variables, so

$$\begin{aligned} y_1^* &= \mathbf{x}'_1 \boldsymbol{\beta}_1 + \varepsilon_1, \\ y_2^* &= \mathbf{x}'_2 \boldsymbol{\beta}_2 + \varepsilon_2, \end{aligned} \quad (16.31)$$

with problems arising in estimating $\boldsymbol{\beta}_2$ if ε_1 and ε_2 are correlated. The Tobit model is clearly the special case where $y_1^* = y_2^*$.

There is no generally accepted name for this model. Heckman (1979) used it to illustrate estimation given sample selection. The model is equivalent to a **Tobit model with stochastic threshold** (Nelson, 1977). Suppose we observe y_2^* if $y_2^* > L^*$, where y_2^* is defined as in (16.31) and the threshold is $L^* = \mathbf{z}' \boldsymbol{\gamma} + v$ rather than $L^* = 0$ in

Section 16.3. Then, equivalently, we observe y_2^* if $y_1^* > 0$, where $y_1^* = y_2^* - L^* = (\mathbf{x}'_2 \beta_2 - \mathbf{z}' \gamma) + (\varepsilon_2 - v) = \mathbf{x}'_1 \beta_1 + \varepsilon_1$ and where \mathbf{x}_1 denotes the union of \mathbf{x}_2 and \mathbf{z} , and β_1 and ε_1 are defined in an obvious manner. Amemiya (1985, p. 384) calls the model a **type 2 Tobit model**. Wooldridge (2002, p. 506) calls the model one with a **probit selection equation**. Others call this model the generalized Tobit model or the sample selection model, though there are many such models.

Estimation by ML is straightforward given the additional assumption that the correlated errors are joint normally distributed and homoskedastic, with

$$\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \sim \mathcal{N} \left[\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right]. \quad (16.32)$$

As for the probit model in Section 14.4.1, the normalization $\sigma_1^2 = 1$ is used since only the sign of y_1^* is observed.

Given (16.29) and (16.30), for $y_1^* > 0$ we observe y_2^* , with probability equal to the probability that $y_1^* > 0$ times the conditional probability of y_2^* given that $y_1^* > 0$. Thus for positive y_2 the density of the observables is $f^*(y_2^* | y_1^* > 0) \times \Pr[y_1^* > 0]$. For $y_1^* \leq 0$ all that is observed is that this event has occurred, and the density is the probability of this event occurring. The bivariate sample selection model therefore has likelihood function

$$L = \prod_{i=1}^n \{\Pr[y_{1i}^* \leq 0]\}^{1-y_{1i}} \{f(y_{2i} | y_{1i}^* > 0) \times \Pr[y_{1i}^* > 0]\}^{y_{1i}}, \quad (16.33)$$

where the first term is the discrete contribution when $y_{1i}^* \leq 0$, since then $y_{1i} = 0$, and the second term is the continuous contribution when $y_{1i}^* > 0$. This likelihood function is applicable to quite general models, not just linear models with joint normal errors.

Specializing to linear models with joint normal errors gives a bivariate density $f^*(y_1^*, y_2^*)$ that is normal, leading to a conditional density in the second term that is univariate normal and easily handled. Amemiya (1985, pp. 385–387) provides details, including the exact form of the likelihood function.

The classic early application of this model was to labor supply, where y_1^* is the unobserved desire or propensity to work, whereas y_2 is actual hours worked. The model is also conceptually more appealing for labor supply than the Tobit model in Section 14.2.1 which required the artifice of “desired” hours of work. This prototypical application does have the complication that data on a key regressor, the offered wage, is missing for those individuals who do not work. This complication is handled by adding an equation for the offered wage and substituting this in, though the model is then strictly speaking not just a bivariate sample selection model. See Mroz (1987) for an excellent application to labor supply.

16.5.3. Conditional Means in the Bivariate Sample Selection Model

In this section we obtain the conditional truncated mean in the bivariate sample selection model. It differs from $\mathbf{x}'_2 \beta_2$, so that OLS regression of y_2 on \mathbf{x}_2 leads to inconsistent parameter estimates. Nonetheless, the expression for the conditional mean can be

used to motivate an alternative estimation procedure given in the subsequent section that relies on weaker distributional assumptions than those of the MLE.

We consider the truncated mean in the sample selectivity model where only positive values of y_2 are used. In general this is

$$\begin{aligned} E[y_2 | \mathbf{x}, y_1^* > 0] &= E[\mathbf{x}'_2 \boldsymbol{\beta}_2 + \varepsilon_2 | \mathbf{x}'_1 \boldsymbol{\beta}_1 + \varepsilon_1 > 0] \\ &= \mathbf{x}'_2 \boldsymbol{\beta}_2 + E[\varepsilon_2 | \varepsilon_1 > -\mathbf{x}'_1 \boldsymbol{\beta}_1], \end{aligned} \quad (16.34)$$

where \mathbf{x} denotes the union of \mathbf{x}_1 and \mathbf{x}_2 . If the errors ε_1 and ε_2 are independent then the last term simplifies to $E[\varepsilon_2] = 0$, and OLS regression of y_2 on \mathbf{x}_2 will give a consistent estimate of $\boldsymbol{\beta}_2$. However, any correlation between the two errors means that the truncated mean is no longer $\mathbf{x}'_2 \boldsymbol{\beta}_2$ and we need to account for selection.

To obtain $E[\varepsilon_2 | \varepsilon_1 > -\mathbf{x}'_1 \boldsymbol{\beta}_1]$ when ε_1 and ε_2 are correlated, Heckman (1979) noted that if the errors $(\varepsilon_1, \varepsilon_2)$ in (16.31) are joint normal as in (16.32) then Equation (16.36) in the following implies that

$$\varepsilon_2 = \sigma_{12} \varepsilon_1 + \xi, \quad (16.35)$$

where the random variable ξ is independent of ε_1 . To obtain this result, note that in general the joint normal distribution

$$\begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{bmatrix} \sim \mathcal{N} \left[\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right],$$

implies the conditional normal distribution

$$\mathbf{z}_2 | \mathbf{z}_1 \sim \mathcal{N} \left[\boldsymbol{\mu}_2 + \Sigma_{21} \Sigma_{11}^{-1} (\mathbf{z}_1 - \boldsymbol{\mu}_1), \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \right],$$

a result that implies that

$$\mathbf{z}_2 = \boldsymbol{\mu}_2 + \Sigma_{21} \Sigma_{11}^{-1} (\mathbf{z}_1 - \boldsymbol{\mu}_1) + \boldsymbol{\xi}, \quad (16.36)$$

where $\boldsymbol{\xi} \sim \mathcal{N}[\mathbf{0}, \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}]$ is independent of \mathbf{z}_1 . For the joint density given in (16.32) we have scalars and $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$ and $\sigma_1^2 = 1$, so (16.36) specializes to (16.35).

By using (16.35), the truncated mean (16.34) becomes

$$\begin{aligned} E[y_2 | \mathbf{x}, y_1^* > 0] &= \mathbf{x}'_2 \boldsymbol{\beta}_2 + E[(\sigma_{12} \varepsilon_1 + \xi) | \varepsilon_1 > -\mathbf{x}'_1 \boldsymbol{\beta}_1] \\ &= \mathbf{x}'_2 \boldsymbol{\beta}_2 + \sigma_{12} E[\varepsilon_1 | \varepsilon_1 > -\mathbf{x}'_1 \boldsymbol{\beta}_1], \end{aligned}$$

where we use independence of $\boldsymbol{\xi}$ and ε_1 . The selection term is similar to that in the simpler Tobit model and again using the expression for $E[z | z > -c]$ in Proposition 16.1 we obtain

$$E[y_2 | \mathbf{x}, y_1^* > 0] = \mathbf{x}'_2 \boldsymbol{\beta}_2 + \sigma_{12} \lambda(\mathbf{x}'_1 \boldsymbol{\beta}_1), \quad (16.37)$$

where $\lambda(z) = \phi(z)/\Phi(z)$ and we have used $\sigma_1^2 = 1$. Similarly, Proposition 16.1(iii) yields the truncated variance

$$V[y_2 | \mathbf{x}, y_1^* > 0] = \sigma_2^2 - \sigma_{12}^2 \lambda(\mathbf{x}'_1 \boldsymbol{\beta}_1)(\mathbf{x}'_2 \boldsymbol{\beta}_2 + \lambda(\mathbf{x}'_1 \boldsymbol{\beta}_1)). \quad (16.38)$$

The preceding analysis specifies no value for y_2 when $y_1^* \leq 0$. In some applications y_2 may equal zero when $y_1^* < 0$. Then it is meaningful to consider the censored mean.

Conditioning the observable y_2 on the unobservables y_1^* and y_2^* and then unconditioning yields

$$\begin{aligned} E[y_2|\mathbf{x}] &= E_{y_1^*}[E[y_2|\mathbf{x}, y_1^*]] \\ &= \Pr[y_1^* \leq 0|\mathbf{x}] \times 0 + \Pr[y_1^* > 0|\mathbf{x}] \times E[y_2^*|\mathbf{x}, y_1^* > 0] \\ &= 0 + \Phi(\mathbf{x}'_1 \boldsymbol{\beta}_1) \{ \mathbf{x}'_2 \boldsymbol{\beta}_2 + \sigma_{12} \lambda(\mathbf{x}'_1 \boldsymbol{\beta}_1) \} \\ &= \Phi(\mathbf{x}'_1 \boldsymbol{\beta}_1) \mathbf{x}'_2 \boldsymbol{\beta}_2 + \sigma_{12} \phi(\mathbf{x}'_1 \boldsymbol{\beta}_1), \end{aligned} \quad (16.39)$$

where the third line uses (16.37) and the last line uses $\lambda(z) = \phi(z)/\Phi(z)$. The censored variance can be shown to be heteroskedastic.

16.5.4. Heckman Two-Step Estimator

An important result is that OLS regression of y_2 on \mathbf{x}_2 alone using just the observed positive values of y_2 leads to inconsistent estimation of $\boldsymbol{\beta}$ unless the errors are uncorrelated so that $\sigma_{12} = 0$. This is clear from the truncated mean formula (16.37), which additionally includes the “regressor” $\lambda(\mathbf{x}'_1 \boldsymbol{\beta}_1)$.

Heckman’s two-step procedure, sometimes called the **Heckit estimator**, augments the OLS regression by an estimate of the omitted regressor $\lambda(\mathbf{x}'_1 \boldsymbol{\beta}_1)$. Thus using positive values of y_2 estimate by OLS the model

$$y_{2i} = \mathbf{x}'_{2i} \boldsymbol{\beta}_2 + \sigma_{12} \lambda(\mathbf{x}'_{1i} \widehat{\boldsymbol{\beta}}_1) + v_i, \quad (16.40)$$

where v is an error term, $\widehat{\boldsymbol{\beta}}_1$ is obtained by first-step probit regression of y_1 on \mathbf{x}_1 since $\Pr[y_1^* > 0] = \Phi(\mathbf{x}'_1 \boldsymbol{\beta}_1)$, and $\lambda(\mathbf{x}'_1 \widehat{\boldsymbol{\beta}}_1) = \phi(\mathbf{x}'_1 \widehat{\boldsymbol{\beta}}_1)/\Phi(\mathbf{x}'_1 \widehat{\boldsymbol{\beta}}_1)$ is the estimated inverse Mills ratio. This regression does not directly provide an estimate of σ_2^2 , but the truncated variance formula (16.38) leads to estimate $\widehat{\sigma}_2^2 = N^{-1} \sum_i [\widehat{v}_i^2 + \widehat{\sigma}_{12}^2 \widehat{\lambda}_i (\mathbf{x}'_1 \widehat{\boldsymbol{\beta}}_1 + \widehat{\lambda}_i)]$, where \widehat{v}_i is the OLS residual from (16.40) and $\widehat{\lambda}_i = \lambda(\mathbf{x}'_{1i} \widehat{\boldsymbol{\beta}}_1)$. The correlation between the two errors in (16.32) can then be estimated by $\widehat{\rho} = \widehat{\sigma}_{12}/\widehat{\sigma}_2$.

A test of whether or not $\sigma_{12} = 0$ or $\rho = 0$ is a test of whether or not the errors are correlated and sample selection correction is needed. One such test is a Wald test based on $\widehat{\sigma}_{12}$, the estimated coefficient of the inverse Mills ratio.

It is important to note that both the usual OLS standard errors and heteroskedasticity-robust standard errors reported from the regression (16.40) are incorrect. Correct formulas for the standard errors take account of two complications in the second-stage regression. First, even if $\boldsymbol{\beta}_1$ were known, the error in (16.40) is heteroskedastic from (16.38). Second, in fact $\boldsymbol{\beta}_1$ is replaced by an estimate, a complication studied in Section 6.6 and analyzed in Section 16.10.2 for the simpler Tobit model. Formulas for the correct standard errors are given in Heckman (1979); see also Greene (1981). Section 16.10.2 derives these formulas for the simpler Tobit model. Implementation is not simple so it is best to use a package that automatically handles this complication or to use the bootstrap.

The resulting estimator of $\boldsymbol{\beta}_2$ is consistent. Despite an efficiency loss compared to the MLE under joint normality of the errors that can be quite large, the estimator is very popular for the following reasons: (1) It is simple to implement; (2) the approach is applicable to a range of selection models including those given in Section 16.7; (3) the estimator requires distributional assumptions weaker than joint normality of ε_1

and ε_2 ; and (4) these distributional assumptions can be weakened even further to permit semiparametric estimation as in Section 16.9.

The **key assumption** needed is (16.35), essentially that

$$\varepsilon_2 = \delta\varepsilon_1 + \xi, \quad (16.41)$$

where ξ is independent of ε_1 . This seems to be a quite sensible model. In the case of expenditures on a durable good, say, this says that the error in the expenditure equation is a multiple of the error in the purchase decision equation, plus some noise that is independent of the purchase decision; essentially a linear regression model for the errors. Given assumption (16.41) the conditional mean (16.34) becomes

$$E[y_2|y_1^* > 0] = \mathbf{x}'_2\beta_2 + \delta E[\varepsilon_1|\varepsilon_1 > -\mathbf{x}'_1\beta_1]. \quad (16.42)$$

If ε_1 is standard normal distributed this leads to (16.37), the basis for the OLS regression (16.40).

More generally, Heckman's two-step method can be applied to (16.42) with distributions for ε_1 other than normal; see, for example, Olsen (1980). One can also use semiparametric methods that do not impose a functional form for $E[\varepsilon_1|\varepsilon_1 > -\mathbf{x}'_1\beta_1]$ (see Section 16.9).

16.5.5. Identification Considerations

The bivariate sample selection model with normal errors is theoretically identified without any restriction on the regressors. In particular, exactly the same regressors can appear in the equations for y_1^* and y_2^* .

The model with normally distributed errors is close to unidentified, however, if exactly the same regressors are used. If $\mathbf{x}_1 = \mathbf{x}_2$ then $E[y_2|y_1^* > 0] \simeq \mathbf{x}'_2\beta_2 + a + b\mathbf{x}'_2\beta_1$, using (16.37) and the observation from Section 16.3.2 that the inverse Mills ratio term $\lambda(\cdot)$ is approximately linear over a wide range of its argument. This leads to obvious multicollinearity problems, discussed in many articles including those by Nawata (1993), Nawata and Nagase (1996), and Leung and Yu (1996). Multicollinearity can be detected using the condition number given in Section 10.4.2, where from (16.40) the regressors are \mathbf{x}_2 and $\lambda(\mathbf{x}'_1\widehat{\beta}_1)$. The problem is less severe the greater the variation in $\mathbf{x}'_1\widehat{\beta}_1$ across observations, that is, the better a probit model can discriminate between participants and nonparticipants.

Semiparametric variants of the Heckman two-step method (see Section 16.9.3) do require an exclusion restriction. So **identification** in the bivariate sample selection model with normal errors is being achieved by functional form assumptions.

For practical purposes therefore, estimation of the bivariate sample selection model may require that at least one regressor in the participation equation (y_1^*) be excluded from the outcome equation (y_2^*). For example, fixed costs of working unrelated to hours worked will affect the decision to work but not hours worked. This can be a great limitation as in many applications, such as that in Section 16.6, it can be very difficult to make defensible exclusion restrictions.

16.5.6. Marginal Effects

The marginal effects in the bivariate sample selection model vary according to whether we consider the latent variable mean or the truncated mean given in (16.37) or the censored mean (if it is appropriate).

It is convenient to define \mathbf{x} to be the vector formed by union of \mathbf{x}_1 and \mathbf{x}_2 and rewrite $\mathbf{x}'_1\beta_1$ as $\mathbf{x}'\gamma_1$ and $\mathbf{x}'_2\beta_2$ as $\mathbf{x}'\gamma_2$. For example, the truncated mean becomes $E[y_2|\mathbf{x}] = \mathbf{x}'\gamma_2 + \sigma_{12}\lambda(\mathbf{x}'\gamma_1)$. Note that γ_1 and/or γ_2 will have some zero entries if $\mathbf{x}_1 \neq \mathbf{x}_2$. Differentiating with respect to \mathbf{x} yields the **marginal effects**

$$\text{uncensored: } \partial E[y_2^*|\mathbf{x}] / \partial \mathbf{x} = \gamma_2, \quad (16.43)$$

$$\text{truncated (at 0): } \partial E[y_2|\mathbf{x}, y_1 = 1] / \partial \mathbf{x} = \gamma_2 - \sigma_{12}\lambda(\mathbf{x}'\gamma_1)(\mathbf{x}'\gamma_1 + \lambda(\mathbf{x}'\gamma_1))$$

$$\begin{aligned} \text{censored (at 0): } \partial E[y_2|\mathbf{x}] / \partial \mathbf{x} &= \gamma_1\phi(\mathbf{x}'\gamma_1)\mathbf{x}'\gamma_2 + \Phi(\mathbf{x}'\gamma_1)\gamma_2 \\ &\quad - \sigma_{12}\mathbf{x}'\gamma_1\phi(\mathbf{x}'\gamma_1)\gamma_1, \end{aligned}$$

where $\lambda(z) = \phi(z)/\Phi(z)$, and we use $\partial\phi(z)/\partial z = -z\phi(z)$ and $\partial\lambda(z)/\partial z = -z\phi(z)/\Phi(z) - \phi(z)^2/\Phi(z)^2 = -\lambda(z)(z + \lambda(z))$. Interpretation of these three derivatives is similar to that discussed in some detail in Section 16.3.5. As already noted, analysis of the censored mean is appropriate only if y_2 takes the value of zero when $y_1 = 0$. In applications such as the log-normal health expenditures example discussed later there is no censored mean.

16.5.7. Selection on Observables and on Unobservables

There are many modeling situations that can be considered a two-part decision problem of first engaging in an activity and then determining the level of the activity. These decisions are intertwined and can be expected to depend on common factors. The natural model for such data is the bivariate selection model (16.29)–(16.31).

After inclusion of regressors any remaining error (ε_1 and ε_2) in the two processes may in some cases be uncorrelated. For example, for models of hospitalization it is possible that, after controlling for observed individual characteristics such as health status, there is no correlation between the error in the equation determining hospital admission and in the error in the equation determining length of hospital stay. In that case analysis is straightforward as selection is only based on observables since, for example, (16.37) simplifies when $\sigma_{12} = 0$. The two pieces can be modeled separately and the simpler two-part model of Section 16.4 can be used.

In other cases the errors may be correlated even after inclusion of the regressors. For example, in labor supply unobserved factors that make someone more likely to work may also make them more likely to work longer hours than would be predicted by the observable regressors. One can test whether there is such correlation between the errors. If there is correlation, then selection is on unobservables and the methods of this chapter come into play. Relatively strong distributional assumptions are needed, even with the Heckman two-step method.

The study by Duan et al. (1983) summarized in Section 16.4.2 was criticized for using the two-part model, which is more restrictive than the sample selection model. This led to considerable debate, with many of the relevant articles referenced in Leung

and Yu (1996), who emphasize the important role of potential correlation of the inverse Mills ratio term with the remaining regressors.

More generally, selection models such as the bivariate selection model permit **selection on both observables and unobservables**, as it permits selection on both observed regressors and unobserved errors. It is often more simply referred to as a model of **selection on unobservables**, with selection on observables implicit. This chapter emphasizes selection on unobservables.

If instead we have only **selection on observables**, analysis becomes much simpler. The two-part model of this chapter is an example. Chapter 25 on treatment evaluation emphasizes selection on observables (see the discussion in Section 25.3.3) and details methods such as propensity score matching.

16.6. Selection Example: Health Expenditures

For illustration we use data from the RAND Health Insurance Experiment (RHIE). The data extract comes from Deb and Trivedi (2002), who modeled the number of outpatient visits to a medical doctor and to all providers using count data models. Section 20.3 summarizes the data and Section 20.7 presents estimates of some standard count models.

Here instead we model annual health expenditures. The regressors are the same regressors as defined in detail in Table 20.4. They can be broken down into health insurance variables (LC, IDP, LPI, and FMDE), socioeconomic characteristics (LINC, LFAM, AGE, FEMALE, CHILD, FEMCHILD, BLACK, and EDUCDEC) and health status variables (PHYSLIM, NDISEASE, HLTHG, HLTHF, and HLTHP). The analysis in Chapter 20 uses four years of data whereas here we use only the second year of data, yielding 5,574 observations with summary statistics similar to but not exactly the same as those given in Table 20.4.

The dependent variable y is annual individual health expenditures. An econometric model needs to take account of two complications: (1) Health expenditures are zero for 23.2% of the sample and (2) the positive health expenditures are very right-skewed with a mean of \$221 that is much larger than the median of \$53. The logarithmic transformation eliminates this skewness, with a mean of 4.07 close to the median of 3.96 and the skewness statistic falls from 24.0 to 0.3. The kurtosis is 3.29, close to the normal value of 3.

We focus on modeling $\ln y$ for those with positive medical expenditures. Possible models include a two-part model, exposited for log medical expenditures in Section 16.4.2, and a bivariate sample selection model (see Section 16.5.2), where y_1 in (16.29) is an indicator for positive expenditures and y_2 in (16.30) is $\ln y$. Note that it is not meaningful to consider the value of y_2 when $y_1 = 0$ because $\ln 0$ is not defined. The two-part model is a special case of the bivariate sample selection model with $\sigma_{12} = 0$ in (16.32).

Table 16.1 presents results for the health insurance variables and health status regressors. Socioeconomic variables also included in the regression are omitted from the table for brevity.

Table 16.1. *Health Expenditure Data: Estimates from Two-Part and Selection Models^a*

Model Equation	Two-Part		Selection Two-Step		Selection MLE	
	DMED	LNMED	DMED	LNMED	DMED	LNMED
LC	-0.119 (-4.41)	-0.016 (-0.52)	-0.119 (-4.41)	-0.028 (-0.70)	-0.107 (-4.03)	-0.076 (2.25)
IDP	-0.128 (-2.45)	-0.079 (-1.28)	-0.128 (-2.45)	-0.028 (-0.70)	-0.109 (-2.13)	-0.150 (-2.26)
LPI	0.028 (3.19)	0.003 (0.28)	0.028 (3.19)	0.005 (0.47)	0.029 (3.42)	0.015 (1.42)
FMDE	0.008 (0.47)	-0.031 (-1.69)	0.008 (0.47)	-0.030 (-1.62)	0.001 (0.05)	-0.024 (1.21)
PHYSLIM	0.273 (3.67)	0.262 (3.81)	0.273 (3.67)	0.281 (3.50)	0.285 (3.94)	0.355 (4.70)
NDISEASE	0.022 (6.25)	0.020 (5.78)	0.022 (6.25)	0.022 (4.29)	0.021 (6.03)	0.029 (7.54)
HLTHG	0.039 (0.88)	0.144 (2.97)	0.039 (0.88)	0.147 (3.01)	0.058 (1.35)	0.156 (2.99)
HLTHF	0.192 (2.29)	0.364 (4.13)	0.192 (2.29)	0.382 (3.98)	0.224 (2.75)	0.445 (4.66)
HLTHP	0.640 (3.01)	0.787 (4.63)	0.640 (3.01)	0.833 (4.22)	0.798 (3.90)	0.999 (5.32)
ρ		0.000		0.168		0.736
σ_2				1.401		1.570
$\sigma_{12} = \rho\sigma_2$		0.000		0.236 (0.47)		1.155 (16.43)
$-\ln L$	10184.1				10170.1	

^a The *t*-statistics are in parentheses. Regressors also include eight socioeconomic characteristics. DMED is an indicator for whether or not medical expenditures are positive and LN MED is the natural logarithm of expenditures if positive. The *t*-statistics for the second step of the two-step selection model are based on errors that correct for the first-step estimation used to obtain the fitted inverse Mills ratio term.

We first compare the two-part model estimates with the two-step estimates of the bivariate sample selection model. The DMED equation estimates are identical as they are obtained by probit regression of DMED on the same regressors. The LN MED equation estimates differ because for two-step sample selection the second-step OLS regression for LN MED additionally includes as a regressor the fitted value of the inverse Mills ratio term. This additional term is statistically insignificant ($t = 0.47$) and low in magnitude with implied $\hat{\rho} = 0.168$ that is close to zero. As a result the two models lead to similar coefficient estimates in the LN MED equation.

As noted in Section 16.4.4 the two-step estimator can perform poorly if the inverse Mills ratio term is highly correlated with the other regressors. Here this does not appear to be the case as there is considerable range in the probit model predicted probabilities from 0.15 to 0.99 and the condition number (see Section 10.4.4) of the second-stage regressors at the second stage, although somewhat high, only doubles from 37 to 82 upon inclusion of the inverse Mills ratio. Although it is still preferable to have some exclusion restrictions, it is not clear in this application which regressors in the DMED equation might be reasonably excluded on a priori grounds from the LN MED equation.

The ML estimates of the bivariate sample selection model differ considerably from the previous estimates, in both DMED and LN MED equations. The errors in the

latent variable models for DMED and LNMED are highly correlated with estimate $\hat{\rho} = 0.736$ that is highly statistically significant ($t = 16.43$). The big difference between the two-step estimates and the ML estimates of σ_{12} (or of ρ) is best viewed as signifying a problem with the bivariate sample selection model. Rejection of the null hypothesis that the estimates have the same probability limit, a Hausman test given in Section 8.4, can be interpreted as rejection of the additional joint normality assumption needed to go from two-step estimation to ML estimation of the bivariate selection model. However, there may be a more fundamental problem that the bivariate sample selection model with the weaker assumption (16.41) and ε_1 iid normal is also not reasonable. Such fragility of the bivariate sample selection model is not unusual, especially if the same regressors are being used in both parts of the model so that identification is being secured through model specification assumptions. It is compounded here by use of health expenditure data, which can have quite large outliers so that errors may not be normal. Even though LNMED has skewness close to 0 and kurtosis close to 3, as already noted, standard tests of heteroskedasticity, skewness, and kurtosis resoundingly reject (with p -value 0.000) the null hypothesis that LNMED is normally distributed.

The regressor of most interest is LC, the natural logarithm of the coinsurance rate where the coinsurance rate equals the percentage of health cost borne by the insured paid by the patient. The most statistically significant effect is in determining whether or not expenditures are positive, rather than on the size of positive expenditures. If all observations were positive then the coefficient of LC in regression on LNMED equals the price elasticity of demand for health care. In fact in predicting the effect of changes in price on the conditional truncated mean of log expenditure we need to control for the effect of those with zero expenditure, as in the second line of (16.43).

In some applications interest lies in prediction rather than estimation of marginal effects. This is complicated in this example by a desire to predict the level rather than the log of expenditure. Assuming log-normality, the expression for the two-part model is given in (16.28). Duan et al. (1983) present a method to make predictions without the log-normality assumption that can be viewed as a variant of a bootstrap. See also Mullaly (1998).

16.7. Roy Model

In the bivariate sample selection model the dependent variable for an individual might not be observed. Thus we observe y_2 for an individual if $y_1 = 1$ but may not observe y_2 at all if $y_1 = 0$. In this section we consider a model in which y_2 is observed for all individuals, but in only one of the two possible states. This important model emphasizes **counterfactuals** and connects with the program evaluation literature presented in Chapter 25.

16.7.1. Roy Model

An often-cited article by Roy (1951) considered the consequences for the occupational distribution of earnings (both mean and variance) when there is individual

heterogeneity in skills and individuals self-select into occupations. The treatment was relatively general and nonmathematical, though it did assume that individual worker output in an occupation is log-normally distributed in the absence of selection, and it did not consider at all estimation of a formal model. During the 1970s a number of authors independently proposed models for similar situations that were estimable with cross-section data and considered selection on both observables and unobservables. Such models have become known as Roy models.

We define the prototypical **Roy model** as follows. A latent variable y_1^* determines whether the outcome observed is y_2^* or y_3^* . Specifically, we observe whether y_1^* is positive or negative,

$$y_1 = \begin{cases} 1 & \text{if } y_1^* > 0, \\ 0 & \text{if } y_1^* \leq 0, \end{cases} \quad (16.44)$$

and observe exactly one of y_2^* and y_3^* according to

$$y = \begin{cases} y_2^* & \text{if } y_1^* > 0, \\ y_3^* & \text{if } y_1^* \leq 0. \end{cases} \quad (16.45)$$

It is customary to specify a linear model with additive errors for the latent variables, with

$$\begin{aligned} y_1^* &= \mathbf{x}'_1 \boldsymbol{\beta}_1 + \varepsilon_1, \\ y_2^* &= \mathbf{x}'_2 \boldsymbol{\beta}_2 + \varepsilon_2, \\ y_3^* &= \mathbf{x}'_3 \boldsymbol{\beta}_3 + \varepsilon_3. \end{aligned} \quad (16.46)$$

A model with additive effect is the specialization $\mathbf{x}'_3 \boldsymbol{\beta}_3 = \mathbf{x}'_2 \boldsymbol{\beta}_2 + \alpha$. The simplest parametric model for correlated errors is the joint normal, with

$$\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix} \sim \mathcal{N} \left[\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix} \right], \quad (16.47)$$

where as usual the normalization $\sigma_1^2 = 1$ is used as only the sign of y_1^* is observed.

The log-likelihood function is similar to that for the bivariate sample selection model of Section 16.5, except that now y_3^* is observed if $y_1^* \leq 0$, so the term $\Pr[y_{1i}^* \leq 0]$ in (16.33) is replaced by $f(y_{3i} | y_{1i}^* \leq 0) \times \Pr[y_{1i}^* \leq 0]$.

It is more common to estimate the model using Heckman's two-step method applied to the truncated means,

$$\begin{aligned} E[y | \mathbf{x}, y_1^* > 0] &= \mathbf{x}'_2 \boldsymbol{\beta}_2 + \sigma_{12} \lambda(\mathbf{x}'_1 \boldsymbol{\beta}_1), \\ E[y | \mathbf{x}, y_1^* \leq 0] &= \mathbf{x}'_3 \boldsymbol{\beta}_3 - \sigma_{13} \lambda(-\mathbf{x}'_1 \boldsymbol{\beta}_1), \end{aligned} \quad (16.48)$$

where $\lambda(z) = \phi(z)/\Phi(z)$ and we have used $\sigma_1^2 = 1$. First-stage probit estimation of whether or not $y_1^* > 0$ yields an estimate of $\boldsymbol{\beta}_1$ and hence $\lambda(\mathbf{x}'_1 \hat{\boldsymbol{\beta}}_1)$. Two separate OLS regressions then lead to direct estimates of $(\boldsymbol{\beta}_2, \sigma_{12})$ and $(\boldsymbol{\beta}_3, \sigma_{13})$. Estimates of σ_2^2 and σ_3^2 can then be obtained using the squared residuals from the regressions, similar to the technique used for the bivariate sample selection model after (16.40). Maddala (1983, p. 225) provides complete details for this model, which he calls a **switching**

regression model with endogenous switching. This is also the **Tobit type 5 model** presented in Amemiya (1985, p. 399).

16.7.2. Variations of the Roy Model

Many models fall into the class of Roy models. Maddala (1983, Chapter 9) gives numerous references to what he calls models with self-selectivity. See also Amemiya (1985, Chapter 10). Here we present a few leading examples.

The bivariate sample selection model can be viewed as a special case where y_3^* is ignored and we only model the truncated moment $E[y_2^*|y_1^* > 0]$. Bivariate sample selection models where $y = 0$ when $y_1^* \leq 0$, such as in labor supply applications, can more directly be viewed as Roy models where we observe either $y = y_2^*$ or $y = 0$, so $y_3^* = 0$.

In the study of L.-F. Lee (1978), y_2^* and y_3^* denote, respectively, union and nonunion wage and y_1^* denotes tendency to be a union member. This adds the additional structure that

$$y_1^* = y_2^* - y_3^* + \mathbf{z}'\gamma + \zeta,$$

where $\mathbf{z}'\gamma + \zeta$ reflect costs of union membership and is very much in the spirit of Roy (1951). Substituting for y_2^* and y_3^* yields a reduced form for y_1^* :

$$y_1^* = (\mathbf{x}_2'\beta_2 - \mathbf{x}_3'\beta_3 + \mathbf{z}'\gamma) + (\varepsilon_2 - \varepsilon_3 + \zeta).$$

This model is now the same as the earlier model, with correction term $\lambda(\mathbf{x}_1'\widehat{\beta}_1)$ obtained by first-step probit regression of y_1 on \mathbf{x}_1 , where \mathbf{x}_1 denotes the unique regressors in \mathbf{x}_2 , \mathbf{x}_3 , and \mathbf{z} .

If only the intercept varies across the two possible outcomes, by an amount α say, then the Roy model reduces to two latent variables

$$\begin{aligned} y_1^* &= \mathbf{x}_1'\beta_1 + \varepsilon_1, \\ y^* &= \mathbf{x}'\beta + \alpha y_1 + \varepsilon, \end{aligned}$$

where $y = y^*$ is always observed and we also observe the binary variable y_1 equal to one if $y_1^* > 0$ and equal to zero otherwise. This model for y can be viewed as one with **dummy endogenous variable** (y_1). It can be estimated using the Heckman two-step estimator applied to the expression for $E[y^*|\mathbf{x}]$. Alternatively, instrumental variables estimation can be used, provided an instrument for y_1 is available. This requires a regressor that does not determine the level of the outcome of interest but does determine which outcome is chosen.

These Roy models are similar to the models studied in the treatment effects literature. There are two potential outcomes, here y_2^* and y_3^* , but we can only observe one of them. The approach in this chapter has been to create the counterfactual by making strong distributional assumptions on the distribution of unobservables. Chapter 25 presents alternative methods. See especially Section 25.3 for connections between the different approaches.

16.8. Structural Models

Regression models for selected samples have the feature that the outcome of interest depends in part on a participation decision that will in turn depend on expected outcomes. The participation decision and outcomes are simultaneous decisions. The preceding presentations simplified this interdependence by giving a **reduced-form** version of the participation equation. In particular, see the exposition of Lee (1978) in Section 16.7.2. This is a valid approach though is less efficient than working with a fully structural version.

In this section we explicitly model the interdependence using **structural** economic models based on utility maximization, and using structural statistical models that extend linear simultaneous equations to cover censoring and truncation, including binary outcomes.

16.8.1. Structural Models Based on Utility Maximization

Initial **structural model** research considered female **labor supply**. The textbook model has consumers maximizing utility, a function of goods consumption and leisure time, subject to a budget constraint and a time constraint that available discretionary time be allocated between leisure time and working time. At an interior solution the marginal rate of substitution (MRS) between leisure and goods consumption equals the wage rate. However, a corner solution where the woman chooses not to work can arise if the MRS exceeds the offered wage. Gronau (1973) and Heckman (1974) presented econometric models consistent with utility maximization that led to Tobit-like models, accounting for the additional complication that the offered wage is not observed for women who do not work. Subsequent advances include incorporation of fixed costs of work, leading to sample selection models, and use of panel data, leading to panel Tobit models. Killingsworth and Heckman (1986) and Blundell and MacCurdy (2001) provide surveys and Mroz (1987) provides an application.

To illustrate the structural approach we summarize the following example. Dubin and McFadden (1984) modelled household **consumption of energy** (electricity or natural gas) and **choice of appliances** (such as electric heater or natural gas heater) as being interrelated decisions coming from the **same utility function**. Specifically, it is assumed that for the j th of m appliance portfolios household **indirect utility** is given by

$$V_j = \{\alpha_0 + \alpha_1/\beta + \alpha_1 p_1 + \alpha_2 p_2 + \mathbf{w}'\gamma + \beta(y - r_j) + \eta\}e^{-\beta p_1} + \varepsilon_j, \quad (16.49)$$

where p_1 and p_2 denote the prices of electricity and gas, y denotes income, and r_j denotes the annualized total life-cycle cost of portfolio j with

$$r_j = p_1 q_{1j} + p_2 q_{2j} + \rho c_j,$$

where q_{1j} and q_{2j} denote the typical electricity and gas consumption by household with appliance portfolio j , c_j is the cost of appliance portfolio j , and ρ is the discount rate. Tastes differ across households owing to observable characteristics \mathbf{w} , unobservable error η , and an appliance portfolio specific error ε_j , which is assumed to be

independent over j but correlated with η . In addition, there is a common appliance specific taste factor α_{0j} .

Electricity demand x_1 given appliance portfolio j equals $-(\partial V_j / \partial p_1) / (\partial V_j / \partial y)$, by **Roy's identity**, yielding

$$x_1 - q_{1j} = \alpha_{0j} + \alpha_1 p_1 + \alpha_2 p_2 + \mathbf{w}' \boldsymbol{\gamma} + \beta(y - r_j) + \eta.$$

To emphasize that choice of appliance portfolio j is **endogenous**, introduce m mutually exclusive indicator variables δ_{jk} , $k = 1, \dots, m$, where

$$\delta_{jk} = \begin{cases} 1 & \text{if } k = j \\ 0 & \text{if } k \neq j. \end{cases}$$

Then electricity demand x_1 given appliance portfolio j is given by

$$x_1 - q_{1j} = \sum_{k=1}^m \alpha_{0k} \delta_{jk} + \alpha_1 p_1 + \alpha_2 p_2 + \mathbf{w}' \boldsymbol{\gamma} + \beta \left(y - \sum_{k=1}^m r_j \delta_{jk} \right) + \eta. \quad (16.50)$$

Even though the model (16.50) is linear, OLS regression yields inconsistent estimates as the result of endogeneity of δ_{jk} . Dubin and McFadden (1984) present two alternative estimation procedures.

An **IV approach** estimates (16.50) using \hat{p}_k and $r_j \hat{p}_k$ as instruments for δ_{jk} and $r_j \delta_{jk}$, $k = 1, \dots, m$, where \hat{p}_k are the predicted probabilities of choosing the various appliance portfolios. Here V_j is being used to denote the indirect utility function. It includes both deterministic and stochastic components of utility and corresponds to U_j in the Section 15.5.1 presentation of the ARUM. A similar approach yields

$$\begin{aligned} p_k &= \Pr[V_k > V_l, l \neq k, l = 1, \dots, m] \\ &= \Pr[\varepsilon_l - \varepsilon_k < \{(\alpha_{0k} - \alpha_{0l}) - \beta(r_k - r_l)\} e^{-\beta p_1}, \text{ all } l \neq k] \\ &= \frac{\exp[(\alpha_{0k} - \beta r_k) e^{-\beta p_1} \pi / \lambda \sqrt{3}]}{\sum_{l=1}^m \exp[(\alpha_{0l} - \beta r_l) e^{-\beta p_1} \pi / \lambda \sqrt{3}]}, \end{aligned}$$

under the assumption that the ε_j , $j = 1, \dots, m$, are iid type II extreme value with cdf $F(\varepsilon) = \exp(-\exp(-\gamma - \varepsilon\pi/\lambda\sqrt{3}))$, where $\gamma \approx 0.5772$ is Euler's constant. Note that here ε_j has mean zero and variance $\lambda^2/2$ that differ from those for the parameterization of the type II extreme value distribution used in Chapters 14 and 15. Estimation of a nonlinear multinomial logit model gives predicted probabilities \hat{p}_k .

An alternative **sample selection approach** notes that $E[\eta | \text{portfolio } j \text{ chosen}] \neq 0$ and uses assumptions on the distribution of η and $\varepsilon_1, \dots, \varepsilon_m$ to obtain this expectation. Specifically, assume that $\eta | \varepsilon_1, \dots, \varepsilon_m$ is iid with mean $(\sqrt{2}\sigma/\lambda) \sum_{k=1}^m R_k \varepsilon_k$ and variance $\sigma^2(1 - \sum_{k=1}^m R_k^2)$, where $\sum_{k=1}^m R_k = 0$ and $\sum_{k=1}^m R_k^2 < 1$ and the distribution of ε_k has already been given. Then performing some algebra given in Dubin and McFadden yields

$$E[\eta | \text{portfolio } j \text{ chosen}] = \sum_{k \neq j}^m (\sigma \sqrt{6} R_k / \pi) \left[\frac{p_k \ln p_k}{1 - p_k} + \ln p_k \right].$$

A Heckman two-step procedure then estimates by OLS

$$x_1 - q_{1j} = \sum_{k=1}^m \alpha_{0k} \delta_{jk} + \alpha_1 p_1 + \alpha_2 p_2 + \mathbf{w}' \boldsymbol{\gamma} + \beta \left(y - \sum_{k=1}^m r_j \delta_{jk} \right) \\ + \sum_{k \neq j}^m \gamma_k \left[\frac{\hat{p}_k \ln \hat{p}_k}{1 - \hat{p}_k} + \ln \hat{p}_k \right] + \xi,$$

where p_k are predicted probabilities from the preceding model for p_k , and ξ is an error with asymptotic mean zero.

Dubin and McFadden estimated these models using data on 3,249 households with two possible appliance portfolios: electric for water and space heating and gas for water and space heating.

Related examples include those of Hanemann (1984), who modeled the consumption level of a branded good where consumers consume only one of the possible branded goods in the choice set, and of Cameron et al. (1988), who modeled health service demand conditional on choice of one of a number of mutually exclusive health insurance policies.

Much creativity, evident in the Dubin and McFadden example, can be required to specify a model that yields analytical solutions for both choice probabilities and demand conditional on choice. The advances in computational methods detailed in Chapters 12 and 13 permit estimation of such models even when analytical solutions are not obtained. Nonetheless, results will still be dependent on the assumed utility function and distribution of unobservables.

16.8.2. Simultaneous Equations Tobit and Probit Models

To illustrate the issues involved in extending the linear SEM approach of Section 2.4 we consider a selection model that depends on two latent variables and introduce **simultaneity** into the models for the latent variables. A quite general model is

$$y_1^* = \alpha_1 y_2^* + \gamma_1 y_1 + \delta_1 y_2 + \mathbf{x}'_1 \boldsymbol{\beta}_1 + \varepsilon_1, \quad (16.51)$$

$$y_2^* = \alpha_2 y_1^* + \gamma_2 y_1 + \delta_2 y_2 + \mathbf{x}'_2 \boldsymbol{\beta}_2 + \varepsilon_2,$$

where y_1^* and y_2^* are not completely observed but do determine the observed variables y_1 and y_2 , and the errors are assumed to be joint normally distributed. For example, we may observe the binary indicator $y_1 = 1$ if $y_1^* > 0$ and observe $y_2 = y_2^*$ if $y_1^* > 0$. Note that in principal either latent variables or observed outcomes or both may appear as regressors, though identification requires restrictions such as those given in the following.

Endogenous Latent Variables

It is simplest to permit only the latent variables to be regressors in (16.51). Then

$$y_1^* = \alpha_1 y_2^* + \mathbf{x}'_1 \boldsymbol{\beta}_1 + \varepsilon_1, \quad (16.52)$$

$$y_2^* = \alpha_2 y_1^* + \mathbf{x}'_2 \boldsymbol{\beta}_2 + \varepsilon_2.$$

The bivariate sample selection model (16.31) is an example that additionally specifies $\alpha_2 = 0$ and directly specifies a reduced form rather than a structural form for the y_1^* equation. Model (16.52) is easily estimated because the reduced form for y_1^* and y_2^* can be obtained in exactly the same way as for regular linear simultaneous equations. This reduced form can then be estimated using methods such as probit or Tobit depending on the way that y_1 and y_2 are determined given y_1^* and y_2^* . The parameters of the structural model (16.52) can then be estimated by replacing the regressors y_2^* and y_1^* by the reduced-form predictions \hat{y}_2^* and \hat{y}_1^* .

Models such as (16.52) are called **simultaneous equations Tobit models**. A **simultaneous equations probit model** arises if the observed dependent variables y_1 and y_2 are binary. Estimators are presented by Nelson and Olson (1978), Amemiya (1979), and Lee, Maddala, and Trost (1980) and a very general treatment for a range of models is given in L-F. Lee (1981). The standard errors of the estimators can be obtained using the results on sequential two-step m-estimators in Section 6.6. However, it is much simpler to obtain them using the bootstrap pairs procedure presented in Section 11.2. Identification requires exclusion restrictions in (16.51) similar to those for linear simultaneous equations.

Endogenous Regressors

A common specialization of the model (16.52) is to a Tobit model with **endogenous regressor** that is **completely observed**. Then y_2^* is fully observed, so $y_2 = y_2^*$, whereas we observe $y_1 = y_1^*$ if $y_1^* > 0$ and $y_1 = 0$ otherwise. The model becomes

$$\begin{aligned} y_1^* &= \alpha_1 y_2 + \mathbf{x}'_1 \boldsymbol{\beta}_1 + \varepsilon_1, \\ y_2 &= \mathbf{x}' \boldsymbol{\pi} + v, \end{aligned} \tag{16.53}$$

where the first equation is the structural equation of interest and the second equation is the reduced form for the endogenous regressor y_2 . Again note that here y_2 is continuous, not discrete. For joint normal errors $\varepsilon_1 = \gamma v + \xi$, where ξ is an independent normal error (see Section 5.1), so $y_1^* = \alpha_1 y_2 + \mathbf{x}'_1 \boldsymbol{\beta}_1 + \gamma v + \xi$.

A two-step estimation procedure calculates predicted residuals $\hat{v} = y_2 - \mathbf{x}' \hat{\boldsymbol{\pi}}$ from OLS regression of y_2 on \mathbf{x} and then obtains Tobit estimates from the model

$$y_1^* = \alpha_1 y_2 + \mathbf{x}'_1 \boldsymbol{\beta}_1 + \gamma \hat{v} + e_1,$$

where the error e_1 is normally distributed. A test for endogeneity of y_2 can be implemented as a Wald test of $\gamma = 0$ using the usual standard errors from a Tobit package. This test is an extension of the auxiliary regression to implement the Hausman endogeneity test in the linear model (see Section 8.4.3). If the null hypothesis is rejected then the aforementioned second-step Tobit regression yields consistent estimates of α_1 and $\boldsymbol{\beta}_1$, but standard errors then need to be adjusted because of first-step estimation of the additional regressor \hat{v} . See Smith and Blundell (1986) for details for the Tobit model and Rivers and Vuong (1988) for a similar procedure that estimates a probit model at the second step.

Endogenous Censored or Binary Variables

Analysis is more complicated if the observed **censored or binary endogenous variables** y_1 or y_2 appear as regressors in (16.51). Heckman (1978) considered the following model:

$$\begin{aligned} y_1^* &= \gamma_1 y_1 + \delta_1 y_2^* + \mathbf{x}'_1 \boldsymbol{\beta}_1 + \varepsilon_1, \\ y_2^* &= \alpha_2 y_1^* + \gamma_2 y_1 + \mathbf{x}'_2 \boldsymbol{\beta}_2 + \varepsilon_2, \end{aligned} \quad (16.54)$$

where we observe $y_1 = 1$ if $y_1^* > 0$ and $y_1 = 0$ if $y_1^* \leq 0$, and we observe $y_2 = y_2^*$ all the time. The complication here is that y_1 appears as a regressor. A meaningful reduced form for y_1^* can depend only on \mathbf{x}_1 and \mathbf{x}_2 and not y_1 . This imposes the restriction that $\delta_1 \gamma_2 + \gamma_1 = 0$, an example of what is called a **coherency condition** in this literature. Then the reduced form of the model becomes

$$\begin{aligned} y_1^* &= \mathbf{x}' \boldsymbol{\pi}_1 + v_1, \\ y_2 &= \gamma_2 y_1 + \mathbf{x}' \boldsymbol{\pi}_2 + v_2. \end{aligned}$$

This is a special case of the Roy model where participation ($y_1 = 1$) leads to only an intercept shift (via γ_2) in the outcome. In general, models with regressors that include censored or truncated endogenous variables are difficult to estimate. See, for example, Blundell and Smith (1989).

Example

Brooks, Cameron, and Carter (1998) applied a simultaneous equations Tobit model to explain the vote by congressional representatives on a pro-sugar amendment. The three observed outcomes y_1 , y_2 , and y_3 were, respectively, the vote (yes or no) and contributions to their campaign funds from sugar interests and (opposing) sweetener-user interests. The first outcome is a binary outcome and the other two outcomes are censored at zero. A simultaneous equations model for the associated latent variables y_1^* , y_2^* , and y_3^* was specified, so the structural model is of the simpler form (16.52).

How reasonable is this specification? Here campaign contributions y_2^* and y_3^* should depend on the latent variable y_1^* since the actual vote y_1 was made at a later date. For y_1^* however, an alternative and more difficult model is that y_1^* , the latent variable for the vote, depends on actual contributions received (y_2 and y_3) rather than on the latent contributions. However, if this is viewed as a game likely to be repeated in the future, a case can be made for using y_2^* and y_3^* . Clearly, the reasonableness of such assumptions will vary with the application. Parameter identification was secured by exclusion restrictions on the exogenous regressors. Consistent estimation relies on errors being joint normally distributed.

16.9. Semiparametric Estimation

Censoring, truncation, and sample selection lead to a sample that differs from the population. This is essentially a missing data problem, one that is complicated because data are missing on the dependent variable(s) rather than on exogenous regressors.

The preceding methods solved this missing data problem by making distributional assumptions to obtain either a likelihood function for the sample data or an appropriate censored, truncated, or selected conditional mean.

These methods are fragile to even very minor misspecification of error distributions. For example, both the MLE and the Heckman two-step estimator in the standard Tobit model are inconsistent if errors are normal but heteroskedastic, or if they are homoskedastic but nonnormal. See, for example, Paarsch (1982) and the references therein.

Considerable efforts have been devoted to developing semiparametric estimators that are consistent under weaker distributional assumptions. Before presenting leading examples, however, we note that an alternative is to continue to take a fully parametric approach that is based on richer, more flexible distributional assumptions.

16.9.1. Flexible Parametric Models

For simplicity begin with the classical Tobit model $y_i^* = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i$. The assumption that $\varepsilon_i \sim \mathcal{N}[0, \sigma_i^2]$ can be relaxed in two ways. First, heteroscedasticity can be incorporated through an explicit model $\sigma_i^2 = \exp(\mathbf{z}'_i \boldsymbol{\gamma})$, where now both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ need to be estimated. Second, more flexible distributions than the normal distribution might be used. For example, one might use a squared polynomial expansion of the normal (see Section 9.7.7).

For the bivariate sample selection model a similar approach may be taken, where now a more flexible joint distribution for $(\varepsilon_1, \varepsilon_2)$ is used. Lee (1983) proposed working with transformations $(\varepsilon_1^*, \varepsilon_2^*)$ of $(\varepsilon_1, \varepsilon_2)$ for which the bivariate normality assumption may be more reasonable.

Bayesian methods can also be applied to such models. Chib (1992) considered the censored Tobit model. The latent variables \mathbf{y}^* are introduced as auxiliary variables and the data augmentation approach (see Section 13.7) is used. The Gibbs sampler cycles among (1) the conditional posterior for $\boldsymbol{\beta}|\mathbf{y}, \mathbf{y}^*, \sigma^2$, (2) the conditional posterior for $\sigma^2|\mathbf{y}, \mathbf{y}^*, \boldsymbol{\beta}$, and (3) the posterior for $\mathbf{y}^*|\mathbf{y}, \boldsymbol{\beta}, \sigma^2$.

A **flexible parametric approach** is particularly advantageous for handling censoring, truncation, and sample selection in nonlinear models such as those for counts and for duration data or mixed types of data, as semiparametric methods are less likely to be available then.

16.9.2. Semiparametric Estimation for Censored Models

We now move on to semiparametric estimation. We consider a linear model for the latent variable $y_i^* = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i$, which is left-censored at zero so that we observe $y_i = y_i^*$ if $y_i^* > 0$ and $y_i = 0$ if $y_i^* \leq 0$. The semiparametric literature usually expresses the model as

$$y_i = \max(\mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i, 0). \quad (16.55)$$

This is the Tobit model (16.11)–(16.13), except the distribution of ε is unspecified. With some adaptation this model also covers left-censoring at known fixed point other than zero and to right-censoring such as for top-coded data. For example, if

$y = \min(\mathbf{x}'\beta + \varepsilon, U)$ then $U - y = \max(U - \mathbf{x}'\beta - \varepsilon, 0)$. The goal is to consistently estimate β without specifying a complete parametric distribution for ε_i . The estimators are called **semiparametric** as the uncensored mean $\mathbf{x}_i'\beta$ is parameterized but the error distribution is not. The methods presented in the following differ in the assumptions made on the distribution of ε .

From (16.8) ML estimation is possible given knowledge of the cdf of y^* and hence of ε . The cdf of ε can be nonparametrically estimated using the Kaplan–Meier product limit estimator for the cdf presented in Chapter 17 for the case of right-censored duration data. Alternatively, the distribution of ε can be nonparametrically determined using the series expansion of Gallant and Nychka (1987); see Section 9.7.7. These semiparametric ML estimation methods are rarely implemented.

Instead, the literature focuses on estimation based on conditional moments. From (16.20) the conditional censored mean $E[y|\mathbf{x}]$ is clearly a single-index model with $E[y|\mathbf{x}] = g(\mathbf{x}'\beta)$, where the function $g(\cdot)$ is unknown if the distribution of ε is not specified. The single-index methods of Section 9.7.4 can therefore be applied, though as noted there β can be estimated only up to location and scale.

A more popular approach considers alternative conditional censored moments that are less altered by censoring. Powell (1984) proposed using the **conditional median**. The key distributional assumption is that $\varepsilon|\mathbf{x}$ has median zero, in which case the conditional median of $y|\mathbf{x}$ equals the conditional mean $\mathbf{x}'\beta$. The intuition for Powell's estimator is most easily obtained by supposing y is iid. If less than half the sample is censored, so that less than half of the observations are zero and more than half are positive, then the censored sample median provides a consistent estimate of the population median. Powell (1984) extended this idea to the regression case, where the same logic follows for those observations for which less than half the observations on $\varepsilon|\mathbf{x}$ are censored, where $\varepsilon = y - \mathbf{x}'\beta$ depends on β , which needs to be estimated. The regression analogue of median estimation is LAD estimation (see Section 4.6). This leads to the **censored least absolute deviations (CLAD) estimator** $\hat{\beta}_{\text{CLAD}}$, which minimizes

$$Q_N(\beta) = N^{-1} \sum_{i=1}^N |y_i - \max(\mathbf{x}_i'\beta, 0)|. \quad (16.56)$$

The essential assumption for consistency of this estimator is that $\varepsilon|\mathbf{x}$ has median zero. Given this assumption the estimator is consistent even if errors are conditionally heteroskedastic. The estimator for β is \sqrt{N} -consistent and asymptotically normal. More efficient estimators can be obtained by weighting the terms in sums by $f(0|\mathbf{x}_i)$, the conditional density of $\varepsilon_i|\mathbf{x}_i$ evaluated at zero. The method can also be extended to conditional quantiles.

An alternative procedure uses a **symmetrically trimmed mean**, rather than the median, that is also unaffected by censoring. Assume that the distribution of $\varepsilon|\mathbf{x}$ is symmetrically distributed. This implies that for observations with positive mean (i.e., $\mathbf{x}'\beta > 0$) $y|\mathbf{x}$ is symmetrically distributed on the interval $(0, 2\mathbf{x}'\beta)$. Then either $\mathbf{x}'\beta + \varepsilon < 0$ and $y = 0$ is observed or, with equal probability, $\mathbf{x}'\beta + \varepsilon > 2\mathbf{x}'\beta$ and the data are artificially set to $2\mathbf{x}'\beta$ to preserve the symmetry about $\mathbf{x}'\beta$. We have shown that

$$E[\mathbf{1}(\mathbf{x}'\beta > 0)[\min(y, 2\mathbf{x}'\beta) - \mathbf{x}'\beta]\mathbf{x}] = \mathbf{0}, \quad (16.57)$$

where $\mathbf{1}(\mathbf{x}'\beta > 0)$ restricts attention to observations with positive mean, and the new dependent variable is $y = 0$, or $0 < y < 2\mathbf{x}'\beta$, or $2\mathbf{x}'\beta$ if $y > 2\mathbf{x}'\beta$. The moment estimator based on (16.57) does not have unique solution for β . Powell (1986b) proposed the **symmetrically censored least squares (SCLS) estimator** that minimizes

$$Q_N(\beta) = N^{-1} \sum_{i=1}^N \{[y_i - \max(y_i/2, \mathbf{x}'_i\beta)]^2 + \mathbf{1}(y_i > 2\mathbf{x}'_i\beta)[y_i^2/4 - \max(0, \mathbf{x}'_i\beta)]^2\}, \quad (16.58)$$

which with some algebra can be shown to yield first-order conditions that are the sample analogue of moment condition (16.57). Chay and Honoré (1998) provide a graphical exposition of the trimming for the SCLS estimator, as well as for the related pairwise difference estimators of Honoré and Powell (1994).

Melenberg and Van Soest (1996), Chay and Honoré (1998), and Chay and Powell (2001) provide applications of some of these estimators. Pagan and Ullah (1999) provide additional methods and theory.

As an empirical example we applied CLAD estimation to the Section 16.2.1 data that were generated from a Tobit model with normal errors. The slope parameter (set to 1000) was estimated to be 956 (standard error 117) using ML compared to 838 (standard error 165) using CLAD. As expected the CLAD robustness to nonnormality comes at the expense of some loss in efficiency.

16.9.3. Semiparametric Estimation for Selection Models

Semiparametric estimation of sample selection models is more challenging. We consider the most commonly studied model, the **bivariate sample selection model** defined in Section 16.5.2, where now we relax the assumption that the errors $(\varepsilon_1, \varepsilon_2)$ are joint normally distributed.

Semiparametric ML estimation is possible. In particular Gallant and Nychka (1987) explicitly considered the bivariate sample selection model as a suitable candidate for their series expansion estimator presented in Section 9.7.7.

The literature instead uses as starting point the expression for the truncated conditional mean, which from (16.34) is given by

$$\begin{aligned} E[y_{2i} | \mathbf{x}_i, y_{1i}^* > 0] &= \mathbf{x}'_{2i}\beta_2 + E[\varepsilon_2 | \varepsilon_1 > -\mathbf{x}'_{1i}\beta_1] \\ &= \mathbf{x}'_{2i}\beta_2 + g(\mathbf{x}'_{1i}\beta_1), \end{aligned} \quad (16.59)$$

where the second equality assumes that $\varepsilon_{2i} | \mathbf{x}_i, \varepsilon_{1i}$ has distribution that depends on just \mathbf{x}_{1i} similar to assumption (16.41). The distribution of $(\varepsilon_1, \varepsilon_2)$ is left unspecified so the function $g(\cdot)$ is unknown, leading to a semiparametric estimation problem. Since it is possible that $g(\mathbf{x}'_1\beta_1) = \mathbf{x}'_1\beta_1$, identification in this model with $g(\cdot)$ unspecified requires an **exclusion restriction** that at least one component of \mathbf{x}_1 does not appear in \mathbf{x}_2 . Moreover, the more uncorrelated $\mathbf{x}'_1\beta_1$ is with \mathbf{x}_2 the better β_2 and $g(\cdot)$ can be distinguished. The model (16.59) is a partially linear model, which can be estimated using methods presented in Section 9.7.3. Popular methods include the Robinson (1988a) differencing estimator and using a series expansion for $g(\mathbf{x}'_1\beta_1)$. Since β_1 is unknown the regression is of y_{2i} on $\mathbf{x}'_{2i}\beta_2 + g(\mathbf{x}'_{1i}\widehat{\beta}_1)$, where $\widehat{\beta}_1$ can be obtained by regression

of the binary outcome y_{1i} on \mathbf{x}_{li} , using one of the semiparametric binary model estimators given in Section 14.7. These methods provide consistent estimates of the slope parameters β_2 . To additionally estimate the intercept, necessary for analysis of the levels rather than changes in y_2 , see Andrews and Schafgens (1998).

Newey, Powell, and Walker (1990) applied this approach to female labor supply. The participation indicator model was estimated using several different methods and the equation for the outcome y_2 was estimated using the method of Robinson (1988a). Melenberg and Van Soest (1996) modeled vacation expenditures using a wide range of semiparametric methods for both the bivariate sample selection and censored regression models. A richer model is provided by Das, Newey and Vella (2003).

Manski (1989) considered **identification** in the bivariate sample selection model under relatively minimal assumptions and provided **bounds** for the mean and for marginal effects, conditional on both regressors and selection.

16.10. Derivations for the Tobit Model

16.10.1. Truncated Moments of Standard Normal

Consider $z \sim \mathcal{N}[0, 1]$, with density $\phi(z) = (1/\sqrt{2\pi}) \exp(-z^2/2)$ and cdf $\Phi(z)$. Since $\Pr[z > c] = 1 - \Phi(c)$, the conditional density of $z|z > c$ is $\phi(z)/(1 - \Phi(c))$. It follows that

$$\begin{aligned} E[z|z > c] &= \int_c^\infty z (\phi(z)/[1 - \Phi(c)]) dz \\ &= \int_c^\infty z (1/\sqrt{2\pi}) \exp(-z^2/2) dz / [1 - \Phi(c)] \\ &= \int_c^\infty \frac{\partial}{\partial z} \left(-(1/\sqrt{2\pi}) \exp(-z^2/2) \right) dz / [1 - \Phi(c)] \\ &= \left[-(1/\sqrt{2\pi}) \exp(-z^2/2) \right]_c^\infty / [1 - \Phi(c)] \\ &= \phi(c)/[1 - \Phi(c)]. \end{aligned}$$

Similarly,

$$\begin{aligned} E[z^2|z > c] &= \int_c^\infty z^2 (\phi(z)/[1 - \Phi(c)]) dz \\ &= \int_c^\infty z \times z \times (1/\sqrt{2\pi}) \exp(-z^2/2) dz / [1 - \Phi(c)] \\ &= \int_c^\infty z \times \frac{\partial}{\partial z} \left(-(1/\sqrt{2\pi}) \exp(-z^2/2) \right) dz / [1 - \Phi(c)] \\ &= \left[z \times (-1/\sqrt{2\pi}) \exp(-z^2/2) \right]_c^\infty / [1 - \Phi(c)] \\ &\quad - \int_c^\infty \frac{\partial}{\partial z} (z) \times \left(-(1/\sqrt{2\pi}) \exp(-z^2/2) \right) dz / [1 - \Phi(c)] \\ &= c\phi(c)/[1 - \Phi(c)] + (1 - \Phi(c))/[1 - \Phi(c)] \\ &= c\phi(c)/[1 - \Phi(c)] + 1. \end{aligned}$$

It follows after a little algebra that

$$\begin{aligned} \text{V}[z|z > c] &= \text{E}[z^2|z > c] - (\text{E}[z|z > c])^2 \\ &= 1 + c\phi(c)/[1 - \Phi(c)] - \phi(c)^2/[1 - \Phi(c)]^2. \end{aligned}$$

16.10.2. Asymptotic Theory for Heckman's Two-Step Estimator in the Tobit Model

The asymptotic variance matrix of the two-step Heckman estimator is complicated by its dependence on first-step parameter estimates. There are several ways to obtain the asymptotic variance, such as that in Amemiya (1985, pp. 369–370). Here we instead apply the general result for sequential two-step m-estimators given in Section 6.6. We consider the simplest case of the Tobit model (see Section 16.3.6). The methods can be adapted to two-step estimators for the bivariate sample selection model (Section 16.5.4) and simultaneous equations Tobit model (Section 16.8.2). A much simpler quite different approach is to use the bootstrap pairs procedure (see Section 11.2).

From (16.26) we wish to estimate the parameters $\gamma = [\beta' \quad \sigma']'$ in the equation for positive y_i :

$$\begin{aligned} y_i &= \mathbf{x}'_i \beta + \sigma \lambda(\mathbf{x}'_i \alpha) + \eta_i \\ &= \mathbf{w}_i(\alpha)' \gamma + \eta_i, \end{aligned}$$

where $\mathbf{w}_i(\alpha) = [\mathbf{x}'_i \quad \lambda(\mathbf{x}'_i \alpha)]'$ and $\eta_i = y_i - \mathbf{x}'_i \beta - \sigma \lambda(\mathbf{x}'_i \alpha)$ is heteroskedastic with variance $\sigma_{\eta_i}^2$ defined in (16.24). The first step of the two-step procedure is to obtain an estimate $\hat{\alpha}$ of the unknown parameter α by probit MLE. It follows that the normal equations for the two parts of the Heckman two-step estimator are

$$\begin{aligned} \sum_{i=1}^N (y_i - \Phi(\mathbf{x}'_i \alpha)) \frac{\phi^2(\mathbf{x}'_i \alpha)}{\Phi(\mathbf{x}'_i \alpha)(1 - \Phi(\mathbf{x}'_i \alpha))} \mathbf{x}_i &= \mathbf{0}, \\ - \sum_{i=1}^N d_i \mathbf{w}_i(\alpha) (y_i - \mathbf{w}_i(\alpha)' \gamma) &= \mathbf{0}, \end{aligned} \tag{16.60}$$

where the first equation gives the probit first-order conditions for α , and the second equation gives first-order conditions for γ for OLS on positive y_i ($d_i = 1$).

These equations can be combined as $\sum_{i=1}^N \mathbf{h}(\mathbf{x}_i, \theta) = \mathbf{0}$ where $\theta = (\alpha', \gamma')'$. By the usual first-order Taylor series expansion $\hat{\gamma} - \gamma \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{G}_0^{-1} \mathbf{S}_0 (\mathbf{G}_0^{-1})']$ where $\mathbf{G}_0 = \lim N^{-1} \mathbb{E}[\sum_{i=1}^N \partial \mathbf{h}(\mathbf{x}_i, \theta) / \partial \theta]$ and $\mathbf{S}_0 = \lim N^{-1} \mathbb{E}[\sum_{i=1}^N \mathbf{h}(\mathbf{x}_i, \theta) \mathbf{h}(\mathbf{x}_i, \theta)']$. We are interested in the subcomponent corresponding to γ . Simplification occurs because $\partial \mathbf{h}(\mathbf{x}_i, \theta) / \partial \theta$ is block triangular because γ does not appear in the first set of equations. Partitioning yields the general result

$$\text{V}[\hat{\theta}_2] = \mathbf{G}_{22}^{-1} \{ \mathbf{S}_{22} + \mathbf{G}_{21} [\mathbf{G}_{11}^{-1} \mathbf{S}_{11} \mathbf{G}_{11}^{-1}] \mathbf{G}'_{21} - \mathbf{G}_{21} \mathbf{G}_{11}^{-1} \mathbf{S}_{12} - \mathbf{S}_{21} \mathbf{G}_{11}^{-1} \mathbf{G}'_{21} \} \mathbf{G}_{22}^{-1},$$

where the matrices are defined in Section 6.6.

Specializing to the problem here, we first consider the terms in \mathbf{G}_0 . Then

$$\begin{aligned}\mathbf{G}_{11} &= \lim \frac{1}{N} \sum_{i=1}^N \frac{\phi^2(\mathbf{x}'_i \boldsymbol{\alpha})}{\Phi(\mathbf{x}'_i \boldsymbol{\alpha})(1 - \Phi(\mathbf{x}'_i \boldsymbol{\alpha}))} \mathbf{x}_i \mathbf{x}'_i, \\ \mathbf{G}_{21} &= \lim \frac{1}{N} \sum_{i=1}^N d_i \mathbf{w}_i \frac{\partial \lambda(\mathbf{x}'_i \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}, \\ \mathbf{G}_{22} &= \lim \frac{1}{N} \sum_{i=1}^N \mathbb{E}[d_i \mathbf{w}_i \mathbf{w}'_i].\end{aligned}$$

The expression for \mathbf{G}_{11} uses knowledge that \mathbf{G}_{11}^{-1} is just the variance of the probit MLE. The expression for \mathbf{G}_{21} uses

$$\begin{aligned}\mathbb{E}\left[\frac{\partial \mathbf{h}_{2i}}{\partial \boldsymbol{\theta}'_1}\right] &= \mathbb{E}\left[-\frac{\partial d_i \mathbf{w}_i(\boldsymbol{\alpha})(y_i - \mathbf{w}_i(\boldsymbol{\alpha})' \boldsymbol{\gamma})}{\partial \boldsymbol{\alpha}}\right] \\ &= \mathbb{E}\left[\mathbf{w}_i \frac{\partial d_i \mathbf{w}_i(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}'}\right] \\ &= \mathbb{E}\left[d_i \mathbf{w}_i \frac{\partial \lambda(\mathbf{x}'_i \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\right].\end{aligned}$$

The expression for \mathbf{G}_{22} uses

$$\frac{\partial \mathbf{h}_{2i}}{\partial \boldsymbol{\theta}'_2} = \frac{\partial d_i \mathbf{w}_i(\boldsymbol{\alpha})(y_i - \mathbf{w}_i(\boldsymbol{\alpha})' \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} = d_i \mathbf{w}_i \mathbf{w}'_i.$$

Turning to \mathbf{S}_0 we have

$$\begin{aligned}\mathbf{S}_{11} &= \mathbf{G}_{11}^{-1}, \\ \mathbf{S}_{21} &= \mathbf{0}, \\ \mathbf{S}_{22} &= \lim \frac{1}{N} \sum_{i=1}^N \mathbb{E}[d_i (y_i - \mathbf{w}_i(\boldsymbol{\alpha})' \boldsymbol{\gamma})^2].\end{aligned}$$

The expression for \mathbf{S}_{11} follows by applying the information matrix equality. Taking expectations and some manipulation leads to $\mathbf{S}_{21} = \mathbf{0}$, and \mathbf{S}_{22} is simply $\mathbb{V}[\eta_i]$.

Combining these results gives the Heckman two-step estimator $\hat{\boldsymbol{\gamma}} \stackrel{d}{\sim} \mathcal{N}(\boldsymbol{\gamma}, \mathbb{V}_{\boldsymbol{\gamma}})$, where

$$\hat{\mathbf{V}}_{\boldsymbol{\gamma}} = (\hat{\mathbf{W}}' \hat{\mathbf{W}})^{-1} (\hat{\mathbf{W}}' \Sigma_{\hat{\eta}} \hat{\mathbf{W}} + \hat{\mathbf{W}}' \hat{\mathbf{D}} \hat{\mathbf{V}}_{\boldsymbol{\alpha}} \hat{\mathbf{D}} \hat{\mathbf{W}}) (\hat{\mathbf{W}}' \hat{\mathbf{W}})^{-1}, \quad (16.61)$$

and where $\hat{\mathbf{W}}' \hat{\mathbf{W}} = \sum_{i=1}^N d_i \hat{\mathbf{w}}_i \hat{\mathbf{w}}'_i$, $\hat{\mathbf{D}} = \text{Diag}[\partial \lambda(\mathbf{x}'_i \boldsymbol{\alpha}) / \partial \boldsymbol{\alpha} \mid \hat{\boldsymbol{\alpha}}]$, $\hat{\mathbf{V}}_{\boldsymbol{\alpha}}$ is the variance matrix for the first-stage probit MLE, and $\Sigma_{\hat{\eta}}$ is a diagonal matrix with i th entry $\hat{\sigma}_{\eta_i}^2$. This estimate is straightforward to obtain if matrix commands are available. The hardest part can be analytically obtaining $\sigma_{\eta_i}^2 = \mathbb{V}[\eta_i]$ given in (16.24). If this is difficult we can instead use $\hat{\sigma}_i^2 = (y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}} + \hat{\sigma} \lambda_i(\mathbf{x}'_i \hat{\boldsymbol{\alpha}}))^2$ following the approach of White (1980).

16.11. Practical Considerations

Most major packages include ML estimation of the Tobit model under normality. The two-part model is easy to estimate as one can separately estimate each part. In principle

the bivariate sample selection model can be estimated by Heckman's two-step procedure using only a probit and OLS routine. However, the standard errors are difficult to compute owing to the two-step nature of the estimator, and it is much easier to obtain standard errors using a package with Heckman's two-step procedure built-in. Implementing semiparametric estimators generally requires specialized code in a programming language such as GAUSS. Some packages also permit ML estimation of censored and truncated variants of other models, such as the Poisson and negative binomial for count data.

Censoring and truncation are easily handled if one views as reasonable the specified distribution. For example, top-coded income data are easily handled if the log-normal distribution fits the data well. Censored LAD, which relies on much weaker distributional assumptions, can also be used in this situation.

Much more problematic is handling models with sample selection. The more parametric versions of these models can rely on distributional assumptions that are felt to be strong. Semiparametric versions still have to struggle with the identification requirement that a variable that determines participation does not also determine the outcome of interest. A more promising route, one often taken in the treatment effects literature, is to limit attention to cases where it may be reasonable to assume that selection is only on observables.

16.12. Bibliographic Notes

The literature on models from selected samples is vast. Book-length treatments are provided by Maddala (1983) and Gouriéroux (2000), and shorter summaries are provided by Amemiya (1984, 1985) and Greene (2003).

- 16.3** Tobit (1958) proposed and applied the Tobit model to expenditure data. Amemiya (1973) formally established its consistency and asymptotic normality. Heckman (1974) provides an excellent female labor supply application with detailed analysis of results.
- 16.4** The many studies of the Rand Health Insurance Experiment, such as that by Duan et al. (1983), are leading applications of the two-part model.
- 16.5** Heckman (1976, 1979) presented the two-step estimator of the bivariate sample selection model that is also the basis for many more recent semiparametric estimation procedures. Mroz (1987) provides an excellent application to female labor supply that places emphasis on the role of assumptions on wage exogeneity.
- 16.7** There are many variants on the ideas of Roy (1951), just as there are many variants of the Tobit model. L-F. Lee (1978) provides a good early application to the union–nonunion wage differential.
- 16.8** The work by Dubin and McFadden (1984) is a leading example of structural microeconometric analysis based on complete specification of utility function and distribution of unobservables.
- 16.9** Semiparametric estimation of binary choice models is presented in detail in the books by M-J. Lee (1996), Horowitz (1997), and Pagan and Ullah (1999) and in surveys by

Vella (1998) and L-F. Lee (2001). Chay and Honoré (1998) and Chay and Powell (2001) provide applications for censored models, and Melenberg and Van Soest (1996) additionally estimate bivariate sample selection models.

Exercises

16–1 This question considers the impact of different degrees of truncation in the Tobit model.

- (a) Generate 200 draws of a latent variable $y^* = k + 3x + u$, where $u \sim \mathcal{N}[0, 3]$ and the regressor $x \sim \text{uniform}[0, 1]$. Choose k such that you generate approximately 30% of y^* to be negative.
- (b) Generate a censored or truncated subsample by excluding observations that correspond to $y^* < 0$.
- (c) Estimate the model using all 2,000 observations, as if the latent variable were observable, by OLS. Evaluate your results in the light of the theoretical properties of OLS, keeping in mind that you have only one replication.
- (d) Using the truncated subsample of $y > 0$ only, estimate the model by OLS.
- (e) Use the truncated maximum likelihood option to estimate the parameters using all observations. Evaluate your results in light of the properties of the truncated MLE. Compare with the least-squares results from the previous two parts.
- (f) Repeat all previous steps using a value of k so as to generate 20, 40, and 50% censored observations. Compare your results with those based on 30% censored observations. Hence suggest what is the consequence on the parameter estimates of higher levels of censoring. Reinforce your arguments using theory where possible.

16–2 Consider a latent variable modeled by $y_i^* = \mathbf{x}_i'\beta + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}[0, \sigma^2]$. Suppose y_i^* is censored from above so that we observe $y_i = y_i^*$ if $y_i^* < U_i$ and $y_i = U_i$ if $y_i^* \geq U_i$, where the upper limit U_i is a known constant for each individual (i.e., data) and may differ over individuals.

- (a) Give the log-likelihood function for this model. [Hint: Note that this differs from the standard case both owing to presence of U_i and because the equalities are reversed with $y_i = y_i^*$ if $y_i^* < U_i$.]
- (b) Obtain the expression for the truncated mean $E[y_i|\mathbf{x}_i, y_i < U_i]$. [Hint: For $z \sim \mathcal{N}[0, 1]$, we have $E[z|z > c] = \phi(c)/(1 - \Phi(c))$. Also, $E[z|z < c] = -E[-z] - z > -c$ and $-z \sim \mathcal{N}[0, 1]$.]
- (c) Hence give Heckman's two-step estimator for this model.
- (d) Obtain the expression for the censored mean $E[y_i|\mathbf{x}_i]$. [Hint: An essential part is the answer in part (b).]

16–3 This question considers the consequences of misspecification in the Tobit model. The starting point is the model of Exercise 16.1.

- (a) Generate y^* with heteroskedasticity by letting $u \sim \mathcal{N}[0, \sigma^2 z]$, where $z > 0$ is chosen to be a suitable positive-valued variable that is correlated with x , though not perfectly so. Again set k to obtain about 30% of censored observations. Use the MLE for censored normal to estimate this model and compare your results with the corresponding homoskedastic case.

- (b)** Now consider the impact of nonnormality in the sample. Use the simulation macro available in some packages to carry out a Monte Carlo evaluation based on a sample of 1,000 observations and 500 replications. In each replication generate a sample with censored observations such that the errors are drawn from a mixture of two normals: $\mathcal{N}[1, 9]$ or $\mathcal{N}[0.4, 1]$ with probabilities 0.4 and 0.6, respectively. Estimate the model using the censored Tobit MLE and compare your results with the normal case. Carry out an analysis of the Monte Carlo output for the two estimators. Draw appropriate conclusions about the impact of nonnormality on the distribution of the Tobit estimator.

16–4 Consider a Poisson regression model where y^* has density $f^*(y^*) = e^{-\mu} \mu^{y^*} / y^*!$, $y_i^* = 0, 1, 2, \dots$, and we have independence over i . Because of coding error we only fully observe y^* when $y^* \geq 2$. When $y^* = 0$ or 1 we only observe that $y^* \leq 1$. Suppose this is coded as $y^* = 1$. Define the observed data $y = y^*$ for $y_i^* \geq 2$ and $y = 1$ for $y_i^* = 0$ or 1.

- (a)** Obtain the density $f(y)$ of the observed y .
- (b)** Obtain $E[y]$. [There is some algebra here.]
Now introduce regressors with $E[y^* | \mathbf{x}] = \exp(\mathbf{x}'\beta)$ and define the indicator variable $d = 1$ for $y^* \geq 2$ and $d = 0$ for $y^* = 0$ or 1.
- (c)** Give the exact formula for this example of the objective function of an estimator that provides a consistent estimator of β using data on y_i , d_i , and \mathbf{x}_i .
- (d)** Give the exact formula for this example of the objective function of an estimator that provides a consistent estimator of β using data on only d_i and \mathbf{x}_i .
- (e)** Is it possible to consistently estimate β using data on only d_i and \mathbf{x}_i ? Explain your answer.

16–5 Using a 50% random subsample of the RAND data on medical expenditure over a 12-month period used in this chapter, and using a similar model specification, we wish to consider the following broad question: Which model is appropriate for modeling the expenditure data?

- (a)** Using the data summary of the expenditure variable, analyze the implications of the high proportion of zero expenditures observed. Is this a violation of the normality assumption? Is there a transformation of expenditure that would make the assumption of normality more appropriate?
- (b)** Three candidate models are considered, each with the same set of covariates. These covariates are the same as in the count data Exercise 20.6. The models are (i) the Tobit model, (ii) the two-part (“hurdle”) model (TPM), and (iii) the selection model. Explain how each one of these will be set up, the relationship and connections among them, and how one might compare and choose among them. If you are likely to encounter any specific specifications or estimation problems, state them and suggest how you might handle them. Pay attention to the choice of exclusion restrictions.
- (c)** Estimate in turn the Tobit model, the TPM, and the selection models. For the TPM you have two equations, and the second is for those who have positive expenditures only. In the case of the selection model, use both the MLE and the two-step (Heckman) estimators. Discuss your reasons underlying

the exclusion restriction required in the estimation of the selection model. Is there evidence that the selection problem is a serious issue?

- (d) How can we compare the statistical fit of the three models? Which model appears to provide the best fit to the data? By what criterion?
- (e) Suppose our main interest is in the impact of two variables on expenditure, log income, and log of $(1 + \text{coinsurance rate})$. Use the results of your estimated Tobit model and TPM to make a comparison between the marginal impact of a change in these variables on expenditure. Given that there is considerable heterogeneity in the sample, suggest how to present the results of your analysis in the most informative manner.
- (f) Briefly explain how quantile regression (see Section 4.6) provides an alternative method of analyzing the same data. What are the main advantages and disadvantages of this approach in the present data situation?

Transition Data: Survival Analysis

17.1. Introduction

Econometric models of durations are models of the length of time spent in a given state before transition to another state, such as duration unemployed or alive or without health insurance. In biostatistics a duration in a state is also known as **lifetime** and the time of transition is referred to as **death**; in operations research where one often studies lifetimes of physical objects such as light bulbs and machines, the end of useful life, that is, transition to useless life, is called **failure time**. In econometrics a **state** is a classification of an individual entity at a point in time, **transition** is movement from one state to another, and a **spell** length or **duration** is the time spent in a given state. A typical regression example is determining the effect of higher unemployment benefit levels on the average length of an unemployment spell or the probability of transition out of unemployment.

The literature on this subject can be quite daunting, for a number of reasons. First, several related distributional functions are of interest and either the duration or probability of transition may be modeled. Second, many different sampling schemes are possible and statistical inference depends on both the duration model and the sampling scheme. For example, sampling methods for data on unemployment duration include **flow sampling** of those entering unemployment in a given month, **stock sampling** of people unemployed in a given month, and population sampling of all people regardless of employment status. Third, the data on spell duration are often censored. This is a major reason for modeling transitions rather than the mean duration, the usual object of regression analysis, as weaker distributional assumptions are needed to consistently estimate models of the transitions. Fourth, transition data can be very rich with several states, such as unemployment, part-time employment, full-time employment, and out-of-the labor force, and data for a given individual may be available on multiple transitions among these states. Fifth, the literature appears in several different applied areas of statistics with different emphases. **Duration analysis** or **transition analysis** is also called **survival analysis** (length of time survived) in biostatistics, **failure time analysis** (length of time to failure of an item such as a light bulb or a machine part)

in operations research, **life table analysis** in demography and actuarial studies (where leaving a state corresponds to death), and **hazard analysis** in insurance and accident theory. In the social sciences applications include recidivism, length of marriages, and interelection duration.

In this chapter we present results for **single-spell duration** data obtained by **flow sampling**. The classic example is modeling survival time, with transition being from alive to dead, and many of the results come from survival analysis and life table analysis. This is the most studied example of transition analysis in statistics, and the survival analysis methods presented in this chapter are implemented in many statistical and microeconometric packages. The chapter begins with a regression example to outline the issues raised with survival data.

Sections 17.3–17.5 present results without regressors, as many new concepts arise even in this case. Section 17.3 introduces basic duration data concepts such as the hazard, cumulative hazard, and survivor functions. Section 17.4 defines various types of censoring, a common complication in duration analysis because the completed spell is not always observed. For example, a clinical trial will usually end before the last subject dies. Section 17.5 presents nonparametric estimators of the hazard, cumulative hazard (Nelson–Aalen estimator), and survivor functions (Kaplan–Meier estimator) that are consistent under independent censoring.

The remainder of the chapter extends analysis to regression models, again under independent censoring. Estimation of fully parametric models, notably the Weibull model, is presented in Section 17.6. The treatment of censoring is similar to that given for fully parametric Tobit models. Some important duration models are given in Section 17.7. An alternative semiparametric approach is to instead model the hazard function, the probability of death conditional on survival to date. In his seminal paper, Cox (1972) proposed a method to consistently estimate a proportional hazards function with independent censoring under relatively weak distributional assumptions. The Cox model, the standard model for survival data, is presented in Section 17.8. Unlike most cross-section models, in survival models regressors such as unemployment benefits in an unemployment duration model may vary for a given person over the period that the subject is observed. Models with time-varying regressors are detailed in Section 17.9. Discrete hazards models are presented in Section 17.10. Section 17.11 presents an empirical example.

Two subsequent chapters consider more complicated aspects of transition modelling that are rarely given a textbook treatment. These include unobserved heterogeneity, multiple spells, and multiple destinations.

17.2. Example: Duration of Strikes

Consider a data set on the duration of strikes that has been used by Kennan (1985), Jaggia (1991c), and others. The variable of interest is the duration of strikes in U.S.

17.2. EXAMPLE: DURATION OF STRIKES

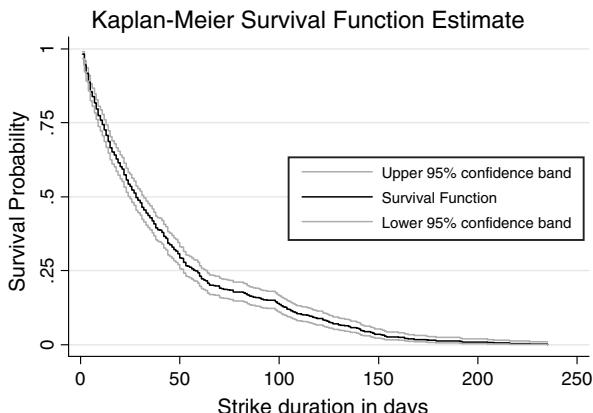


Figure 17.1: Strike duration: Kaplan-Meier estimate of survival function. Data on completed spells for 566 strikes in the U.S. during 1968–76.

manufacturing, measured in number of days from the start of the strike. The sample has 566 complete (uncensored) observations on strike duration. The average duration of strike (dur) is 43.6 days, and the median is about 28 days. However, 90 days after the start of the strike 88 strikes are still in progress.

We can show the strike duration information graphically as an empirical **survival function**. Figure 17.1 shows on the vertical axis the proportion of strikes started that are still in progress after a stated number of days. Calender time is ignored in this figure, meaning that the different start date of different strikes plays no role in the construction of the figure. As expected, the function starts at one and monotonically declines to zero, indicating that all strikes must eventually end.

Now introduce a regressor variable (z) that measures the deviation of output from its trend level, an indicator of the business cycle position of the economy. Positive values of z indicate above-trend growth period and negative values indicate the converse. Suppose that our main interest lies in testing whether average strike duration is pro cyclical (i.e., $\partial(dur)/\partial z > 0$) or anticyclical (i.e., $\partial(dur)/\partial z < 0$). A simple way to proceed might be to model the conditional expectation of $\ln(dur)$ by a linear regression of $\ln(dur)$ on z . This may serve the purpose if one is testing for the presence of a positive or negative association between dur and z .

Possibly we might instead be interested in modeling the conditional probability of a strike. Such a goal could be achieved by a binomial regression with a 0/1 outcome variable. However, suppose that our interest is in modeling the probability that a strike that has been in progress for t days will end on day $t + 1$, or in modeling the conditional probability of the strike in progress ending, as a function of the length of the strike, controlling for z ; then the previously mentioned regression approaches will be less direct and less efficient than survival analysis, which also has the additional advantage that it can handle censored durations. In the next section we will consider statistical concepts that are used in survival analysis.

17.3. Basic Concepts

Duration in a state is a nonnegative random variable, denoted T , which in economic data is often a discrete random variable. For explaining the basic concepts we focus on the continuous case, followed by the discrete case later in the chapter.

17.3.1. Survivor, Hazard, and Cumulative Hazard Functions

The **cumulative distribution function** of T is denoted $F(t)$ and the **density function** is $f(t) = dF(t)/dt$. Then the probability that the duration or spell length is less than t is

$$\begin{aligned} F(t) &= \Pr[T \leq t] \\ &= \int_0^t f(s)ds. \end{aligned} \tag{17.1}$$

A complementary concept to the cdf is the probability that duration equals or exceeds t , called the **survivor function**, which is defined by

$$\begin{aligned} S(t) &= \Pr[T > t] \\ &= 1 - F(t). \end{aligned} \tag{17.2}$$

The definition of the cdf in (17.1) equals the usual definition, following Kalbfleisch and Prentice (2002). In the duration analysis literature other authors, such as Lancaster (1990) instead define $F(t) = \Pr[T < t]$ and hence $S(t) = \Pr[T \geq t]$ because hazard functions, defined below, condition on $T \geq t$ rather than $T > t$. The particular definition used will make a difference in the discrete case, considered in Section 17.3.2, at the exact time that a transition occurs.

The survivor function is monotonically declining from one to zero since the cdf is monotonically increasing from zero. If all individuals at risk of leaving the state eventually do so then $S(\infty) = 0$. Otherwise, $S(\infty) > 0$ and the duration distribution is called defective. The sample mean of a completed spell length is the integral $\int_0^\infty S(u)du$. To obtain this result, use

$$\int_0^\infty uf(u)du = \int_0^\infty u dF(u) = uF(u)|_0^\infty - \int_0^\infty F(u)du.$$

Since $F(\infty) = 1$ and $F(0) = 0$, it follows that

$$\mathbb{E}[T] = \int_0^\infty (1 - F(u))du = \int_0^\infty S(u)du. \tag{17.3}$$

The mean duration equals the area under the survival curve.

Another key concept is the **hazard function**, which is the instantaneous probability of leaving a state conditional on survival to time t . This is defined as

$$\begin{aligned} \lambda(t) &= \lim_{\Delta t \rightarrow 0} \frac{\Pr[t \leq T < t + \Delta t | T \geq t]}{\Delta t} \\ &= \frac{f(t)}{S(t)}. \end{aligned} \tag{17.4}$$

Table 17.1. Survival Analysis: Definitions of Key Concepts

Function	Symbol	Definition	Relationships
Density	$f(t)$		$f(t) = dF(t)/dt$
Distribution	$F(t)$	$\Pr[T \leq t]$	$F(t) = \int_0^t f(s)ds$
Survivor	$S(t)$	$\Pr[T > t]$	$S(t) = 1 - F(t)$
Hazard	$\lambda(t)$	$\lim_{h \rightarrow 0} \frac{\Pr[t \leq T < t+h T \geq t]}{h}$	$\lambda(t) = f(t)/S(t)$
Cumulative hazard	$\Lambda(t)$	$\int_0^t \lambda(s)ds$	$\Lambda(t) = -\ln S(t)$

It is easily verified that the hazard equals the change in log-survivor function,

$$\lambda(t) = -\frac{d \ln(S(t))}{dt}.$$

The hazard $\lambda(t)$ specifies the distribution of T . In particular, integrating $\lambda(t)$ and using $S(0) = 1$ we can show that

$$S(t) = \exp\left(-\int_0^t \lambda(u)du\right). \quad (17.5)$$

In regression analysis of transitions the conditional hazard rate, $\lambda(t|\mathbf{x})$, is of central interest. This contrasts with more standard regression approaches in which the conditional mean function, $E[T|\mathbf{x}]$, is of chief interest. The latter approach has the disadvantage that in practice the durations are often censored.

A final related function is the **cumulative hazard function** or **integrated hazard function**

$$\begin{aligned} \Lambda(t) &= \int_0^t \lambda(s)ds \\ &= -\ln S(t), \end{aligned} \quad (17.6)$$

where the last equality uses (17.5). If $S(\infty) = 0$ then $\Lambda(\infty) = \infty$. The cumulative hazard is of interest as it can be more precisely estimated than the hazard function.

For any choice of distribution of T , it can be shown that the transformation $\Lambda(T)$ is unit exponentially distributed and $\ln \Lambda(T)$ is extreme value distributed, providing the basis for model specification tests, see Section 18.7.2.

Various related functions for the nonnegative continuous random variable T are summarized in Table 17.1.

Other functions are also used at times, most notably the **Laplace transform** $L(s) = E[\exp(-sT)]$, $s > 0$, which is a variant of the moment-generating function for random variable T restricted to be positive.

17.3.2. Discrete Data

It is very common for a duration to be measured as an interval. For example, data may indicate that a transition occurred in a particular week, but the exact time in the week

is not given. In such cases the transition times are said to be grouped and it is assumed that the hazard within the interval is constant. Discrete-time hazard models deal with such data.

The starting point is to define the **discrete-time hazard function** as the probability of transition at discrete time t_j , $j = 1, 2, \dots$, given survival to time t_j :

$$\begin{aligned}\lambda_j &= \Pr [T = t_j | T \geq t_j] \\ &= f^d(t_j) / S^d(t_{j-}),\end{aligned}\tag{17.7}$$

where the superscript d denotes discrete, and where $S^d(a_-) = \lim_{t \rightarrow a_-} S^d(t)$, an adjustment made because formally $S^d(t)$ equals $\Pr[T > t]$ rather than $\Pr[T \geq t]$, and the superscript d denotes discrete.

The **discrete-time survivor function** is obtained recursively from the hazard function as

$$\begin{aligned}S^d(t) &= \Pr [T \geq t] \\ &= \prod_{j|t_j \leq t} (1 - \lambda_j).\end{aligned}\tag{17.8}$$

For example, $\Pr[T > t_2]$ equals the probability of no transition at time t_1 times the probability of no transition at time t_2 conditional on surviving to just before t_2 , so that $\Pr[T > t_2] = (1 - \lambda_1) \times (1 - \lambda_2)$. The function $S^d(t)$ is a decreasing step function with steps at t_j , $j = 1, 2, \dots$.

The **discrete-time cumulative hazard function** is

$$\Lambda^d(t) = \sum_{j|t_j \leq t} \lambda_j.\tag{17.9}$$

Using (17.7), we have that the discrete probability that the spell ends at t_j is $\lambda_j S^d(t_j)$.

The continuous and discrete cases can be combined. The survivor function is then defined using the **product integral**, which reduces to the regular product (17.8) in the discrete case and to the exponential of the regular integral (17.5) in the continuous case. See Kalbfleisch and Prentice (2002, p. 10) or Lancaster (1990, pp. 10–12).

Discrete duration data may arise because the process generating transitions is intrinsically discrete. More often, however, the underlying process is continuous but the data are observed discretely. For example, one may know the week or month in which a spell ends, but not the day or hour. Such data are sometimes known as **grouped data**. The discrete data formulas can be used as follows. Let time be divided into $k + 1$ intervals $[a_0, a_1], [a_1, a_2], \dots, [a_{k-1}, a_k], [a_k, a_\infty)$. The discrete time duration $T = t_j$ indicates a transition in the interval $[a_{j-1}, a_j)$, that is, transition at time a_{j-1} or later. It is customary to treat discrete data as resulting from grouping, so that transitions are modeled in continuous time and then necessary adjustments are made for grouping. Further discussion is given in Section 17.10.

17.4. Censoring

Survival data are usually censored, as some spells are incompletely observed. That is, the lifetimes are only known to lie in certain intervals. As an example, instead of observing the length of completed spell of unemployment, data may come from a survey of the currently unemployed, so that only the length of an incomplete spell of unemployment is observed.

17.4.1. Censoring Mechanisms

In practice data may be right-censored, left-censored, or interval-censored. For **right-censoring** or censoring from above, we observe spells from time 0 until a censoring time c . Some spells will have ended by this time anyway (completed spells), but others will be incomplete and all we know is that they will end some time in the interval (c, ∞) . **Left-censoring** or censoring from below occurs when spells are known to end at some time in the interval $(0, c)$ but the exact time is unknown. The classical Tobit model is an example, where data on some spells are lost and the censoring time is unknown. **Interval-censoring** occurs when the completed spell length is observed but only in interval form such as in $[t_1^*, t_2^*)$.

The survival analysis literature has focused on right-censoring. Even with this restriction there are a variety of possible reasons for censoring, including random censoring, type I censoring, and type II censoring.

Random censoring or exogenous censoring means that each individual in the sample has a completed duration T_i^* and censoring time C_i^* that are independent of each other. We observe the completed duration T_i^* if the spell ends before the censoring time and the censoring time C_i^* if the spell ends after the censoring time. In addition it is known whether or not censoring has occurred. The observed data $(t_1, \delta_1), (t_2, \delta_2), \dots, (t_N, \delta_N)$ are realizations of the random variables

$$\begin{aligned} T_i &= \min(T_i^*, C_i^*), \\ \delta_i &= \mathbf{1}[T_i^* < C_i^*], \end{aligned} \tag{17.10}$$

where the indicator function $\mathbf{1}[A]$ equals one if event A occurs and equals zero otherwise. Note that δ_i equals one if a completed spell is observed and equals zero otherwise. Random censoring may result from causes such as random failure to follow up a case, individuals randomly dropping out of the study, or termination of the study.

Type I censoring occurs when durations are censored above a certain fixed known censoring time, say t_{ci} . For example, a sample of light bulbs may be tested for no more than 5,000 hours, with a common starting time for all items. Thus at the termination of the study the failure times or durations of some items will be known but other objects will still not have “failed.” Their lifetimes are said to be right-censored. This is a special case of random sampling, with $C_i^* = t_{ci}$. The classic Tobit model is an example of type I censoring from below for a random variable continuous on $(-\infty, \infty)$.

17.4.2. Independent (Noninformative) Censoring

For standard survival analysis methods to be valid in the presence of censoring the censoring mechanism needs to be one with **independent (noninformative) censoring**. This means that parameters of the distribution of C^* are not informative about the parameters of the distribution of the duration T^* . Then one may treat the censoring indicator δ as exogenous, and it is then not necessary to model the censoring mechanism if interest lies in the duration model parameters.

For censored data (t, δ) the uncensored observations are observed with probability

$$\Pr[T = t, \delta = 1] = \Pr[T = t | \delta = 1] \times \Pr[\delta = 1].$$

If the censoring mechanism is independent then $\Pr[T = t | \delta = 1] = \Pr[T = t]$. If the censoring is noninformative then the term $\Pr[\delta = 1]$ can be dropped from the likelihood function as it does not involve parameters of the distribution for T . Similarly, for censored observations,

$$\Pr[T = t, \delta = 0] = \Pr[T \geq t | \delta = 0] \times \Pr[\delta = 0]$$

with $\Pr[T \geq t | \delta = 0] = \Pr[T \geq t]$ under independent censoring and $\Pr[\delta = 0]$ being ignored under noninformative censoring. Combining, the density of interest reduces to $\Pr[T = t]$ when $\delta = 1$ and $\Pr[T \geq t]$ when $\delta = 0$.

When regressors \mathbf{x} are introduced it is possible for T^* and C^* to vary with the same regressors. Again what matters is that C^* parameters are not informative about the T^* parameters. Even more simply, at any given point in time, censoring must not occur because a subject has unusually high or low risk of failure given \mathbf{x} .

Type II censoring occurs when observation on N subjects ceases after the p th failure. Then only the durations for the p shortest spells are completely observed, and the remaining $N - p$ are censored at $C_i^* = t_{(p)}$, the duration of the p th shortest complete spell. For example, a clinical trial may end after p patients have died.

Random, type I, and type II censoring are all examples of independent censoring. A more formal treatment is given in Kalbfleisch and Prentice (2002, pp. 194–196).

17.5. Nonparametric Models

This section deals with nonparametric estimation of survival functions. These methods are very useful for descriptive purposes. It is often insightful to know the shape of the raw (unconditional) hazard or survival function before considering introducing regressors. The strike duration example illustrates the point.

We present estimators of the survivor, hazard, and cumulative hazard functions in the presence of independent censoring. Nonparametric estimation of the density itself is not considered because of the difficulty introduced by censoring; more importantly the survivor and hazard functions are more interpretable than the density.

No regressors are included. If interest lies in just a few key values of regressor(s), such as different treatment regimes or levels of treatment, then one can obtain separate nonparametric estimates at each key value and compare them. In economics

applications this is rarely the case and more structural models with regressors, presented in Sections 17.6–17.10, are needed.

We focus on discrete durations, such as life table data, so that the discrete-time formulation of Section 17.3.3 is used. Consider, for example, a cohort of N_0 individuals of specific age and gender, which is subsequently tracked for a number of years. At the end of year 1, there are N_1 individuals in the cohort, and $N_1 - N_0$ individuals from the original cohort have either died or been lost for other reasons (censored). A year later the size of the cohort is $N_2 - N_1$, and so forth. Such life table data can be used to construct a discrete-time survivor function without any prior parametric assumptions.

17.5.1. Nonparametric Estimation

With no censoring the obvious estimator of the survivor function is one minus the sample cumulative distribution function. Then $\widehat{S}(t)$ equals the number of spells in the sample of duration greater than t , divided by the sample size N . This is a step function with jump at each discrete failure time; see Figure 17.1. An alternative equivalent representation of this estimator, given momentarily in (17.13), maintains consistency in the presence of independent censoring.

Let $t_1 < t_2 < \dots < t_j < \dots < t_k$ denote the observed **discrete failure times** of the spells in a sample of size N , $N \geq k$. Define d_j to be the number of spells that end at time t_j . Since the data are discrete d_j may exceed one. Some spells may be incompletely observed. Define m_j to be the number of spells **right-censored** in the interval $[t_j, t_{j+1})$. The censoring mechanism is assumed to be independent censoring, so the only thing known about a spell censored in $[t_j, t_{j+1})$ is that the failure time is greater than t_j . Spells are **at risk** of failure if they have not yet failed or been censored. Define r_j to equal the number of spells at risk at time t_{j-} , that is, just before time t_j . Then $r_j = (d_j + m_j) + \dots + (d_k + m_k) = \sum_{l|l \geq j} (d_l + m_l)$. Note that $r_1 = N$. In summary,

$$\begin{aligned} d_j &= \# \text{ spells ending at time } t_j, \\ m_j &= \# \text{ spells censored in } [t_j, t_{j+1}), \\ r_j &= \# \text{ spells at risk at time } t_{j-} = \sum_{l|l \geq j} (d_l + m_l). \end{aligned} \tag{17.11}$$

The discrete-time formulation of Section 17.3.2 is used. Since $\lambda_j = \Pr[T = t_j | T \geq t_j]$, an obvious estimator of the hazard function is the number of spells ending at time t_j divided by the number at risk of failure at time t_{j-} , or

$$\widehat{\lambda}_j = \frac{d_j}{r_j}. \tag{17.12}$$

The discrete-time survivor function is defined in (17.8). The **Kaplan–Meier estimator** or product limit estimator of the survivor function is the sample analogue

$$\widehat{S}(t) = \prod_{j|t_j \leq t} (1 - \widehat{\lambda}_j) = \prod_{j|t_j \leq t} \frac{r_j - d_j}{r_j}. \tag{17.13}$$

Table 17.2. Hazard Rate and Survivor Function Computation: Example^a

j	r_j	d_j	m_j	$\widehat{\lambda}_j = d_j/r_j$	$\widehat{\Lambda}(t_j)$	$\widehat{S}(t_j)$
1	80	6	4	6/80	6/80	(1–6/80)
2	70	5	3	5/70	6/80 + 5/70	(1–6/80) × (1–5/70)
3	62	2	1	2/62	$\widehat{\Lambda}(t_2) + 2/62$	$\widehat{S}(t_2) \times (1 - 2/62)$
4	–	–	–	–		

^a At time t_j , r_j is the number of observations at risk, d_j is the number of deaths (failures), m_j is the number of missing spells (censored), $\widehat{\lambda}_j$ is the estimated hazard rate, $\widehat{\Lambda}(t_j)$ is the estimated cumulative hazard, and $\widehat{S}(t_j)$ is the estimated survivor function.

This is a decreasing step function with jump at each discrete failure time. The Kaplan–Meier estimator can be shown to be the nonparametric MLE (see Kalbfleisch and Prentice, 2002, pp. 14–16).

In the case of no censoring $\widehat{S}(t)$ in (17.13) simplifies to $\widehat{S}(t) = r/N$, the number still at risk at time t divided by the sample size, which is one minus the empirical cdf. To see this note that $r_j - d_j = r_{j+1}$, if $m_j = 0$, since then the number at risk at time j less the number of deaths at time j equals the number at risk at time $j + 1$. Then (17.13) becomes $\widehat{S}(t) = \prod_{j|t_j \leq t} r_{j+1}/r_j$, which simplifies to r/r_1 where $r_1 = N$.

The discrete-time cumulative hazard function is defined in (17.9). The **Nelson–Aalen estimator** of the cumulative hazard function is the obvious sample analogue

$$\widehat{\Lambda}(t) = \sum_{j|t_j \leq t} \widehat{\lambda}_j = \sum_{j|t_j \leq t} \frac{d_j}{r_j}. \quad (17.14)$$

This estimator can also be used to estimate the survival function by $\widetilde{S}(t_j) = \exp(-\widehat{\Lambda}(t))$, using the continuous case equality $S(t) = \exp(-\Lambda(t))$.

As an illustration, suppose that there are initially 80 observations, with 6 failures at time t_1 , 4 spells censored in $[t_1, t_2]$, 5 failures at time t_2 , 3 spells censored in $[t_2, t_3]$, 2 failures at time t_3 , 1 spell censored in $[t_3, t_4]$, and so on. Then the estimates for the cumulative hazard and survivor function for $t \leq t_3$ are given in Table 17.2.

Tied data arise when multiple failures occur at a particular point in time. It is common to assume that ties occur because of grouping, rather than because the process generates true discrete ties. The hazard estimate $\widehat{\lambda}_j = d_j/r_j$ assumes that all deaths occur simultaneously at time t_j . In fact deaths may occur progressively over the interval $[t_j, t_{j+1})$ and censoring may also occur progressively over this interval. Then r_j overstates the number of subjects at risk on average over the interval $[t_j, t_{j+1})$. A standard correction in life table analysis is to replace $\widehat{\lambda}_j = d_j/r_j$ by $d_j/(r_j - m_j/2)$, with similar changes in the formulas for $\widehat{S}(t)$, $\widehat{\Lambda}(t)$, and so on. Other corrections have also been proposed.

Most survival analysis programs do a good job of producing basic Kaplan–Meier plots and tables. Table 17.3 provides an abstract of such output for the strike data and complements Figure 17.1 given earlier.

Table 17.3. Strike Duration: Kaplan–Meier Survivor Function Estimates

Day	Beginning Total	Failures	Survivor Function	Standard Error
1	566	10	0.9823	0.0055
2	556	21	0.9452	0.0096
3	535	16	0.9170	0.0116
4	519	17	0.8869	0.0133
5	502	18	0.8551	0.0148
6	484	9	0.8392	0.0154
7	475	12	0.8180	0.0162
8	463	12	0.7968	0.0169
:	:	:	:	:
13	411	11	0.7067	0.0191
14	400	11	0.6873	0.0195

17.5.2. Confidence Bands for Nonparametric Estimates

The estimate $\hat{\lambda}_j = d_j/r_j$ of the hazard function is very discontinuous, especially for t large as then r_j becomes small relative to d_j/r_j . It can be visually useful to first smooth the hazard estimates, using nonparametric regression methods, see Section 9.5, before plotting them against time.

The survivor and cumulative hazard functions are much smoother, and it is standard to plot these against time, along with confidence bands that do reflect sampling variability. There are several ways to estimate these confidence bands. The formulas we give are those used in STATA.

For the Kaplan–Meier estimate of the survivor function it is common to use the Greenwood estimate of the variance

$$\widehat{V}[\widehat{S}(t)] = \widehat{S}(t)^2 \sum_{j|t_j \leq t} \frac{d_j}{r_j(r_j - d_j)}.$$

Reported confidence intervals for $S(t)$ are often based on $\ln(-\ln \widehat{S}(t))$ rather than on $\widehat{S}(t)$, as this transformation ensures the confidence interval lies in the range of the survivor function, which is between zero and one. The transformation yields the $100(1 - \alpha)\%$ confidence interval

$$S^d(t) \in (\widehat{S}(t) \exp(-z_{\alpha/2} \widehat{\sigma}(t)), \widehat{S}(t) \exp(z_{\alpha/2} \widehat{\sigma}(t))), \quad (17.15)$$

where $\sigma(t)$ denotes the standard deviation of $\ln(-\ln \widehat{S}(t))$, which is estimated using

$$\widehat{\sigma}_s^2(t) = \frac{\sum_{j|t_j \leq t} d_j/(r_j(r_j - d_j))}{\left[\sum_{j|t_j \leq t} \ln((r_j - d_j)/d_j) \right]^2}.$$

Table 17.4. Exponential and Weibull Distributions: pdf, cdf, Survivor Function, Hazard, Cumulative Hazard, Mean, and Variance

Function	Exponential	Weibull
$f(t)$	$\gamma \exp(-\gamma t)$	$\gamma \alpha t^{\alpha-1} \exp(-\gamma t^\alpha)$
$F(t)$	$1 - \exp(-\gamma t)$	$1 - \exp(-\gamma t^\alpha)$
$S(t)$	$\exp(-\gamma t)$	$\exp(-\gamma t^\alpha)$
$\lambda(t)$	γ	$\gamma \alpha t^{\alpha-1}$
$\Lambda(t)$	γt	γt^α
$E[T]$	γ^{-1}	$\gamma^{-1/\alpha} \Gamma(\alpha^{-1} + 1)$
$V[T]$	γ^{-2}	$\gamma^{-2/\alpha} [\Gamma(2\alpha^{-1} + 1) - \Gamma(\alpha^{-1} + 1)^2]$
γ, α	$\gamma > 0$	$\gamma > 0, \alpha > 0$

For the Nelson–Aalen estimator of the cumulative hazard function one variance estimate is

$$\widehat{V}[\widehat{\Lambda}(t)] = \sum_{j|t_j \leq t} \frac{d_j}{r_j^2}.$$

The transformation $\ln \widehat{\Lambda}(t)$, yields the $100(1 - \alpha)\%$ confidence interval for the cumulative hazard

$$\Lambda(t) \in [\widehat{\Lambda}(t) \exp(-z_{\alpha/2} \widehat{\sigma}_\Lambda(t)), \widehat{\Lambda}(t) \exp(z_{\alpha/2} \widehat{\sigma}_\Lambda(t))], \quad (17.16)$$

where $\widehat{\sigma}_\Lambda(t)$ denotes the standard deviation of $\ln \widehat{\Lambda}(t)$, which is estimated using

$$\widehat{\sigma}_\Lambda^2(t) = \widehat{V}[\widehat{\Lambda}(t)] / [\widehat{\Lambda}(t)^2].$$

17.6. Parametric Regression Models

We begin by outlining the properties of two distributions that perform a benchmark role. Then some standard regression models for duration data are considered.

17.6.1. Exponential and Weibull Distributions

The natural parametric starting point is the exponential, because a pure Poisson point process has durations that are exponentially distributed, see Lancaster (1990, p. 86). The **exponential duration distribution** has a constant hazard rate γ that does not vary with t , the memoryless property of the exponential. It follows from (17.5) that $S(t) = \exp(-\int_0^t \gamma du) = \exp(-\gamma t)$. The density is $f(t) = -S'(t) = \gamma \exp(-\gamma t)$, and the cumulative hazard $\Lambda(t) = -\ln S(t) = \gamma t$ is linear in t .

The exponential is a one-parameter distribution that is too restrictive in practice. A generalization commonly used in econometrics is the **Weibull distribution**. Table 17.4 presents the density and other distributional functions and moments for the Weibull and the exponential, which is the special case $\alpha = 1$. The function $\Gamma(\cdot)$ given in the Table 17.5 is the gamma function.

Table 17.5. Standard Parametric Models and Their Hazard and Survivor Functions^a

Parametric Model	Hazard Function	Survivor Function	Type
Exponential	γ	$\exp(-\gamma t)$	PH, AFT
Weibull	$\gamma \alpha t^{\alpha-1}$	$\exp(-\gamma t^\alpha)$	PH, AFT
Generalized Weibull	$\gamma \alpha t^{\alpha-1} S(t)^{-\mu}$	$[1 - \mu \gamma t^\alpha]^{1/\mu}$	PH
Gompertz	$\gamma \exp(\alpha t)$	$\exp(-(\gamma/\alpha)(e^{\alpha t} - 1))$	PH
Log-normal	$\frac{\exp(-(\ln t - \mu)^2/2\sigma^2)}{t\sigma\sqrt{2\pi}[1-\Phi((\ln t - \mu)/\sigma)]}$	$1 - \Phi((\ln t - \mu)/\sigma)$	AFT
Log-logistic	$\alpha \gamma^\alpha t^{\alpha-1} / [(1 + (\gamma t)^\alpha)]$	$1 / [1 + (\gamma t)^\alpha]$	AFT
Gamma	$\frac{\gamma(\gamma t)^{\alpha-1} \exp[-(\gamma t)]}{\Gamma(\alpha)[1-I(\alpha, \gamma t)]}$	$1 - I(\alpha, \gamma t)$	AFT

^a All the parameters are restricted to be positive, except that $-\infty < \alpha < \infty$ for the Gompertz model.

The Weibull has hazard $\lambda(t) = \gamma \alpha t^{\alpha-1}$, which is monotonically increasing if $\alpha > 1$ and monotonically decreasing if $\alpha < 1$. This is a special case of the proportional hazards (PH) family, see Section 17.7.1, in which $\lambda(t)$ factors into a baseline component that depends only on t , $\lambda_0(t)$, and a second term (e.g., γ) that can be parameterized as a function of covariates only. Figure 17.2 presents properties of the Weibull distribution with $\gamma = 0.01$ and $\alpha = 1.5$. The density is right-skewed, as is usually the case with duration data. The shape of the survivor curve is one common for many different distributions, making visual comparison of different estimated survivor curves difficult. The hazard is increasing for this Weibull example, since

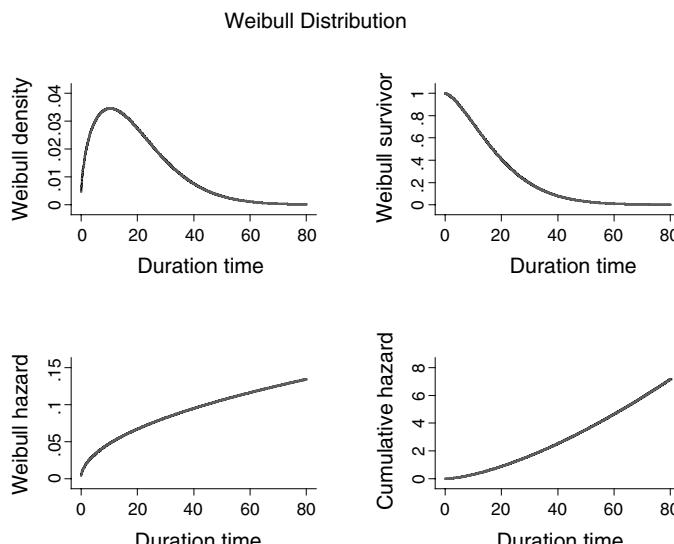


Figure 17.2: Weibull distribution: density, survivor, hazard and cumulative hazard functions plotted against time for $\gamma = 0.01$ and $\alpha = 1.5$.

$\alpha > 1$. Other parametric models can have quite different shaped hazard functions, including monotonically increasing, monotonically decreasing, *U*-shaped and inverse *U*-shaped.

The hazard function is often imprecisely estimated in practice, especially in the right tail. The cumulative hazard $\Lambda(t)$ is more precisely estimated and permits some discrimination across models. Even better is $\ln \Lambda(t)$ plotted against $\ln t$, since for the Weibull model $\ln \Lambda(t) = \ln \gamma + \alpha \ln t$ is linear in $\ln t$ with slope α .

17.6.2. Some Parametric Models

Popular choices for parametric models include the exponential, Weibull, Gompertz, log-normal, log-logistic, and the gamma. The hazard and survivor functions for these models are in Table 17.5.

For the gamma, $\Gamma(\alpha) = \int_0^\infty e^{-t} t^{\alpha-1} dt$, is the **gamma function** and $I(\alpha, \gamma t)$ is the **incomplete gamma function**, where $I(a, x) = \int_0^x e^{-t} t^{\alpha-1} dt / \Gamma(a)$, $0 < I(a, x) < 1$.

The generalized Weibull model was suggested by Mudholkar, Srivastava, and Kollia (1996). Through the introduction of additional shape parameter μ in the Weibull, it overcomes an important restriction of that model and allows the hazard function to have a more flexible shape. The Weibull model is obtained in the limit as $\mu \rightarrow 0$. From Table 17.5 note that

$$\ln \lambda(t) = \ln(\gamma\alpha) + (\alpha - 1)\ln t - \mu \ln S(t).$$

Because $\partial \ln S(t) / \partial t < 0$, the right-hand side of this equation is increasing in t if $\mu > 0$ and $\alpha > 1$. If $\alpha \leq 1$ and $\mu < 0$, then the hazard function is monotonically decreasing. If $\alpha > 1$ and $\mu < 0$, then the hazard function has two components, one of which is a decreasing function and the other an increasing function in t . Hence the two together can generate a unimodal or U-shaped hazard function. Therefore, the generalized Weibull is a potentially flexible and useful functional form.

The Gompertz is similar to the Weibull as it has hazard function that can be monotonically increasing (if $\alpha > 0$) or monotonically decreasing if ($\alpha < 0$), with the exponential as a special case ($\alpha = 0$). The Gompertz is a good model for mortality data and is used more in biostatistics than econometrics.

The log-normal distribution has an inverted bathtub hazard that first increases with t and then decreases with t . So too does the log-logistic, for $\alpha > 1$. These models are clearly more appropriate than exponential, Weibull, and Gompertz for duration data with this property.

Other parametric models include models based on the Rayleigh and Makeham distributions, inverse-Gaussian piecewise continuous hazards model, and the generalized gamma model (Lawless, 1982), which nests the gamma and Weibull models as special cases. Many parametric models are presented in detail in Kalbfleisch and Prentice (1980, chapter 3) and Lancaster (1990, chapter 3).

The distributions are generally two-parameter distributions. **Regressors** are introduced by letting $\gamma = \exp(\mathbf{x}'\boldsymbol{\beta})$ with α left as a constant, but for the log-normal $\mu = \mathbf{x}'\boldsymbol{\beta}$ and σ^2 is left as a constant.

The main issues in parametric modeling are the dependence on correct model specification for consistent parameter estimates and the wide range of parametric models that are available. Most models can be classified as either a PH model (the first four in Table 17.5) or an accelerated failure time model (the first two and the last three models in Table 17.5). The Weibull model, a member of both classes, is widely used in economics applications. Another widely used model, particularly for economics applications in which many observations are available, is the piecewise constant hazard model, which is a special case of the PH model.

17.6.3. Maximum Likelihood Estimation

We now consider fully parametric analysis with independent or noninformative censoring, with estimation by ML and by least squares. The continuous duration formulation is used since parametric models are based on continuous distributions. The regressors are assumed to be time-invariant, with time-varying regressors deferred to Section 17.9.

Let T^* denote durations without censoring, with conditional density $f(t|\mathbf{x}, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a $q \times 1$ parameter vector and \mathbf{x} are regressors that can vary across subjects but do not vary over a spell for a given subject. Estimation is complicated by the presence of censoring. Then the observed duration t is the length of a possibly incomplete spell, and the data are augmented by a variable indicating the presence of censoring, which is assumed to be noninformative.

From Section 17.4.2, the treatment is similar to that for the Tobit model. For uncensored observations the contribution to the likelihood is $f(t|\mathbf{x}, \boldsymbol{\theta})$. For right-censored observations we know only that the duration exceeded t , so the contribution is

$$\begin{aligned}\Pr[T > t] &= \int_t^\infty f(u|\mathbf{x}, \boldsymbol{\theta})du \\ &= 1 - F(t|\mathbf{x}, \boldsymbol{\theta}) = S(t|\mathbf{x}, \boldsymbol{\theta}),\end{aligned}$$

where $S(\cdot)$ is the survivor function. The density for the i th observation can be written as

$$f(t_i|\mathbf{x}_i, \boldsymbol{\theta})^{\delta_i} S(t_i|\mathbf{x}_i, \boldsymbol{\theta})^{1-\delta_i},$$

where δ_i is a right-censoring indicator with

$$\delta_i = \begin{cases} 1 & (\text{no censoring}), \\ 0 & (\text{right-censoring}). \end{cases}$$

Taking logs and summing, we have that the MLE $\hat{\boldsymbol{\theta}}$ maximizes the log-likelihood

$$\ln L(\boldsymbol{\theta}) = \sum_{i=1}^N [\delta_i \ln f(t_i|\mathbf{x}_i, \boldsymbol{\theta}) + (1 - \delta_i) \ln S(t_i|\mathbf{x}_i, \boldsymbol{\theta})], \quad (17.17)$$

where independence over i has been assumed. The first term in the sum corresponds to completed spells and the second term to right-censored spells. Since $\ln S(t) = \Lambda(t)$ and $\ln f(t) = \ln(\lambda(t)S(t)) = \ln \lambda(t) + \ln S(t)$, this log-likelihood can alternatively be

written in terms of the hazard and integrated hazard functions:

$$\ln L(\boldsymbol{\theta}) = \sum_{i=1}^N [\delta_i \ln \lambda(t_i | \mathbf{x}_i, \boldsymbol{\theta}) + \Lambda(t_i | \mathbf{x}_i, \boldsymbol{\theta})]. \quad (17.18)$$

This result is useful if the parametric model is defined by specifying the hazard rate rather than the pdf.

The usual estimation theory applies. The MLE will be distributed as $\hat{\boldsymbol{\theta}} \stackrel{a}{\sim} \mathcal{N}[\boldsymbol{\theta}, (-E[\partial^2 \ln L / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'])^{-1}]$ if the density is correctly specified, see Section 5.7.3. If the density is incorrectly specified, however, the MLE is inconsistent. The one notable exception is the exponential duration model in the absence of censoring, for which consistency requires only that the conditional mean function be correctly specified; see Section 5.7.3. However, inconsistency under misspecification arises even for the exponential model if censoring is introduced, and it arises for other parametric duration models even without censoring. This lack of robustness is the major weakness of the parametric approach, just as in the Tobit model case.

The ML approach can be adapted to permit other types of censoring. With **left-censoring**, the spell is known to be of length at most t , and the likelihood contribution is $\Pr[T^* < t] = \int_0^t f(s | \mathbf{x}, \boldsymbol{\theta}) ds = F(t | \mathbf{x}, \boldsymbol{\theta})$.

With **interval-censoring** the data are known to lie in $[t_a, t_b]$ and the likelihood contribution is $\Pr[t_a \leq T^* < t_b] = \int_{t_a}^{t_b} f(s | \mathbf{x}, \boldsymbol{\theta}) ds = S(t_b | \mathbf{x}, \boldsymbol{\theta}) - S(t_a | \mathbf{x}, \boldsymbol{\theta})$.

Duration data in economics applications are often interval-censored. For example, unemployment durations may be grouped into weeks and months, yet the parametric model is a continuous distribution such as the Weibull. It is usually assumed that the effect of interval-censoring is sufficiently minor so that the interval-censoring can be ignored. For example, a person who is unemployed after two months but no longer unemployed after three months may be treated as having an unemployment spell of exactly three months, rather than a spell in the range of two to three months.

17.6.4. Components of Likelihood

Given a mix of data, with durations that may be complete, truncated, or censored in one of the aforementioned ways, maximum likelihood of a parametrically specified model requires one to set up the likelihood function. (Lancaster (1979) displays different likelihood expressions appropriate for three different data setups for unemployment durations.) Each type of observation contributes a term to the likelihood function, and the full likelihood is formed by taking appropriate products of terms such as the following (see Klein and Moeschberger, 1997, p. 66):

complete durations:	$f(t),$
left-truncated at t_L ($t \geq t_L$):	$f(t) / S(t_L),$
left-censored at t_{CL} :	$1 - S(t_{CL}),$
right-censored at t_{CR} :	$S(t_{CR}),$
right-truncated at t_{CR} ($t \leq t_R$):	$f(t_R) / [1 - S(t_R)],$
interval-censored at t_{CL}, t_{CR} :	$S(t_{CL}) - S(t_{CR}).$

17.6.5. Weibull MLE Example

The Weibull distribution is presented in detail in Section 17.6.1. The hazard function is $\lambda(t) = \gamma\alpha t^{\alpha-1}$, where $\alpha > 0$ and $\gamma > 0$.

Regressors can be introduced in many possible ways, but the usual specification is to let $\gamma = \exp(\mathbf{x}'\beta)$, which ensures $\gamma > 0$, while α does not vary with regressors. (Some programs instead specify $\gamma = \exp(-\mathbf{x}'\beta)$, which leads to a reversal in the signs of the estimates of β .) Then

$$\begin{aligned}\ln f(t|\mathbf{x}, \beta, \alpha) &= \ln [\exp(\mathbf{x}'\beta)\alpha t^{\alpha-1} \exp(-\exp(\mathbf{x}'\beta)t^\alpha)] \\ &= \mathbf{x}'\beta + \ln \alpha + (\alpha - 1) \ln t - \exp(\mathbf{x}'\beta)t^\alpha\end{aligned}$$

and

$$\begin{aligned}\ln S(t|\mathbf{x}, \beta, \alpha) &= \ln [\exp(-\exp(\mathbf{x}'\beta)t^\alpha)] \\ &= -\exp(\mathbf{x}'\beta)t^\alpha.\end{aligned}$$

The likelihood function (17.17) becomes

$$\ln L = \sum_i [\delta_i \{\mathbf{x}'_i \beta + \ln \alpha + (\alpha - 1) \ln t_i - \exp(\mathbf{x}'_i \beta)t_i^\alpha\} - (1 - \delta_i) \exp(\mathbf{x}'_i \beta)t_i^\alpha]. \quad (17.19)$$

The first-order conditions for β and α are

$$\begin{aligned}\frac{\partial \ln L}{\partial \beta} &= \sum_i (\delta_i - \exp(\mathbf{x}'_i \beta)t_i^\alpha) \mathbf{x}_i = \mathbf{0}, \\ \frac{\partial \ln L}{\partial \alpha} &= \sum_i \delta_i (1/\alpha + \ln t_i) - \ln t_i \exp(\mathbf{x}'_i \beta)t_i^\alpha = 0.\end{aligned}$$

Consistency clearly requires strong assumptions. For example, even with no censoring $E[\partial \ln L / \partial \beta] = 0$ requires $E[T^\alpha | \mathbf{x}] = \exp(-\mathbf{x}'\beta)$.

17.6.6. Use of Model Estimates

The usual way to interpret estimates of nonlinear regression models is to consider the effect of regressors on the conditional mean. If $\gamma = \exp(\mathbf{x}'\beta)$ then from Table 17.4 the completed Weibull durations have mean $E[T^* | \mathbf{x}] = \exp(-\mathbf{x}'\beta/\alpha)\Gamma(\alpha^{-1} + 1) = \exp(-\mathbf{x}'\beta/\alpha)\Gamma(\alpha^{-1})/\alpha$. One can calculate the expected length of completed spells at various values of \mathbf{x} . For example, the length of completed unemployment for a person of given age, gender, and education level, say, can be predicted postestimation.

Parametric regression models also permit prediction of aspects of durations other than just the sample mean. For example, interest may lie in what fraction of population total time in completed unemployment spells is due to spells in excess of a given length or is experienced by individuals in a given socioeconomic group. The econometrics of duration models focuses on the role of covariates but it is especially concerned with the shape of the hazard function, notably because some economic theories make explicit predictions about the shape of the hazard function.

Despite these possibilities, interpretation of estimates of parametric duration models often focuses on the Weibull hazard rate $\lambda(t) = \gamma\alpha t^{\alpha-1}$ and how it changes over time and with changes in regressors. As noted in Section 17.3.2, this hazard rate is increasing if $\alpha > 1$ and is decreasing if $\alpha < 1$ so that one-sided tests of $\alpha = 1$ are obviously of interest. For changes in regressors

$$d\lambda(t)/dx = \exp(x'\beta)\alpha t^{\alpha-1}\beta = \lambda(t)\beta,$$

so that changes in regressors have the effect of a multiplicative change in the hazard function. A positive coefficient β_j therefore implies an increase in the hazard rate as a component of \mathbf{x} increases. Thus if $\beta_j > 0$ an increase in x_j leads to an increase in the hazard of failure and hence to a decrease in the expected duration.

17.6.7. Least-Squares Estimation

Estimation of fully parametric models can be by least squares rather than MLE, similar to the censored Tobit model. We present results, although least-squares regression sees little use in practice because the methods still rely on correct specification of the density and yet are less efficient than the MLE.

We begin with the exponential duration regression model. Then $E[T|\mathbf{x}] = 1/\gamma = \exp(-\mathbf{x}'\beta)$, so that NLS regression of t_i on $\exp(-\mathbf{x}'_i\beta)$ gives a consistent though inefficient estimator for β . Alternatively, the exponential duration model can be written as $\ln t = \mathbf{x}'\beta + u$, where u is extreme value distributed (see Section 17.7.2). Then $E[\ln T|\mathbf{x}] = \mathbf{x}'\beta - c$, where $c \approx 0.5722$ is Euler's constant. So β can be consistently estimated by linear regression of $\ln t_i$ on \mathbf{x}_i . With right-censoring we need to obtain analytical censored moments, which is possible for the exponential.

Extensions can be made using the more general results of Kiefer (1988, p. 665). He considers the PH model (17.21) with $\phi(\mathbf{x}'\beta) = \exp(\mathbf{x}'\beta)$. Then

$$\lambda(t|\mathbf{x}) = \lambda_0(t, \alpha) \exp(\mathbf{x}'\beta).$$

Then an expression for the baseline integrated hazard can be derived as follows:

$$\begin{aligned} \int_0^t \lambda(s|\mathbf{x}) ds &= \int_0^t \lambda_0(s, \alpha) \exp(\mathbf{x}'\beta) ds, \\ \Lambda(t|\mathbf{x}) &= \Lambda_0(t, \alpha) \exp(\mathbf{x}'\beta), \\ \ln \Lambda(t|\mathbf{x}) &= \ln \Lambda_0(t, \alpha) + \mathbf{x}'\beta, \\ -\ln \Lambda_0(t, \alpha) &= \mathbf{x}'\beta - \ln \Lambda(t|\mathbf{x}) \\ &= \mathbf{x}'\beta + u, \end{aligned} \tag{17.20}$$

where the error term $u = -\ln \Lambda(t|\mathbf{x})$ is type I extreme value distributed.

This result holds regardless of the choice of baseline hazard. We interpret this result in the following way. For a particular choice of baseline hazard $\lambda_0(t, \alpha)$, a convenient transformation of the dependent variable t is $-\ln \Lambda_0(t, \alpha)$, since it can be expressed as a linear regression model with error term that is type 1 extreme value distributed. For the exponential, already discussed, $\ln \Lambda_0(t, \alpha) = \ln t$ whereas for the Weibull $\ln \Lambda_0(t, \alpha) = \alpha \ln t$. In censored samples we obtain $E[\ln \Lambda_0(T, \alpha)|T > t^*]$ using

results for the censored type 1 extreme value, and then follow a Heckman two-step procedure. These results can also be used as the basis for simple diagnostics; this topic is discussed in the next chapter.

17.7. Some Important Duration Models

Perhaps the most widely used formulation used in regression analysis of durations is the proportional hazard model. However, familiarity with some of its variants and with the accelerated failure time (AFT) models, discussed in Section 17.7.2, is also helpful.

17.7.1. Proportional Hazards Model

In a **proportional hazard model**, as previously mentioned, the conditional hazard rate $\lambda(t|\mathbf{x})$ can be factored into separate functions of

$$\lambda(t|\mathbf{x}) = \lambda_0(t, \alpha)\phi(\mathbf{x}, \beta), \quad (17.21)$$

where $\lambda_0(t, \alpha)$ is called the **baseline hazard** and is a function of t alone, and $\phi(\mathbf{x}, \beta)$ is a function of \mathbf{x} alone. Usually $\phi(\mathbf{x}, \beta) = \exp(\mathbf{x}'\beta)$. **Polynomial baseline hazards** are popular in the literature.

All hazard functions $\lambda(t|\mathbf{x})$ of form (17.21) are proportional to the baseline hazard, with scale factor $\phi(\mathbf{x}, \beta)$ that is not an explicit function of t . The PH model is widely used as the parameters β can be consistently estimated without specification of the functional form for $\lambda_0(\cdot)$ (see Section 17.8).

The exponential, Weibull, and Gompertz regression models are all PH models, since their hazards are, respectively, $\exp(\mathbf{x}'\beta)$, $\exp(\mathbf{x}'\beta)\alpha t^{\alpha-1}$, and $\exp(\mathbf{x}'\beta)\exp(\alpha t)$.

Another example of the PH model, used especially in applications to unemployment durations, is the **piecewise constant hazard model**, which lets $\lambda_0(t, \alpha)$ be a step function with k segments so that

$$\lambda_0(t, \alpha) = e^{\alpha_j}, \quad c_{j-1} \leq t < c_j, \quad j = 1, \dots, k, \quad (17.22)$$

where $c_0 = 0$, $c_k = \infty$, the other breakpoints c_1, \dots, c_{k-1} are specified, and the parameters $\alpha_1, \dots, \alpha_k$ are to be estimated. These parameters are exponentiated to ensure $\lambda_0(t, \alpha) > 0$. This model has more baseline parameters to estimate than models such as the Weibull, which has only one baseline hazard parameter, but can still be practical with a sufficiently large data set.

The identifiability of the PH model in the presence of unobserved heterogeneity is discussed in Section 18.3.

17.7.2. Accelerated Failure Time Model

An AFT model arises by first modeling $\ln t$ rather than t . A regression model is specified for

$$\ln t = \mathbf{x}'\beta + u, \quad (17.23)$$

and different distributions for u lead to different AFT models. Since $\ln t$ can take values on $(-\infty, \infty)$ the distribution for u can be any continuous distribution on $(-\infty, \infty)$.

The term **accelerated failure time** arises because $t = \exp(\mathbf{x}'\boldsymbol{\beta})v$, where $v = e^u$, has hazard rate $\lambda(t|\mathbf{x}) = \lambda_0(v)\exp(\mathbf{x}'\boldsymbol{\beta})$, where the baseline hazard $\lambda_0(v)$ does not depend on t . Substituting $v = t \exp(-\mathbf{x}'\boldsymbol{\beta})$ yields the hazard

$$\lambda(t|\mathbf{x}) = \lambda_0(t \exp(-\mathbf{x}'\boldsymbol{\beta})) \exp(\mathbf{x}'\boldsymbol{\beta}). \quad (17.24)$$

This is an acceleration of the baseline hazard $\lambda_0(t)$ if $\exp(-\mathbf{x}'\boldsymbol{\beta}) > 1$ and a deceleration if $\exp(-\mathbf{x}'\boldsymbol{\beta}) < 1$.

The log-normal model for t results if $u \sim \mathcal{N}[0, \sigma^2]$; the log-logistic model is obtained by specifying u to be logistic distributed. The gamma model can also be obtained as an AFT model, by letting u have density $f(u) = \exp(\alpha u - e^u)/\Gamma(\alpha)$.

The Weibull and exponential models are unique in being of both PH form and AFT form. The latter form is obtained by letting u be αw , where w is extreme value distributed with density $f(w) = e^w \exp(-e^w)$.

Additional duration models can be obtained by considering $g(t) = \mathbf{x}'\boldsymbol{\beta} + u$, for transformations other than $g(t) = \ln t$. This is a member of the class of transformation models, which includes, for example, the Box–Cox regression model.

17.7.3. Flexible Hazard Models

Some models begin with specification of the hazard rate, rather than the pdf. For example, the hazard may be specified to be quadratic in t , such as $\lambda(t) = \mathbf{x}'\boldsymbol{\beta} + a_1t + a_2t^2$. This permits a U-shaped hazard function. The corresponding integrated hazard is $\Lambda(t) = (\mathbf{x}'\boldsymbol{\beta})t + (a_1/2)t^2 + (a_2/3)t^3$. Given $\lambda(t)$ and $\Lambda(t)$ we can directly form the log-likelihood, using the earlier result.

The weaknesses of this approach are that negative values of λ and Λ may occur and that the hazard rate may be defective as the corresponding pdf may not necessarily integrate to unity.

17.8. Cox PH Model

Fully parametric models for single-spell duration data are relatively simple to estimate in the presence of censoring but produce inconsistent parameter estimates if any part of the parametric model is misspecified. One way of resolving this impasse is to choose parametric functional forms that are flexible and hence provide some protection against misspecification. Although this is a valid approach in principle, identification and estimation of such flexible functional forms is not always straightforward. An example is the generalized gamma model, which many users find difficult to estimate.

Fortunately, there is a semiparametric method that requires less than complete distributional specification. The method differs considerably from semiparametric methods proposed for Tobit models, where similar issues of model robustness under censoring arise, as it is based on a model for the hazard rate that has no meaningful physical interpretation in the Tobit case. In addition, unlike the Tobit case, the method

is viewed as empirically so successful that it has become the standard method for survival data.

17.8.1. Proportional Hazards Model

The starting point is to propose a particular functional form for the hazard rate, the proportional hazard model, introduced in Section 17.7.1, with conditional hazard rate $\lambda(t|\mathbf{x})$ factored into separate functions of

$$\lambda(t|\mathbf{x}, \boldsymbol{\beta}) = \lambda_0(t)\phi(\mathbf{x}, \boldsymbol{\beta}). \quad (17.25)$$

As before, the function $\lambda_0(t)$ is called the baseline hazard and is a function of t alone. The function $\phi(\mathbf{x}, \boldsymbol{\beta})$ is a function of \mathbf{x} alone, where initially we consider time-invariant regressors \mathbf{x} but later relax this assumption. A semiparametric model is considered, with the functional form for $\lambda_0(t)$ unspecified and the functional form for $\phi(\mathbf{x}, \boldsymbol{\beta})$ fully specified.

The most common choice of $\phi(\mathbf{x}, \boldsymbol{\beta})$ is the exponential form

$$\phi(\mathbf{x}, \boldsymbol{\beta}) = \exp(\mathbf{x}'\boldsymbol{\beta}). \quad (17.26)$$

This permits coefficients to be easily interpretable, in addition to ensuring $\phi(\mathbf{x}, \boldsymbol{\beta}) > 0$. Suppose the j th regressor x_j increases by one unit and other regressors are unchanged; then

$$\begin{aligned} \lambda(t|\mathbf{x}_{\text{new}}, \boldsymbol{\beta}) &= \lambda_0(t) \exp(\mathbf{x}'\boldsymbol{\beta} + \beta_j) \\ &= \exp(\beta_j) \lambda(t|\mathbf{x}, \boldsymbol{\beta}). \end{aligned} \quad (17.27)$$

Thus the new hazard is $\exp(\beta_j)$ times the original hazard, and the change in the hazard is $1 - \exp(\beta_j)$ times the original hazard. If one instead uses calculus methods, the change in the hazard is β_j times the original hazard, since

$$\partial \lambda(t|\mathbf{x}, \boldsymbol{\beta}) / \partial x_j = \lambda_0(t) \exp(\mathbf{x}'\boldsymbol{\beta}) \beta_j = \beta_j \lambda(t|\mathbf{x}, \boldsymbol{\beta}). \quad (17.28)$$

This is consistent with the noncalculus result as $\exp(\beta_j) \simeq 1 + \beta_j$. Statistical packages often report estimates and associated confidence intervals for both β_j and $\exp(\beta_j)$.

For more general forms of $\phi(\mathbf{x}, \boldsymbol{\beta})$, changes in regressors can again be interpreted as having a multiplicative effect on the original hazard, since

$$\begin{aligned} \partial \lambda(t|\mathbf{x}, \boldsymbol{\beta}) / \partial \mathbf{x} &= \lambda_0(t) \partial \phi(\mathbf{x}, \boldsymbol{\beta}) / \partial x_j \\ &= \lambda(t|\mathbf{x}, \boldsymbol{\beta}) \times [\partial \phi(\mathbf{x}, \boldsymbol{\beta}) / \partial x_j] / \phi(\mathbf{x}, \boldsymbol{\beta}). \end{aligned} \quad (17.29)$$

This requires knowledge of $\boldsymbol{\beta}$ but not of the baseline hazard $\lambda_0(t)$.

An important issue is the identification of the PH model. This is discussed in the next chapter in a more general setting that allows for the presence of unobserved heterogeneity in the model.

17.8.2. Partial Likelihood Estimation

Cox (1972, 1975) proposed a method to estimate β in the PH model that does not require simultaneous estimation of the baseline hazard function $\lambda_0(t)$. If desired an estimate of the baseline hazard can be recovered after estimation of β . The results presented here accommodate independent censoring and tied data.

The setup resembles that in Section 17.5, with failure times ordered and categorization of observations into those that die or are at risk at each failure time. Let $t_1 < t_2 < \dots < t_j < \dots < t_k$ denote the observed **discrete failure times** of the spells in a sample of size N , $N \geq k$. The **risk set** $R(t_j)$ is defined to be the set of individuals who are at risk of failing just before the j th ordered failure, $D(t_j)$ is the set of subjects that die at time t_j , and d_j denotes the number that die at time t_j . To summarize, we have

$$\begin{aligned} R(t_j) &= \{l : t_l \geq t_j\} && \text{set of spells at risk at } t_j, \\ D(t_j) &= \{l : t_l = t_j\} && \text{set of spells completed at } t_j, \\ d_j &= \sum_l \mathbf{1}(t_l = t_j) && \text{number of spells completed at } t_j. \end{aligned} \quad (17.30)$$

The risk set at time t_j includes all spells that are not yet completed or not yet censored. Tied data are possible, in which case $d_j > 1$.

Now consider the probability of a particular at-risk spell ending at time t_j . The probability that spell j is the actual spell that ends equals the conditional probability of failure for spell j divided by the conditional probability that a spell of any individual in the risk set $R(t_j)$ fails. This latter probability is the sum of the conditional probability of failure for each individual in $R(t_j)$. Then

$$\begin{aligned} \Pr [T_j = t_j | R(t_j)] &= \frac{\Pr [T_j = t_j | T_j \geq t_j]}{\sum_{l \in R(t_j)} \Pr [T_l = t_l | T_l \geq t_j]} \\ &= \frac{\lambda_j(t_j | \mathbf{x}_j, \beta)}{\sum_{l \in R(t_j)} \lambda_l(t_j | \mathbf{x}_l, \beta)} \\ &= \frac{\phi(\mathbf{x}_j, \beta)}{\sum_{l \in R(t_j)} \phi(\mathbf{x}_l, \beta)}, \end{aligned}$$

where in the last line the baseline hazard factor $\lambda_0(t_j)$ has dropped out, as a consequence of the PH assumption. (As a result the intercept in this model is not identified.) The preceding result that the baseline hazard can be eliminated provides a basis for estimating β . However, we must control for tied durations that are likely to occur when durations are grouped.

Ties are more likely when durations are grouped. If the data include ties (i.e., there is more than one failure at a given time), an adjustment is needed. For example, suppose there are two tied values at time t_j , for individuals j_1 and j_2 with regressors \mathbf{x}_{j1} and \mathbf{x}_{j2} . If j_1 fails before j_2 then the probability is

$$\phi(\mathbf{x}_{j1}, \beta) / \sum_{l \in R(t_j)} \phi(\mathbf{x}_l, \beta) + \phi(\mathbf{x}_{j2}, \beta) / \sum_{l \in R_1(t_j)} \phi(\mathbf{x}_l, \beta),$$

where $R_1(t_j)$ equals $R(t_j)$ with subject j_1 excluded. A similar term arises if j_2 fails before j_1 , and the likelihood contribution is the sum of these two possibilities. The

exact likelihood becomes quite complicated with many tied values. A standard approximation, due to Breslow and Peto, see Cox and Oakes (1984), is to let

$$\Pr[T_j = t_j | j \in R(t_j)] \simeq \frac{\prod_{m \in D(t_j)} \phi(\mathbf{x}_m, \boldsymbol{\beta})}{\left[\sum_{l \in R(t_j)} \phi(\mathbf{x}_l, \boldsymbol{\beta}) \right]^{d_j}}, \quad (17.31)$$

where $D(t_j)$ denotes the set of subjects that die at time t_j and d_j denotes the number that die at time t_j . This approximation works well if the number of failures at time t_j is small relative to the number at risk.

Cox defined the **partial likelihood function** to be the joint product of $\Pr[T_j = t_j | j \in R(t_j)]$ over the k ordered failure times. Then

$$L_p(\boldsymbol{\beta}) = \prod_{j=1}^k \frac{\prod_{m \in D(t_j)} \phi(\mathbf{x}_m, \boldsymbol{\beta})}{\left[\sum_{l \in R(t_j)} \phi(\mathbf{x}_l, \boldsymbol{\beta}) \right]^{d_j}}. \quad (17.32)$$

Cox proposed estimation of $\boldsymbol{\beta}$ by minimizing the log partial likelihood function

$$\ln L_p = \sum_{j=1}^k \left[\sum_{m \in D(t_j)} \ln \phi(\mathbf{x}_m, \boldsymbol{\beta}) - d_j \ln \left(\sum_{l \in R(t_j)} \phi(\mathbf{x}_l, \boldsymbol{\beta}) \right) \right]. \quad (17.33)$$

Censored spells appear only in the second term of $\ln L_p$ because they do not contribute to observed deaths but, until they are censored, affect the size of the risk set. Equation (17.33) can be rewritten as

$$\ln L_p(\boldsymbol{\beta}) = \sum_{i=1}^N \delta_i \left[\ln \phi(\mathbf{x}_i, \boldsymbol{\beta}) - \ln \left(\sum_{l \in R(t_i)} \phi(\mathbf{x}_l, \boldsymbol{\beta}) \right) \right], \quad (17.34)$$

where the indicator variables $\delta_i = 1$ for uncensored observation and equal zero otherwise.

For the usual specification of $\phi(\mathbf{x}, \boldsymbol{\beta}) = \exp(\mathbf{x}'\boldsymbol{\beta})$, so that $\ln \phi(\mathbf{x}, \boldsymbol{\beta}) = \mathbf{x}'\boldsymbol{\beta}$, the resulting first-order conditions become

$$\frac{\partial \ln L_p(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^N \delta_i [\mathbf{x}_i - \mathbf{x}_i^*(\boldsymbol{\beta})] = \mathbf{0},$$

where $\mathbf{x}_i^*(\boldsymbol{\beta}) = \sum_{l \in R(t_i)} \mathbf{x}_l \exp(\mathbf{x}_l'\boldsymbol{\beta}) / \sum_{l \in R(t_i)} \exp(\mathbf{x}_l'\boldsymbol{\beta})$ is a weighted average of the regressors \mathbf{x}_l for subjects at risk at failure time t_i .

The partial likelihood is a limited information likelihood, as the baseline hazard $\lambda_0(t)$ has dropped out, but is neither a conditional likelihood nor a marginal likelihood. Whether $L_p(\boldsymbol{\beta})$ is a valid likelihood function has given rise to much discussion in the statistics literature. It can be shown (Andersen et al., 1993) that even though $\ln L_p$ is not the full likelihood function, the estimator of $\boldsymbol{\beta}$ that maximizes $\ln L_p$ is consistent. See also Kalbfleisch and Prentice (2002, pp. 99–101) and Lancaster (1990, chapter 9).

The Chapter 5 results on extremum estimation apply, with the simplification that $\mathbf{A}(\boldsymbol{\beta}) = -\mathbf{B}(\boldsymbol{\beta})$ similar to the ML case, so that

$$\widehat{\boldsymbol{\beta}} \stackrel{a}{\sim} \mathcal{N} \left[\boldsymbol{\beta}, \left(-E \left[\frac{\partial^2 \ln L_p(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \right] \right)^{-1} \right]. \quad (17.35)$$

The estimator is inefficient, though comparisons of the partial likelihood estimator with the MLE for fully parametric PH models such as the Weibull reveal relatively small efficiency loss.

17.8.3. Survivor Function for the Cox PH Model

Many studies stop at estimation of β , being content to measure the impact of changes in regressors on the baseline hazard using (17.28) or (17.29). Other studies are additionally interested in the shape of the baseline hazard function. For the PH model it is possible to obtain a nonparametric estimate of the baseline hazard or survivor function, once β is obtained by maximizing the partial likelihood. The estimates are analogous to the Kaplan–Meier estimator of Section 17.5.1.

We obtain the PH hazard function's associated survivor function

$$S(t|\mathbf{x}, \beta) = S_0(t)^{\phi(\mathbf{x}, \beta)},$$

using $S(t|\mathbf{x}, \beta) = \exp\left[-\int_0^t \lambda_0(s)\phi(\mathbf{x}, \beta)ds\right]$ and defining $S_0(t) = \exp\left[-\int_0^t \lambda_0(s)ds\right]$.

Now assume a discrete time formulation with baseline hazard rate $1 - \alpha_j$ at discrete failure time t_j , $j = 1, \dots, k$. Some considerable algebra given in the next section yields estimate $\hat{\alpha}_j$ that is the solution to

$$\sum_{l \in D(t_j)}^k \frac{\phi(\mathbf{x}_l, \hat{\beta})}{1 - \hat{\alpha}_j^{\phi(\mathbf{x}_l, \hat{\beta})}} = \sum_{m \in R(t_j)} \phi(\mathbf{x}_m, \hat{\beta}), \quad j = 1, \dots, k, \quad (17.36)$$

where $\hat{\beta}$ is the partial likelihood estimator of β , $D(t_j)$ denotes the subjects that die at time t_j , and $R(t_j)$ denotes the subjects at risk at time t_j . From the discussion of discrete time hazard in Section 17.3.3, the baseline survivor function $S_0(t) = \prod_{j|t_j \leq t} \alpha_j$, the cumulative product of the instantaneous conditional survival probabilities. The estimated baseline survival function is then

$$\hat{S}_0(t) = \prod_{j|t_j \leq t} \hat{\alpha}_j. \quad (17.37)$$

If there are no regressors then $\hat{S}_0(t)$ reduces to the Kaplan–Meier estimator – normalize $\phi(\mathbf{x}_l, \beta) = 1$ and the expression yields hazard rate $1 - \hat{\alpha}_j = d_j/r_j$. If there are regressors but no ties then the expression yields baseline hazard rate $1 - \hat{\alpha}_j = \phi(\mathbf{x}_j, \hat{\beta}) / \sum_{m \in R(t_j)} \phi(\mathbf{x}_j, \hat{\beta})$.

The survivor function for individuals with regressors $\mathbf{x} = \mathbf{x}^*$ can be estimated using

$$\hat{S}(t|\mathbf{x}^*, \beta) = \hat{S}_0(t)^{\phi(\mathbf{x}^*, \hat{\beta})}.$$

Linear transformations of regressors do not change the estimates of β , but they do change the baseline hazard function. For example,

$$\begin{aligned} \lambda(t|\mathbf{x}, \beta) &= \lambda_0(t) \exp(\mathbf{x}'\beta) \\ &= \lambda_0(t) \exp(\bar{\mathbf{x}}'\beta) \exp((\mathbf{x} - \bar{\mathbf{x}})'\beta) \\ &= \lambda_0^*(t) \exp((\mathbf{x} - \bar{\mathbf{x}})'\beta), \end{aligned}$$

where the new baseline hazard is $\lambda_0^*(t \exp((\mathbf{x} - \bar{\mathbf{x}})'\beta))$. Hence subtracting the sample mean from each regressor will change the baseline hazard, and care is needed in interpretation of the baseline hazard or survivor function.

Also, although the estimated baseline hazard is useful for computing and comparing hazard rates for specific groups of individuals, it may have a very choppy appearance, so some smoothing may be applied for ease of interpretation.

17.8.4. Derivation for the Survivor Function

We obtain the estimating equations for α_j given in (17.36), following Kalbfleisch and Prentice (2002, pp. 114–118).

A subject with duration time t_j has likelihood contribution equal to the probability of survival time $t > t_{j-1}$ less the probability of survival time $t > t_j$. This is

$$\begin{aligned} S(t_j | \mathbf{x}, \beta) - S(t_{j+1} | \mathbf{x}, \beta) &= S_0(t_j)^{\phi(\mathbf{x}, \beta)} - S_0(t_{j+1})^{\phi(\mathbf{x}, \beta)} \\ &= (\alpha_j^{-1} S_0(t_{j+1}))^{\phi(\mathbf{x}, \beta)} - S_0(t_{j+1})^{\phi(\mathbf{x}, \beta)} \\ &= (\alpha_j^{-\phi(\mathbf{x}, \beta)} - 1) S_0(t_{j+1})^{\phi(\mathbf{x}, \beta)} \end{aligned}$$

using $S_0(t_{j+1}) = \prod_{l=1}^j \alpha_l = \alpha_j S_0(t_j)$.

For those subjects that are censored at time t_j the likelihood contribution is the probability of survival $t > t_j$, or $S_0(t_{j+1})^{\phi(\mathbf{x}, \beta)}$. So subjects that either die or are censored in $[t_j, t_{j+1})$ contribute probability $S_0(t_{j+1})^{\phi(\mathbf{x}, \beta)} = \prod_{l=1}^j \alpha_l^{\phi(\mathbf{x}, \beta)}$ with an additional multiplier $(\alpha_j^{-\phi(\mathbf{x}, \beta)} - 1)$ for subjects that die. Then over all failure times the likelihood is

$$L(\alpha, \beta) = \prod_{j=1}^k \left[\prod_{l \in D(t_j)} (\alpha_j^{-\phi(\mathbf{x}_l, \beta)} - 1) \prod_{m \in R(t_j)} \alpha_j^{-\phi(\mathbf{x}_m, \beta)} \right].$$

The log-likelihood is

$$\ln L(\alpha, \beta) = \sum_{j=1}^k \left[\sum_{l \in D(t_j)} \ln(\alpha_j^{-\phi(\mathbf{x}_l, \beta)} - 1) + \sum_{m \in R(t_j)} -\phi(\mathbf{x}_m, \beta) \ln \alpha_j \right].$$

Then $\partial \ln L(\alpha, \beta) / \partial \alpha_j = 0$ can be re-expressed as (17.36).

17.9. Time-Varying Regressors

The preceding results have been restricted to models where regressors are variables such as gender that vary across individuals but for given individual do not vary over time. This is standard in other standard cross-section models such as logit and Tobit models. For survival data, however, individuals may be observed at several stages during a spell and relevant regressors may take different values over the spell. For example, in a medical survival study dosage levels of a medication may vary over time for a given individual. During an unemployment spell the rate of unemployment

benefits may change, perhaps in a discrete manner. During a job search the marital status of a person may change.

Time-varying covariates pose two kinds of problems. First, it is clearly a misspecification to treat a time-varying covariate as a fixed variable. The entire history of the covariate over the spell may be relevant, a consideration that may require us to incorporate lagged values of some regressors as determinants of the hazard rate. Second, a time-varying covariate may exhibit **feedback** and hence may not be strictly exogenous as is often assumed in a duration model. For example, the duration of an unemployment spell may depend on the job search strategy of an individual, but the latter may change as the duration of unemployment lengthens. A second example is that the dosage level of the treatment may be varied in response to the deteriorating or improving condition of the patient. Deterministic time variation is easier to handle and hence standard analysis considers only the first of these issues, requiring the assumption that the covariates are weakly exogenous; that is, whatever the process, stochastic or deterministic, that underlies the time variation, we do not need to take account of the parameters of that process in estimating the hazard model under consideration. Some authors (e.g., Kalbfleish and Prentice, 2002, pp. 196–200) refer to such time variation as **external**. Endogenous time-varying covariates are then called **internal**.

One rather simple solution, especially if the software cannot handle time-varying covariates, is to replace the time-varying covariate by its average value during the spell. Good software, however, allows greater flexibility.

Consider an individual spell of (say) unemployment that lasts from the origin to time T , at which time a transition to employment is observed. Let $0 < t_1 < t_2 < t_3 < T$, where t_1 , t_2 , and t_3 are intermediate points within the spell. Suppose that there are two covariates x_1 and $x_2(t)$ that are, respectively, time-invariant and time-varying. For simplicity assume that x_1 is binary and x_2 takes the values $x_2(t_1)$, $x_2(t_2)$, and $x_2(t_3)$ in a step fashion in the intervals $[0, t_1]$, $[t_1, t_2]$, and $[t_2, T]$, respectively. Also assume that the time-varying regressor is exogenous and/or that the pattern of time variation is deterministic. Then for this particular spell the data can be written as a three-line record, rather than a one line record, as follows:

Observation	Duration	x_1	$x_2(t)$	Censoring Indicator
1	t_1	1	$x_2(t_1)$	0
1	t_2	1	$x_2(t_2)$	0
1	T	1	$x_2(T)$	1

The interpretation of this information is that we can split the total observed duration into three segments. During the first and the second segment the covariate values are $(1, x_2(t_1))$ and $(1, x_2(t_2))$, respectively, and no transition is observed (hence the censoring indicator is 0, and then in the third segment the covariate values are $(1, x_2(T))$ and a transition is observed. This is akin to having three observations, in two of which the duration is censored and in the third duration is complete.

Suppose now that both the current and one lagged value of $x_2(t)$ are thought to be appropriate covariates. That is, the hazard rate at a point in time may depend on changes in a covariate earlier in the spell. Then the data array can be written as follows:

Observation	Duration	x_1	$x_2(t)$	$x_2(t - 1)$	Censoring Indicator
1	t_1	1	$x_2(t_1)$	0	0
1	t_2	1	$x_2(t_2)$	$x_2(t_1)$	0
1	T	1	$x_2(T)$	$x_2(t_2)$	1

Here we have assumed that the value of the $x_2(t)$ prior to the commencement of the spell was zero. Notice that in both of these examples, the covariate $x_2(t)$ varies at discrete points in time.

Although one could have multiline entries in a data set, in a large data set this is potentially tedious and confusing if the software ends up treating the entries as different observations. Fortunately, computer software can usually allow the user to identify a time-varying covariate as a part of the definition of the regression model. One can accommodate step functions or continuous functions in terms of the elapsed duration of the spell.

17.9.1. Extended Cox Model

The fixed regressor analysis of the Cox model in Section 17.8 is readily extended to time-varying regressors.

In general the hazard function depends on the complete time path of regressors $\mathbf{x}(t)$, so that

$$\lambda(t|\mathbf{x}(t)) = \lim_{\Delta t \rightarrow 0} \frac{\Pr[t \leq T < t + \Delta t | \mathbf{x}(t), T \geq t]}{\Delta t}.$$

We consider the PH form

$$\lambda(t|\mathbf{x}(t)) = \lambda_0(t, \boldsymbol{\alpha})\phi(\mathbf{x}(t), \boldsymbol{\beta}),$$

where the restriction is made that only the current value $\mathbf{x}(t)$ of the covariate matters, rather than the entire history of $\mathbf{x}(t)$.

It is clear from Section 17.8.2 on the Cox partial likelihood approach that what matters at each failure time t_j is the value of regressors $\mathbf{x}(t_j)$ for those observations in the risk set $R(t_j)$. Thus for the i th subject \mathbf{x}_i is replaced by $\mathbf{x}_i(t_j)$. The partial likelihood has similar changes, and

$$\ln L_p = \sum_{j=1}^k \left[\sum_{m \in D(t_j)} \ln \phi(\mathbf{x}_m(t_j), \boldsymbol{\beta}) - d_j \ln \left(\sum_{l \in R(t_j)} \phi(\mathbf{x}_l(t_j), \boldsymbol{\beta}) \right) \right].$$

Note that the form of the data is more complicated now, as there may be multiple observations for each subject. For example, suppose time is in discrete integer values,

there is only one regressor, and observation one has completed duration 25 and regressor x_1 , which takes value 50 in $[0, 5]$, 100 in $[6, 15]$, and then 200 in $[16, 25]$. Further, suppose the first five ordered failure times are 3, 8, 13, 18, and 25. Then $x_1(t_1) = 50$, $x_1(t_2) = 100$, $x_1(t_3) = 100$, $x_1(t_4) = 200$, and $x_1(t_5) = 200$.

17.10. Discrete-Time Proportional Hazards

Grouped duration models are more appropriate when failure times are observed or recorded at aggregated time intervals like a week or a month.

A simple method is to form a panel and estimate a stacked logit or probit model of the probability of individual failure in each period, with separate intercept for period. This is presented in Section 17.10.3. However, first we present the discrete-time variant of a continuous-time PH model, considered by several authors including Kalbfleisch and Prentice (1980), Fahrmeir and Tutz (1994), Kiefer (1988), and Meyer (1990). Our exposition follows Blake, Lunde, and Timmermann (1999).

17.10.1. Discrete-Time Proportional Hazards

For grouped data, with grouping points t_a , $a = 1, \dots, A$, the discrete-time hazard function is defined by

$$\lambda^d(t_a | \mathbf{x}) = \Pr [t_{a-1} \leq T < t_a | T \geq t_{a-1}, \mathbf{x}(t_{a-1})], \quad a = 1, \dots, A.$$

Time-varying regressors are permitted. The associated discrete-time survivor function is

$$S^d(t_a | \mathbf{x}) = \Pr [T \geq t_{a-1} | \mathbf{x}] = \prod_{s=1}^{a-1} (1 - \lambda^d(t_s | \mathbf{x}(t_s))).$$

We first obtain the general relationship between the discrete- and continuous-time hazards. The discrete-time hazard is the probability of failure in $[t_{a-1}, t_a)$ divided by the probability of surviving to at least time t_{a-1} , so can be rewritten as

$$\lambda^d(t_a | \mathbf{x}) = \frac{S(t_{a-1} | \mathbf{x}) - S(t_a | \mathbf{x})}{S(t_{a-1} | \mathbf{x})}, \quad (17.38)$$

where $S(t | \mathbf{x})$ is the survivor function. In the continuous case $S(t | \mathbf{x}) = \exp(-\int_0^t \lambda(s) ds)$, and after some algebra (17.38) becomes

$$\lambda^d(t_a | \mathbf{x}) = 1 - \exp\left(-\int_{t_{a-1}}^{t_a} \lambda(s) ds\right). \quad (17.39)$$

Now specialize to the discrete-time hazard associated with the continuous PH model

$$\lambda(t) = \lambda_0(t) \exp(\mathbf{x}(t_{a-1})' \boldsymbol{\beta}),$$

for t in $[t_{a-1}, t_a)$. Note that the regressors are constant within the interval but can vary

across intervals, and $\lambda_0(t)$ can vary within the interval. Then (17.39) becomes

$$\begin{aligned}\lambda^d(t_a | \mathbf{x}) &= 1 - \exp(-\exp(\mathbf{x}(t_{a-1})' \boldsymbol{\beta}) \times \int_{t_{a-1}}^{t_a} \lambda_0(s) ds) \\ &= 1 - \exp(-\lambda_{0a} \exp(\mathbf{x}(t_{a-1})' \boldsymbol{\beta})) \\ &= 1 - \exp(-\exp(\ln \lambda_{0a} + \mathbf{x}(t_{a-1})' \boldsymbol{\beta})),\end{aligned}\quad (17.40)$$

where $\lambda_{0a} = \int_{t_{a-1}}^{t_a} \lambda_0(s) ds$. The associated discrete-time survivor function is

$$S^d(t_a | \mathbf{x}) = \prod_{s=1}^{a-1} \exp(-\exp(\ln \lambda_{0s} + \mathbf{x}(t_{s-1})' \boldsymbol{\beta})). \quad (17.41)$$

The density for the i th subject is the product of the survivor function in each period that the subject survives times the hazard at the time of failure. It follows from (17.40) and (17.41) that the likelihood is

$$\begin{aligned}L(\boldsymbol{\beta}, \lambda_{01}, \dots, \lambda_{0A}) &= \prod_{i=1}^N \left[\prod_{s=1}^{a_i-1} \exp(-\exp(\ln \lambda_{0s} + \mathbf{x}_i(t_{s-1})' \boldsymbol{\beta})) \right] \\ &\quad \times (1 - \exp(-\exp(\ln \lambda_{0a_i} + \mathbf{x}_i(t_{a-1})' \boldsymbol{\beta}))),\end{aligned}\quad (17.42)$$

where censoring is ignored for simplicity and failure is assumed to occur at time t_{a_i} for the i th observation. At least one failure is assumed to occur in each interval $[t_{a-1}, t_a]$.

The MLE maximizes (17.42) with respect to $\boldsymbol{\beta}$ and $\lambda_{01}, \dots, \lambda_{0A}$. In a special case partial likelihood is asymptotically equivalent to the MLE, though in general they differ. More parsimonious models place some structure on the $\lambda_{01}, \dots, \lambda_{0A}$, such as a polynomial in time. Even more structure is placed by a fully parametric model such as the Weibull, which sets $\lambda_{0s} = \int_{t_{a-1}}^{t_a} \alpha s^{\alpha-1} ds$.

17.10.2. Han and Hausman Approach

Han and Hausman (1990) suggested a flexible approach to recovering the baseline hazard that is relatively easy to implement and that predates the work of Blake et al. (1999) but has similarities with the work of Meyer (1990) and Sueyoshi (1992). It allows for considerable flexibility in the specification of the baseline hazard, $\lambda_0^d(t)$, while maintaining a parametric form (e.g., $\exp(\mathbf{x}' \boldsymbol{\beta})$) for the function of covariates. It also has the merit of explicitly dealing with discrete duration data and of providing a framework that can more easily accommodate features of discrete data such as tied observations and unobserved heterogeneity. Tied observations can be a major problem with discrete data; for example, with unemployment durations the termination of many unemployment spells is likely to coincide with the end of the period of unemployment benefits (usually 26 weeks in the United States).

The starting point is the hazard rate for observation i , $\lambda_i(\tau)$, denoting the conditional probability that a spell terminates in the interval $(\tau, \tau + \Delta)$ written in the PH form:

$$\lambda_i(\tau) = \lambda_0(\tau) \exp(-\mathbf{x}_i' \boldsymbol{\beta}),$$

where $\lambda_0(\tau)$ denotes the baseline hazard. Then (as shown in (17.20)) taking logs after integration and then rearranging yields

$$\Lambda_0(t) - \mathbf{x}'_i \boldsymbol{\beta} = \varepsilon_i, \quad (17.43)$$

where $\Lambda_0(t) = \ln \int_0^t \lambda_0(\tau) d\tau$ denotes the log of the integrated baseline hazard, and $\varepsilon_i = \ln \int_0^t \lambda_i(\tau) d\tau$. Then the probability is given by

$$\Pr[\text{failure in period } t] = \int_{\Lambda_0(t-1) - \mathbf{x}'_i \boldsymbol{\beta}}^{\Lambda_0(t) - \mathbf{x}'_i \boldsymbol{\beta}} f(\varepsilon) d\varepsilon.$$

Let $y_{it} = 1$ if the i th person experiences failure in period t , and $y_{it} = 0$ otherwise. Then the joint likelihood of N observations is given by

$$\ln L(\boldsymbol{\beta}, \Lambda_0(1), \dots, \Lambda_0(T)) = \sum_{i=1}^N \sum_{t=1}^T y_{it} \ln \left[\int_{\Lambda_0(t-1) - \mathbf{x}'_i \boldsymbol{\beta}}^{\Lambda_0(t) - \mathbf{x}'_i \boldsymbol{\beta}} f(\varepsilon) d\varepsilon \right], \quad (17.44)$$

and the baseline hazard parameters $(\Lambda_0(1), \dots, \Lambda_0(T))$ are estimated along with $\boldsymbol{\beta}$ in a flexible manner (i.e., without imposing a specific functional form).

The integral in the log-likelihood is of course the difference in the cdf $[\Lambda_0(t-1) - \mathbf{x}'_i \boldsymbol{\beta}, \Lambda_0(t) - \mathbf{x}'_i \boldsymbol{\beta}]$. The precise form of this expression depends on the functional form of the cdf. If the random ε_i are assumed to be standard normal distributed, the log-likelihood takes the ordered probit form; under the assumption of extreme value distribution the log-likelihood takes the ordered logit form. To be specific, under normality the integral in the i th term is of the form

$$\Pr[\Lambda_0(t) < \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i \leq \Lambda_0(t+1)] = \Phi(\Lambda_0(t+1) - \mathbf{x}'_i \boldsymbol{\beta}) - \Phi(\Lambda_0(t) - \mathbf{x}'_i \boldsymbol{\beta}).$$

In contrast to the partial likelihood approach, which treats the baseline hazard as a nuisance function and eliminates it, the approach of Han and Hausman (1990) estimates all the unknown parameters simultaneously at a modest computational cost. Their Monte Carlo results show that the method is flexible and can well approximate arbitrary hazard function, eliminating the need for strong functional form assumptions.

17.10.3. Discrete-Time Binary Choice

An alternative approach for discrete duration data is to use a binary choice model for transitions, since in each discrete time interval two outcomes are possible – the spell either ends or it does not.

A general formulation of a **discrete-time transition model** is

$$\Pr[t_{a-1} \leq T < t_a | T \geq t_{a-1} | \mathbf{x}] = F(\lambda_a + \mathbf{x}'(t_{a-1}) \boldsymbol{\beta}), \quad a = 1, \dots, A. \quad (17.45)$$

This specification restricts the coefficients of regressors to be constant over time, whereas the intercept λ_a , $a = 1, \dots, A$, can vary over time. The obvious choices of the function F are the standard normal cdf or the logistic cdf. Then the parameters λ_a and $\boldsymbol{\beta}$ can be estimated by a stacked logit or stacked probit model in which a separate intercept is permitted for each duration interval. This method is very appealing because of its simplicity.

The resulting likelihood function is

$$L(\beta, \lambda_1, \dots, \lambda_A) = \prod_{i=1}^N \left[\prod_{s=1}^{a_i-1} (1 - F(\lambda_s + \mathbf{x}'_i(t_{s-1})\beta)) \right] \times F(\lambda_{a_i} + \mathbf{x}'(t_{a_i-1})\beta).$$

This is similar to (17.42), the log-likelihood for discrete time PH model, aside from the choice of function F . The hazard (17.40) is the extreme value cdf evaluated at $\ln \lambda_{0a} + \mathbf{x}(t_{a-1})'\beta$, so (17.40) yields the complementary log-log model binary choice model (see Table 14.3) rather than the more commonly used logit or probit model.

17.11. Duration Example: Unemployment Duration

The following empirical application uses the data of McCall (1996), generously provided to us by the author Brian McCall. The data set is derived from the January Current Population Survey's Displaced Workers Supplements (DWS) for the years 1986, 1988, 1990, and 1992. We refer to the duration measure (spell) in this example as unemployment duration, though more accurately it represents joblessness duration since DWS does not provide information as to whether a person is looking for job or not.

For this application, information on the part-time or full-time status of the first postdisplacement job is required. To determine whether the first postdisplacement job was part-time or full-time, the following method is adopted. The first postdisplacement job is designated as part-time if a subject was still in that job at the time of the survey and if the subject was working less than 35 hours per week in that job in the previous week.

Table 17.6 defines the key economic covariates used to explain joblessness duration. The number of covariates in the models estimated is quite large, but in the interest of brevity only a subset is listed. McCall (1996) provides a fuller description.

Table 17.6. Unemployment Duration: Description of Variables

Variable Name	Variable Label	Mean
spell	periods jobless: two-week interval	6.248
CENSOR1	1 if reemployed at full-time job	0.321
CENSOR2	1 if reemployed at part-time job	0.102
CENSOR3	1 if reemployed but left job: pt–ft status unknown	0.172
CENSOR4	1 if still jobless	0.375
UI	1 if filed UI claim	0.553
RR	eligible replacement rate	0.454
DR	eligible disregard rate	0.109
TENURE	tenure years in lost job	4.114
LOGWAGE	log weekly earnings	5.693

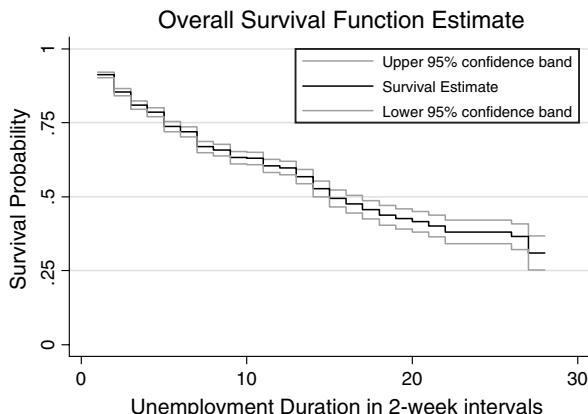


Figure 17.3: Unemployment duration: Kaplan-Meier estimate of survival function. U.S. data from 1986–92 on 3343 spells, some incomplete.

Unemployment durations have been measured in two-week intervals. Four binary variables (CENSOR1, CENSOR2, CENSOR3, and CENSOR4) have been introduced to indicate the status of the first postdisplacement job. For the analysis in this chapter we use CENSOR1. Thus a spell is complete if person is re-employed at a full-time job. Another indicator variable UI is used to denote whether the subject filed an unemployment claim or not. Replacement rate, which is the weekly benefit amount divided by the amount of weekly earnings in the lost job, is represented by the variable RR. “Disregard” is defined to be the threshold amount up to which recipients of unemployment insurance who accept part-time work can earn without any reduction in unemployment benefits. Disregard rate is the disregard divided by weekly earnings in the lost job. It is described by the variable DR in this example. As we can see, all the other variables are self-explanatory.

We begin with a descriptive analysis of the duration data. The simplest first step is to plot the Kaplan–Meier survival curve, which is shown in Figure 17.3 by the dark line. The lighter lines around the estimated Kaplan–Meier survival curve represent 95% confidence intervals developed in Section 17.5.2. As expected, the estimated survival curve declines rapidly at first and then slowly.

As we see from Table 17.7, after the first period the survival probability is 0.91, indicating that roughly 9% of the sampled individuals have terminated their spell within the first two weeks of beginning joblessness spell.

In Figure 17.4, we plot the survival function by UI, that is, by whether the subject claims unemployment insurance or not. Again, as one can expect, it shows that those who claim unemployment insurance are more likely to remain unemployed than those who do not claim unemployment insurance.

The Nelson–Aalen cumulative hazard in Figure 17.5 shows little variation in the hazard rate, which translates into an approximately linear hazard. If the crude hazard rate varies a lot, then the cumulative hazard would appear nonlinear.

Table 17.7. Unemployment Duration: Kaplan–Meier Survival and Nelsen–Aalen Cumulated Hazard Functions

Time	Survivor Function	Cumulative Hazard
1	0.9121	0.0879
2	0.8541	0.1514
3	0.8103	0.2027
4	0.7864	0.2322
5	0.7376	0.2943
⋮	⋮	⋮
12	0.5974	0.5005
13	0.5680	0.5496
14	0.5270	0.6219
⋮	⋮	⋮
26	0.3651	0.9809
27	0.3098	1.1325
28	0.3098	1.1325

The cumulated hazard functions by UI recipiency, shown in Figure 17.6, exhibit the expected pattern: The hazard increases at a higher rate for those who do not claim unemployment insurance than it does for those who do.

Next we consider four parametric regression models using the covariates UI, RR, DR, and LOGWAGE and the interaction terms RRUI and DRUI. The four models are exponential, Weibull, Gompertz, and Cox PH. Writing the hazard function as

$$\lambda(t|\mathbf{x}) = \lambda_0(t, \alpha)\phi(\mathbf{x}, \boldsymbol{\beta}) = \lambda_0(t, \alpha) \exp(\mathbf{x}'\boldsymbol{\beta}),$$

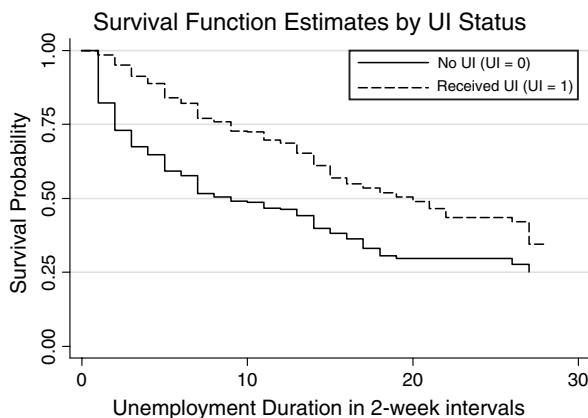


Figure 17.4: Unemployment duration: estimated survival functions by whether or not subjects receive unemployment insurance. Same data as Figure 17.3.

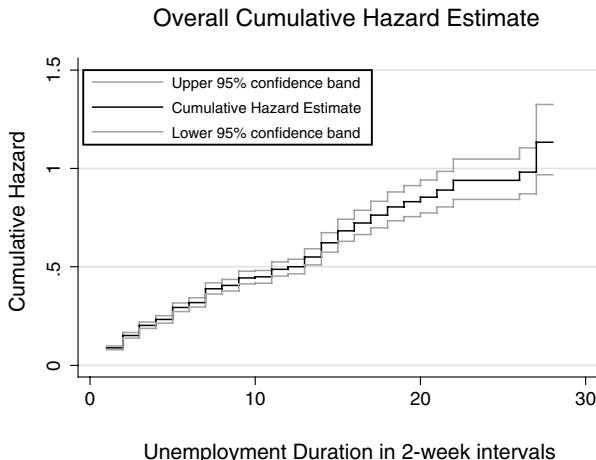


Figure 17.5: Unemployment duration: Nelson-Aalen estimate of cumulative hazard function. Same data as Figure 17.3.

recall that exponential hazard assumes $\lambda_0(t, \alpha) = \text{constant} = \exp(a)$ for some constant a , the Weibull model assumes $\lambda_0(t, \alpha) = \exp(a)\alpha t^{\alpha-1}$ (i.e., monotonic hazards), Gompertz assumes $\lambda_0(t, \alpha) = \exp(a)\exp(\gamma t)$, and the Cox PH model has no intercept and makes no assumption about the shape of the baseline hazard. Recall also that the formulation here is of the proportional hazard type and can also be interpreted either as a parametric regression model or as an AFT model. In this parameterization of the likelihood function, the parameters (α, β) are estimated. These are given in Table 17.8 with the associated t -statistics. We also list the negative of the log-likelihood, but recall that for the Cox PH model it is the partial log-likelihood. Both exponential and Gompertz models fit equally well. The Weibull model provides the

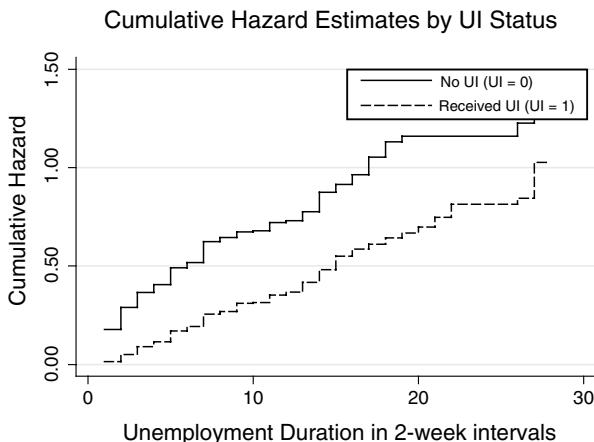


Figure 17.6: Unemployment duration: estimated cumulative hazard functions by whether or not receive unemployment insurance. Same data as Figure 17.3.

Table 17.8. Unemployment Duration: Estimated Parameters from Four Parametric Models

Var	Exponential		Weibull		Gompertz		Cox PH	
	coeff.	t	coeff.	t	coeff.	t	coeff.	t
RR	0.472	0.79	0.448	0.70	0.472	0.78	0.522	0.91
DR	-0.576	-0.75	-0.427	-0.53	-0.563	-0.74	-0.753	-1.04
UI	-1.425	-5.71	-1.496	-5.67	-1.428	-5.69	-1.317	-5.55
RRUI	0.966	0.92	1.105	1.57	0.969	1.58	0.882	1.52
DRUI	-0.199	-0.20	-0.299	-0.28	-0.211	-0.21	-0.095	-0.10
LOGWAGE	0.35	3.03	0.37	2.99	0.35	3.03	0.34	3.03
CONS	-4.079	-4.65	-4.358	-4.74	-4.097	-4.65	-	-
α		1.129						
$-\ln L$	2700.7		2687.6		2700.6		-	

best fit. As we see from Table 17.8, the fit of the Weibull model exhibits positive state dependence ($\alpha = 1.129 > 1$); that is, the probability of the spell terminating increases as the spell lengthens.

For all the models considered, only UI and LOGWAGE are significant whereas other covariates are not. The estimated coefficient of UI is negative for all models, implying that the joblessness spell of those who claim unemployment insurance terminates slower. There is little variation of the estimates of UI across different models: This estimate in Weibull and Gompertz models is approximately 5% and 0.2% higher in absolute value than that in the exponential model, whereas it is 8% lower in the Cox PH model. Similarly, the estimate of the coefficient of LOGWAGE is positive for all the models and exhibits very little variation across models.

Whereas in the econometric literature it is common to report the estimate of (α, β) coefficients of the hazard function in AFT metric, in the biostatistics literature a different parameterization is often used based on the PH metric. Note that the hazard ratio $\lambda(t|\mathbf{x})/\lambda_0(t, \alpha) = \phi(\mathbf{x}, \beta) = \exp(\mathbf{x}'\beta)$. For a categorical 0/1 scalar variable x , the impact of a change from 0 to 1 is given by $\exp(\beta) - 1$, which measures impact relative to the baseline hazard. Numerous packages give the users an option to estimate the model in either or both metrics. The relative merits of the two parameterization are discussed in Cleves, Gould, and Guitirrez (2002).

Consider the exponential specification in Table 17.9 where the coefficients are exponentials of the corresponding ones Table 17.8. Here UI has hazard ratio 0.241. This means that belonging to the category of subjects that claims unemployment insurance decreases the hazard by nearly 76% over the baseline hazard. Similarly, for Weibull, Gompertz, and Cox PH models, the hazard decreases by about 78%, 76%, and 73%, respectively.

For this example, we have taken into account right-censoring and have ignored the role of unobserved heterogeneity. Hence the results obtained from the three models are qualitatively similar. However, the relatively few included variables with significant

Table 17.9. Unemployment Duration: Estimated Hazard Ratios from Four Parametric Models

Var	Exponential		Weibull		Gompertz		Cox PH	
	β	t	β	t	β	t	β	t
RR	1.603	0.63	1.565	0.57	1.604	0.62	1.686	0.71
DR	0.562	-1.02	0.653	-0.66	0.570	-0.99	0.471	-1.55
UI	0.241	-12.65	0.224	-13.12	0.240	-12.65	0.268	-11.53
RRUI	2.626	1.01	2.760	0.99	2.635	1.01	2.416	1.01
DRUI	0.819	-0.22	0.742	-0.33	0.810	-0.23	0.909	-0.10
LOGWAGE	1.420	2.56	1.441	0.08	1.42	2.55	1.40	2.57
α			1.129					
$-\ln L$	2700.7		2687.6		2700.6		-	

coefficients probably indicates that large unexplained variation (perhaps caused by unobserved heterogeneity) may be a serious problem. This issue is considered further in the next chapter.

17.12. Practical Considerations

Most computer packages offer a good selection of computer programs for parametric survival analysis. Standard nonparametric Kaplan–Meier survival function estimates, with or without confidence intervals, with both numeric and graphic output are widely available. In some cases survival analysis modules are sufficiently detailed to warrant a special manual. For example, Allison (1995) offers a practical guide to survival analysis in the SAS system; Cleves et al. (2002) provide a tutorial style guide to survival analysis in STATA. Not only do these guides explain the mechanics of implementing particular program commands, but in many cases they provide insightful expositions of the subtleties arising from specific features of data, alternative parameterizations, and interpretation of results. A convenient way to learn about duration data analysis is by using the examples in econometrics or statistical packages such as LIMDEP, STATA, SAS, or S-Plus. The program manuals are themselves excellent sources of information for standard models.

17.13. Bibliographic Notes

17.3–17.7 Kalbfleisch and Prentice (1980, 2002) is the classic statistical reference for survival analysis, with emphasis on the Cox model. Other useful sources include Lawless (1982) and Cox and Oakes (1984) and the considerable number of statistical texts on survival analysis that now exist. For a Bayesian treatment see Ibrahim, Chen, and Sinha (2001). Recent statistical work has increasingly emphasized the counting process approach, detailed in Fleming and Harrington (1991) and Andersen et al. (1993).

These references are very challenging, especially the latter. Lancaster (1990) provides a thorough treatment of survival analysis, though the presentation is quite technical and the book is oriented more to the general topic of transitions and material presented in the subsequent two chapters. For social scientists, Allison's (1984) excellent exposition, like that of Lancaster, covers much more than single-spell survival analysis. For practitioners in microeconomics the survey article by Kiefer (1988) is a good start.

- 17.8** Lancaster (1990) provides a thorough discussion of the partial likelihood approach.
- 17.10** Meyer (1990), Han and Hausman (1990), and Blake et al., (1999) are helpful references on discrete hazard models. These articles generally allow for unobserved heterogeneity, a topic discussed in the next chapter.
- 17.11** Economics applications are cited in Kiefer (1988) and in Greene (2003). Good examples of parametric reduced-form type duration analysis are given by Lancaster (1979), Narendranathan, Nickell, and Stern (1985), Jaggia (1991c), and Gritz (1993). More recently the emphasis has shifted to computationally more complex structural duration models. Examples are found in Van den Berg (1990) and Ferall (1997). Most applications of duration analysis are reduced-form models. Economists have proposed structural duration models; references include Lancaster (1990) and Van den Berg (2001). Van den Berg also provides an interesting discussion of the economic theoretical foundations of the PH model. Duration data can often be analyzed using different concepts of waiting time. Tunali and Pritchett (1997) use three alternatives: calendar-time, age, and duration.

Exercises

- 17–1** (Adapted from Sapra, 1998) Show that the duration data model with Pareto density of the first kind $f(t) = \alpha k^\alpha / t^{\alpha+1}$, $\alpha > 0$, $t \geq k \geq 0$, is an accelerated failure time duration model but is not a proportional hazards model. [Hint: Show that $\ln t$ can be expressed as a linear regression in $k = \exp(\mathbf{x}'\beta)$ with an additive homoskedastic error.]
- 17–2** (Based on Lancaster, 1979). For each of the following situations develop an appropriate expression for the joint likelihood of N observations in terms of the duration density $f(t|\mathbf{x}, \theta)$ and survivor function $S(t|\mathbf{x}, \theta)$.
- (a) A sample of independent completed durations, t_i , $i = 1, \dots, N$, is available.
 - (b) The sample is generated as follows. Initially, individuals are selected from a pool of unemployed and interviewed. Subsequently, they are reinterviewed after h periods. Selected individuals have been unemployed for t weeks on selection. Between selection and interview some find jobs, and others do not. For those who have jobs the time of termination of unemployment spells is known.
 - (c) The situation is the same as in (b) except that it is not known when the unemployment spell ended.
- 17–3** (a) Using a 50% random sample of the McCall data set estimate the Kaplan–Meier nonparametric survival and integrated hazard function estimates by type of censoring, that is, by whether transition is to full-time or part-time

employment. Does the survival function look significantly different for the two groups?

- (b) Ignoring the censoring variable for type of spell termination, estimate the hazard model for unemployment duration under the following parametric distributional assumptions: (i) exponential, (ii) Weibull, (iii) log-logistic, and (iv) Cox PH. Use the same covariates as in this chapter.
- (c) Compare models (i)–(iii) and discuss which one you think provides the best fit to the data. What does each model imply regarding the duration dependence (shape of the hazard function) of a spell of unemployment?

Mixture Models and Unobserved Heterogeneity

18.1. Introduction

There is a large statistical and econometric literature concerning the topic of unobserved heterogeneity. Observed heterogeneity refers to interindividual differences that are measured by regressors, and unobserved heterogeneity refers to all other differences. Both factors affect survival times. In the presence of unobserved heterogeneity even individuals with the same values of all covariates may have different hazards out of a given state. When unobserved heterogeneity is ignored, its impact is confounded with that of the baseline hazard.

To motivate further study consider a well-known empirical example. The aggregate hazard rate out of unemployment is known to be a declining function of the length of unemployment spell. If all individuals were identical then this would imply negative duration dependence, that is, a falling probability of escaping unemployment the longer an individual has remained unemployed. However, suppose that there are two types of individuals in the unemployed population, type F (fast), who have a constant hazard rate of 0.4, and type S (slow), whose constant hazard rate is 0.1. The population is a 50/50 mixture of the two types. Then for 100 type F people we observe 40 transitions in the first period, 24 transitions in the second period, and 14.4 in the third. For the type S, we observe 10, 9, and 8.1 transitions in the first, second, and third periods, respectively. Hence the aggregate proportion of transitions will be $(40 + 10)/200 = 0.25$, $(24 + 9)/150 = 0.22$, and $(14.4 + 8.1)/117 = 0.192$. This shows that the declining aggregate hazard is a consequence of aggregation across heterogeneous groups, which themselves have constant but different hazard rates. Accurate statements about duration dependence require that models incorporate unobserved heterogeneity.

In linear regression models there are no complications caused by unobserved heterogeneity if the heterogeneity is independent of regressors. In that case the conditional mean is unchanged, the unobserved heterogeneity is absorbed into the error term, and there is no omitted variables bias. In contrast, unobserved heterogeneity usually causes problems in durations models. In the simplest models, such as the exponential model, it is possible to specify multiplicative unobserved heterogeneity uncorrelated

with regressors that leaves the conditional mean duration unchanged. However, even in this simple case the conditional hazard function does change, and it is the hazard that is modeled out of necessity, given the presence of censoring and given also, for example, the interest of policy makers in determining how exit rates from unemployment vary with length of unemployment spell.

The role of unobserved heterogeneity lies at the heart of numerous empirical puzzles and conundrums. Although our focus in this chapter is in the context of duration models, most of the issues are of more general relevance. The material and techniques used here are also relevant to all econometric models, since all econometric models omit some individual-specific unobservable variables from the model. Leading examples in other chapters include random parameters logit (Section 15.7), sample selection (Section 16.4), finite mixture for counts (Section 20.4) and fixed and random effects models for panel data (Chapters 21–23). These factors go under the collective heading of unobserved heterogeneity. In biostatistics the term **frailty** is also used. In actuarial studies (multiplicative) unobserved heterogeneity measures proportional increase or decrease in the hazard rate (“force of mortality”) operating on a given individual relative to that on an average individual. Individual-specific heterogeneity need not be time-invariant, but in cross section models it is convenient to assume it is.

It is important to consider the consequences of such an unavoidable misspecification. From ordinary linear multiple regression analysis it is known that such an omission in general can lead to an omitted variable bias. In duration models, which are nonlinear, the analysis of unobserved heterogeneity is more complex. Introduction of unobserved heterogeneity leads to an important class of models called **mixture models**, this being one of the many names for this class. The subject matter of this chapter concerns both the techniques for generating and analyzing mixture models and the substantive consequences of omitted heterogeneity.

Distinguishing between heterogeneity and true state dependence has been a long-standing issue that can be traced back in history to discussions concerning true and apparent **contagion**. Neyman has been credited for his early insight that longitudinal data may be essential to make this distinction empirically possible. When, however, only cross-section data are available, there is a tendency to rely heavily on strong parametric assumptions. The emphasis in the recent literature has been to free empirical analysis from such assumptions and on testing the validity of maintained model assumptions.

The first part of this chapter, Sections 18.2–18.4, deals with mixture models based on continuous distribution of heterogeneity. Section 18.5 presents models based on discrete heterogeneity. Section 18.6 considers relationships among different duration concepts from flow and stock data. Tests of misspecification and neglected heterogeneity are dealt with in Section 18.7. An empirical example in Section 18.8 illustrates several of the ideas developed in the chapter.

18.2. Unobserved Heterogeneity and Dispersion

In this section we focus on unobserved heterogeneity in the exponential and Weibull models. We consider a form of multiplicative unobserved heterogeneity that, after

being integrated out, leaves the conditional mean unchanged but does inflate the conditional variance and, more importantly, changes the conditional hazard function. The popular Weibull model with gamma distributed heterogeneity is also presented.

18.2.1. Mixtures

The simplest model to consider is the exponential duration model. In an exponential regression without heterogeneity the distribution of complete spells, t_i , is specified conditional on observable weakly exogenous covariates \mathbf{x}_i . This is equivalent to specifying the conditional mean function as nonstochastic: $E[T|\mathbf{x}] = \exp(\mathbf{x}'\boldsymbol{\beta})$. In mixture models we instead specify the distribution of $(t_i|\mathbf{x}_i, v_i)$, where the additional v_i denotes an unobserved heterogeneity term for observation i . Simply, individuals are assumed to differ randomly in a manner not fully accounted for by the observed covariates. The marginal distribution of t_i is obtained by averaging with respect to v_i .

The precise functional form linking t_i and (\mathbf{x}_i, v_i) must be specified. A commonly used functional form is the exponential mean with a *multiplicative* error. For example, consider the PH model with unobserved heterogeneity. From Section 17.8 we have the proportional hazards model, (17.25) and (17.26), which can be extended to include a multiplicative term v . That is,

$$\lambda(t|\mathbf{x}, v) = \lambda_0(t) \exp(\mathbf{x}'\boldsymbol{\beta})v, \quad v > 0,$$

and hence we can obtain an expression for integrated baseline hazard as follows:

$$\begin{aligned} \lambda_0(t) &= \lambda(t|\mathbf{x}, v) \exp(-\mathbf{x}'\boldsymbol{\beta})v^{-1}, \\ \int \lambda_0(u)du &= \exp(-\mathbf{x}'\boldsymbol{\beta})v^{-1} \int \lambda(u|\mathbf{x}, v) du, \\ \ln \left[\int \lambda_0(u)du \right] &= -\mathbf{x}'\boldsymbol{\beta} - \ln v + \varepsilon, \end{aligned} \tag{18.1}$$

where $\varepsilon = \ln \int \lambda(u|\mathbf{x}, v) du$, and v is assumed to be independent of the regressors and of censoring time. A common normalization restriction is $E[v] = 1$. When $v > 1$, the hazard rate is greater than for the average subject; it is less than that for the average subject if $v < 1$. The independence assumption is strong and not necessarily realistic. The multiplicative heterogeneity assumption is also rather special, but it is mathematically convenient and more attractive than an additive error, which could violate nonnegativity of t_i . A standard approach involves postulating a distribution for v_i , and then deriving the marginal distribution of t_i .

Multiplicative heterogeneity has two important and related consequences. Not surprisingly, the variance of the mixture (conditional on the observable variables) exceeds the variance of the parent distribution (conditional on both the observables and heterogeneity). That is, the variance gets inflated. Consider the exponential mean case.

Replace $\mu_i = \exp(\mathbf{x}'_i \boldsymbol{\beta})$ by

$$\begin{aligned}\mu_i^* &= E[t_i | \mathbf{x}_i, v_i] \\ &= \exp(\mathbf{x}'_i \boldsymbol{\beta}) v_i \\ &= \exp(\mathbf{x}'_i \boldsymbol{\beta}) \exp(\varepsilon_i) \\ &= \exp(\beta_0 + \varepsilon_i + \mathbf{x}'_{1i} \boldsymbol{\beta}_1),\end{aligned}\tag{18.2}$$

where the unobserved heterogeneity term v_i is redefined as $\exp(\varepsilon_i)$ in the third line, and the term $\mathbf{x}'_i \boldsymbol{\beta}$ is broken into the intercept and slope terms in the last line. The last line has an interpretation as a conditional mean with a randomly varying intercept ($\beta_0 + \varepsilon_i$). It is usually assumed that v_i s are iid, possibly with a known parametric distribution, and that they are independent of the \mathbf{x}_i .

Assume that v_i is iid with $E[v_i] = 1$ and $V[v_i] = \sigma_v^2$. The assumption that $E[v_i] = 1$ permits identification of the intercept. For the exponential density, the moments of t_i can be derived as $E[t_i | \mathbf{x}_i, v_i] = \mu_i v_i$, and using Section A.8 result on variance decomposition,

$$\begin{aligned}V[t_i | \mathbf{x}_i] &= V_v[E_{t|v,\mathbf{x}}(t_i | v_i, \mathbf{x}_i)] + E_v[V_{t|v,\mathbf{x}}(t_i | v_i, \mathbf{x}_i)] \\ &= \mu_i^2 V(v_i) + \mu_i^2 (V(v_i) + 1) \\ &= \mu_i^2 [1 + 2\sigma_v^2] \\ &> \mu_i^2.\end{aligned}\tag{18.3}$$

The unconditional variance is inflated by unobserved heterogeneity.

18.2.2. Choice of Heterogeneity Distribution

Consider how the distribution of t is affected by heterogeneity. This requires us to look at the marginal distribution of t_i by integrating out the heterogeneity term, v , from $S(t | \mathbf{x}, v)$. A parametric distribution of v is usually specified. What considerations apply to choosing this distribution?

To respect the property $v_i > 0$, we may specify a distribution with support on the positive line. Examples are gamma, inverse Gaussian, and log-normal.

The **gamma density** is

$$g(v; \delta, k) = \frac{\delta^k v^{k-1} \exp(-\delta v)}{\Gamma(k)}, \quad v > 0,\tag{18.4}$$

which has $E[v] = k/\delta$ and $V[v] = k/\delta^2$. Normalization sets $k = \delta$, $E[v] = 1$, and $V[v] = 1/\delta$. The gamma assumption is mathematically convenient. It is also employed in a number of popular software packages for duration modeling.

The **inverse-Gaussian density** is

$$g(v; \delta, \theta) = \delta \pi^{-1/2} \exp(2\delta\theta^{1/2}) v^{-3/2} \exp(-\theta v - \delta^2/v), \quad v > 0,\tag{18.5}$$

which has $E[v] = \delta\theta^{-1/2}$ and $V[v] = \delta\theta^{-3/2}/2$. Normalization $\theta = \delta^2$ yields $E[v] = 1$, and $V[v] = 1/2\theta$. Relative to the gamma the inverse-Gaussian distribution has more tail probability.

These will not necessarily produce an analytically tractable marginal distribution of t . As we will show, some combinations such as exponential and gamma, or Weibull and gamma, lead to closed-form marginals, whereas others do not. However, this consideration is one of mathematical and computational convenience only and hence is not necessarily compelling on its own. Unfortunately, one rarely has guidance from economic theory on this aspect of duration modeling.

A second consideration is generality and flexibility. The gamma model is quite flexible and has many attractive properties. However, the inverse-Gaussian may better handle heavy-tailed distributions. Both of these are one-parameter families (after normalization). Hougaard (1986) introduced a more flexible two-parameter family that has gamma and inverse-Gaussian as special cases. Later in this chapter we consider a discrete (nonparametric) representation that also affords considerable flexibility.

18.2.3. Weibull–Gamma Mixture

Next we consider the popular **Weibull–gamma mixture**, which can be specialized to the exponential–gamma case. This model is a leading special case of a **mixed proportional hazards** (MPH) model. The Weibull–gamma mixture is, of course, of independent interest because of its greater generality, and especially because it will be shown to encompass both increasing and decreasing hazards.

The survivor function conditional on multiplicative v for the Weibull model is

$$S(t|v) = \exp(-\mu t^\alpha v), \quad \lambda > 0, \alpha > 0, \quad (18.6)$$

where μ replaces α used in Chapter 17.

The unconditional survivor function is given by the average survivor function. Averaging across the heterogeneous population using the density of v , $g(v)$, as the weighting function yields,

$$S(t) = E_v[S(t|v)] = \int S(t|v)g(v)dv. \quad (18.7)$$

Different choices of $g(v)$ lead to different mixtures. With appropriate changes in interpretation both continuous and discrete distributions are valid. The integral in (18.7) may not have an analytical solution. For example, if $g(v)$ is the log-normal density the integral does not have an analytical solution but if it is a gamma distribution it does. For mathematical convenience we work with the gamma case in what follows.

Given gamma heterogeneity the unconditional survivor function is

$$\begin{aligned} S(t) &= \int_0^\infty \exp(-\mu t^\alpha v) \frac{\delta^k v^{k-1} \exp(-\delta v)}{\Gamma(k)} dv \\ &= \frac{\delta^k}{\Gamma(k)} \int_0^\infty v^{k-1} \exp(-v(\mu t^\alpha + \delta)) dv. \end{aligned} \quad (18.8)$$

To obtain the mixture density we solve the integral. Letting $\mu t^\alpha + \delta = \beta$, we get

$$S(t) = \frac{\delta^k}{\Gamma(k)} \int_0^\infty \frac{(v\beta)^{k-1}}{\beta^{k-1}} \exp(-v\beta) dv.$$

Define $y = \nu\beta$, so that $d\nu = \beta^{-1}dy$ and

$$\begin{aligned} S(t) &= \frac{\delta^k}{\Gamma(k)\beta^k} \int_0^\infty y^{k-1} \exp(-y) dy \\ &= \frac{\delta^k}{\Gamma(k)} \frac{\Gamma(k)}{(\mu t^\alpha + \delta)^k} \\ &= \delta^k (\mu t^\alpha + \delta)^{-k} \\ &= [1 + (\mu t^\alpha / \delta)]^{-k}, \end{aligned} \tag{18.9}$$

where the second line is obtained using the definition of $\Gamma(k)$ and substituting for β .

The unconditional duration density function is obtained by differentiating with respect to t and multiplying by -1 , which yields

$$f(t) = \frac{k}{\delta} \mu \alpha t^{\alpha-1} [1 + (\mu t^\alpha / \delta)]^{-(k+1)}. \tag{18.10}$$

The unconditional hazard function $\lambda(t) = f(t)/S(t)$ is given by

$$\lambda(t) = \frac{k}{\delta} \mu \alpha t^{\alpha-1} [1 + (\mu t^\alpha / \delta)]^{-1}. \tag{18.11}$$

These general expressions can be specialized by setting the mean of ν at 1; that is, set $k = \delta$, which normalizes $E[\nu] = 1$, and leads to the following expressions for the *Weibull–gamma mixture*:

$$S(t) = [1 + (\mu t^\alpha / \delta)]^{-\delta}, \tag{18.12}$$

$$f(t) = -\frac{\partial S(t)}{\partial t} = \mu \alpha t^{\alpha-1} [1 + (\mu t^\alpha / \delta)]^{-(\delta+1)}, \tag{18.13}$$

$$\lambda(t) = -\frac{\partial \ln S(t)}{\partial t} = \mu \alpha t^{\alpha-1} [1 + (\mu t^\alpha / \delta)]^{-1}, \tag{18.14}$$

which tends to the Weibull hazard as the variance $1/\delta$ goes to zero.

The Weibull model permits either increasing or decreasing hazards but somewhat restrictively assumes conditionally monotonic hazards at the individual level. Yet this mixture distribution has been popular in the econometrics literature, mainly because of its convenient properties; see Lancaster (1979) and Narendranathan, Nickell, and Stern (1985).

To specialize the results to the **exponential–gamma mixture** set $\alpha = 1$. This yields $S(t) = [1 + (\mu t / \delta)]^{-\delta}$, $f(t) = \mu [1 + (\mu t / \delta)]^{-(\delta+1)}$, and $\lambda(t) = \mu [1 + (\mu t / \delta)]^{-1}$. The exponential–gamma mixture, also known as the **Pareto distribution** of the second kind, has more mass in the tails relative to the exponential. The difference between the two depends on the variance, $1/\delta$. The r th moment exists only if $\delta > r$.

18.2.4. Interpreting the Mixture Hazard Function

An important issue in economic applications is whether positive or negative duration dependence is present in duration data. For example, does the probability of exiting from unemployment increase (e.g., owing to worker reservation wage falling) or decrease (e.g., owing to the worker being viewed as damaged goods) as the length of the unemployment spell increases? In the iid case this can be easy to establish by nonparametric estimation methods. With non-iid data, however, a decreasing hazard in the raw data may be due to aggregating across different individuals, each of whom has a different constant hazard rate, or to a decreasing hazard for each individual. Distinguishing between the two can be difficult.

Consider the problem of interpreting the hazard function in the presence of unobserved heterogeneity in the exponential–gamma mixture. Notice that even if individual hazard (i.e., hazard conditional on v) is constant at μ , the average or aggregate hazard $\lambda(t)$ is declining in t . This does not mean that there is negative duration dependence in the individual hazard rate. Rather, it is the effect induced by aggregation across individuals who differ randomly in their hazard rates. A similar erroneous interpretation can occur in the Weibull–gamma case. In that case the actual slope of the hazard function depends on α , but the slope of the average or aggregate hazard function is affected by the presence of heterogeneity. Thus the neglect of unobserved heterogeneity may lead to underestimation of the slope of the hazard function. This result seems fairly general (see Lancaster, 1990). Salant (1977) provided an early extensive discussion of this phenomenon.

This result is the basis of the claim (see, for example, Lancaster, 1979; Heckman and Singer, 1984a) that the estimation of hazard function in the presence of neglected unobserved heterogeneity may lead to serious biases. Our discussion motivates tests of unobserved heterogeneity in hazard models. Let us examine the argument in the context of the Weibull mixture model for which $S(t) = \int \exp(-\mu t^\alpha v) g(v) dv$. The aggregate hazard function is

$$\begin{aligned}\lambda(t) &= - \int \frac{\partial \ln S(t|v)}{\partial t} g(v) dv \\ &= \alpha \mu t^{\alpha-1} \int \frac{v \exp(-\mu t^\alpha v)}{S(t|v)} g(v) dv \\ &= \alpha \mu t^{\alpha-1} E[v|T \geq t].\end{aligned}$$

Because $E[v|T \geq t]$ is the average of v over those surviving at time t , it must decrease with time as individuals with higher values of v leave the state sooner than individuals with low values of v . This changes the slope of the aggregate hazard function. This phenomenon can also be thought of as a form of **selectivity bias** (Chapter 16.5). Formally, the average of v over time can be written as

$$E[v|T \geq t] = \int \frac{v \exp(-\mu t^\alpha v)}{S(t|v)} g(v) dv.$$

Therefore, for the Weibull mixture model

$$\begin{aligned}
 \frac{\partial E[v|T \geq t]}{\partial t} &= -\alpha \mu t^{\alpha-1} \left[\int \frac{v^2 \exp(-\mu t^\alpha v)}{S(t|v)} g(v) dv \right] \\
 &\quad + \alpha \mu t^{\alpha-1} \left[\int \frac{v \exp(-\mu t^\alpha v)}{S(t|v)} g(v) dv \right]^2 \\
 &= -\alpha \mu t^{\alpha-1} \{ E[v^2|T \geq t] - (E[v|T \geq t])^2 \} \\
 &= -\alpha \mu t^{\alpha-1} V[v|T \geq t] \\
 &< 0.
 \end{aligned} \tag{18.15}$$

Hence, neglecting heterogeneity results in an estimated hazard rate that is falling faster or rising more slowly than the actual hazard rate.

Another interesting comparison between models with and without heterogeneity is the proportional impact of a change in a covariate on the hazard rate. In the absence of heterogeneity

$$\ln \lambda(t|\mu) = \ln(\mu t^{\alpha-1}) + \ln \alpha,$$

and the proportional impact of a change in x_j on μ is

$$\frac{\partial \ln \lambda(t|\mu)}{\partial x_j} = \beta_j,$$

which is a property of the proportional hazard model.

Allowing for unobserved heterogeneity

$$\begin{aligned}
 \ln \lambda(t|\mu) &= \ln(\mu t^{\alpha-1}) + \ln \alpha + \ln E[v|T \geq t] \\
 &= \ln \alpha + \ln \mu + (\alpha - 1) \ln t + \ln E[v|T \geq t],
 \end{aligned}$$

whence, noting that $\ln \mu = \mathbf{x}'\boldsymbol{\beta}$ and $\partial E[v|T \geq t]/\partial x_j = -\mu t^\alpha V[v|T \geq t] \beta_j$, it follows that for the Weibull mixture model

$$\begin{aligned}
 \frac{\partial \ln \lambda(t|\mu, v)}{\partial x_j} &= \beta_j \left[1 - \frac{\mu t^\alpha V[v|T \geq t]}{E[v|T \geq t]} \right] \\
 &< \beta_j.
 \end{aligned} \tag{18.16}$$

The result shows that given heterogeneity the proportional impact of a change in x_j is smaller and depends on t and is no longer of the proportional hazard type. Thus, the estimates derived from the model may be misleading even if the unobserved heterogeneity term is uncorrelated with the included covariates.

Similar consequences of unobserved heterogeneity for models more general than the Weibull are discussed in Lancaster and Nickell (1980).

18.3. Identification in Mixture Models

Associated with mixture models is a general **identification problem**. This issue concerns the logical possibility of decomposing the individual contributions to the average

survival probability of the baseline hazard, the unobserved heterogeneity, and the covariates, given the observed data (t, \mathbf{x}) pertaining to a single spell. More specifically, if the PH model were not identified, then it would be logically impossible to separate the individual contributions of duration dependence and unobserved heterogeneity. As in most discussions of identification, some restrictions are placed on the formulation. In econometric literature the case of (mixed) proportional hazards has been investigated in detail. Heckman and Singer (1984b) and Elbers and Ridder (1982) have established the identification of the MPH model under certain conditions. Van den Berg (2001) provides an excellent discussion of these earlier proofs as well as later contributions.

Discussions of identifiability of the MPH model begin with the **average or aggregate survivor function**

$$\begin{aligned} S(t|\mathbf{x}) &= E_v [S(t|\mathbf{x}, v)] \\ &= \int \exp(-v \Lambda_0(t)\phi(\mathbf{x})) g(v) dv, \end{aligned} \tag{18.17}$$

which assumes proportionality of hazards as in (18.1), uses the PH formulation of Section 17.8, but does not make parametric assumptions on Λ_0 , ϕ , or g . Here $\Lambda_0(t) = \int_0^T \lambda_0(s) ds$. The model is said to be nonparametrically identified if, given the data, the functions λ_0 , g , and ϕ are unique. We add the qualifier “nonparametrically” because of the absence of functional form assumptions.

Variations in observed survival times are due to variations in the covariates \mathbf{x} , in v , and in the duration dependence function (baseline hazard). Identifiability means a unique decomposition of the variation. A proof of identifiability must show that these separate contributions are in principle identifiable. Most of the available proofs use advanced mathematical tools to show that the likelihood function can be uniquely decomposed. Melino and Sueyoshi (1990) provide a simpler proof.

The conditions required for nonparametric identification include the following:

- (i) The heterogeneity term v is assumed to be time-invariant and independently distributed of \mathbf{x} .
- (ii) $g(v)$ is nondegenerate and has finite mean (i.e., $E[v] < \infty$).
- (iii) $\phi(\mathbf{x}) > 0$ for all \mathbf{x} .
- (iv) $\lambda_0(t)$ is continuous and positive on $[0, \infty)$.
- (v) Observed explanatory variables \mathbf{x} are linearly independent and have sufficient variation. Different proofs have some subtle variation on these conditions but we will not delve into these here.

Whereas the issue of nonparametric identification involves considerable mathematical subtleties, the problem is also relevant in the context of parametric models. If one specifies parametric forms such as $\lambda_0(t|\alpha)$, $\phi(\mathbf{x}|\beta)$, and $g(v|\gamma)$, then are these functions unique given the data? The answer, unfortunately, may be “no” in many cases. This means that one investigator may estimate a particular mixture model with no computational problems, and apparently “nice” results and meaningful coefficients. However, this representation may not be unique. Another investigator may produce equally nice results under different parametric assumptions and with different implications. That is, the observed survivor function may be consistent with other choices of the baseline hazard and heterogeneity distributions (Lancaster, 1990, chapter 4). In the terminology of Section 2.2, different structural models, with substantively different policy implications, may have the same reduced form. This clearly poses a problem for

parametric applied work. One appealing solution is to choose flexible parametric forms for hazard and heterogeneity, or else to take the semiparametric approach of partial likelihood analysis. The discussion of this issue continues in the next section.

18.4. Specification of the Heterogeneity Distribution

The sensitivity of coefficient estimates to alternative assumptions about the heterogeneity has been extensively discussed in the literature. Two apparently contradictory positions may be discerned:

1. Parametric specifications of unobserved heterogeneity are often somewhat arbitrary. They may seriously distort inferences about the hazard function. Hence a parametrically flexible or nonparametric specification is desirable. See Heckman and Singer (1984a).
2. Parametric specifications of unobserved heterogeneity are relatively innocuous if the baseline hazard function is correctly specified. When the specification of the hazard function is in doubt and/or is incorrect, the estimates produced using different parametric assumptions for heterogeneity may lead to different estimates of the marginal distribution of the data. See Manton, Stallard, and Vaupel (1986).

The apparent contradiction between the two positions may be resolved as follows. The specification of the hazard function affects the first moment of the distribution of $f(t)$, whereas that of heterogeneity affects its second moment, assuming that it is uncorrelated with the observed covariates. If the hazard function is correctly specified, then the main impact of the heterogeneity distribution would be on the relative efficiency of the estimator.

18.4.1. Discrete-Time PH with Gamma Heterogeneity

The preceding considerations suggest that a proportional hazard formulation with an arbitrary hazard function makes an attractive model with which to combine a specific heterogeneity assumption. Han and Hausman (1990) and Meyer (1990) combine the gamma heterogeneity assumption with the discrete-time proportional hazard model developed in Section 17.10. They report that when the baseline hazard is not parameterized estimates show little sensitivity to alternative functional forms for $g(v)$.

For specificity reconsider (17.43) after including a heterogeneity term:

$$\varepsilon_i = \ln \left(\int \lambda_0(\tau) d\tau \right) - \mathbf{x}'_i \boldsymbol{\beta} - v_i,$$

which can be substituted into the expression for log-likelihood (17.44). The heterogeneity term needs to be integrated out. Han and Hausman give a closed-form expression under the gamma heterogeneity assumption and report results that indicate relatively minor sensitivity to parametric assumptions given their flexible hazard specification.

18.4.2. Some Other Models for Heterogeneity

The preceding discussion emphasized the computational convenience of Weibull–gamma model, which has a closed form.

If the tail of the observed marginal distributions is thicker than is consistent with the gamma or log-normal, one may consider a member of the Mandelbrot **stable family** of distributions. Hougaard (1986) proposed a very general family that nests, for example, the gamma and inverse-Gaussian families (also see Jaggia, 1991b). A strictly stable distribution obeys the condition that the sum of p independent realizations should have the same distribution as a scale factor times the distribution. Hougaard (2000, appendix 3.3) provides a summary of its properties.

Although a more highly parameterized heterogeneity distribution looks attractive because of its greater generality, it may lead to two kinds of problems. The first problem is that the available data may not be sufficiently rich to allow us to identify or precisely estimate the parameters. Often this situation cannot be recognized without attempting estimation in the first place.

The second problem is computational. If the mixture density does not have a closed form, it is then left in the form of an integral. The resulting likelihood function has terms that are also integrals. Estimation requires the use of computer-intensive numerical methods such as numerical or Monte Carlo integration that were discussed in Chapter 12. An example of a mixture model that requires such estimation techniques is the Weibull–log-normal mixture in which unobserved heterogeneity has a log-normal distribution. Simulation-based estimation of heterogeneity models is discussed by Gouriéroux and Monfort (1991, 1996) and considered as an example in Section 12.2.

18.5. Discrete Heterogeneity and Latent Class Analysis

The preceding analysis assumed a continuous distribution of unobserved heterogeneity and concentrated on estimation of the parameters of that distribution.

An alternative approach assumes that the sample of individuals is drawn from a population that consists of a finite number of **latent classes**, say q , and that each element in the sample can be regarded as a draw from one of these q latent sub-populations or strata. This model is known variously as the finite mixture model, **semiparametric heterogeneity model** (Heckman and Singer, 1984a), and **latent class model** (Aitken and Rubin, 1985). Its attractive feature is that it leads to a flexible parametric distribution. In duration modeling the model has been analyzed, advocated, and applied by Heckman and Singer (1984a).

Although these popular models are presented in the context of duration models, a general notation is used to emphasize the potential for application elsewhere; see, for example, Section 20.4.

18.5.1. Finite Mixture Model

Consider the following two-component *finite mixture* model. If the sample is a probabilistic mixture from two subpopulations with pdf $f_1(t|\mu_1(\mathbf{x}))$ and $f_2(t|\mu_2(\mathbf{x}))$, then

$\pi f_1(\cdot) + (1 - \pi) f_2(\cdot)$, where $0 \leq \pi \leq 1$, defines a two-component finite mixture. That is, observations are draws from f_1 and f_2 , with probabilities π and $1 - \pi$, respectively. The parameters to be estimated are (π, μ_1, μ_2) . The parameter π may be treated as constant or may be further parameterized using, for example, the logit function. Thus $\pi = \exp(\lambda)/[1 + \exp(\lambda)]$ and λ in turn may be parameterized in terms of further observable covariates. Thus we think of two types of individuals, those that come from $f_1(\cdot)$ and those that come from $f_2(\cdot)$. Perhaps there may be an a priori case for thinking along these lines, for example if there is some latent characteristic that partitions the sampled population in this way. An alternative interpretation is simply that the linear combination of densities makes a good approximation to the observed distribution of t .

Generalization to additive mixtures with three or more components is in principle straightforward but subject to potential problems of the identifiability of the components. This is discussed further later in the chapter. Therefore, it is very helpful in empirical application if the components have a natural interpretation. At the simplest level we think of each subpopulation as a “type,” but in many situations a more informative interpretation may be possible (Lindsey, 1995).

Another interpretation of the finite mixture model is in terms of a discrete representation of population heterogeneity. Suppose the population consists of m homogeneous subpopulations, usually called **components**. A parametric model, such as the Weibull or exponential, is supposed to apply to each component. Assume that the j th component is a fraction π_j of the total population, $\sum \pi_j = 1$.

Formally, the problem is formulated as follows: In all previous examples the distribution of the unobserved heterogeneity term has infinite points of support. If the continuous mixing distribution $g(v_i)$ can be approximated by a discrete distribution, denoted by π_j ($j = 1, \dots, m$) with a finite number, m , of support points then the marginal (mixture) distribution is

$$h(t_i | \mathbf{x}_i, \pi_j, \boldsymbol{\beta}) = \sum_{j=1}^m f(t_i | \mathbf{x}_i, v_j, \boldsymbol{\beta}) \pi_j(v_j), \quad (18.18)$$

where v_j is an estimated support point and π_j is the associated probability. This semi-parametric representation of unobserved heterogeneity was examined by Heckman and Singer (1984a) in duration modeling. Closely related work is that of Wedel et al. (1993), where the latent class interpretation is favored. If the mixing distribution π_j is not subject to any parametric assumptions, then the mixture model is called a **semiparametric mixture model** for t .

The estimation of the finite mixture model may be carried out under the assumption of either known or unknown number of components. If the fractions π_j are known, maximum likelihood estimates of the component distributions can be estimated. More usually the proportions π_j , $j = 1, \dots, m$, are unknown and the estimation involves both the π_j and the component parameters. The maximum likelihood estimator for the latter case is called nonparametric maximum likelihood estimator (NPMLE). Here the nonparametric component is the number of classes, but it is strictly a semiparametric method because it is combined with parametric models for the components. If the number of components is unknown, as is usually the case, then some delicate issues of inference arise. See Section 18.5.4 for details.

An obvious motivation for the finite mixture class is that it is a natural and simple way to treat population heterogeneity. In many situations it is simpler to think of unobserved heterogeneity in terms of a small number of latent classes rather than a continuum of “types” as in Section 18.2.

18.5.2. Latent Class Interpretation

The finite mixture model is related to **latent class analysis** (Aitkin and Rubin, 1985; Wedel et al., 1993). Let $d_i = (d_{i1}, \dots, d_{im})$ define an indicator (dummy) variable such that $d_{ij} = \mathbf{1}(\sum_j d_{ij} = 1)$ indicates that t_i was drawn from the j th (latent) group or class for $i = 1, \dots, N$. That is, each observation may be regarded as a sample from one of the m latent subpopulations, classes, or “types.” In the discussion that follows we assume that the model is identified.

The model specifies that $(t_i | d_i, \boldsymbol{\mu}, \boldsymbol{\pi})$ are independently distributed with densities

$$\sum_{j=1}^m d_{ij} f(t_i | \mu_j) = \prod_{j=1}^m f(t_i | \mu_j)^{d_{ij}}, \quad (18.19)$$

where $\mu_j = \mu(\mathbf{x}_j, \boldsymbol{\beta}_j)$, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_m)$, and $(d_i | \boldsymbol{\mu}, \boldsymbol{\pi})$ are iid with multinomial distribution

$$\prod_{j=1}^m \pi_j^{d_{ij}}, \quad 0 < \pi_j < 1, \quad \sum_{j=1}^m \pi_j = 1. \quad (18.20)$$

The last two relations imply that

$$(t_i | \boldsymbol{\mu}, \boldsymbol{\pi}) \stackrel{iid}{\sim} \sum_{j=1}^m \pi_j^{d_{ij}} f_j(t_i | \mu_j)^{d_{ij}},$$

which leads to the likelihood function

$$L(\boldsymbol{\beta}, \boldsymbol{\pi} | \mathbf{t}) = \prod_{i=1}^N \sum_{j=1}^m \pi_j^{d_{ij}} f_j(t_i | \mu_j)^{d_{ij}}. \quad (18.21)$$

18.5.3. EM Algorithm

This likelihood function may be maximized directly or by applying the EM algorithm in which the variables $\mathbf{d} = (d_1, \dots, d_n)$ are treated as missing data; see Section 10.3. If the \mathbf{d} were observable the log-likelihood of the model would be

$$\ln L(\boldsymbol{\mu}, \boldsymbol{\pi} | \mathbf{t}, \mathbf{d}) = \sum_{i=1}^N \sum_{j=1}^m d_{ij} \ln f_j(t_i | \mu_j) + \sum_{i=1}^N \sum_{j=1}^m d_{ij} \ln \pi_j. \quad (18.22)$$

If π_j , $j = 1, \dots, m$, are given, the posterior probability that observation t_i belongs to the population j , $j = 1, 2, \dots, m$, denoted z_{ij} , is given by

$$z_{ij} \equiv \Pr[y_i \in \text{population } j] = \frac{\pi_j f_j(y_i | \mathbf{x}_i, \boldsymbol{\beta}_j)}{\sum_{j=1}^m \pi_j f_j(y_i | \mathbf{x}_i, \boldsymbol{\beta}_j)}. \quad (18.23)$$

The average value of z_{ij} over i is the probability that a randomly chosen individual belongs to subpopulation j . This equals π_j :

$$E[z_{ij}] = \pi_j.$$

Suppose we have available an estimate \hat{z}_{ij} of $E[d_{ij}]$. Then, conditional on this estimate we have

$$EL(\beta_1, \dots, \beta_m, \pi | \mathbf{t}, \hat{\mathbf{z}}, \mathbf{x}_1, \dots, \mathbf{x}_m) = \sum_{i=1}^N \sum_{j=1}^m \hat{z}_{ij} \ln f_j(t_i, \mu(\mathbf{x}_j, \beta_j)) + \sum_{i=1}^N \sum_{j=1}^m \hat{z}_{ij} \ln \pi_j, \quad (18.24)$$

which constitutes the E-step of the EM algorithm. The M-step of the algorithm maximizes EL by solving the first-order conditions

$$\hat{\pi}_j - N^{-1} \sum_{i=1}^m \hat{z}_{ij} = 0, \quad j = 1, \dots, m, \quad (18.25)$$

$$\sum_{i=1}^N \sum_{j=1}^m \hat{z}_{ij} \frac{\partial \ln f_j(t_i | \beta_j)}{\partial \beta_j} = \mathbf{0}. \quad (18.26)$$

Next we can use (18.23) to get new values of \hat{z}_{ij} and iterate through the E- and M-steps. Once the process converges the variances can be computed using either the information matrix or the robust formula.

18.5.4. Choosing the Number of Latent Classes

The first important issue concerns the choice of m , the number of components. Often there is no guiding prior theory and the choice is usually made on pragmatic grounds. Because the dimension of parameters to be estimated is $m \dim[\beta] + m - 1$, the number of parameters can be quite large. This number can be decreased somewhat if some elements of β are restricted to be equal. One popular method involves allowing the intercept to vary but restricting the slope parameters to be the same across groups (as in (18.18)). However, there is clearly an incentive to keep m small if all parameters are allowed to vary across classes. Even when only the intercepts are allowed to vary, many applications use $m = 2$. A sensible strategy is to start with $m = 2$, and then check the fit of the model using diagnostic tests. An additional component is added if the fit is poor. Adding components that cannot be reliably differentiated is problematic. When intergroup differences are small, the finite mixture representation is not needed. The most desirable situation is one in which the components have an interpretation. Choice between models of different dimensions can be made using the penalized likelihood criterion (AIC or BIC), see Section 8.5.1. The likelihood ratio test is not appropriate because of the parameter boundary hypothesis problem. Baker and Melino (2000) describe a Monte Carlo experiment that dramatically reveals the potential pitfalls of over-parameterization in a model in which both duration dependence and heterogeneity are flexibly specified owing to a desire to avoid misspecification. For model selection they recommend comparing a penalized likelihood criterion across candidate latent class models, with a high penalty for more parameters.

When the model is overparameterized the parameters cannot be identified. The problem may manifest itself by the presence of multiple optima or a flat likelihood surface. The computational algorithm may converge to different points depending on the starting values.

A model selected from competing models using the penalized likelihood criterion may not necessarily describe the sample data well. This can only be ascertained by a suitable goodness-of-fit test and model diagnostics. Essentially one compares the actual and fitted distribution of durations; a significantly large deviation between the two indicates that the systematic component of the model does not adequately explain the observed sample variation. Some possibilities are considered in the next section.

Computational Considerations

A second issue concerns the choice of computer algorithm. Whereas the EM algorithm is very helpful in understanding the computational structure of the problem, in practice it often tends to be slow. The authors have found many instances in which the Newton–Raphson algorithm based on numerical derivatives has produced satisfactory results. See Haughton (1997) for a survey of alternatives. No matter which algorithm is used, if the intergroup differences are small, the likelihood surface will tend to exhibit several local maxima. In any case, a single unique maximum is not guaranteed.

All finite mixture models are unidentified in the sense that the distribution of the data is unchanged if the subpopulation labels are permuted. That is, relabeling “component 1” as “component 2,” or vice versa, makes no difference. This problem can be dealt with by specifying either the π_j or μ_j to be nondecreasing. It is desirable that the component labels have some behavioral interpretation.

One potential limitation of the finite mixture model is that additional components may simply reflect the presence of outliers. Though this is not necessarily a bad thing, it is useful to be able to identify the outlying observations that are responsible for one or more components. Equation (18.23) can be useful in this regard. Postestimation one could calculate the posterior probability. For outliers these probabilities will be large with respect to one component and small with respect to the rest.

18.6. Stock and Flow Sampling

In many practical situations the following question arises: What is the relationship between two or more different average duration measures that are available? From demography comes the well-known distinction between average age and expected life span. In real estate there is the distinction between the average period that a property offered for sale has remained unsold and the expected period before which a newly added property for sale will be sold. Often the first concept is used in popular discussions when the second may be more relevant. In economics there is a similar question about the relationship between different measures of unemployment duration that are published by government statistical agencies. The issue of unobserved heterogeneity, as it pertains to the pool of the unemployed, and to the flow into that pool, is closely

involved in these discussions. One of the earlier influential discussion of these issues was given in Salant (1977).

For specificity, let us focus on the familiar example of unemployment duration. One statistic that measures the unemployment experience of an already unemployed individual, published by statistical agencies in many countries, is the **average interrupted duration** (AID), which is the average period for which members of the current stock of unemployed have been unemployed. It is an estimate of the **expected elapsed duration**, a period for which a newly unemployed individual can expect to remain unemployed, often referred to as average duration of a complete spell of unemployment (ACD), a measure that features prominently in the job search literature and is the one that the current and previous chapters have concentrated on. This is an estimate of the expected length of a **completed duration**. We may think of AID as a stock-based measure and ACD as a flow-based measure; the former is analogous to average age in a population and the latter to the expected life span. The question of interest is the relationship between the two.

The appropriate statistical tool for handling issues such as these is **renewal theory**. The stationary Poisson process with constant intensity parameter is an example of a **renewal process**. The number of renewals in a time interval dt refers to the number of events. Duration is the time between successive occurrences of events (i.e., renewals). For an individual in a given state the **backward recurrence time** refers to the elapsed duration since renewal, and **forward recurrence time** refers to the duration from current state to a transition. The expected number of events, denoted $E[N(t)]$, in the time interval $(0, t]$ is called the **renewal function** and $\lim_{dt \rightarrow 0} dE[N(t)]/dt$ is the **renewal intensity**, which determines the relationship between ACD and the average backward recurrence time. In what follows, we concentrate on some well-known results.

Salant (1977) showed that heterogeneity in hazard rates provides a key to understanding the differences between AID and ACD. His diagrammatic representation provides intuition into the two key factors that affect the calculated averages. In Figure 18.1 the vertical axis measures calendar time and the horizontal axis represents the date of the survey. **Stock sampling** refers to sampling in the survey period from the stock of individuals who are then in a given state. In contrast, **flow sampling** means that we sample those who enter the state during a particular interval. The lengths of spells in progress are shown as vertical lines. For illustration nine realizations of spells are shown and four of these (S6, S7, S8, and S9) are in progress on the survey date. Five spells (S1, S2, S3, S4, and S5) are completed during the 12-month survey period. If u_j denotes the length of the j th in-progress spell sampled by the survey, then for our example, $AID = 1/4(\sum_j u_j)$. If t_i denotes the length of the i th completed spell sampled by the survey, then $ACD = 1/5(\sum_i t_i)$.

Now observe that the survey is more likely to capture longer spells than shorter spells, and this leads to an upward bias that is the result of **length-biased sampling**. This type of bias is likely to lead to $AID > ACD$. However, because the survey measures only incomplete durations, the average of such incomplete durations is likely to be shorter than the average of the completed durations. This is the phenomenon of **interruption bias**. The answer to the question of which source of bias dominates

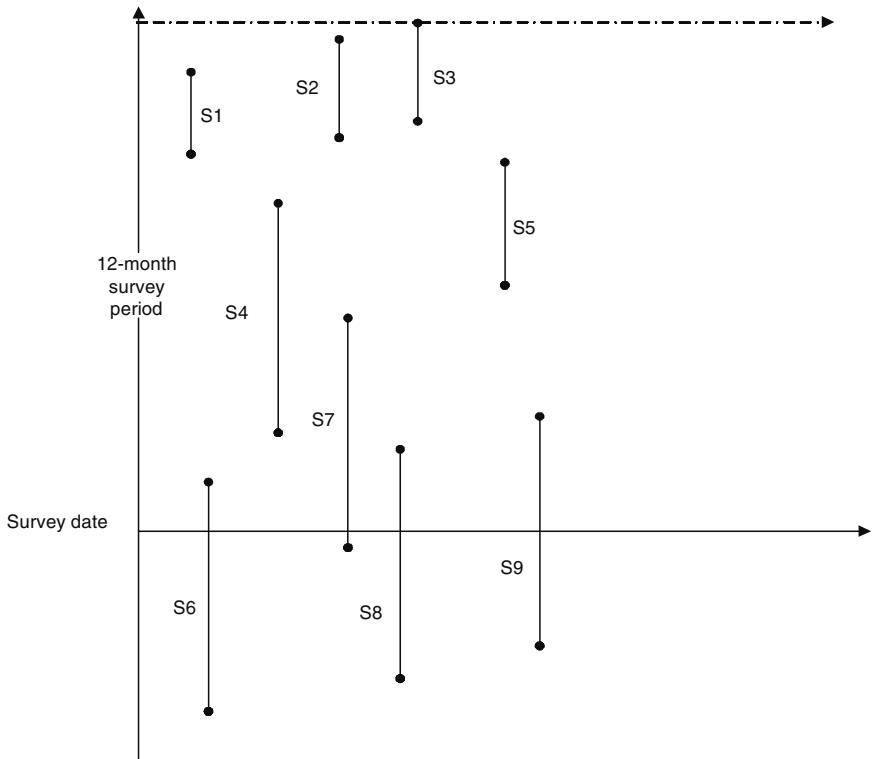


Figure 18.1: Length-biased sampling under stock sampling: examples.

depends on the distribution of spell lengths, and this in turn depends on the distribution of hazard rates. Heterogeneous hazard rates provide a key to understanding the relationship between the two.

The key assumption is that of a stationary environment which refers to a situation in which inflows into and the outflow from the state are equal. Let $f(u)$ denote the density of interrupted spells and $g(t)$ denote the density of completed spells. Then, the distribution of u is given by

$$f(u) = \frac{\bar{G}(u)}{\int \bar{G}(u) du} = \frac{\bar{G}(u)}{E[t]}, \quad (18.27)$$

where

$$\bar{G}(u) = \int g(x) dx$$

is the survivor function corresponding to be density $g(u)$, and $E[t]$ is the mean of the distribution of completed durations. For a full derivation of this result and the underlying assumptions, see Salant (1977) or Lancaster (1990, Section 5.3).

An implication of this result is that if $g(t)$ is exponential, so that the stochastic process for the event is the Poisson process, then $f(u)$ is also exponential, and the mean of both duration measures is equal.

Given (18.27), the general relationship between moments of the distributions of u and t can be derived. One useful result links the mean of u to the mean and variance of t :

$$E[u] = \frac{1}{2} \left(E[t] + \frac{V[t]}{E[t]} \right). \quad (18.28)$$

Another interesting result concerns the relationship between $E[t]$ and the mean completed duration of the *constant population* with spells in progress (i.e., the average across the stock of spells in progress). In line with intuition based on length-based sampling, the relation is

$$E[t^{(S)}] = E[t] + \frac{V[t]}{E[t]} > E[t], \quad (18.29)$$

which says that the mean duration for the constant stock, denoted $E[t^{(S)}$], exceeds the average expected duration of a new spell. If $f(t)$ is exponential, then $E[t^{(S)}] = 2E[t]$, and $E[u] = 1/2E[t^{(S)}$]; on average the sampled interrupted spell will be halfway to completion.

What if the hazard rate is not constant? If the hazard rate is increasing in spell length (i.e., positive state dependence) then $E[u] < E[t]$, and if it is decreasing (i.e., negative state dependence) then $E[u] > E[t]$.

Although these results have been obtained under the assumption of a constant population, they have proved very useful in interpreting and clarifying the connections among various average measures of duration that are commonly employed. The results given here are valid regardless of the reason for spell occurrence. They also motivate a more careful investigation of the shape of the hazard function.

18.7. Specification Testing

Tests of specification in duration models take several different forms, including the following:

- inclusion and exclusion tests for covariates,
- tests of functional forms of the survival function,
- tests of unobserved heterogeneity, and
- joint tests of state dependence and unobserved heterogeneity.

The first type of specification test does not raise new problems and can be handled by Wald-type tests.

Tests of restrictions on functional form are the same as tests of unobserved heterogeneity if the restriction is the absence of unobserved heterogeneity. Because the latter can bias the estimation of the hazard rate, as shown in the Section 18.2, diagnostic testing for unobserved heterogeneity is desirable.

The standard formulation for this is to test whether the heterogeneity (variance) parameter is zero. If this hypothesis is tested using the restricted model that assumes zero heterogeneity, a score test is appropriate. The use of the likelihood ratio or Wald test

based on the unrestricted model will be problematic if the hypothesis is a boundary hypothesis. For example, in the Weibull–gamma model (18.9), the restriction $1/\delta = 0$ will specialize the model to the Weibull, but this is a boundary hypothesis. The standard one-degree-of-freedom chi-square test has a weighted chi-squared distribution under the null.

18.7.1. Hypothesis Tests

One type of specification test is a score test of unobserved heterogeneity based on the exponential null model. Because of possible confounding between heterogeneity and duration dependence it is desirable to carry out a joint rather than a separate test. This can be done using the framework of a locally heterogeneous Weibull model (Lancaster, 1985).

A **locally heterogeneous density** is generated by considering a Taylor expansion of an arbitrary density around $\nu = 1$ of the Weibull density with multiplicative heterogeneity ν , yielding

$$\begin{aligned} S(t|\nu) &= e^{-\mu t^\alpha \nu} = e^{-\varepsilon \nu} \\ &= e^{-\varepsilon} [1 + (-\varepsilon)(\nu - 1) + (\varepsilon^2/2)(\nu - 1)^2 + O(\varepsilon^3)], \end{aligned}$$

where $\varepsilon = \mu t^\alpha$. From the second line

$$E[e^{-\varepsilon \nu}] = e^{-\varepsilon} [1 + (\varepsilon^2 \sigma^2/2)] \equiv S_m(t),$$

where the term σ^2 is the variance of the heterogeneity distribution.

Then

$$\begin{aligned} f_m(t) &= -\frac{\partial S_m(t)}{\partial t} \\ &= \alpha \mu t^{\alpha-1} e^{-\varepsilon} [1 + (\varepsilon^2 \sigma^2/2)] - e^{-\varepsilon} [2\varepsilon(\alpha \mu t^{\alpha-1}) \sigma^2/2] \\ &= \alpha \mu t^{\alpha-1} e^{-\varepsilon} [1 + \sigma^2(\varepsilon^2 - 2\varepsilon)/2]. \end{aligned}$$

Using the last result and allowing for censored observations, the log-likelihood is given by

$$\begin{aligned} \ln L(\alpha, \beta, \sigma^2) &= \sum_{i=1}^N \ln \left\{ [f_m(t_i)]^{\delta_i} [S_m(t_i)]^{1-\delta_i} \right\} \\ &= \sum_{i=1}^N \delta_i \left[\ln \alpha + (\alpha - 1) \ln t_i + \ln \mu_i + \ln (1 + \sigma^2 (\varepsilon_i^2 - 2\varepsilon_i)/2) - \varepsilon_i \right. \\ &\quad \left. + (1 - \delta_i) \ln (1 + \sigma^2 \varepsilon_i^2/2) \right], \end{aligned}$$

where δ_i is the censoring indicator, which takes the value one for uncensored durations and zero otherwise, $\ln \mu_i = \beta_0 + \mathbf{x}'_i \beta_1$, and $\varepsilon_i = \mu_i t_i^\alpha$ is the **generalized error** (Section 18.7.2).

The null hypothesis of interest is $H_0 : \sigma^2 = 0$ and $\alpha = 1$. This is a joint test of zero unobserved heterogeneity and the exponential distribution specification. Let $\theta = (\theta'_1, \theta'_2)$, $\theta'_1 = (\sigma^2, \alpha)$, and $\theta'_2 = (\beta_0, \beta_1)$, and let $\theta'_0 = (0, 1, \beta_0, \beta_1)$ denote the restricted vector.

For simplicity consider only the case of uncensored data. Then the joint score test statistic is

$$\text{LM}_{\text{HD}} = \frac{1}{d} \mathbf{s}' \begin{bmatrix} \Psi'(1) & 1 \\ 1 & 1 \end{bmatrix} \mathbf{s}, \quad (18.30)$$

where $\mathbf{s}' = [\frac{1}{2} \sum_i (\varepsilon_i^2 - 2\varepsilon_i), \sum_i (1 + (1 - \varepsilon_i) \ln t_i)]$, and $\Psi'(r)$ denotes the first derivative of the digamma function $d \ln \Gamma(r)/dr$ and $d = 1/(N(\Psi'(1) - 1))$. To implement the test, LM_{HD} is evaluated at the null (i.e., replacing all quantities by their estimates under the null of exponential distribution). This test statistic has an asymptotic $\chi^2(2)$ distribution (Jaggia and Trivedi, 1994).

Notice that the matrix of the quadratic form in the LM_{HD} statistic is not diagonal. That is, the two components of the joint test are correlated. A separate test of heterogeneity (duration dependence) has power against duration dependence (heterogeneity). More explicitly, suppose we consider two separate score tests for heterogeneity and duration dependence. They are

$$\text{LM}_H = \frac{1}{4N} (\sum_i (\varepsilon_i^2 - 2\varepsilon_i))^2, \quad (18.31)$$

$$\text{LM}_D = \frac{1}{d} (\sum_i (1 + (1 - \varepsilon_i) \ln t_i))^2, \quad (18.32)$$

each of which has a $\chi^2(1)$ distribution under the null. The separate test of zero unobserved heterogeneity has power against the other null hypothesis because the tests are correlated, see (18.30). Consequently, inferring the direction of misspecification on the basis of a separate test can be misleading.

Because the specification of unobserved heterogeneity and state dependence are closely related, testing hypotheses about them separately can produce misleading results (Jaggia and Trivedi, 1994). Formally speaking, tests of state dependence in the presence of incorrectly neglected heterogeneity are biased, and the reverse is also true. Jaggia (1991c) reanalyzes strike duration data that have been analyzed in a misleading manner in the econometrics literature. Jaggia and Trivedi (1994) develop some joint tests for a class of parametric models. See also Bera and Yoon (1993) who consider the more general issue of hypothesis testing when the model is misspecified.

Useful as these tests are in simple parametric models, the starting point of an investigation might be a Weibull, Weibull–gamma, or proportional hazard model. In such cases testing for unobserved heterogeneity, or any other specification error, can be carried out using the integrated hazard function because in the absence of heterogeneity integrated hazard is a unit exponential random variable. We now discuss some graphical methods for evaluating the fit of the model based on integrated hazard.

18.7.2. Graphical Tools for Detecting Misspecification

In Section 8.7.2 we developed the concept of generalized residuals. In nonlinear models a clear-cut choice of such a measure is difficult. In the present context there is a good choice.

Generalized Residuals

A useful type of test is a nonparametric graphical test of fit of the duration model. The test uses the generalized residual, which is defined as a certain function of data and estimated parameters. For a correctly specified model the residuals should behave approximately like an iid sample from a known distribution. The integrated hazard turns out to have such a property and hence functions as an ingredient for a residual-based specification test. In the context of duration models where from Section 17.3.1

$$S(t|\mu) = \exp[-\Lambda(t|\mu)],$$

$$f(t|\mu) = \lambda(t|\mu) \exp[-\Lambda(t|\mu)],$$

consider the distribution of the **generalized residual**

$$\begin{aligned} \epsilon &= \Lambda(t|\mu) \\ &= -\ln(S(t|\mu)). \end{aligned} \tag{18.33}$$

The Jacobian of this transformation is

$$\begin{aligned} |J| &= dt/d\epsilon \\ &= \frac{1}{d\Lambda(t|\mu)/dt} \\ &= 1/\lambda(t|\mu). \end{aligned}$$

Given $f(t|\mu)$, the transformation in (18.33), and the Jacobian of transformation, the density of ϵ is given by

$$\lambda(t|\mu) \exp(-\epsilon) \frac{1}{\lambda(t|\mu)} = \exp(-\epsilon), \tag{18.34}$$

which does not depend on μ ; the density is the unit exponential distribution. This result was referred to in Sections 17.3.1 and 17.6.7.

Diagnostic Test Based on Integrated Hazard

A diagnostic test can be constructed by exploiting the unit exponential property of the generalized residual ϵ under the null of correct specification. The survivor function of the generalized residual is $S(\epsilon) = \exp(-\epsilon)$. Hence $-\ln S(\epsilon) = \Lambda(\epsilon) = \epsilon$. For a correctly specified model, a graphical comparison of the estimated integrated hazard with $\hat{\epsilon}$ should yield an approximately linear positive relationship with 45° slope. If the plot deviates significantly from the 45° line a misspecification could be indicated.

For example, the **estimated integrated hazard** for the Weibull model is $\hat{\epsilon} = \hat{\mu}t^{\hat{\alpha}}$. Its survivor function is $\hat{S}(\hat{\epsilon}) = N^{-1}$ (number of sample observations $\geq \hat{\epsilon}$).

A small formalization of this is to regress $-\ln \hat{S}(\hat{\epsilon})$ on $\hat{\epsilon}$ and an intercept and test whether the intercept is zero and the slope equals one.

The technique may be applied to any parametric model for which the integrated hazard expression is available. For example, the *generalized error* for the Weibull–gamma mixture (easily specialized to an exponential–gamma mixture by setting $\alpha = 1$)

is $\epsilon = k \ln [(k + \mu t^\alpha)/k]$. To apply the test, compute $\widehat{\epsilon}$ given estimates of (μ, α, k) , and then plot $\widehat{\epsilon}$ against $-\ln \widehat{S}(\widehat{\epsilon})$.

Censored Data

In the case of censored observations the observed duration $t = \min[T, L]$, where L denotes the right-censoring limit. If the observation exceeds L it is censored at L . Then the generalized error $\epsilon(t)$ is not unit exponential distributed. The following derivation leads to a relationship that suggests an adjustment for censoring:

$$\begin{aligned} E[\epsilon(T)|T \geq L] &= \int_{\epsilon(L)}^{\infty} \frac{\epsilon f(\epsilon)}{S(\epsilon(L))} d\epsilon \\ &= \frac{1}{e^{-\epsilon(L)}} \left[\int_{\epsilon(L)}^{\infty} \epsilon e^{-\epsilon} d\epsilon \right] \\ &= \frac{1}{e^{-\epsilon(L)}} \left[1 + \epsilon(L) e^{-\epsilon(L)} + e^{-\epsilon(L)} - 1 \right] \\ &= 1 + \epsilon(L), \end{aligned} \tag{18.35}$$

upon integration by parts and simplification.

This suggests that one might estimate the generalized error as $\widetilde{\epsilon}(t) = \widehat{\epsilon}(t)$ if data are not censored, and as $\widetilde{\epsilon}(t) = 1 + \widehat{\epsilon}(L)$ if the observations are censored. Available results suggest that this technique works reasonably well in the censored exponential model when the proportion of censoring is not too heavy (Jaggia and Trivedi, 1994; Jaggia, 1997).

18.7.3. Conditional Moment Tests

The **conditional moment** framework (see Section 8.2) applied to the generalized residuals provides a fruitful approach to specification testing. The idea can be illustrated in the context of tests of unobserved heterogeneity.

The integrated hazard function was shown previously to be distributed as a unit exponential random variable with mean 1 and variance 1. In this case the conditional second-moment restriction of interest is $E[(\epsilon - 1)^2] = V[\epsilon] = 1$, or equivalently

$$E[\epsilon^2 - 2] = 0.$$

Higher order moment restrictions can also be generated and tested jointly or separately. For details see Jaggia (1991a).

18.8. Unobserved Heterogeneity Example: Unemployment Duration

In this section, we rework the empirical example of Section 17.11 under the assumption that unobserved heterogeneity is present and can be parameterized within an analytically tractable parametric model.

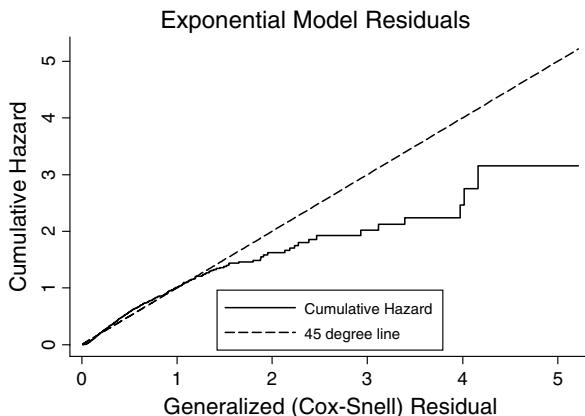


Figure 18.2: Unemployment duration: generalized residuals from the exponential model. U.S. data from 1986–92 on 3343 spells, some incomplete.

As discussed in Section 18.7.2, we can use a graphical tool to examine the possible presence of unobserved heterogeneity by looking at the estimated fit of the model. For a correctly specified model, the residuals should follow the unit exponential distribution. One can evaluate the model fit informally by computing and plotting the empirical cumulated hazard function against the generalized residual. For a correctly specified model the plot should exhibits an approximate straight line with slope one.

Figures 18.2 and 18.3 show the generalized residual plots for the exponential model without and with (gamma) heterogeneity, respectively. As we can see from the two graphs, the fit of the model improves only marginally after we introduce unobserved heterogeneity.

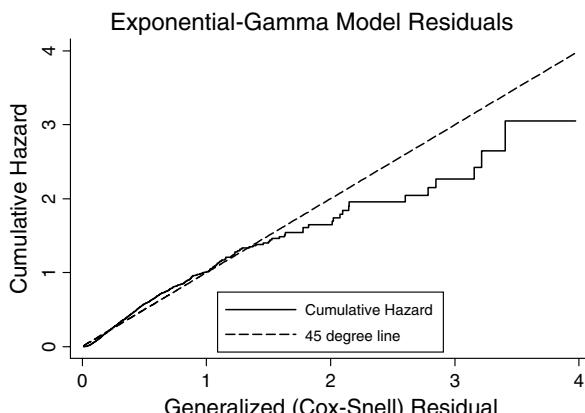


Figure 18.3: Unemployment duration: generalized residuals from the exponential-gamma model. Same data as Figure 18.2.

Table 18.1. *Unemployment Duration: Exponential Model with Gamma and IG Heterogeneity*

Variable	Exponential–Gamma		Exponential–IG	
	Coeff.	t	Coeff.	t
RR	0.501	0.817	0.504	0.821
DR	−0.882	−1.118	−0.807	1.032
UI	−1.585	−6.043	−1.545	−5.994
RRUI	1.091	1.725	1.057	1.686
DRUI	0.057	0.055	−0.013	−0.012
LNWAGE	0.379	3.184	0.373	3.156
CONS	−4.095	−4.507	−4.097	−4.545
σ^2	0.232	3.178	0.207	2.925
−ln L	2695.35		2696.48	

This result can be verified by the actual estimates shown in Table 18.1, which also presents the estimates of the exponential model with inverse-Gaussian (IG) heterogeneity. Although there is evidence of significant unobserved heterogeneity, the estimates of coefficients in these two settings do not differ much from what we have obtained earlier without the presence of unobserved heterogeneity. It is expected that the presence of unobserved heterogeneity will have a large impact on the duration dependence parameter, as this factor is absent from the exponential model.

However, a more interesting case arises when we consider a model with duration dependence and unobserved heterogeneity. Without presuming that it is the “correct” model, we consider the Weibull distribution–inverse Gaussian mixture model. For ease of comparison, we present these estimates in Table 18.2 along with the estimates when unobserved heterogeneity is neglected.

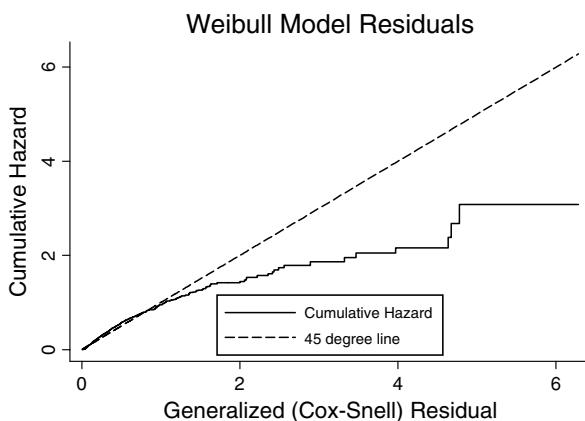
The introduction of unobserved heterogeneity has a substantial impact on the duration dependence parameter, which increases from 1.129 in Table 17.8 to 1.753 in Table 18.2. The latter implies a more steeply rising hazard rate out of unemployment than was the case when unobserved heterogeneity was ignored. Recall from Section 18.2.4 that one of the consequences of neglected heterogeneity in proportional hazards model is to underestimate the hazard rate; so the aforementioned empirical finding is consistent with theory. Second, note that the evidence for unobserved heterogeneity is very strong; the estimated variance parameter σ^2 has a t-ratio exceeding 11. Third, the fit of the model, as reflected in the log-likelihood, has also improved (from −2687.6 to −2616.6). Although there is not much qualitative change in the estimates of the coefficients, the effects of the significant coefficients (UI, LNWAGE, and CONS) have become more pronounced after unobserved heterogeneity is introduced.

The improvement in the fit of the model notwithstanding, the new mixture model could still be misspecified. Once again we use the graphical device as an informal

Table 18.2. Unemployment Duration: Weibull Model with and without IG Heterogeneity

Variable	Weibull-IG		Weibull	
	Coeff.	t	Coeff.	t
RR	0.736	0.812	0.448	0.70
DR	-1.073	-0.933	-0.427	-0.53
UI	-2.575	-6.698	-1.496	-5.67
RRUI	1.734	1.857	1.105	1.57
DRUI	-0.061	-0.039	-0.299	-0.28
LNWAGE	0.576	3.259	0.37	2.99
CONS	-5.303	-3.953	-4.358	-4.74
α	1.753	44.19	1.129	51.44
σ^2	6.377	11.149	—	—
-ln L	2616.6		2687.6	

specification test. Figures 18.4 and 18.5 plot the generalized residuals from the Weibull model with and without unobserved heterogeneity. The plots suggest that the mixture model, despite being more general than the exponential-IG model, appears to be misspecified. To reiterate, although a simpler model that allows for neither duration dependence nor unobserved heterogeneity shows little graphical evidence of misspecification, an “improved” specification that generalizes the model in both directions appears to be misspecified, a result similar to that of Jaggia (1991c). The apparent puzzle may be resolved by the argument that the interaction between heterogeneity and duration dependence accounts for the result. The Weibull model assumes monotonic hazards. However, McCall (1996) provides evidence based on the same data that

**Figure 18.4:** Unemployment duration: generalized residuals from the Weibull model. Same data as Figure 18.2.

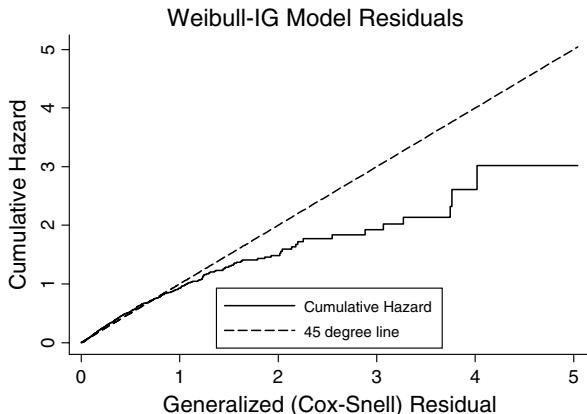


Figure 18.5: Unemployment duration: generalized residuals from the Weibull-Inverse Gaussian model. Same data as Figure 18.2.

a bathtub-shaped hazard function is more appropriate. He specifies a polynomial baseline hazard function that is less restrictive than the monotonic function used here. Thus a reasonable interpretation of our results is that a model that simultaneously allows for both unobserved heterogeneity and duration dependence makes it easier to detect misspecification than a model that ignores both.

Finally, we implement a parametric test for the presence of unobserved heterogeneity. The purpose is to illustrate some of the theory discussed in Section 18.7. The score test for neglected heterogeneity developed in Section 18.7.1 assumed uncensored data. Because the data used here include right-censored observations we implement the score test for the censored sample developed by Jaggia (1997).

We wish to test for zero unobserved heterogeneity, $H_0 : \sigma^2 = 0$, in the exponential duration model. Let the parameter set be denoted by $\theta = (\sigma^2, \beta)$ and let $s(\theta_0)$ and $\mathcal{I}(\theta_0)$ be, respectively, the score and the information matrix calculated under the null. Using the log-likelihood derived in Section 18.7.1, we can write $s(\theta_0) = (s_1(\theta_0), s_2(\theta_0))$, where $s_1(\theta_0) = \frac{\partial \mathcal{L}}{\partial \sigma^2} \Big|_{H_0} = \frac{1}{2} \sum (\epsilon_i^2 - 2C_i \epsilon_i)$ and $\mathcal{I}(\theta_0) = -E \left[\frac{\partial^2 \mathcal{L}}{\partial \theta \partial \theta'} \right] \Big|_{H_0}$. The score test for unobserved heterogeneity is then given by

$$LM = s'_1(\tilde{\theta}_0) \mathcal{I}^{11}(\tilde{\theta}_0) s_1(\tilde{\theta}_0) \sim \chi^2(1), \quad (18.36)$$

where $\mathcal{I}^{11} = [\mathcal{I}_{11} - \mathcal{I}_{12} (\mathcal{I}_{22})^{-1} \mathcal{I}_{21}]^{-1}$ is the first diagonal component of the partitioned inverse of $\mathcal{I}(\theta)$, given in Jaggia (1997), and the tilde superscript is used to denote restricted maximum likelihood estimates.

For our sample, we found that $LM = 44.25$, which far exceeds the critical value of $\chi^2(1)$ and hence we reject the null of $\sigma^2 = 0$. This result is consistent with that from the Weibull-gamma and Weibull-IG models where a significant improvement in the fit of the model resulted from introduction of unobserved heterogeneity. As previously noted, this test has power against a test of misspecified duration dependence also.

18.9. Practical Considerations

The issue of interaction between hazard function and unobserved heterogeneity has generated a huge literature. One point of view that is well documented states essentially that if the hazard function is well specified then the precise parametric specification of the heterogeneity distribution is relatively innocuous (Manton et al., 1986). This view implies that rather than parametrically modeling unobserved heterogeneity we can simply use robust variance estimates, given that the hazard function is well specified. Other studies suggest that parametric specification of the heterogeneity distribution is not innocuous (Heckman and Singer, 1984a) and that it is desirable to use a nonparametric specification. Some highly influential work has advocated use of a discrete hazard model with a very flexible specification of the hazard function, combined with a parametric assumption about heterogeneity (Meyer, 1990; Han and Hausman, 1990). Finally, as a compromise between all the foregoing positions, some researchers use the Han–Hausman discrete-time approach, or a high-order polynomial hazard function, and combine it with the Heckman–Singer approach of nonparametric heterogeneity. However, as Baker and Melino (2000) have pointed out, this may lead to overparameterization that is far from innocuous. Hence it seems sensible to approach this issue with caution, and use parsimonious models in preference to models saturated with heterogeneity parameters.

The Cox PH model has a central place in the biometrics literature. When there is no intrinsic interest in the baseline hazard function then this seems an attractive choice of functional form. It is often a good place to start modeling. However, unobserved heterogeneity is important in most econometric specifications and should not be ignored.

Many statistical packages offer a choice of standard parametric duration models that can be combined with any of the standard (gamma, inverse-Gaussian, or log–normal) heterogeneity (“frailty”) specifications. Although this is a very convenient option to use, discrete hazard models hold greater appeal as they provide greater flexibility and a better match with economic data.

The implementation of the EM algorithm for the latent class model often suffers from slow computational speed. Direct maximization of the likelihood is often both feasible and efficient.

18.10. Bibliographic Notes

18.2 There are many papers that discuss the specification of the heterogeneity distribution and consequences of misspecification. Vaupel et al. (1979) provide a good discussion of the properties of the gamma model. Hougaard (1984) considers several alternatives to the gamma. Hougaard (1995) gives a survey of heterogeneity models. Heckman and Singer (1984a) advocate nonparametric specification and emphasize the sensitivity to misspecification. Manton et al. (1986) attempt to disentangle the relative importance of misspecifying the hazard and heterogeneity, suggesting that the former is critical.

18.3 Van den Berg (2001) provides a thorough and accessible treatment of and further references on the identification of the MPH model.

- 18.4** Han and Hausman (1990) and Meyer (1990) offer good empirical examples that combine flexible hazard specifications with parametric assumptions about heterogeneity.
- 18.5** The paper by Heckman and Singer (1984a) is an early discussion of the discrete heterogeneity model. The finite mixture model of unobserved heterogeneity is also commonly referred to as the “nonparametric heterogeneity” model. Baker and Melino (2000) describe a Monte Carlo study of duration dependence and nonparametric heterogeneity. They consider models with very flexible specification of duration dependence with nonparametric heterogeneity. Their results suggest that, when both are present, the strategy of having many finite mixture components in likelihood generates large biases and unreliable results. Using the BIC or the Hannan–Quinn criterion, which penalizes overparameterization, can be helpful.
- 18.6** Lancaster (1990) and Salant (1977) are excellent references on length-biased sampling. Lancaster provides foundational material on renewal theory that forms the basis of several key results. Also see Taylor and Karlin (1994).
- 18.7** There are many papers on specification testing in duration models, most of them handling the easier case of no censoring. Kiefer (1988) provides an overview. Jaggia (1991a) offers a brief but clear introduction to the conditional moment approach to specification testing (which is also summarized in Greene (2003)). As yet untried in the context of duration models is a very general, but computationally demanding, approach to specification testing due to Andrews (1997). Model selection issues for finite mixture models are discussed in Cameron and Trivedi (1998, chapter 6), in the context of count models. A good introduction to model diagnostics based on different types of residuals for duration models is given in Hosmer and Lemeshow (1999, pp. 196–240).
- 18.8** Lancaster’s (1979) classic empirical paper analyzes unemployment duration in the context of a Weibull–gamma mixture model. Jaggia (1991c) studies misspecification in a strike duration model using a generalized gamma model that nests several popular specifications. His paper also highlights the difficulty of making inferences from overly restrictive models. A number of other applications of duration models are covered in Chapter 19.

Exercises

- 18–1** (Adapted from Sapra, 2002) The analysis of Section 18.2 shows the effects of unobserved heterogeneity on the unconditional or averaged hazard function. The result that neglected heterogeneity leads to under-estimation of the slope of the average hazard function is emphasized. Let the conditional hazard function be $\lambda_C(t|v) = v\lambda_0(t)$, where λ_0 denotes the baseline or unconditional hazard function. Show that (i) the unconditional hazard $\lambda_U(t) < \lambda_0(t)$ and (ii) $\partial\lambda_U(t)/\partial t < 0$ in each of the following examples.

- (a) $v \sim \text{Uniform}[0, 1]$ and $\lambda_0(t) = 1 \forall t$.
- (b) v follows a unit exponential distribution with pdf $g(v) = e^{-v}$ and $\lambda_0(t) = \rho \exp(\gamma t)$, $\rho > 0$, $\gamma < 0$.

- 18–2** Reconsider the Weibull–gamma model of Section 18.2.3 after replacing the gamma distributed heterogeneity assumption by the assumption that heterogeneity is distributed according to the log-normal distribution with unit mean.
- (a) Verify that in this case an analytical expression for the unconditional hazard function is not obtainable.
 - (b) Substitute the integral expression for unconditional hazard into the log-likelihood given in Section 17.6.3. Using the simulation-based maximum likelihood approach of Section 12.4, describe an estimation algorithm that details the various steps involved in likelihood maximization.
- 18–3** Consider the exponential–gamma mixture. This model is a special case of a MPH model. The survivor function, conditional on a multiplicative heterogeneity factor ν , for the exponential model is $S(t|\nu) = \exp(-\mu t\nu)$, $\lambda > 0$. The unconditional survivor function is given by the average survivor function. Averaging is across the heterogeneous population using $g(\nu)$, the density of ν , as the weighting function, so $S(t) = \int_0^\infty S(t|\nu)g(\nu)d\nu$. Assume that ν is (two-parameter) gamma distributed with $g(\nu) = \delta^k \nu^{k-1} \exp(-\delta\nu)/\Gamma(k)$.
- (a) Show that, given gamma heterogeneity, $S(t) = (1 + \mu t/\delta)^{-k}$.
 - (b) Derive expressions for the unconditional duration density function $f(t)$ and the unconditional hazard function $\lambda(t)$. These general expressions can be specialized by setting the mean of ν at 1; that is, set $k = \delta$, which leads to the exponential–gamma mixture. Compare the mean and variance properties of this mixture distribution with those of the original exponential distribution.
 - (c) Suppose that the random variable ν has a two-point distribution such that with probability π it takes the value ν_1 and with probability $(1 - \pi)$ it takes the value ν_2 . What are the implications of this assumption for the specification of the unconditional survivor function? Explain your answer.
- 18–4** Using the sample of the McCall data set from the empirical exercise in the previous chapter, reestimate the Weibull model for those transiting to full-time employment ($CENSOR1 = 1$) under the assumption that unobserved heterogeneity (also called frailty in some computer packages, which may also have a subcommand for specifying it) has gamma distribution.
- (a) Using generalized residuals as in Section 18.7.2 test the hypothesis of model misspecification.
 - (b) Does the new model display a duration dependence property? Does it provide a better fit to the data? Explain the results by reference to the interaction between unobserved heterogeneity and duration dependence.
 - (c) Repeat the exercise of part (a) under the assumption of log-normal heterogeneity. Are the results about duration dependence significantly different from those for the gamma heterogeneity?

Models of Multiple Hazards

19.1. Introduction

This chapter deals with several different duration models that can be interpreted broadly as multivariate models, a category that covers both parallel and repeated transitions. Any transition model that involves more than one destination state can be regarded as a multivariate model because the analysis will involve joint distribution of two or more durations. The models we consider arise in a variety of ways and apply to several different types of data. Despite their differences, they are grouped in this chapter for reasons of organizational convenience.

To be concrete consider some examples. A familiar model from labor economics involves a transition from unemployment to employment or out of the labor force. The first transition can be further broken into return to the old job or to a new job. These destinations are mutually exclusive. An unemployment spell may end by a transition to any one of the destinations. A variant of this example considers an unemployed individual who could find either a new full-time or part-time job or remain unemployed. Thus there are three possible states (destinations). The models of Chapters 17 and 18 dealt with transitions between two states. One can still use the two-state methods to handle such data. For example, state 1 could be that of full-time employment and state 0 could be any other state. This would, as before, involve modeling one hazard rate. However, one could also characterize this situation in terms of a model with three states and two transitions and hence two hazard functions, one specific to each destination state. More generally, there will be a number of failure types and we may wish to model the transition from a given state to any one of the failure types. In this chapter we wish to extend the conceptual tools developed in the previous two chapters to deal with multiple hazards (failures) or a multivariate duration model.

The important issues are as follows:

1. How does one model the relation between covariates and failures of different types?
2. How does one model interaction between failure types under a specific set of study conditions?

3. How does one estimate failure rates for certain types of failures given the “removal” of some or all other failure types?

A **multivariate duration model** involves simultaneous modeling of all transitions, that is, joint specification and estimation of two or more hazard functions. There are several possible frameworks for analyzing multivariate duration data; the **competing risks** framework is one of the most popular. McCall (1996) provides an empirical application of the competing risks framework to unemployment data with focus on the role of unemployment insurance. Using an approach similar to McCall’s, Deng, Quigley, and Van Order (2000) study the transitions of mortgage holders to the states of prepayment or termination of mortgages.

What is the motivation for and the gains from joint modeling of hazards? If the different hazards are essentially independent then separate and joint modeling will produce the same results. However, different hazards may be linked; for example, there may be present a common unobserved heterogeneity term in each hazard function. Alternatively, each hazard may include an unobserved heterogeneity term with one or more common shared components, leading to correlated hazards.

A second class of examples involves a case of parallel events in which one analyzes the joint distribution of durations to destinations. For example, the pair (T_1, T_2) could be the duration of unemployment and duration without health insurance. Here the motivation for joint estimation of the hazards could be similar to that previously outlined.

A third example involves joint distribution of lengths of **repeat spells** in the same state (e.g., unemployment, or without health insurance). That is, for a given individual, one wants to simultaneously model the hazards of terminating a spell. If the spells in question are independent, then they can be analyzed by single-spell methods of earlier chapters. If the researcher wants to study the dependence structure of the transitions, then joint modeling of spells in a given state is appropriate. New models and methods are called for when the spells are dependent. This last example is potentially more complex than the preceding ones because of possible dependence between events separated by time intervals. For example, the length and type of a previous spell, or more generally the past history of spells, may affect the probability and length of a succeeding spell; or the unobserved characteristics of the individual may persist over successive spells. Such serially correlated unobserved heterogeneity creates a link between repeat spells. Even the occurrence probability of an event may depend on previous occurrence of the same event. Heckman and Borjas (1980) characterize several structural types of state dependence for an individual using concepts such as **occurrence dependence** and (Markovian) **lagged duration dependence**.

Corresponding to these different data situations are a variety of models in the literature. However, though they might appear to be a disparate selection they are linked by several common threads. After introducing the basic concepts, in Section 19.2 we examine the popular competing risks model. In Section 19.3 we consider a multivariate model based on marginal distributions of a set of survival times and introduce the copula approach to joint modeling of survival times. Multiple-spell modeling is considered in Section 19.4.

19.2. Competing Risks

First, we introduce some concepts that are used to in the **competing risks model** (CRM) and in other multivariate formulations. Often these are extensions of concepts already introduced in Chapter 17. The basic CRM formulation is applicable to modeling time in one state when exit is to a number of competing states, such as different causes of death. The CRM is attractive because it is relatively straightforward to implement if the model is a PH model.

19.2.1. Basic Concepts

We now consider the CRM in which there are m latent duration or failure times, one for each distinct competing cause of failure.

Latent Durations

The setup of the model is as follows. Each subject has an underlying failure time, which is subject to censoring. Failure time may be one of m different types, given by the set $J = \{1, \dots, m\}$. We may think of this as a situation with m distinct causes of transition from a given state (“death”). However, the occurrence of a failure of one kind of event removes the individual from risks of other kinds of events. Therefore, given censoring of the remaining $(m - 1)$ durations for each individual, we observe at most one complete duration.

In a CRM with m types of failures, there are $m + 1$ states $\{0, 1, \dots, m\}$, where 0 represents the initial state and $\{1, \dots, m\}$ are possible destination states. For the i th individual the data vector is of the form $(\mathbf{x}_i, t_i, d_{1i}, \dots, d_{mi}, d_{ci})$, where \mathbf{x}_i is a vector of weakly exogenous covariates that measure the characteristics of i , $t_i = \min(t_{1i}, \dots, t_{mi}, t_{ci})$, where t_{ki} denotes the time to transition to the k th destination, t_{ci} denotes the time of censoring, and $d_{ji} \equiv \mathbf{1}(t_{ji} = t_i)$, $j = 1, \dots, m, c$ are dummy variables that take the value one if $t_{ji} = t_i$. Because we only observe one of the t_{ji} , the remaining are interpreted as latent variables.

Censoring may be regarded as a competing risk. It operates on individuals according to a probability distribution. In this chapter the censoring variable is assumed to be independent of the (t_1, \dots, t_m) .

Unobserved characteristics of i are subsumed under unobserved heterogeneity, denoted as v . If v varies with cause of exit, then we write it as v_j , $j = 1, \dots, m$.

Competing Causes

A standard example of competing risks is death from competing causes. Consider an individual who has had a kidney transplant operation and is “at risk” of transitioning to the healthy state, or to rejection, or to some other unhealthy condition such as a liver complication. Succumbing to any one condition means that transition to other states is not possible. So in an m -event setup, each event provides one complete

duration and $m - 1$ censored durations. Thus we have a situation of “competing risks” in which there is competition to determine the transplant patient’s destination state.

Although discrete-time models are often required in empirical applications, our exposition of the joint hazard formulation uses the continuous-time framework and generally follows the exposition given in Mealli and Pudney (1996). We also assume that we have single-spell data.

The model provides the joint distribution of the **spell duration**, denoted τ , and the **exit route** r , which is an integer variable that takes one of the values in the set $(1, 2, \dots, m)$.

We ignore censoring for simplicity and assume that there exist latent variables (t_1, \dots, t_m) , one for each destination, that correspond to the spell duration for each possible exit route by which the spell may terminate if there were no other risk factors that might cause the spell to end sooner. Destination-specific covariates are denoted by \mathbf{x}_j ($j = 1, \dots, m$). We observe one duration, τ , where

$$\begin{aligned}\tau &= \min(t_1, \dots, t_m) \\ &= \min_j(t_j), \quad t_j > 0,\end{aligned}\tag{19.1}$$

at the termination of the spell; that is, only the shortest duration is observed and the rest are censored. Censoring owing to factors other than exit are not considered. Then

$$\begin{aligned}\Pr[\tau > t] &= \Pr[t_1 > t, \dots, t_m > t] \\ &= S_\tau(t),\end{aligned}\tag{19.2}$$

which is the joint survivor function. If the risks are independent then

$$\Pr[\tau > t] = \Pr[t_1 > t] \times \Pr[t_2 > t] \times \dots \times \Pr[t_m > t].\tag{19.3}$$

The corresponding exit route r is given by

$$r = \arg \min_{j \in J} (t_j).\tag{19.4}$$

Let $g_j(t)dt$ denote the probability of succumbing to risk j in the interval $(t, t + dt)$; then the total hazard rate applicable to all causes is

$$\lambda_\tau(t) \equiv -d/dt \ln S_\tau(t) = \sum_{j=1}^m g_j(t).$$

In biostatistics this is referred to as the **total force of mortality** (David and Moeschberger, 1978). If risks are independent, then the hazard rate for a specific cause j is $\lambda_j(t) = g_j(t)$. This means that probability of failure from cause j in $(t, t + dt)$, conditional on survival to t , is the same whether j is one of the risks or the only risk.

The probability of surviving the risk j in the interval (T_1, T_2) conditional on surviving to T_1 is

$$\begin{aligned}\int_{T_1}^{T_2} \lambda_j(t) dt &= \int_0^{T_2} \lambda_j(t) dt - \int_0^{T_1} \lambda_j(t) dt \\ &= \ln S(T_2) - \ln S(T_1) \\ &= -\ln \frac{\Pr[t_j > T_2]}{\Pr[t_j > T_1]},\end{aligned}\tag{19.5}$$

or equivalently

$$\exp\left(-\int_{T_1}^{T_2} \lambda_j(t) dt\right) = \frac{\Pr[t_j > T_2]}{\Pr[t_j > T_1]}. \tag{19.6}$$

One minus the left-hand side expression is referred to as the *net probability of death from cause j* in the interval (T_1, T_2) . The expression in (19.6) is useful for building up the likelihood function for estimation.

Independent Risks

We can now explicitly bring into the picture covariates that affect the hazard rate. We assume **independent risks** (as opposed to correlated risks) and consider the distribution of t_j . The hazard rate for failure of j th type is defined by

$$\lambda_j(t_j | \mathbf{x}_j) = \lim_{\Delta t_j \rightarrow 0} \frac{\Pr[t_j \leq T \leq t_j + \Delta t, |T \geq t_j, \mathbf{x}_j]}{\Delta t_j},$$

and the integrated hazard $\Lambda_j(t_j | \mathbf{x}_j)$ for the j th type risk is defined by

$$\Lambda_j(t_j | \mathbf{x}_j) = \int_0^{t_j} \lambda_j(s | \mathbf{x}_j) ds.$$

Then the duration density is

$$\begin{aligned}f_j(t_j | \mathbf{x}_j, \boldsymbol{\beta}_j) &= \lambda_j(t_j | \mathbf{x}_j, \boldsymbol{\beta}_j) S_j(t_j | \mathbf{x}_j, \boldsymbol{\beta}_j), \\ &= \lambda_j(t_j | \mathbf{x}_j, \boldsymbol{\beta}_j) \exp[-\Lambda_j(t_j | \mathbf{x}_j, \boldsymbol{\beta}_j)],\end{aligned}$$

using the relation between survivor and integrated hazard functions. Defining $\mathbf{x} = [\mathbf{x}_1, \dots, \mathbf{x}_m]'$ and $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_m]'$ gives the joint density of τ and r :

$$\begin{aligned}f_j(\tau, r | \mathbf{x}, \boldsymbol{\beta}) &= f_r(\tau | \mathbf{x}_r, \boldsymbol{\beta}_r) \prod_{j \neq r} \exp[-\Lambda_j(\tau | \mathbf{x}_j, \boldsymbol{\beta}_j)] \\ &= \lambda_r(\tau | \mathbf{x}_r, \boldsymbol{\beta}_r) \exp[-\Lambda_r(\tau | \mathbf{x}_r, \boldsymbol{\beta}_r)] \\ &\quad \times \prod_{j \neq r} \exp[-\Lambda_j(\tau | \mathbf{x}_j, \boldsymbol{\beta}_j)] \\ &= \lambda_r(\tau | \mathbf{x}_r, \boldsymbol{\beta}_r) \prod_{j=1}^m \exp[-\Lambda_j(\tau | \mathbf{x}_j, \boldsymbol{\beta}_j)].\end{aligned}\tag{19.7}$$

The first line follows from the product of conditional and marginal probabilities. The second term on the right-hand side is the product of survival probabilities for all exit routes other than r , which uses the independence of risks assumption.

Equation (19.7) implies that

$$\begin{aligned} \lambda_j(\tau | \mathbf{x}_j, \boldsymbol{\beta}_j) & \exp \left[\sum_{j=1}^m -\Lambda_j(\tau | \mathbf{x}_j, \boldsymbol{\beta}_j) \right] \\ & = \lambda_j(\tau | \mathbf{x}_j, \boldsymbol{\beta}_j) \exp [-\Lambda^a(\tau | \mathbf{x}, \boldsymbol{\beta})], \end{aligned} \quad (19.8)$$

where $\Lambda^a(\tau | \mathbf{x}, \boldsymbol{\beta}) = \sum_{j=1}^m \Lambda_j(\tau | \mathbf{x}_j, \boldsymbol{\beta}_j)$ is the aggregate or overall integrated hazard. This last equation simply says that the total hazard of leaving the origin state is the sum of hazards for all destinations. The overall survivor function is

$$S(t) = \exp (-\Lambda^a(t)). \quad (19.9)$$

The likelihood function given independent risks is the product over all observations of terms like (19.7). This likelihood can be written explicitly if all relevant functional forms are specified. Many issues that were previously relevant, such as flexibility of functional form, unobserved heterogeneity, and so forth, remain relevant in the context of CRM. Instead of keeping the discussion at a general level, we now consider specific functional forms. The proportional hazard specification is popular in the literature and will be used here.

19.2.2. CRM with Proportional Hazards

The goal here is to derive the joint density of spell length and reason for exit, and this can be done by aggregating the integrated hazard over reasons for exit.

Consider PH models of the form

$$\lambda_j(t; \mathbf{x}) = \lambda_{0j}(t) \exp[\mathbf{x}'(t)\boldsymbol{\beta}_j], \quad j = 1, \dots, m,$$

where both the baseline hazard λ_{0j} and $\boldsymbol{\beta}_j$ are specific to type j hazard, and $t_{j1} < \dots < t_{jk_j}$ denote the k_j ordered failures of type j . For example, if $m = 2$, then k_1 refers to the number of individuals who registered a failure of type 1, and k_2 to the number of individuals who registered a failure of type 2.

The likelihood function for the **Cox CRM** given is then

$$\begin{aligned} L(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_m) & = \prod_{j=1}^m \prod_{i=1}^{k_j} \frac{\exp[\mathbf{x}'_{ji}(t_{ji})\boldsymbol{\beta}_j]}{\sum_{l \in R(t_{ji})} \exp[\mathbf{x}'_l(t_{ji})\boldsymbol{\beta}_j]}, \\ & = \prod_{j=1}^m L_j(\boldsymbol{\beta}_j), \end{aligned} \quad (19.10)$$

where

$$L_j(\boldsymbol{\beta}_j) = \prod_{i=1}^{k_j} \frac{\exp[\mathbf{x}'_{ji}(t_{ji})\boldsymbol{\beta}_j]}{\sum_{l \in R(t_{ji})} \exp[\mathbf{x}'_l(t_{ji})\boldsymbol{\beta}_j]}. \quad (19.11)$$

Notice the following four features of this likelihood: (1) $L_j(\boldsymbol{\beta}_j)$ is the partial likelihood developed in Section 17.8.2. The baseline hazard function is absent, and the asymptotic distribution results stated previously also apply. (2) $L(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_m)$ can be

jointly maximized by maximizing each individual factor $L_j(\beta_j)$, given the independence of risks; hence joint and separate maximizations are equivalent. Estimation and comparison of the β_j s can be made by applying standard asymptotic techniques to each individual factor in the m -term likelihood. (3) The ideas of Sections 17.7 and 17.8 can be extended directly. If a discrete-time (dummy variable) formulation is used for each type of hazard, then the identifiable components of the hazard function can be estimated for each type of hazard jointly with the β_j . (4) Unobserved heterogeneity can be introduced exactly as in the single-spell, two-state proportional hazards model in Chapter 18.

19.2.3. Identification of CRM

Cox (1962a) and Tsiatis (1975) showed that when the CRM has no covariates, the model is not identified. More precisely, this means that any CRM with dependent risks is observationally equivalent to a CRM with independent risks. However, Heckman and Honoré (1989) showed that under certain assumptions a CRM that has the mixed PH form with covariates is identified. Van den Berg (2001, pp. 3438–3441) provides an exposition of the underlying assumptions. Assumptions additional to those discussed in Chapter 17 are needed. For example, the covariates must show “sufficient variation” and should not be perfectly collinear. We also require that the baseline hazards for different risks should not be perfectly related.

19.2.4. Interpretation of Regression Coefficients

In the proportional hazards type formulation of CRM, the impact of a change in a covariate on the hazard rate for transition from a given state is analogue to the PH model in Chapter 17, but the direct interpretation of regression coefficients faces an interpretation problem similar to that discussed for the multinomial logit in Section 15.4.3.

However, one may also be interested in the impact of change in a covariate on the probability of exit via route r . This is harder to calculate. To see this note that the expression for the probability of exiting a given state via route r is given by

$$\Pr[r|\tau, \mathbf{x}, \boldsymbol{\beta}] = \frac{\lambda_r(\tau|\mathbf{x}_r, \beta_r)}{\sum_{j=1}^m \lambda_j(\tau|\mathbf{x}_j, \beta_j)}. \quad (19.12)$$

Because covariates appear in both the numerator and the denominator, and moreover the denominator is the sum of all hazards, the sign of the partial derivative $\partial \Pr[r|\tau, \mathbf{x}, \boldsymbol{\beta}] / \partial x_{rk}$ depends on all the parameters in the model. It is then not true that the sign of β_{rk} is also the sign of the partial. (The situation is exactly analogous to that discussed in Chapter 15 on multinomial models.) However, the following result is available if the competing risk is of the proportional hazard type (Thomas, 1996, p. 31). If $\beta_{rk} > \beta_{jk}$, $\forall j \neq r$, then the sign of $\partial \Pr[r|\tau, \mathbf{x}, \boldsymbol{\beta}] / \partial x_{rk}$ is positive. In words, an increase in x_k will increase the conditional probability of exit via route r if its estimated coefficient in $\lambda_r(\cdot)$ is larger than the corresponding coefficients in all other hazard functions.

19.2.5. CRM with Unobserved Heterogeneity

If the competing risks are of the proportional hazards type, then the methods of the previous chapter can be extended to include unobserved heterogeneity. A general specification of unobserved heterogeneity allows for a state-specific random component. Let $\boldsymbol{\nu} = (\nu_1 \dots \nu_m)$ be the vector of unobserved multiplicative heterogeneity terms that are assumed to have a joint distribution function $G(\boldsymbol{\nu})$; then,

$$\begin{aligned} f_j(\tau, r | \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\nu}) &= \lambda_j(\tau | \mathbf{x}_j, \boldsymbol{\beta}_j, \nu_j) \exp \left[\sum_{j=1}^m -\Lambda_j(\tau | \mathbf{x}_j, \boldsymbol{\beta}_j, \nu_j) \right] \\ &= \lambda_j(\tau | \mathbf{x}_j, \boldsymbol{\beta}_j) \nu_j \exp \left[\sum_{j=1}^m -\Lambda_j(\tau | \mathbf{x}_j, \boldsymbol{\beta}_j) \nu_j \right], \end{aligned}$$

where the second line follows from assumption of multiplicative heterogeneity.

This is an example of a competing risks model with state-specific random effects. The distribution marginal with respect to $\boldsymbol{\nu}$ is obtained by integrating out $\boldsymbol{\nu}$,

$$f_j(\tau, r | \mathbf{x}, \boldsymbol{\beta}) = \int \dots \int \lambda_j(\tau | \mathbf{x}_j, \boldsymbol{\beta}_j) \nu_j \exp \left[\sum_{j=1}^m -\Lambda_j(\tau | \mathbf{x}_j, \boldsymbol{\beta}_j) \nu_j \right] dG(\boldsymbol{\nu}),$$

which involves an m -fold integral.

A manageable case is one in which the m elements of $\boldsymbol{\nu}$ are independent gamma distributed random variables. In this case the m -fold integral decomposes into a product of m integrals. An example is the case in which we have a Weibull–gamma mixture for each cause-specific hazard function. In this case the competing risks are independent.

If we allow the elements of $\boldsymbol{\nu}$ to be correlated, then we get a more interesting case in which the competing risks are dependent. Indeed, this is a very widely used “trick” for generating dependence among competing risks. Specifically, suppose we have a multivariate log-normal distribution for $\boldsymbol{\nu}$, that is, $[\ln \nu_1 \dots \ln \nu_m]' \sim \mathcal{N}[\mathbf{0}, \boldsymbol{\Sigma}]$. This has two consequences. First, it induces dependence in the competing risks through heterogeneity; second, it makes computation of maximum likelihood estimates considerably more difficult. The reason for the latter is that the m -fold integral does not have an analytic expression. Consequently, Monte Carlo integration will have to be used. If m equals two or three as in many applied examples, this is still manageable but far from trivial. To reduce the dimensionality of the integral it may be useful to restrict the structure of the covariance matrix. For example, we may use a factor structure in which each term ν_j may be specified to be a linear function of (say) two iid random variables, with unknown weights (factor loadings). For identifiability, normalization restrictions on the weight coefficients may be necessary.

19.2.6. CRM with Dependent Competing Risks

The independent CRM has an important computational advantage over the model in which dependence is induced through heterogeneity variables correlated across competing hazards. However, the latter yields valuable additional information about the

structure of heterogeneity, such as the association parameter(s). Nonetheless, there remains the practical issue of how restrictive a specification of correlated heterogeneity one should choose. For exposition let us view the problem in a bivariate regression-like setting using the following setup similar to that in (17.20):

$$\begin{aligned}\ln \left[\int \lambda_1(u)du \right] &= -\mathbf{x}'\boldsymbol{\beta}_1 - \nu_1 + \varepsilon, \\ \ln \left[\int \lambda_2(u)du \right] &= -\mathbf{x}'\boldsymbol{\beta}_2 - \nu_2 + \varepsilon.\end{aligned}$$

Now we could assume $\nu_1 = \nu_2 = \nu$, that is, exactly the same unobserved heterogeneity term in both hazard models. The assumption is that the same unobserved factors affect both spells but their impact may differ. This amounts to perfectly correlated heterogeneity across the two hazards. Less restrictively, we could assume that, for example, ν_1 and ν_2 are correlated and estimate an association parameter. We can think of these as one- and two-factor models of heterogeneity, respectively. Whether the more restrictive approach is empirically desirable depends on the context. For example, if the two hazards pertain to the same individual, and we think of ν_1 and ν_2 as reflecting individual-specific factors, then the one-factor model has justification. If, however, we think of the two factors as hazard-specific, then the two-factor model is more appealing. There is some theoretical and Monte Carlo evidence that the use of the one-factor model when the two-factor model is the correct specification causes significant distortions (Lindeboom and Van den Berg, 1994).

19.3. Joint Duration Distributions

In this section we consider the case of nonmutually exclusive or parallel spells that are dependent. Survival times are assumed to be continuous. The exposition is at a general level and for simplicity it is restricted to the case where the spells are not censored and have parametric distributions.

In applied work on jointly distributed survival times a natural starting point would be a particular functional form for the joint survival or the joint density function that may be used. Are there some “standard” functional forms available? Or is there a general method for generating the multivariate counterparts of the models considered in the previous chapters? We consider these issues in the following.

19.3.1. Extending Survival Concepts to a Multivariate Setting

It is helpful to begin by extending the definitions and concepts of the two previous chapters to the multivariate case.

A multivariate survival function $S(\mathbf{t})$ is defined by

$$\begin{aligned}S(\mathbf{t}) &= S(t_1, \dots, t_q) \\ &= \Pr [T_1 > t_1, \dots, T_q > t_q],\end{aligned}\tag{19.13}$$

where T_1, \dots, T_q are q survival times with univariate survival functions $S_j(t_j)$. By definition,

$$\begin{aligned} S_j(t_j) &= \Pr[T_j > t_j] \\ &= S(T_1 \geq 0, \dots, T_j \geq t_j, \dots, T_q \geq 0) \\ &= S(0, \dots, t_j, \dots, 0). \end{aligned} \tag{19.14}$$

Unlike the case of the univariate survival function

$$S(t_1, \dots, t_q) \neq 1 - F(t_1, \dots, t_q).$$

For example, $S(t_1, t_2) = 1 - F(t_1) - F(t_2) + F(t_1, t_2)$.

The joint density of (t_1, \dots, t_q) is denoted by $f(t_1, \dots, t_q)$; if $F(t_1, \dots, t_q)$ is continuous then

$$f(t_1, \dots, t_q) = (-1)^q \frac{\partial^q F(t_1, \dots, t_q)}{\partial t_1 \dots \partial t_q}. \tag{19.15}$$

Analogous to the univariate case the **joint hazard function** is $\lambda(t_1, \dots, t_q)$ and is defined by

$$\lambda(t_1, \dots, t_q) = \frac{f(t_1, \dots, t_q)}{S(t_1, \dots, t_q)}. \tag{19.16}$$

The joint integrated hazard $\Lambda(t_1, \dots, t_q)$ is the q -fold integral of $\lambda(t_1, \dots, t_q)$. However, there is no simple relationship between $\Lambda(t_1, \dots, t_q)$ and $S(t_1, \dots, t_q)$ analogous to the univariate case.

Given these definitions, is it possible to derive joint survival functions? Clayton and Cuzick (1985) consider a bivariate model that illustrates the definitions given here. The starting point in their analysis is an assumption about the “**cross-hazard ratio**” function, defined as a function of two conditional hazard functions of t_1 , given $T_2 = t_2$ and $T_2 \geq t_2$. This leads to a nonlinear, second-order partial differential equation whose solution generates a joint survival function in which the cross-hazard ratio function plays an important role. We refer to the original sources for detail but note that this approach requires assumptions that may be difficult to extend beyond dimension higher than two.

19.3.2. Bivariate Distributions Based on Marginals

This section briefly reviews some approaches for generating bivariate duration models. The approach builds on assumptions about marginal survival functions. This may be useful if the researcher has a good feel for such marginal distributions and wants to use them as building blocks. Of course, choice of the building blocks places restrictions on the form of the resulting joint distribution.

One approach, which is due to Marshall and Olkin (1990), considers a model with multiplicative unobserved heterogeneity in the marginal distributions of both failure times in the following way. Let $f_i(t_i | \mathbf{x}_i, v)$, $i = 1, 2$, denote the marginal distributions of t_1, t_2 , given covariates $\mathbf{x}_1, \mathbf{x}_2$; here v is the common unobserved heterogeneity term

in the two marginals and is the source of association between the two hazards. In survival analysis such a model might be referred to as “shared frailty” model; it is the (only) source of correlation between t_1 and t_2 . Assume that $\nu, \nu > 0$, has probability distribution with density $g(\nu)$. The bivariate distribution of t_1, t_2 is formally defined as

$$f(t_1, t_2 | \mathbf{x}_1, \mathbf{x}_2) = \int_0^\infty f_1(t_1 | \mathbf{x}_1, \nu) f_2(t_2 | \mathbf{x}_2, \nu) g(\nu) d\nu, \quad (19.17)$$

where distribution parameters are suppressed for notational simplicity.

This bivariate distribution generated as a **mixture** may or may not have a closed-form solution, so without a specific parametric specification one cannot say whether the result will be computationally convenient to use. It is also the case that the resulting bivariate distribution will restrict the correlation between t_1 and t_2 to be positive. In some cases this may not be desirable.

This general approach, applicable to any type of data, can be specialized to the present case by replacing the marginal distributions with **marginal survivor functions** and deriving the **joint survivor function** by integrating out the variable ν ; thus,

$$S(t_1, t_2 | \mathbf{x}_1, \mathbf{x}_2) = \int_0^\infty S_1(t_1 | \mathbf{x}_1, \nu) S_2(t_2 | \mathbf{x}_2, \nu) g(\nu) d\nu. \quad (19.18)$$

An example of the application of this idea is provided by Clayton and Cuzick (1985), who use such a formulation to obtain a **bivariate survivor function** under the assumption of marginal proportional hazards with gamma heterogeneity.

As illustrated this approach for generating bivariate survivor model is somewhat restrictive. One source of restriction is the assumption of one-factor unobserved heterogeneity. In principle this restriction is easy to remove. For example, we could replace ν by $(\nu_1 \nu_2)$, $\nu_1 > 0, \nu_2 > 0$, which represents a vector of two correlated elements, one specific to each survivor function, with a joint probability distribution $g(\nu_1, \nu_2)$. Then

$$S(t_1, t_2 | \mathbf{x}_1, \mathbf{x}_2) = \int_0^\infty \int_0^\infty S_1(t_1 | \mathbf{x}_1, \nu_1) S_2(t_2 | \mathbf{x}_2, \nu_2) g(\nu_1, \nu_2) d\nu_1 d\nu_2. \quad (19.19)$$

For concreteness suppose that

$$\begin{aligned} \nu_1 &= \omega_{11}\varepsilon_1 + \omega_{12}\varepsilon_2, \\ \nu_2 &= \omega_{21}\varepsilon_1 + \omega_{22}\varepsilon_2, \\ \varepsilon_j &\sim \mathcal{G}[1, \sigma_j^2], \quad j = 1, 2, \end{aligned}$$

where $\{\omega_{ij}, i, j = 1, 2\}$ are unknown parameters, frequently referred to as “**factor loadings**.” This says that heterogeneity components (ν_1, ν_2) are correlated linear combinations of iid random components ε_1 and ε_2 if factor loadings are not zero. Other popular assumptions in empirical work are (i) that $(\ln \varepsilon_1, \ln \varepsilon_2)$ have a standard bivariate normal distribution or (ii) that ν_1, ν_2 have a discrete (finite-mixture) distribution. So the model (19.19) has a bivariate mixture form. Additional identifying restrictions (e.g., the normalization $\omega_{11} = 1$) are necessary also. The Pearson correlation coefficient between ν_1 and ν_2 , $\text{Cov}[\nu_1, \nu_2]/[\text{V}[\nu_1]\text{V}[\nu_2]]^{1/2}$, depends on $\{\omega_{ij}, \sigma_j^2, i, j = 1, 2\}$ and it is straightforward to verify that here this quantity would not have the usual -1 and $+1$ as the lower and upper bounds. (Also note that the

corresponding association parameter for failure times is $\text{Cov}[t_1, t_2]/[\text{V}[t_1]\text{V}[t_2]]^{1/2}$, which is distinct from that given.) Van den Berg (1997) derives sharp bounds on $\text{Cor}[t_1, t_2|\mathbf{x}]$, specifically $-1/3 < \text{Cor}[t_1, t_2|\mathbf{x}] < 1/2$, for a mixed proportional hazard model with constant baseline hazard, and shows that these bounds do not depend on the covariates x nor on the distribution of heterogeneity. If baseline hazard is not constant, the correlation bounds also depends on it.

The factor loading specification has computational advantages relative to that in which the unobserved heterogeneity components enter in an unrestricted manner. Although a one-factor model is likely to be too restrictive, an unrestricted model gives rise to a potentially high dimensional integral. From a computational viewpoint, the resulting distribution may or may not be easy to handle, depending in part on whether or not the integration produces a closed-form expression for the joint survivor function. If it does not, a simulation-based approach will be needed for estimation. At present estimation of such a model would require going beyond standard packages.

The factor loading specification does place restrictions on the model (Van den Berg, 2001; Lindenboom and Van den Berg, 1994). For example, if one of the marginal models does not indicate the presence of unobserved heterogeneity, then $\text{Cov}[\nu_1, \nu_2]$ must be zero; if $\text{V}[\nu_1] > 0$ and $\text{V}[\nu_2] > 0$, then $\text{Cov}[\nu_1, \nu_2] \neq 0$. Hence if $\text{Cov}[\nu_1, \nu_2] = 0$, then one of the marginals has no unobserved heterogeneity.

From an applied perspective an attractive multivariate survivor function should be flexible. The approach just outlined has some limitations. There are alternative approaches that have been proposed. One such approach that holds some promise is the use of copula functions. Hougaard (2000, pp. 435–437) provides an introduction in the context of survival analysis.

19.3.3. The Copula Approach

Copulas, originally introduced by Sklar in a 1959 article in French (see also Sklar, 1973), have been suggested as a useful method for deriving joint distributions given the marginals, especially when one wants to work with nonnormal distributions. Although we introduce this idea in the context of joint survival models, where it has found ready applications, it can also be used to study the joint distributions of any set of discrete, continuous, or mixed discrete/continuous variables.

The approaches already discussed (e.g., the **Marshall–Olkin method**) generate dependence between variables through unobserved heterogeneity components. This seems attractive in most applications because it is impossible for observed covariates to cover all relevant aspects of an economic event.

Properties of Copulas

To define a copula we begin with possibly dependent uniform random variables U_1, \dots, U_q on the $[0, 1]$ interval. The dependence relationship is described through their joint cdf

$$C(u_1, \dots, u_q) = \Pr[U_1 \leq u_1, \dots, U_q \leq u_q], \quad (19.20)$$

where the function $C(\cdot)$ is the **copula**, and u_j is a particular realization of U_j , $j = 1, \dots, q$.

The right-hand side is the joint cdf, $F(\cdot)$, and the q arguments of the copula can be replaced by q marginal cdfs $F_1(\cdot), \dots, F_q(\cdot)$. That is,

$$C(F_1(u_1), \dots, F_q(u_q)) = F(u_1, \dots, u_q)$$

defines a joint cdf. With a copula-based construction of a joint cdf we select a set of marginals and combine them to generate a joint cdf. A given copula is a functional form for combining selected marginals and different choices of $C(\cdot)$ lead to different joint cdfs. **Sklar's Theorem** established that any multivariate distribution function can be written in the form (19.20) and that given continuous marginals the copula representation is unique.

As specialized to a multivariate survival function, Sklar's Theorem says that a q -dimensional multivariate survival function $S(t_1, \dots, t_q)$ has a corresponding copula representation $C(S_1(t_1), \dots, S_q(t_q))$.

Consider the case $q = 2$. Then,

$$\begin{aligned} F(t_1, t_2) &= \Pr[T_1 \leq t_1, T_2 \leq t_2] \\ &= 1 - \Pr[T_1 > t_1] - \Pr[T_2 > t_2] + \Pr[T_1 > t_1, T_2 > t_2] \end{aligned}$$

and

$$\begin{aligned} S(t_1, t_2) &= \Pr[T_1 > t_1, T_2 > t_2] \\ &= 1 - F(t_1) - F(t_2) + F(t_1, t_2) \\ &= S_1(t_1) + S_2(t_2) - 1 + C(1 - S_1(t_1), 1 - S_2(t_2)), \end{aligned}$$

where $C(\cdot)$ is called the **survival copula**. Notice now that $S(t_1, t_2)$ is now a function of the marginal survival functions only.

Copulas have a certain symmetry property that allows one to work with copulas or survival copulas (Nelsen, 1999). Joe (1997) defines a bivariate copula associated with $F(\cdot)$, denoted by $C(u, v)$, as a two-dimensional probability distribution function defined on the unit square $[0, 1]^2$, with univariate marginals uniform on $[0, 1]$. For all $(u, v) \in [0, 1]$, $C(u, 0) = C(0, v) = 0$, $C(u, 1) = u$, and $C(1, v) = v$. In the context of survival copulas we replace u by the marginal survivor function $S(t_1)$ and v by the second marginal survivor function $S(t_2)$. In this notation Sklar's Theorem states that there exists a copula function C such that

$$F(u, v) = C(F_u(u), F_v(v)), \quad (19.21)$$

where $F(u, v) = \Pr[U < u, V < v]$ is a bivariate distribution function of random variables U and V , and $F_u(u)$ and $F_v(v)$ denote the marginal distribution functions.

If F is continuous, and if the univariate marginals have corresponding quantile functions F_u^{-1} and F_v^{-1} , then the unique copula in Equation (19.21) can be expressed as

$$C(u_1, u_2) = F(F_u^{-1}(u_1), F_v^{-1}(v)).$$

The copula approach involves specifying marginal distributions of each random variable along with a function (copula) that binds them together. The copula function

can be parameterized to include measures of dependence between the marginal distributions. If no dependence is detected, the two marginals are independent, and estimation can be performed on each variable separately. However, if dependence is present, improved estimates may be obtained by recovering a joint distribution by way of a copula function. Since a copula can capture dependence structures regardless of the form of the margins, a copula approach to modeling related variables is potentially very useful to econometricians. **Frechet bounds** make it possible to study the extent of dependence permitted by any copula. Despite apparent differences we see that the mixture approach of Section 19.3.2 for deriving the bivariate survival function leading to (19.19) is fundamentally similar to that based on the copula approach as both begin with marginals.

We now consider an example with q durations (T_1, \dots, T_q) that are conditionally independent given common neglected unobserved heterogeneity v ; covariates are excluded for simplicity. Then the conditional joint survivor function is

$$\begin{aligned}\Pr [T_1 > t_1, \dots, T_q > t_q | v] &= \Pr [T_1 > t_1 | v] \times \dots \times \Pr [T_q > t_q | v] \\ &= S_1 [(t_1) | v] \dots S_q [(t_q) | v]\end{aligned}$$

and the multivariate survival function is defined as

$$\Pr [T_1 > t_1, \dots, T_q > t_q] = E_v [S_1(t_1) | v, \dots, S_q(t_q) | v]. \quad (19.22)$$

Measuring Dependence

The functional form of the copulas itself does not depend on the form of the univariate margins. Copulas are usually specified with parameters that generate a measure of the dependence between the univariate margins. Usually dependence is parameterized as a scalar measure. Here we concentrate on bivariate copulas for simplicity.

The copula representation for discrete random variables is not necessarily unique (Joe, 1997, p. 14). This is not a major problem in practical application where the concern is to approximate the unknown joint distribution. The key modeling issue is to choose a sufficiently flexible parametric form for the copula function.

The dependence parameters from copulas can be difficult to interpret because they are not necessarily in the $[0, 1]$ interval. Therefore, it is customary to convert the **dependence parameter** to a familiar measure of association such as **Kendall's tau** or **Spearman's rho**; see Joe (1997). Schweizer and Wolff (1981) showed that Spearman's correlation coefficient can be expressed solely in terms of the copula function; thus,

$$\rho(t_1, t_2) = 12 \iint \{C(u, v) - uv\} dudv.$$

Consider any bivariate joint cdf $F(t_1, t_2)$ with univariate marginal cdfs $F_1(t_1)$ and $F_2(t_2)$. By definition, $0 \leq F_1(t_1), F_2(t_2) \leq 1$, because each marginal distribution takes a value in the range $[0, 1]$. The joint cdf is bounded below and above by the Frechet

Table 19.1. Some Standard Copula Functions

Copula Type	Function $C(u, v)$	θ -Domain
Product	uv	na ^a
FGMS ^b	$uv(1 + \theta(1 - u)(1 - v))$	$-1 < \theta < +1$
Normal ^c	$\Phi[\Phi^{-1}(u)\Phi^{-1}(v); \theta]$	$-1 < \theta < +1$
Clayton	$(u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$	$\theta \in (0, \infty)$
Frank	$-\theta^{-1} \ln(\eta - (1 - e^{-\theta u})(1 - e^{-\theta v})/\eta),$ $\eta = 1 - e^{-\theta}$	$\theta \in (-\infty, \infty)$

^a na, not applicable.

^b Farlie–Gumble–Morgenstern copula.

^c Φ denotes bivariate normal cdf.

lower and upper bounds, F^- and F^+ , defined as

$$F(t_1, t_2) \geq F^-(t_1, t_2) \equiv \max [F_1(t_1) + F_2(t_2) - 1, 0],$$

$$F(t_1, t_2) \leq F^+(t_1, t_2) \equiv \min[F_1(t_1), F_2(t_2)].$$

Since copulas are joint cdfs, they are also subject to the **Frechet bounds**. Knowledge of Frechet bounds is important in selecting an appropriate copula. Every copula places bounds on permissible values for its **dependence parameter** θ . A desirable feature of a bivariate copula is that as θ approaches the lower (upper) bound of its permissible range, the copula approaches the Frechet lower (upper) bound. However, the parametric form of a copula may impose restrictions such that one or both Frechet bounds are not included in the permissible range. Therefore, a particular copula may be a better choice for one data set than for another.

Examples

Table 19.1 gives examples of some bivariate copula functions that have been used in the literature. Joe (1997) discusses the properties of these copulas.

The Normal and the Frank copulas include both Frechet bounds in their permissible ranges. The Clayton copula belongs to the **Archimedean family**, with the representation $C(u, v) = \phi(\phi^{-1}(1 - u) + \phi^{-1}(1 - v))$; see Smith (2003).

Suppose we want to choose the Clayton copula to model the bivariate survival times (t_1, t_2) . Then the bivariate distribution, expressed in terms of marginal survival models $S(t_1)$ and $S(t_2)$, will be

$$(S(t_1))^{-\theta} + (S(t_2))^{-\theta} - 1)^{-1/\theta}.$$

We assume that the marginal survival functions are specified up to unknown parameters. As before these marginal survival functions can be written to capture dependence on covariates and unobserved heterogeneity. For example, these could be based on the proportional hazards model. For estimation we can apply maximum likelihood based on the resulting bivariate copula.

This approach is not without limitations. Two in particular are noteworthy. First, extension to three or more dimensions is not trivial. Second, one needs not only to choose a particular functional form for the copula but also to be aware of its potential restrictiveness in capturing dependence for a given data set. For example, only positive correlation may be supported.

Likelihoods Derived from Copulas

To fit a model derived from a copula (defined in terms of the cdfs) the first step is to select a copula and the second is to derive the likelihood (defined in terms of the pdfs) from it. Having chosen a copula consider the derivation of the likelihood for the special case of a bivariate model with uncensored failure times (t_1, t_2) . Define $f_j(t_j) = \partial F_j(t_j)/\partial t_j$ and $C_j(F_1, F_2)/\partial t_j$ for $j = 1, 2$, define $C_{12}(F_1, F_2) = \partial C(F_1, F_2)/\partial t_1 \partial t_2$. Then the probability density

$$f(t_1, t_2) = f_1(t_1)f_2(t_2)C_{12}(F_1(t_1), F_2(t_2)), \quad (19.23)$$

where $f(t_1, t_2) = \partial^2 F(t_1, t_2)/\partial t_1 \partial t_2$, is used to construct the likelihood function. If censored observations are present in the data, the likelihood must be appropriately modified (Frees and Valdez, 1998, pp. 15–16; Georges et al., 2001).

Using different copulas generates nonnested models. As in other similar instances, **penalized log-likelihood** values can be used to choose among them.

19.4. Multiple Spells

A distinction between parallel states and recurrent states, introduced early in this chapter, is helpful. Parallel states involve parallel events such as being employed and having health insurance; recurrent states involve sequential events such as the first birth, the second birth, and so forth. The term multiple spells refer to the durations between **recurrent spells** of the same event. Joint modeling of such data has similarities with joint modeling of parallel states as both involve multivariate concepts, but there are also important differences because sequential events may generate dynamic dependence in hazards.

Consider some examples of recurrent events. Individuals in the labor market may experience a succession of transitions between employment and unemployment. Young workers, for example, may record a succession of spells of unemployment. Newman and McCulloch (1984) consider the timing of births within a hazard framework. If one wants to model the hazard rate for each birth in a series of births, consideration must be given to the correlation between interbirth durations. Trivedi and Alexander (1989) analyze multiple spells of youth unemployment in Australia. In the literature on fertility, the duration between successive births is of interest (Heckman, Hotz, and Walker, 1985). Mealli and Pudney (1996) analyze the positive association between the duration in employment and pensionable status using data from a retirement survey in the United Kingdom. Engle and Russell (1998) study the time series of durations between successive transactions of a particular stock traded on the stock

market. Stevens (1999) analyzes the persistence of poverty over individuals' lifetimes taking account of multiple spells of poverty.

The aforementioned examples have several noteworthy features. Whether the hazard rate of an event depends on a previous event, conditional on a previous event, is an important modeling issue. Second, the form of dependence is of interest. The duration of a previous spell may enter as a covariate in determining the hazard of a later event; the occurrence of a previous event may affect the baseline hazard for a later spell; and, finally, unobserved heterogeneity may show serial dependence. Each of these raises an important modeling issue.

Multiple spells generate longitudinal or panel data that can potentially help to resolve the important identification issue concerning the influence of dynamic dependence ("the hand of past") relative to that of heterogeneity in the hazard function. Under some assumptions multiple observations make it easier to control for heterogeneity and to make inferences about dynamic dependence.

In general, survival models with unobserved heterogeneity and dependence between spells can be expected to be difficult to estimate. However, multiple-spell data create opportunities to study issues that can be studied only if panel data are available. Occurrence dependence, lagged duration dependence, and serially correlated unobserved heterogeneity are examples. Both lagged duration and occurrence dependence refer to dependence of the termination probability of the spell in progress on either the number or the duration of previous spells. Given such dependence, it is not appropriate to study spells individually, ignoring their interdependence.

In considering the choice of a suitable econometric framework for multiple spells, one possibility is to model dependence using joint survival functions, as discussed in the preceding section. This approach takes care of the multivariate nature of the data. A second possibility is to use the panel data framework with the time subscript replaced by the spell subscript, without ignoring the possibility that calendar time still may have relevance. Spell dependence introduces issues that will be discussed under the topic of dynamic panel models in Sections 22.5 and 23.6. In both these cases an important difference arises from the possibility of censoring because of panel attrition or because the most recent spell is incomplete.

19.4.1. A Model with Two Spells

A proportional hazards model with two spells can illustrate a number of features of multiple-spell models. In econometrics such models have been analyzed by Honoré (1993) and Horowitz and Lee (2003).

Honoré (1993) considers a proportional hazards model of the form

$$\lambda_s(t|\mathbf{x}, \nu) = \lambda_{0,s}(t)\phi(\mathbf{x}, \boldsymbol{\beta})\nu, \quad s = 1, 2. \quad (19.24)$$

Note that in this specification the baseline hazard is spell-specific, but the heterogeneity component, which enters multiplicatively (a key assumption), is not; that is, ν represents the fixed or permanent characteristics of an individual, and hence we have a **fixed effects** model. Under conditions similar to those for the mixed PH discussed in Chapter 18, he shows that the model is identified. He also shows that neither the assumptions

about the distribution of v nor the presence of the covariates is essential for identification.

In a second model Honoré considers spell-specific multiplicative heterogeneity components v_1 and v_2 , with a joint bivariate pdf $g(v_1, v_2)$. The correlation between v_1 and v_2 could reflect serially correlated heterogeneity. This is a **random effects** model. The joint survival function $S(t_1, t_2 | \mathbf{x})$ is derived by the bivariate mixing approach as shown in (19.19) using the mixing distribution $g(v_1, v_2)$. If the marginal survival functions are identified, then the joint survival function is also identified. The identification conditions are essentially those for identifiability of the PH model.

Honoré also considers the **lagged duration dependence** specification of the second-spell model under the assumption that the duration of the first spell, denoted t_1 , enters the hazard for the second-spell multiplicatively. He provides sufficient conditions for identifiability of the parameters in the second-spell conditional model, given covariates and t_1 . These conditions are not discussed here. However, under these conditions, a multiple-spells version of the proportional hazards model has the form

$$\begin{aligned}\lambda_1(t_1 | \mathbf{x}_1, v_1) &= \lambda_{0,1}(t)\phi(\mathbf{x}_1, \beta_1)v_1, \\ \lambda_2(t_2 | \mathbf{x}_2, v_2) &= \lambda_{0,2}(t)\phi(\mathbf{x}_2^a, \beta_2)v_2,\end{aligned}\tag{19.25}$$

where $\mathbf{x}_2^a = (\mathbf{x}_2, t_1)$ is the augmented vector of covariates. Note that there is an endogeneity problem here if v_1 and v_2 are correlated since, in that case, t_1 and v_2 cannot be independent.

The previous occurrence of a spell may not simply shift the hazard function in the succeeding spell. It may also alter the specification of the hazard by bringing in new covariates. For example, an unemployment spell may induce enrollment into a training program, which plausibly could impact the hazard of a later spell of unemployment. If the training variable were treated as weakly exogenous, identification of the model would be under threat. This point is relevant even for the analysis of a single-spell model: The assumption that covariates and unobserved heterogeneity are uncorrelated is not innocuous.

In some cases it may be desirable to model not only multiple spells in one state but also those in other related states. For example, there may be two states, employed or not employed, and we may be interested in not just how length of last unemployment spell affects the length of current unemployment spell but also in the effect of the intervening employment spell on the hazard out of unemployment. Further, we might observe data on individuals when they are in one state but not another. For example, administrative data may cover people when on welfare but not when off welfare.

19.4.2. A More General Model of Multiple Spells

To illustrate the potential computational complexity of multiple-spell models, we describe briefly the model of Mealli and Pudney (1996).

Let $\tau = (\tau_1, \dots, \tau_k)$ denote the k -dimensional vector of complete spells, r_{k-1} the index of origin state, and r_k the index of destination state. Assume independence of durations across spells after controlling for possible lagged duration dependence. Let

$\lambda_j(\mathbf{x}_j, \boldsymbol{\beta}_j)$ denote the destination-specific hazard function, and let $\mathbf{x} = [\mathbf{x}_1, \dots, \mathbf{x}_k]$, $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_k]$.

The joint density of spells and exit routes is given by

$$\begin{aligned}
 & f(\tau_1, r_1, \tau_2, r_2, \dots, \tau_k | \mathbf{x}_1, \dots, \mathbf{x}_k, r_0, \boldsymbol{\beta}) \\
 &= f(\tau_1, r_1 | \mathbf{x}_1, r_0; \boldsymbol{\beta}) \dots f(\tau_{k-1}, r_{k-1} | \mathbf{x}_{k-1}, r_0, r_1, \dots, r_{k-2}, \boldsymbol{\beta}) \\
 &\quad \times S(\tau_k | \mathbf{x}_k, r_0, r_1, \dots, r_{k-1}, \boldsymbol{\beta}) \\
 &= \prod_{j=1}^{k-1} \lambda_{r_j}(\tau_j | \mathbf{x}_j, \boldsymbol{\beta}_{r_j}) \exp\left(-\sum_{l=1}^k \Lambda_0(\tau_l | \mathbf{x}_l, \boldsymbol{\beta})\right),
 \end{aligned} \tag{19.26}$$

where it has been assumed that the k th spell is censored (in progress) and we use relationships (17.4) and (17.6). The covariates include some that vary across spells and possibly lagged durations. This formulation may be compared with the single-spell CRM formulation (19.7).

Mealli and Pudney (1996) build an elaborate model using this formulation as the basis. Because they allow for unobserved heterogeneity with even more complex structure than that considered in this chapter, their computational procedure is also more complicated. They use the method of simulated maximum likelihood (see Section 12.4).

19.5. Competing Risks Example: Unemployment Duration

The duration examples used in Chapters 17 and 18 focused on the time in an unemployment spell, ignoring the destination state after transition. Here we implement a competing risk analysis of the same data used in McCall (1996). The data distinguish three different destination states: full-time employment in the first postdisplacement job, part-time employment in the first postdisplacement job, and either full-time or part-time status in the first postdisplacement job the employee had left by the time of the survey. One can therefore relax the assumption that the hazard function does not depend on the destination state and consider instead the competing risks formulation in which independent competing risks determine the duration of unemployment.

For the McCall data set there are 1073, 339, and 574 transitions, respectively, to each of the three states mentioned. The third destination state lacks a clear interpretation, so the results for that case are not discussed in detail. For each transition we estimated four parametric duration models, exponential and Weibull, with and without inverse-Gaussian heterogeneity. Gamma heterogeneity was also considered but this model was computationally unstable. Because of the assumption of independent competing risks, estimation can be carried out one equation at a time. Selected extracts of the computer output, with focus only on a limited number of variables as in Chapters 17 and 18, are given in Tables 19.2 and 19.3.

Table 19.2. Unemployment Duration: Competing and Independent Risk Estimates of Exponential Model with and without IG Frailty

Risk Coefficient Transitions	No Heterogeneity			IG Heterogeneity		
	Risk 1 1,073	Risk 2 339	Risk 3 574	Risk 1 1,073	Risk 2 339	Risk 3 574
RR	.472 (.601)	-.092 (.976)	-.600 (.725)	.504 (.614)	-.185 (1.025)	-.562 (.744)
DR	-.575 (.762)	-.959 (1.247)	1.122 (.901)	-.806 (.781)	-1.051 (1.295)	1.078 (.921)
UI	-1.424 (.249)	-1.047 (.524)	-.966 (.449)	-1.544 (.258)	-1.092 (.544)	-.963 (.456)
RRUI	.966 (.612)	-.669 (1.192)	-.432 (1.014)	1.057 (.627)	-.742 (1.23)	-.482 (1.033)
DRUI	-.198 (1.019)	1.987 (1.727)	2.102 (1.303)	-.012 (1.041)	2.18 (1.788)	2.158 (1.323)
LNWAGE	.351 (.116)	-.257 (.179)	.003 (.145)	.373 (.118)	-.321 (.191)	-.007 (.147)
TENURE	0 (.006)	.005 (.013)	-.047 (.012)	.0006 (.007)	.007 (.014)	-.047 (.012)
-ln L		5,693.63			5,687.64	

19.5.1. Estimates under Competing Risks Framework

Pairwise comparison of exponential models with and without heterogeneity shows an improvement in the log-likelihood results from the introduction of unobserved heterogeneity. This result is similar to the pattern reported in Section 18.8. However, the Weibull model without heterogeneity has a significantly higher log-likelihood than the exponential model, $-5,666$ against $-5,693$. The Weibull model with inverse-Gaussian heterogeneity has the highest log-likelihood, $-5,543$, and seems to be the best of the four models. This should not be interpreted to mean that it is a satisfactory model for inference – that issue remains open. Henceforth we shall discuss the results in Table 19.3.

Introduction of unobserved heterogeneity in the Weibull model leads to a substantial increase in estimate of the hazard function slope coefficient in all three hazard functions. This coefficient increases from 1.29 to 1.75 for risk 1, and from 1.08 to 1.65 for risk 2. That is, the introduction of unobserved heterogeneity leads to a stronger indication of decreasing duration dependence or steeply rising hazard out of unemployment. These changes are along the lines predicted by the analysis of Section 18.5. In the Weibull model the impact of adding unobserved heterogeneity on the coefficient of unemployment insurance (UI) is also quite substantial, becoming substantially larger in absolute magnitude. The coefficients of RR, DR, RRUI, and DRUI remain imprecisely determined. The coefficient of LNWAGE is significant and positive in the first hazard function, but not in the second. That is, the increase in LNWAGE accelerates

Table 19.3. Unemployment Duration: Competing and Independent Risk Estimates of Weibull Model with and without IG Frailty

Risk Coefficient Transitions	No Heterogeneity			IG Heterogeneity			Cox Model		
	Risk 1 1,073	Risk 2 339	Risk 3 574	Risk 1 1,073	Risk 2 339	Risk 3 574	Risk 1 1,073	Risk 2 339	Risk 3 574
	.448 (.638)	-.085 (.992)	-.694 (.763)	.736 (.906)	-.379 (1.452)	-.432 (1.111)	.522 (-.752)	-.071 (.951)	-.469 (.715)
DR	-.427 (.809)	-.938 (1.279)	1.361 (.969)	-1.072 (1.149)	-1.689 (1.78)	1.167 (1.378)	-.571 (.721)	-1.023 (1.193)	.875 (.878)
UI	-1.496 (.264)	-1.109 (.527)	-1.097 (.46)	-2.574 (.384)	-2.063 (.747)	-1.761 (.623)	-1.317 (.237)	-.906 (.510)	-.905 (.444)
RRUI	1.015 (.646)	-.616 (1.204)	-.305 (1.047)	1.734 (.933)	-.301 (1.702)	-.515 (1.418)	.882 (.582)	-.781 (1.166)	-.539 (1.002)
DRUI	-.299 (1.065)	1.973 (1.757)	1.991 (1.37)	-.06 (1.538)	3.263 (2.47)	3.669 (1.935)	-.095 (.977)	2.031 (1.671)	2.293 (1.274)
LNWAGE	.366 (.122)	-.243 (.183)	.043 (.153)	.576 (.177)	-.494 (.261)	-.006 (.216)	.335 (.110)	-.280 (.173)	-.0140 (.141)
TENURE	-.001 (.007)	.005 (.013)	-.049 (.013)	-.0009 (.01)	.017 (.019)	-.067 (.017)	.000 (.006)	.005 (.012)	-.046 (.011)
α	1.29 (.022)	1.08 (.033)	1.17 (.028)	1.75 (.04)	1.65 (.06)	1.79 (.048)	— —	— —	— —
$-\ln L$	5,666.13				5,543.33				

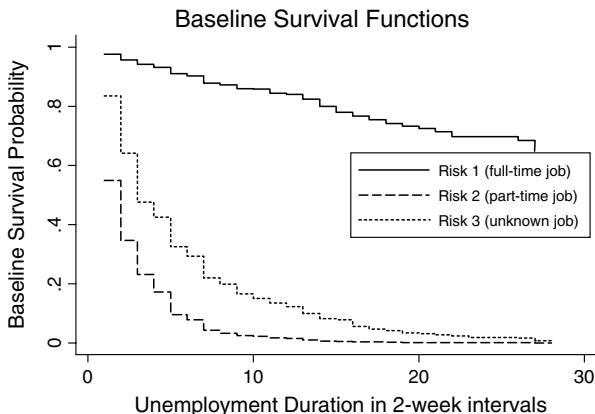


Figure 19.1: Unemployment duration: estimated baseline survival functions from the Cox Competing Risks model. U.S. data from 1986–92 on 3343 spells, some incomplete.

transition out of unemployment of those seeking full-time employment but has a negligible impact on those who transit to part-time employment. This exemplifies how the competing risks framework may allow us to distinguish between the different role of a variable in different hazard functions.

Also consider the Cox model specification of the competing risks model given in Section 19.2. In this specification unobserved heterogeneity is ignored and the baseline hazard is not parametrically specified, but it can be estimated as explained in Section 17.8.3. The point estimates, comparable to those for the exponential model in Table 19.2, are given in the last three columns of Table 19.3, but the standard errors are much larger, as the Cox specification is less restrictive than the exponential. The estimated coefficient of unemployment insurance is closer to that in the exponential model than to that in the Weibull–IG model; the latter is almost twice as large. The LNWAGE coefficient is also larger in the Weibull–IG model. However, given that unobserved heterogeneity is ignored, identification of the baseline hazard is not possible. Figures 19.1 and 19.2 show, respectively, the computed baseline survival functions and the cumulated hazard functions for the three destinations, but these are better interpreted as reflecting some unknown mixture of unobserved heterogeneity and duration dependence. These estimates show that the baseline survival function for those transiting to full-time employment is the lowest and lies below the other two, and that for those transiting to part-time employment it is the flattest and the highest. Correspondingly, the cumulated hazard function for those transiting to full-time employment is the steepest of the three.

The discussion and analysis presented here is only illustrative, not final in any sense. Indeed, there remain good reasons to suggest that the Weibull hazard function is a misspecification. McCall's (1996) analysis of the same data set allows for a more flexible polynomial hazard function and comes up with evidence supporting a bathtub-shaped hazard, which implies decreasing hazard at low durations, then fairly constant and eventually rising hazard at high durations. The monotonic Weibull hazard function does not capture this possibility. The experience of other researchers modeling

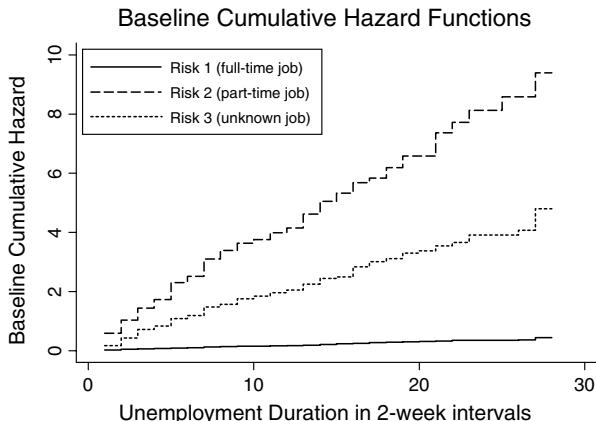


Figure 19.2: Unemployment duration: estimated baseline cumulative hazards from the Cox Competing Risks model. Same data as Figure 19.1.

unemployment duration using the U.S. data has revealed that when the hazard function is flexibly specified, the introduction of unobserved heterogeneity does not have a large impact on the results (Meyer, 1990; Han and Hausman, 1990). The fact that we do not see that result here should motivate the use of a more flexible specification such as the one analyzed in Section 17.10.

19.6. Practical Considerations

In modeling multivariate survival models it is practical to begin with marginal models before undertaking simultaneous estimation. Such a strategy can be helpful in assessing the statistical adequacy of the initial specification.

At the time of this writing, the statistical implementation of multivariate survival and hazard models will in most cases require one's own programming, a task that can be partially eased by the use of supporting software such as optimization programs for maximization or minimization of user-defined functions using functions and programming language offered by many programs and programming platforms.

The CRM with independent risks reduces to estimation of a series of survival models for which practical use information was given in Section 17.12. Programs for general multivariate CRM are not easy to find in commercial software. Some multivariate survival models with special dependence structure are supported. For example, STATA supports computation of the **shared frailty model**. A shared frailty model is a random effects model where the components of unobserved heterogeneity are common to, or shared among, groups of individuals or spells and are randomly distributed across groups.

If the main interest is in modeling the dependence structure among durations, the copula approach, because it does not require numerical integration, is potentially attractive relative to maximum simulated likelihood for the bivariate case. For dimensions higher than two, as in the case of multiple-spell models, it is feasible but there

are relatively few examples in the published literature. Marginal models can be fitted and tested using standard univariate survival models, and the dependence parameter can be estimated in a sequential second-stage procedure. Even if all parameters are to be estimated simultaneously the estimated marginal models provide a useful set of starting values for the iterative computation. We are unaware of statistical software that supports the estimation of these models.

19.7. Bibliographic Notes

- 19.2** Han and Hausman (1990) give an empirical example of CRM in which the specification is generalized to allow for unobserved heterogeneity. Within the framework of the CRM with state-specific random effects, McCall (1996) analyzes the impact of some policy variables on the behavior of the insured unemployed seeking part-time work using the CRM model with correlated risks. In Butler, Anderson, and Burkhauser (1989) the hazards of accepting a job and of dying are modeled using a CRM with correlated risks.
- 19.3** Sklar's pioneering article on copulas appeared in 1959 in French, but Sklar (1973) is a good substitute in English. Radulović and Wegkamp (undated) provide a proof of Sklar's Theorem. A very helpful guided tour of the copula literature with an annotated bibliography is given by Frees and Valdez (1998).
- 19.4** Multiple spells are studied by Mealli and Pudney (1996) and by Flinn and Heckman (1982). Mealli and Pudney (1996) analyze transitions among pensionable jobs, nonpensionable jobs, and other labor market states using simulation-based estimation methods.

Exercises

- 19–1** (Adapted from Sapra, 2000; 2001). This problem involves an example that illustrates the Cox–Tsiatis nonidentification of the competing risks result mentioned in Section 19.2. Consider the following *dependent* competing risks model in which we observe $T = \min(T_1, T_2)$ and δ , where $\delta = 1$ if $T = T_1$, and $\delta = 2$ if $T = T_2$. Here T_1 and T_2 are latent durations of risks 1 and 2, respectively. Suppose that the bivariate joint survivor function is $S(t_1, t_2) = \exp[-(\lambda_1 t_1 + \lambda_2 t_2)^\alpha]$, $0 < \alpha \leq 1$, $\lambda_1, \lambda_2 > 0$. Construct an independent CRM that is equivalent to the specified dependent competing risks model.
- 19–2** For the model specified in the preceding problem, write down the log-likelihood function for each model in terms of hazard rates and integrated hazard rates, if both T and δ are observed. Examine the information matrix of the parameters, and show that all the parameters are locally identified because it is nonsingular.
- 19–3** Consider two parallel durations, say duration of unemployment, T_1 , and the duration of a spell without private health insurance, T_2 . Assume that conditional on unobserved heterogeneity the durations are independent and exponentially distributed with means $\beta_0 + \beta_1 x$ and $\gamma_0 + \gamma_1 x$, respectively. Suppose that multiplicative unobserved heterogeneity terms for the two duration models are v_1 and v_2 , with $E[v_1] = E[v_2] = 1$.
 - (a)** For parameter values of your choice, write an algorithm to generate correlated realizations for (v_1, v_2) such that unconditionally on (v_1, v_2) , but conditionally on x , the two durations will be correlated. You are free to

make distributional assumptions for the joint distribution of (v_1, v_2) that are appealing on grounds of mathematical convenience or other considerations. Explain how you can control the extent of correlation between the two durations.

- (b) Using the technique for obtaining a bivariate joint distribution given in Section 19.3.2, derive the joint distribution of durations.
- (c) Describe how you might extend the analysis of part (b) to allow for the presence of right-censored durations.
- 19–4** Using the same subsample of the McCall data set as in Chapter 18, estimate using a two-state model with unemployment and employment as the two states, (i.e., ignoring the distinction between part-time and full-time employment as two alternative destinations).
- (a) Fit the single-equation Weibull model and compare the results with those for independent CRM with the Weibull specification.
- (b) Evaluate the improvement in goodness of fit resulting from the CRM specification.
- (c) Evaluate and compare the fitted values of the hazard out of unemployment, evaluated at sample averages of the explanatory variables, from the single equation and the CRM models.

CHAPTER 20

Models of Count Data

20.1. Introduction

In many economic contexts the dependent or response variable of interest is a non-negative integer or count that we wish to explain or analyze in terms of a set of regressors. Unlike the classical regression model, the response variable is discrete, with a distribution that places probability mass at nonnegative integer values only. Several models discussed earlier in the book, such as the binary outcome model and the duration model, can be shown to be closely related to the count data regression model. Regression models for counts, like other limited or discrete dependent variable models such as the logit and probit, are nonlinear with many properties and special features intimately connected to discreteness and nonlinearity.

Let us consider some examples from microeconomics, beginning with sample data that are independent cross-section observations. Fertility studies often model the number of live births over a specified age interval of the mother, with interest in analyzing its variation in terms of, say, mother's schooling, age, and household income (Winkelmann, 1995). In some models of family decisions the number of children may appear as an explanatory variable with the acknowledgment that the variable is endogenous. Accident analysis studies model airline safety as measured by the number of accidents experienced by an airline over some period and seek to determine its relationship to airline profitability and other measures of the financial health of the airline (Rose, 1990). Recreational demand studies seek to place a value on natural resources such as national forests by modeling the number of trips to a recreational site (Gurmu and Trivedi, 1996). Health demand studies model data on the number of times that individuals consume a health service, such as visits to a doctor or days in the hospital in the past year (Cameron et al., 1988). If we wish to analyze the relation between this variable and factors such as health status and health insurance, again a count regression is relevant.

The main modeling approaches are presented in Sections 20.2–20.5. Section 20.2 details the Poisson regression model. Section 20.3 gives an application to data from the famous RHIE. The Poisson regression model is often too restrictive and other, more

Table 20.1. Proportion of Zero Counts in Selected Empirical Studies

Study	Variable	Sample Size	Proportion of Zeros
Cameron et al. (1988)	Doctor visits	5,190	0.798
Pohlmeier and Ulrich (1995)	Specialist visits	5,096	0.678
Grootendorst (1995)	Prescription drugs	5,743	0.224
Deb and Trivedi (1997)	Number of hospital stays	4,406	0.806
Gurmu and Trivedi (1996)	Recreational trips	659	0.632
Geil et al. (1997)	Hospitalizations	30,590	0.899
Greene (1997)	Major derogatory reports	1,319	0.803

commonly used, fully parametric count models are presented in Section 20.4. Less-used alternative parametric approaches for counts, such as discrete choice models, are also presented in this section. The partially parametric approach of modeling the conditional mean and conditional variance is detailed in Section 20.5. Multivariate count models and models with endogenous regressors are given an introductory treatment in Section 20.6. Section 20.7 illustrates various models by application to the RHIE data. This is followed by a discussion of some practical issues. For pedagogical reasons the Poisson regression model for cross-section data is presented in some detail. The other models, many superior to Poisson, are presented in less detail for space reasons. For more complete treatment see Cameron and Trivedi (1998) and the Bibliographic Notes.

20.2. Basic Count Data Regression

In some cases, such as number of births, the count is the variable of ultimate interest. In other cases, such as medical demand and results of research and development expenditure, the variable of ultimate interest is continuous, often expenditures or receipts measured in dollars, but the best data available are instead a count. In many cases, the sample is concentrated on a **few small discrete values**, say 0, 1, and 2. Table 20.1 illustrates this point by reference to the proportion of zero counts observed in several published econometric models; this proportion can be as high as 90% in some cases. Also, the data can be **skewed to the right**. Finally, the data are intrinsically **heteroskedastic** with variance increasing with the mean.

20.2.1. Poisson Regression

The Poisson is the starting point for count data analysis, though it is often inadequate. In Sections 20.2.1–20.2.3 we present the Poisson regression model, which was previously introduced in Section 5.2, and estimation by maximum likelihood, interpretation of the estimated coefficients, and extensions to truncated and censored data. In Section 20.2.3 we also present the quasi-MLE based on the Poisson distribution with

Table 20.2. Summary of Data Sets Used in Recent Patent–R&D Studies

Study	Sample Size	Mean	Std. Error	Maximum Patents	Proportion of Zeros
Cincera (1997)	181	60.8	721.6	925	<0.19
Crepon and Duguet (1997b)	698	11.6	na ^a	na	0.441
Crepon and Duguet (1997a)	451	2.73	11.45	na	0.729
Hausman et al. (1984)	346	32.1	66.36	515	0.220
Wang et al. (1998)	70	23.46	39.10	173	0.186

^a na, not available.

correctly specified conditional mean, but with possibly misspecified conditional variance function. Limitations of the Poisson model, notably its property of equidispersion, are presented in Section 20.2.4.

There is a qualification: In some cases a high proportion of zeros in the sample may coexist with very large values of counts, creating a difficult modeling challenge. Table 20.2 illustrates this feature using information from five studies that have investigated the relationship between patent counts and research and development (R&D) expenditure. Observe how large the maximum observed value of the count is relative to the sample mean. The modeling challenge is to select a functional form that can adequately capture the large mean and the high proportion of zeros. In many other examples, such as number of births, virtually all the data are restricted to single digits, and the mean number of events is quite low.

These features motivate the application of special methods and models for count regression. There are two ways to proceed.

The first approach is a **fully parametric** one that completely specifies the distribution of the data, fully respecting the restriction of y to nonnegative integer values. This approach was taken in early applications, mostly in biostatistics, where count regression was seen as an extension and generalization of a vast literature on the distribution of independent and identically distributed counts. It was also taken in the influential econometrics study by Hausman et al. (1984).

The second approach is a **mean–variance approach**, which specifies the conditional mean to be nonnegative and specifies the conditional variance to be a function of the conditional mean. This models well the nonnegativity and heteroskedasticity but does not address the discreteness of the data. This approach, in a framework not limited to only count data, was introduced by Nelder and Wedderburn (1972), leading to the generalized linear model approach widely used in statistics (McCullagh and Nelder, 1989). In econometrics this approach was introduced by Gouriéroux, Monfort, and Trognon (1984a,b) and is best viewed as a specialization of generalized methods of moments.

20.2.2. Poisson MLE and QMLE

The Poisson MLE and quasi-MLE (QMLE) were introduced and studied in Chapter 5 as an example of m -estimation. Here we give a more complete treatment.

The natural stochastic model for counts is a Poisson point process for the occurrence of the event of interest. This implies a **Poisson distribution** for the number of occurrences of the event, with density, or more formally probability mass function,

$$\Pr[Y = y] = \frac{e^{-\mu} \mu^y}{y!}, \quad y = 0, 1, 2, \dots, \quad (20.1)$$

where μ is the intensity or rate parameter. We refer to the distribution as $\mathcal{P}[\mu]$. The first two moments are

$$\begin{aligned} E[Y] &= \mu, \\ V[Y] &= \mu. \end{aligned} \quad (20.2)$$

This shows the well-known **equidispersion** (equality of mean and variance) property of the Poisson distribution.

By introducing the observation subscript i , attached to both y and μ , the iid framework is extended to the regression case. The **Poisson regression model** is derived from the Poisson distribution by parameterizing the relation between the mean parameter μ and covariates (regressors) \mathbf{x} . The standard assumption is to use the exponential mean parameterization,

$$\mu_i = \exp(\mathbf{x}'_i \boldsymbol{\beta}), \quad i = 1, \dots, N, \quad (20.3)$$

where by assumption there are K linearly independent covariates, usually including a constant. Because $V[y_i | \mathbf{x}_i] = \exp(\mathbf{x}'_i \boldsymbol{\beta})$, by (20.2) and (20.3), the Poisson regression is intrinsically heteroskedastic.

Given (20.1) and (20.3) and the assumption that the observations $(y_i | \mathbf{x}_i)$ are independent, the most natural estimator is maximum likelihood. The log-likelihood function is

$$\ln L(\boldsymbol{\beta}) = \sum_{i=1}^N \{y_i \mathbf{x}'_i \boldsymbol{\beta} - \exp(\mathbf{x}'_i \boldsymbol{\beta}) - \ln y_i!\}. \quad (20.4)$$

The **Poisson MLE**, denoted $\hat{\boldsymbol{\beta}}_P$, is the solution to K nonlinear equations corresponding to the first-order condition for maximum likelihood,

$$\sum_{i=1}^N (y_i - \exp(\mathbf{x}'_i \boldsymbol{\beta})) \mathbf{x}_i = \mathbf{0}. \quad (20.5)$$

If \mathbf{x}_i includes a constant term then the residuals $y_i - \exp(\mathbf{x}'_i \boldsymbol{\beta})$ sum to zero by (20.5). The log-likelihood function is globally concave; hence solving these equations by a Gauss–Newton or Newton–Raphson iterative algorithm yields unique parameters estimates.

In the econometrics literature **pseudo-ML** (PML) or **quasi-ML** (QML) estimation refers to estimating by ML, under misspecification of the specified density (Gourieroux et al., 1984a). The terms PML and QML are often used interchangeably. The distribution of the estimator is obtained under weaker assumptions about the data-generating process than those that led to the specified likelihood function; see Section 5.7. In the statistics literature QML often refers to nonlinear generalized least-squares estimation.

For the Poisson regression, QML in the latter sense is equivalent to standard maximum likelihood.

From (20.5), the Poisson PML estimator, $\widehat{\beta}_P$, has first-order conditions $\sum_{i=1}^N (y_i - \exp(\mathbf{x}'_i \beta)) \mathbf{x}_i = \mathbf{0}$. As already noted, the summation on the left-hand side has expectation zero if $E[y_i | \mathbf{x}_i] = \exp(\mathbf{x}'_i \beta)$. Hence the Poisson PML is consistent under the weaker assumption of correct specification of the conditional mean; that is, the data need not be Poisson distributed. Using results given in Section 5.2.3, the variance matrix is of the sandwich form, with

$$\text{V}_{\text{PML}}[\widehat{\beta}_P] = \left(\sum_{i=1}^N \mu_i \mathbf{x}_i \mathbf{x}'_i \right)^{-1} \left(\sum_{i=1}^N \omega_i \mathbf{x}_i \mathbf{x}'_i \right) \left(\sum_{i=1}^N \mu_i \mathbf{x}_i \mathbf{x}'_i \right)^{-1} \quad (20.6)$$

and $\omega_i = V[y_i | \mathbf{x}_i]$ is the conditional variance of y_i .

By standard ML theory if the stronger assumption is made that the Poisson regression is parametrically correctly specified, so that $\omega_i = \mu_i$, the estimator $\widehat{\beta}_P$ is consistent for β and asymptotically normal with the sample covariance matrix

$$\text{V}[\widehat{\beta}_P] = \left(\sum_{i=1}^N \mu_i \mathbf{x}_i \mathbf{x}'_i \right)^{-1}, \quad (20.7)$$

in the case where μ_i is of the exponential form (20.3).

The Poisson ML and PML estimators are identical but have different variances. The empirical implementation of the more robust estimate (20.6) is presented in Section 20.5.1.

20.2.3. Interpretation of Regression Coefficients

For linear models, with $E[y|\mathbf{x}] = \mathbf{x}'\beta$, the coefficients β are readily interpreted as the effect of a one-unit change in regressors on the conditional mean. For nonlinear models this interpretation needs to be modified; see the general discussion given in Section 5.2.4. For any model with exponential conditional mean, differentiation yields

$$\frac{\partial E[y|\mathbf{x}]}{\partial x_j} = \beta_j \exp(\mathbf{x}'\beta), \quad (20.8)$$

where the scalar x_j denotes the j th regressor. For example, if $\widehat{\beta}_j = 0.25$ and $\exp(\mathbf{x}'_i \widehat{\beta}) = 3$, then a one-unit change in the j th regressor increases the expectation of y by 0.75 units. This partial response depends on $\exp(\mathbf{x}'_i \widehat{\beta})$, which is expected to vary across individuals. It is easy to see that β_j measures the relative change in $E[y|x]$ induced by a unit change in x_j . If x_j is measured on a logarithmic scale, β_j is an elasticity.

For purposes of reporting a single response value, a good candidate is an estimate of the average response, $N^{-1} \sum_i \partial E[y_i | \mathbf{x}_i] / \partial x_{ij} = \widehat{\beta}_j \times N^{-1} \sum_i \exp(\mathbf{x}'_i \widehat{\beta})$. For Poisson regression models with intercept included, this can be shown to simplify to $\widehat{\beta}_j \bar{y}$.

Another consequence of (20.8) is that if, say, β_j is twice as large as β_k , then the effect of changing the j th regressor by one unit is twice that of changing the k th regressor by one unit.

20.2.4. Overdispersion

The Poisson regression model is usually too restrictive for count data, leading to alternative models presented in Sections 20.3 and 20.4. The fundamental problem is that the distribution is parameterized in terms of a single scalar parameter (μ) so that all moments of y are a function of μ . By contrast the normal distribution has separate parameters for location (μ) and scale (σ^2). For the same reason the one-parameter exponential is too restrictive for duration data and more general two-parameter distributions such as the Weibull are superior. Note that this complication does not arise with binary data. Then the distribution is clearly the one-parameter Bernoulli, because if the probability of success is p then the probability of failure must be $1 - p$. For binary data the issue is instead how to parameterize p in terms of regressors.

One way this restrictiveness manifests itself is that in many applications a Poisson density predicts the probability of a zero count to be considerably less than is actually observed in the sample. This is termed the **excess zeros** problem, as there are more zeros in the data than the Poisson predicts.

A second and more obvious deficiency of the Poisson model is that for count data the variance usually exceeds the mean, a feature called **overdispersion**. The Poisson model instead implies equality of the variance and the mean (see (20.2)), a property called equidispersion.

Overdispersion has qualitatively similar consequences to the failure of the assumption of homoskedasticity in the linear regression model. Provided the conditional mean is correctly specified, that is, (20.3) holds, the Poisson MLE is still consistent. This is clear from inspection of the first-order conditions (20.5), since the left-hand side of (20.5) will have expected value of zero if $E[y_i | \mathbf{x}_i] = \exp(\mathbf{x}'_i \boldsymbol{\beta})$. This consistency property applies more generally to the quasi-MLE when the specified density is in the LEF. Both Poisson and normal are members of the LEF discussed earlier in Section 5.7.3. It is nonetheless important to control for overdispersion. First, in more complicated settings such as with truncation and censoring, overdispersion leads to the more fundamental problem of inconsistency. Second, even in the simplest settings large overdispersion leads to grossly deflated standard errors and grossly inflated t -statistics in the usual ML output, and hence it is important to use the previously given robust variance estimator. Third, if one wants to estimate probabilities of number of events, rather than merely the conditional mean, these depend on additional parameters of the dgp.

Overdispersion may signal a presence of a more basic misspecification, especially in data settings that involve truncation and censoring if they are ignored in estimation. In such a case the conditional mean is incorrectly specified and the simultaneous presence of overdispersion then leads to inconsistency, not only inefficiency, of the MLE.

A statistical test of overdispersion is therefore highly desirable after running a Poisson regression. Most count models with overdispersion specify overdispersion to be of the form

$$V[y_i | \mathbf{x}_i] = \mu_i + \alpha g(\mu_i), \quad (20.9)$$

where α is an unknown parameter and $g(\cdot)$ is a known function, most commonly $g(\mu) = \mu^2$ or $g(\mu) = \mu$. It is assumed that under both null and alternative hypotheses the mean is correctly specified as, for example, $\exp(\mathbf{x}_i'\beta)$, whereas under the null hypothesis $\alpha = 0$ so that $V[y_i|\mathbf{x}_i] = \mu_i$. A simple **overdispersion test statistic** for $H_0 : \alpha = 0$ versus $H_1 : \alpha \neq 0$ or $H_1 : \alpha > 0$ can be computed by estimating the Poisson model, constructing fitted values $\hat{\mu}_i = \exp(\mathbf{x}_i'\hat{\beta})$, and running the auxiliary OLS regression (without constant)

$$\frac{(y_i - \hat{\mu}_i)^2 - y_i}{\hat{\mu}_i} = \alpha \frac{g(\hat{\mu}_i)}{\hat{\mu}_i} + u_i, \quad (20.10)$$

where u_i is an error term. The reported t -statistic for α is asymptotically normal under the null hypothesis of no overdispersion (Cameron and Trivedi, 1990) even though here generated regressors are used. This test can also be used for **underdispersion**, $\alpha < 0$, in which case the conditional variance is less than the conditional mean. See also Gurmu and Trivedi (1992).

20.3. Count Example: Contacts with Medical Doctor

For illustration we use some of the data from the RAND Health Insurance Experiment previously used by Deb and Trivedi (2002). They estimated a more complete set of models and carried out a deeper analysis of the data than is possible or desirable here. The experiment, conducted by the RAND Corporation from 1974 to 1982, has been the longest running and largest controlled social experiment in medical care research. The main goal of the experiment was to assess how the patient's use of health services is affected by types of randomly assigned health insurance, including both fee-for-service and health maintenance organizations (HMOs). In the experiment the data were collected from about 8,000 enrollees in 2,823 families, from six sites across the country. Each family was enrolled in one of 14 different health insurance plans for either three or five years. The plans ranged from free care to 95% coinsurance below a maximum dollar expenditure (MDE), and also included assignment in a prepaid group practice.

The key point is that because insurance plans are randomly assigned, not freely chosen by the participants, we do not face the problem of endogenous treatment effect, which is the central causal parameter of interest in the study.

Data were collected from the enrollee's use of medical care services and health status throughout the randomly assigned term of enrollment for either three or five years. For additional details of the data see Manning et al. (1987), Newhouse et al. (1993), and Deb and Trivedi (2002). The sample used in this study consists of individuals in the fee-for-service plans only.

The data file consists of utilization, expenditures, demographic characteristics, health status, and insurance status variables. The expenditure data were analyzed in Section 16.6. The coinsurance rate in this sample assumes four different values. Yet, following the RAND studies, we treat it as a continuous variable. The final sample consists of 20,186 observations; each observation represents data for an experimental

Table 20.3. Contacts with Medical Doctor: Frequency Distribution

Contacts	0	1	2	3	4	5	6	7	8	9	10
Relative Frequency	31.2	18.9	13.8	9.3	6.7	4.8	3.4	2.6	2.0	1.4	1.0
Contacts	11	12	13	14	15	16	...	>21	Max		
Relative Frequency	0.9	0.6	0.5	0.4	0.3	0.3		1.0	77		

subject in a given year. For simplicity of exposition the resulting clustering in the data, see Section 24.5, is ignored here.

In the present illustration the measure of utilization analyzed is the number of contacts with a medical doctor (MDU). The relative frequency distribution of MDU, given in percentages, is given in Table 20.3. MDE denotes maximum dollar expenditure, the medical expenditure liability limit set in the experiment above which the participant would not be responsible for cost-sharing. Observe that about 31% of the observations are zeros. The long right tail and variance greatly exceeding the mean indicates that the counts are (unconditionally) overdispersed.

For the purposes of discussion here we consider the regression to be estimated by Poisson ML and by Poisson PML. Other specifications are considered later. The included covariates in all cases are those in Table 20.4.

Table 20.4. Contacts with Medical Doctor: Variable Descriptions

Variable	Definition	Mean	Std. Dev.
MDU	Number of outpatient visits to an MD	2.861	4.505
LC	$\ln(\text{coinsurance} + 1)$, $0 \leq \text{coinsurance} \leq 100$	1.710	1.962
IDP	1 if individual deductible plan, 0 otherwise	0.220	0.414
LPI	$\ln(\max(1, \text{annual participation incentive payment}))$	4.709	2.697
FMDE	0 if IDP = 1 $\ln(\max(1, \text{MDE}/(0.01 \text{ coinsurance})))$ otherwise	3.153	3.641
LINC	$\ln(\text{family income})$	8.708	1.228
LFAM	$\ln(\text{family size})$	1.248	0.539
AGE	Age in years	25.718	16.768
FEMALE	1 if person is female	0.517	0.500
CHILD	1 if age is less than 18	0.402	0.490
FEMCHILD	FEMALE * CHILD	0.194	0.395
BLACK	1 if race of household head is black	0.182	0.383
EDUCDEC	Education of the household head in years	11.967	2.806
PHYSLIM	1 if the person has a physical limitation	0.124	0.322
NDISEASE	Number of chronic diseases	11.244	6.742
HLTHG	1 if self-rated health is good	0.362	0.481
HLTHF	1 if self-rated health is fair	0.077	0.267
HLTHP	1 if self-rated health is poor	0.015	0.121
Omitted category is excellent self-rated health			

Table 20.5. Contacts with Medical Doctor: Count Model Estimates

Model	Poisson		PPML		NB2-PML	
	Coeff.	t-ratio	t-ratio		Coeff.	t-ratio
LC	−.0427	−7.030	−2.835	−0.0504	−3.228	
IDP	−.1613	−13.881	−5.773	−0.1475	−4.889	
LPI	0.0128	6.999	2.912	0.0158	3.574	
FMDE	−.0206	−5.803	−2.319	−0.0213	−2.351	
PHYSLIM	0.2684	21.711	8.240	0.2751	8.068	
NDISEASE	0.0231	38.124	13.487	0.0259	15.324	
HLTHG	0.0394	4.109	1.699	0.0065	0.275	
HLTHF	0.2531	15.613	5.894	0.2368	5.425	
HLTHP	0.5216	19.150	6.966	0.4256	6.205	
α	—	—	—	1.1822	8.926	
−ln L	60087			42777		

A selection of interesting coefficients and their t -ratios are given in Table 20.5, along with log-likelihood and information criteria. To save space we do not reproduce all the output. The coefficients of variables associated with insurance variables (LC, IDP, LPI, and FMDE) are clearly of interest since they reflect the price sensitivity of utilization. Also of interest are the coefficients of the five health status variables (PHYSLIM, NDISEASE, HLTHG, HLTHF, and HLTHP).

Consider the coefficient of the coinsurance rate, here measured on the log scale, LC. This variable is of major interest as it provides information about the price effect. The higher the coinsurance rate, the greater will be the extent of cost sharing by the patient, and hence the lower will be the average number of visits. The estimated coefficient from the Poisson regression (see column 1 in Table 20.5) is negative (−.042), with a t -ratio of 2.835, indicating that the price effect is significantly negative as predicted by standard theory. The elasticity of the number of doctor visits with respect to LC is −.042. However, care should be exercised in interpreting this value as the coinsurance rate only takes a few values and does not vary continuously. Subject to this qualification, the coefficient can be interpreted as elasticity. A similar value for log of income (LINC) is 0.174, indicating that increase in income raises the average number of visits.

How well does the Poisson regression fit the data? One simple way to judge this is to compare the actual and fitted frequencies for different number of doctor visits. Table 20.6 provides such a comparison for up to nine visits, ignoring the higher frequencies that collectively account for less than 10% of the visits. To calculate the fitted value $\Pr[y_i | \mathbf{x}_i' \hat{\beta}]$ for $y_i = 0, 1, \dots, 9$, we plug $\hat{\mu}_i$ into (20.1) and then average over all the observations. Observe that the Poisson regression seriously underpredicts the proportion of zero visits and overestimates the proportion of positive number of visits up to seven. Thus we conclude that the Poisson regression is deficient. This pattern in the lack of fit can be shown to be associated with the neglect of overdispersion in the data (Cameron and Trivedi, 1998, chapter 4).

Table 20.6. Contacts with Medical Doctor: Observed and Fitted Frequencies

Contact frequency	0	1	2	3	4	5	6	7	8	9
Relative frequency	31.2	18.9	13.8	9.3	6.7	4.8	3.4	2.6	2.0	1.4
Poisson fitted	10.6	19.2	20.9	17.6	12.6	7.99	4.69	2.64	1.46	0.8
NB2 fitted	30.9	19.6	13.6	9.67	6.97	5.07	3.70	2.72	2.0	1.47

In the presence of neglected overdispersion it is to be expected that the t -ratios of the Poisson MLE will be inflated. A comparison with the robust t -ratios in column 3 (PPML) of Table 20.5 shows that this is indeed so. For example, robustification causes the t -ratio of LC to drop from -7.03 to -2.83 . Tables 20.5 and 20.6 include results for the NB2 model that are discussed in Section 20.7. The NB2 model is a better parametric model for these data.

20.4. Parametric Count Regression Models

Poisson regression is often too restrictive. In this section we present a number of more flexible parametric alternatives to the Poisson.

First, overdispersion in count data may be due to unobserved heterogeneity. In such a case counts are viewed as being generated by a Poisson process (in which case the events are serially independent), but the researcher is unable to correctly specify the rate parameter of this process. Instead, the rate parameter is itself a random variable. This mixture approach, presented in Sections 20.4.1 and 20.4.2, leads to the widely used negative binomial model.

Second, overdispersion, and in some cases underdispersion, may arise because the process generating the first event may differ from that determining later events. For example, an initial doctor consultation may be solely a patient's choice, whereas subsequent visits are also determined by the doctor. This leads to the modified count models presented in Section 20.4.5.

Third, overdispersion in count data may be due to failure of the assumption of independence of events, which is implicit in the Poisson process. One can postulate dependence so that, for example, the occurrence of one doctor visit makes subsequent doctor visits more likely. (This approach has not been widely used in count data analysis. In duration data analysis this is called true state dependence.) Particular assumptions about unobserved heterogeneity or dependence again lead to the negative binomial; see Winkelmann (1995). A discrete choice model that progressively models $\Pr[y = j | y \geq j - 1]$ is presented in Section 20.4.6.

Fourth, one can refer to the extensive and rich literature on univariate iid count distributions, such as the logarithmic series and hypergeometric distribution (Johnson, Kotz, and Kemp, 1992). New regression models can be developed by letting one or more distribution parameters be a specified function of regressors. Such models are not presented here. The approach has less motivation than the first three approaches and the resulting models may not be any better.

Although overdispersion has been emphasized, underdispersion may also arise. For example, a sample in which the counted outcome is largely 0 or 1, with a very small number of 2s, and hence close to a binomial model, will show underdispersion. Members of the Katz family of distributions, or other distributions based on the series expansion methods such as those developed in Cameron and Johansson (1997), can be used; see also Cameron and Trivedi (1998, chapter 12).

20.4.1. Negative Binomial Model

The negative binomial model, a specific example of a continuous mixture model, can be obtained in many different ways. The following justification using a mixture distribution is one of the oldest and has wide appeal.

Suppose the distribution of a random count y is Poisson, conditional on the parameter λ , so that $f(y|\lambda) = \exp(-\lambda)\lambda^y/y!$. Suppose now that the parameter λ is random, rather than being a completely deterministic function of regressors \mathbf{x} . In particular, let $\lambda = \mu v$, where μ is a deterministic function of \mathbf{x} , for example $\exp(\mathbf{x}'\beta)$, and $v > 0$ is iid with density $g(v|\alpha)$. This is an example of unobserved heterogeneity, as different observations may have different λ (heterogeneity) but part of this difference is due to a random (unobserved) component v . Note that $E[\lambda|\mu] = \mu$ if $E[v] = 1$, so the interpretation of the slope parameters stays as in the Poisson model.

The marginal density of y , unconditional on the random parameter v but conditional on the deterministic parameters μ and α , is obtained by integrating out v . This yields

$$h(y|\mu, \alpha) = \int f(y|\mu, v)g(v|\alpha)dv, \quad (20.11)$$

where $g(v|\alpha)$ is called the mixing distribution and α denotes the unknown parameter of the mixing distribution. The integration defines an “average” distribution. For some specific choices of $f(\cdot)$ and $g(\cdot)$, the integral will have an explicit or closed-form solution.

If $f(y|\lambda)$ is the Poisson density and $g(v) = v^{\delta-1}e^{-v\delta}\delta^\delta/\Gamma(\delta)$, $v, \delta > 0$, is the gamma density with $E[v] = 1$ and $V[v] = 1/\delta$, we obtain the **negative binomial** as a mixture density as follows:

$$\begin{aligned} h[y|\mu, \delta] &= \int_0^\infty \frac{e^{-\mu v} (\mu v)^y}{y!} \frac{v^{\delta-1} e^{-v\delta} \delta^\delta}{\Gamma(\delta)} dv \\ &= \int_0^\infty \frac{e^{-(\mu+\delta)v} \mu^y}{y!} \frac{v^{y+\delta-1} \delta^\delta}{\Gamma(\delta)} dv \\ &= \frac{\mu^y \delta^\delta}{\Gamma(\delta) y!} \int_0^\infty e^{-(\mu+\delta)v} v^{y+\delta-1} dv \\ &= \frac{\mu^y \delta^\delta \Gamma(y+\delta)}{\Gamma(\delta) y! (\mu+\delta)^{y+\delta}} \\ &= \frac{\Gamma(\alpha^{-1}+y)}{\Gamma(\alpha^{-1})\Gamma(y+1)} \left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu} \right)^{\alpha^{-1}} \left(\frac{\mu}{\mu+\alpha^{-1}} \right)^y, \end{aligned} \quad (20.12)$$

where $\alpha = 1/\delta$, $\Gamma(\cdot)$ denotes the gamma integral which specializes to a factorial for an integer argument, and the fourth line follows after some algebra and use of the definition of the gamma function. Special cases of the negative binomial include the Poisson ($\alpha = 0$), ‘the advantage of reparametrization from δ to α ,’ and the geometric ($\alpha = 1$).

As in the case of many mixture distributions, the negative binomial also has independent justification; see Cameron and Trivedi (1998, chapter 4). It can arise in many different ways and one need not always think of it as a mixture distribution.

The algebraic derivation of the negative binomial as a **Poisson–gamma mixture** can be given a Bayesian interpretation. The prior distribution of μ is gamma, given α , and the results on conjugate priors for exponential families in Section 13.2.4. It is expected that the posterior distribution has a closed form. Therefore, the MLE and the Bayesian posterior mean, under the further assumption of a vague prior on α , would coincide.

The first two moments of the negative binomial distribution are

$$\begin{aligned} E[y|\mu, \alpha] &= \mu, \\ V[y|\mu, \alpha] &= \mu(1 + \alpha\mu). \end{aligned} \tag{20.13}$$

The variance therefore exceeds the mean, since $\alpha > 0$ and $\mu > 0$. Indeed, it can be shown easily that overdispersion always arises if $y|\lambda$ is Poisson and unobserved heterogeneity is of the multiplicative form $\lambda = \mu\nu$, where $E[\nu] = 1$. Note also that the overdispersion is of the form (20.9) discussed in Section 20.2.4.

Two standard variants of the negative binomial are used in regression applications. Both variants specify $\mu_i = \exp(\mathbf{x}'_i\boldsymbol{\beta})$. The most common variant lets α be a parameter to be estimated, in which case the conditional variance function, $\mu + \alpha\mu^2$ from (20.13), is quadratic in the mean.

The other variant of the negative binomial model has a linear variance function, $V[y|\mu, \alpha] = (1 + \gamma)\mu$, obtained by replacing α by γ/μ throughout (20.12). Estimation by ML is again straightforward. Sometimes this variant is called negative binomial 1 (NB1) in contrast to the variant with a quadratic variance function which has been called the negative binomial 2 (NB2) model (Cameron and Trivedi, 1998). The log-likelihood is easily obtained from (20.12). Both variants of the model are easily estimated by ML, with details given in, for example, Cameron and Trivedi (1998). In both variants the coefficients have the same interpretation since $E[y|\mathbf{x}] = \exp(\mathbf{x}'\boldsymbol{\beta})$. The NB2 variant is the most often used, as in the application in Section 20.7.

The NB2 model has been found to be very useful in applied work. It appears to have the flexibility necessary for providing a good fit to many types of count data. It does so in part because the quadratic variance specification is a good approximation in many empirical situations. An unfortunate consequence of the fact that NB2 often provides a good fit is that if the Poisson assumption fails there is a tendency to jump to the negative binomial alternative, ignoring other possibilities. Such a mechanical approach should be avoided because poor performance of the Poisson can also be due to a poor specification of the conditional mean function, and observe that using the negative binomial model maintains the same conditional mean.

The negative binomial model is less robust to distributional misspecification than the Poisson. Even if the conditional mean is correctly specified the MLE in negative binomial models is inconsistent, except for the special case of the NB2 model, whereas the MLE for β (but not α) is still consistent.

For mixture models for counts, the Poisson is the natural choice for the initial density $f(y|\mu, \nu)$ in (20.12) since a Poisson process is a natural model for counts. The choice of the gamma for the mixing distribution $g(\nu)$ in (20.12) is more arbitrary. Its use raises issues discussed in Section 18.2–18.4. Other possible choices include the lognormal distribution and the inverse-Gaussian distribution. See Willmot (1987) and Guo and Trivedi (2002). In these cases the marginal distribution cannot be expressed in a closed form, as it is the gamma that is conjugate to the Poisson. Of course, this does not mean that the resulting model cannot be estimated by maximum likelihood. It means simply that one may have to use numerical quadrature or simulated maximum likelihood to estimate the model. These methods are entirely feasible with currently available computing power. If one is prepared to use the simulation-based estimation methods discussed in chapter 12, the scope for using mixed-Poisson models of various types becomes very extensive.

20.4.2. Simulated Maximum Likelihood

Purely for purposes of illustration we now illustrate how we might estimate the NB2 model by **maximum simulated likelihood**. The reader should understand that in practice this is unnecessary because we already have an analytical expression for that model. Suppose we pretend that we do not and tackle estimation by simulation.

Note that $h(y|\alpha, \mu)$ in (20.12) can be approximated by

$$\frac{1}{S} \sum_{s=1}^S \frac{e^{-\mu v_s} (\mu v_s)^y}{y!},$$

where v_s ($s = 1, \dots, S$) are pseudo-random draws from the distribution $g(\nu|\alpha)$, and S is the number of simulation replications used. Drawing from a gamma distribution with mean 1 and variance α is straightforward. One draws from a uniform distribution and then applies a transformation to it. Let u_s denote the uniform random variables and let $v_s = -\ln u_s/\alpha$, and then define the simulator

$$\tilde{f}(y|v_s, \alpha, \mu) = \frac{e^{-\mu(-\ln u_s/\alpha)} (\mu(-\ln u_s/\alpha))^y}{y!}.$$

Then the MSL estimator $\hat{\theta}_{MSL}$ maximizes

$$Q_N(\boldsymbol{\theta}) = \sum_{i=1}^N \ln \left(\frac{1}{S} \sum_{s=1}^S \tilde{f}(y_i|x_i, u_i^s, \boldsymbol{\theta}) \right), \quad (20.14)$$

where $\mu_i = \exp(\mathbf{x}'\boldsymbol{\beta})$ and $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta})$.

Of course, this method is computer intensive but otherwise straightforward. A fuller discussion of the properties of MSL was given earlier in Chapter 12.4. Here we just

remind the reader that when $S, N \rightarrow \infty$, $S/\sqrt{N} \rightarrow 0$ then $\widehat{\theta}_{\text{MSL}}$ and $\widehat{\theta}_{\text{ML}}$ are asymptotically equivalent.

20.4.3. Finite Mixture Models

The mixture model in the previous section was a continuous mixture model, because the mixing random variable v was assumed to have continuous distribution. An alternative approach instead uses a discrete representation of unobserved heterogeneity, which generates a class of models called **finite mixture** models; see Section 18.5. This class of models is a particular subclass of **latent class models**. Some variants and special cases of this model are also known as **discrete factor models**.

In empirical work the more commonly used alternative to the continuous mixture is found in the class of modified count models discussed in the next section. However, it is more natural to follow up the preceding section with a discussion of finite mixtures. Further, the subclass of modified count models can be viewed as a special case of finite mixtures.

We suppose that the density of y is a linear combination of m different densities, where the j th density is $f_j(y|\boldsymbol{\theta}_j)$, $j = 1, 2, \dots, m$. Thus an m -component finite mixture is

$$f(y|\boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{j=1}^m \pi_j f_j(y|\boldsymbol{\theta}_j), \quad 0 \leq \pi_j \leq 1, \quad \sum_{j=1}^m \pi_j = 1. \quad (20.15)$$

In the given formulation the components of the mixture are assumed, for generality, to differ in all their parameters. More restrictive formulations assume that only some parameters differ across the components (e.g., the intercepts) and the remaining parameters are all common to the mixture components. Assumptions at some intermediate level of generality may also be made.

For further insight consider this approach for the $m = 2$ case. Suppose that the sampled population contains two “types” of cases, whose y -outcomes are characterized by distributions $f_1(y|\boldsymbol{\theta}_1)$ and $f_2(y|\boldsymbol{\theta}_2)$, which we assume have different moments. Suppose type-1 subpopulation has mean $\mu(\boldsymbol{\theta}_1)$, and type-2 subpopulation has mean $\mu(\boldsymbol{\theta}_2)$, where $\mu(\boldsymbol{\theta}_2) < \mu(\boldsymbol{\theta}_1)$. For example, in a study of the use of medical services, the first subpopulation corresponds to frequent users of the service and the second to relatively infrequent users. Assume that the fractions of the two types in the populations are π_1 and $\pi_2 (= 1 - \pi_1)$, respectively. Then a random sample drawn from the population will contain proportions π_1 and π_2 of the two types, although one cannot observe which case belongs to which subpopulation. That is, the “types” are **latent classes**.

The goal of the researcher who uses this model is to estimate the unknown parameters $\boldsymbol{\theta}_j$, $j = 1, \dots, m$. It is easy to develop regression models based on (20.15). For example, if NB2 models are used then $f_j(y|\boldsymbol{\theta}_j)$ is the NB2 density (20.12) with parameters $\mu_j = \exp(\mathbf{x}'\boldsymbol{\beta}_j)$ and α_j , so $\boldsymbol{\theta}_j = (\boldsymbol{\beta}_j, \alpha_j)$. If the number of components, m , is given, then under some regularity conditions maximum likelihood estimation of the parameters $(\pi_j, \boldsymbol{\theta}_j)$, $j = 1, \dots, m$, is possible.

The pros and cons of the finite mixture representation have also been given earlier and will only be briefly mentioned here. Further discussion in the context of duration models is in Section 18.5. First, a finite mixture is a flexible and parsimonious method of modeling the data. Each mixture component provides a local approximation to some part of the true distribution. Second, the finite mixture approach is in a sense **semiparametric** because it does not require any distributional assumptions for the mixing variable. Finally, in many cases the results are easy to interpret. The finite mixture representation is attractive if the investigator is especially interested in the behavior of a subpopulation from the viewpoint of public policy. If latent classes are ignored, so $m = 1$, then the estimated parameters will be weighted sums of the latent class parameters.

There are several potential difficulties also. First, we may have very little theoretical guidance on specifying the number of components, and we may not be able to reliably distinguish among some of the components if they are not sufficiently different. The usual practice is to start with a few components and then add additional components if the fit of the model is significantly improved by doing so. In some cases only the intercepts may be allowed to differ and the slopes may be constrained to equality across components. Caution is necessary in this process because the sampling properties of the maximum likelihood estimator are not fully understood for the case in which m is unknown.

There are several studies that indicate that finite mixture models work quite well for count data models of medical care (Deb and Trivedi, 1997; 2002). One possible reason for this is that the population might be split by the latent health status of individuals. Those who are healthy, perhaps the majority, might account for low average demand, whereas those who are ill may account for high average demand. When the observed health status is imperfectly observed, the finite mixture model may do a good job of separating subpopulations.

20.4.4. Truncation and Censoring

In some studies, inclusion in the sample requires that sampled individuals have been engaged in the activity of interest. Then the count data are **truncated**, as the data are observed only over part of the range of the response variable. Examples of truncated counts include the number of bus trips made per week in surveys taken on buses, the number of shopping trips made by individuals sampled at a mall, and the number of unemployment spells among a pool of unemployed. In all these cases we do not observe zero counts, so the data are said to be **zero-truncated**, or more generally left-truncated. Right-truncation results from loss of observations greater than some specified value.

A general treatment of truncated and censored models, using ML estimation, is given in Section 16.2. Here we specialize to count data.

Truncation leads to inconsistent parameter estimates unless the likelihood function is suitably modified. Consider the case of zero truncation. Let $f(y|\boldsymbol{\theta})$ denote the density function and $F(y|\boldsymbol{\theta}) = \Pr[Y \leq y]$ denote the cumulative distribution function of the discrete random variable, where $\boldsymbol{\theta}$ is a parameter vector. If realizations of y less

than the positive integer 1 are omitted, the ensuing zero-truncated density is given by

$$f(y|\theta, y \geq 1) = \frac{f(y|\theta)}{1 - F(0|\theta)}, \quad y = 1, 2, \dots. \quad (20.16)$$

This specializes in the **zero-truncated Poisson** case, for example, to $f(y|\mu, y \geq 1) = e^{-\mu} \mu^y / [y!(1 - \exp(-\mu))]$. It is straightforward to construct a log-likelihood based on this density and to obtain maximum likelihood estimates.

Censored counts most commonly arise from aggregation of counts greater than some value. This is often done in survey design when the total probability mass over the aggregated values is relatively small. An important difference between truncation and censoring is that in the case of the latter, covariates corresponding to the censored counts are observed; in the truncation case neither the counted outcomes nor the covariates are observed. Censoring, like truncation, leads to inconsistent parameter estimates if the uncensored likelihood is mistakenly used. See also Section 16.2.

For example, the number of events greater than some known value c might be aggregated into a single category. Then some values of y are incompletely observed; the precise value is unknown but it is known to equal or exceed c . The observed data has density

$$g(y|\theta) = \begin{cases} f(y|\theta) & \text{if } y < c, \\ 1 - F(c-1|\theta) & \text{if } y \geq c, \end{cases} \quad (20.17)$$

where c is known.

A related complication is that of **sample selection** (Terza, 1998). Then the count y is observed only when another random variable, potentially correlated with y , crosses a threshold. For example, to see a medical specialist one may first need to see a general practitioner.

20.4.5. Modified Count Models

The leading motivation for the modified count models of this section is to solve the so-called problem of **excess zeros**, the presence of more zeros in the data than predicted by count models such as the Poisson, or even NB2.

Hurdle or Two-Part Models

The **hurdle model** or **two-part model** (see Section 16.4) relaxes the assumption that the zeros and the positives come from the same data-generating process. The zeros are determined by the density $f_1(\cdot)$, so that $\Pr[y = 0] = f_1(0)$. The positive counts come from the truncated density $f_2(y|y > 0) = f_2(y)/(1 - f_2(0))$, which is multiplied by $\Pr[y > 0] = 1 - f_1(0)$ to ensure that probabilities sum to unity. Thus

$$g(y) = \begin{cases} f_1(0) & \text{if } y = 0, \\ \frac{1 - f_1(0)}{1 - f_2(0)} f_2(y) & \text{if } y \geq 1. \end{cases} \quad (20.18)$$

This reduces to the standard model only if $f_1(\cdot) = f_2(\cdot)$. Thus in the modified model the two processes generating the zeros and the positives are not constrained to be

the same. Although the motivation for this model is to handle excess zeros, it is also capable of modeling too few zeros.

Maximum likelihood estimation of the hurdle model involves separate maximization of the two terms in the likelihood, one corresponding to the zeros and the other to the positives. This is straightforward.

A hurdle model has the interpretation that it reflects a two-stage decision-making process. For example, a patient may initiate the first visit to a doctor, but the second and subsequent visits may be determined by a different mechanism (Pohlmeier and Ulrich, 1995).

Regression applications use hurdle versions of the Poisson or negative binomial, obtained by specifying $f_1(\cdot)$ and $f_2(\cdot)$ to be the Poisson or negative binomial densities given earlier. In application the covariates in the hurdle part that models the zero/one outcome need not be the same as those that appear in the truncated part, although in practice they are often the same. The hurdle model is widely used, and the hurdle negative binomial model is quite flexible. Drawbacks are that the model is not very parsimonious, typically the number of parameters is doubled, and parameter interpretation is not as easy as in the same model without hurdle.

The choice of the distribution in the hurdle specification is important. Using a more flexible distribution gives the negative binomial obvious advantages over the Poisson. The conditional mean in the hurdle model is the product of the probability of positives and the conditional mean of the zero-truncated density. Therefore, using a Poisson regression when the hurdle model is the correct specification implies a misspecification, which will lead to inconsistent estimates. Because of the form of the conditional mean specification, the calculation of marginal effects is more complicated, with similarities to the two-part model used in Section 16.4.

With-Zeros or Zero-Inflated Model

A second modified count model is the **with-zeros model** or zero-inflated model. This supplements a count density $f_2(\cdot)$ with a binary process with density $f_1(\cdot)$. If the binary process takes value 0, with probability $f_1(0)$, then $y = 0$. If the binary process takes value 1, with probability $f_1(1)$, then y takes count values $0, 1, 2, \dots$ from the count density $f_2(\cdot)$. This lets zero counts occur in two ways: as a realization of the binary process and as a realization of the count process when the binary random variable takes value 1. The density is

$$g(y) = \begin{cases} f_1(0) + (1 - f_1(0))f_2(0) & \text{if } y = 0, \\ (1 - f_1(0))f_2(y) & \text{if } y \geq 1. \end{cases} \quad (20.19)$$

Regression models let $f_1(\cdot)$ be a logit model and $f_2(\cdot)$ be a Poisson or negative binomial density. This model is used much less than the hurdle model. It is capable of modeling too few zeros.

The zero-inflated count model is used less frequently in econometrics than in other statistical disciplines.

20.4.6. Discrete Choice Models

Count data can be modeled by discrete choice model methods, possibly after some grouping of counts to limit the number of categories. For example, the categories may be 0, 1, 2, 3, and 4 or more if few observations exceed four. Unordered models such as multinomial logit, discussed in Section 15.4, are not parsimonious and more importantly are inappropriate. Instead, a sequential model that recognizes the ordering of the data should be used.

One such model is an **ordered model**. This defines an unobserved latent variable, $y^* = \mathbf{x}'\beta + u$, with values of $y = 0, 1, 2, \dots$ being observed as y^* crosses progressively higher thresholds, which are also parameters to be estimated. An ordered logit (or probit) model arises when u is logistic (or standard normal) distributed. Ordered models (see Section 15.9) are particularly useful when the count can also take negative values as may occur when modeling a net change, such as the net change in the number of firms in an industry.

Another possible sequential model, although less parsimonious, is obtained by specifying a sequence of binary models for $\Pr[y = 1|y \geq 0]$, $\Pr[y = 2|y \geq 1]$, and so on.

Finally, in some cases durations may be available in addition to counts. For example, if the dates of doctor visits are known, one can model a count, the number of visits in a month, say, or the duration of time between visits. In general, the latter approach is more efficient, since it uses more detailed data, but the count regression can still provide useful information about the role of covariates (Dean and Balshaw, 1997).

20.5. Partially Parametric Models

By partially parametric models we mean that we focus on modeling the data via the conditional mean and variance, and even these may not be fully specified. In Section 20.5.1 we consider models based on specification of the conditional mean and variance. In Section 20.5.2 we consider and critique the use of least-squares methods that do not explicitly model the heteroskedasticity inherent in count data. In Section 20.5.3 we consider models that are even more partially parametric, such as those giving an incomplete specification of the conditional mean.

The approach is similar in flavor to NLS, except that here we allow for heteroskedasticity that is well modeled as a function of the conditional mean.

20.5.1. Quasi-ML Estimation

As discussed in Section 20.2.1, when using PML or QML, the distribution of the estimator is obtained under weaker assumptions about the dgp than those that lead to a specific likelihood function.

Let us reconsider (20.6). Given an assumption for the functional form for ω_i , and a consistent estimate $\hat{\omega}_i$ of ω_i , one can consistently estimate this covariance matrix. We could use the Poisson assumption, $\omega_i = \mu_i$, but as already noted the data are often overdispersed, with $\omega_i > \mu_i$. Common variance functions used are $\omega_i =$

$(1 + \alpha\mu_i)\mu_i$, that of the NB2 model discussed in Section 20.4.2, and $\omega_i = (1 + \alpha)\mu_i$, that of the NB1 model. Note that in the latter case (20.6) simplifies to $V_{PML}[\hat{\beta}_P] = (1 + \alpha) (\sum_i \mu_i \mathbf{x}_i \mathbf{x}'_i)^{-1}$, so with overdispersion ($\alpha > 0$) the usual ML variance matrix given in (20.7) underestimates the true variance.

If $\omega_i = E[(y_i - \mathbf{x}'_i \beta)^2 | \mathbf{x}_i]$ is instead unspecified, a consistent estimate of $V_{PML}[\hat{\beta}_P]$ can be obtained by adapting the Eicker–White robust sandwich variance estimate formula to this case. The middle sum in (20.6) needs to be estimated. If $\hat{\mu}_i \xrightarrow{P} \mu_i$ then $N^{-1} \sum_i (y_i - \hat{\mu}_i)^2 \mathbf{x}_i \mathbf{x}'_i \xrightarrow{P} \lim N^{-1} \sum_i \omega_i \mathbf{x}_i \mathbf{x}'_i$. Thus a consistent estimate of $V_{PML}[\hat{\beta}_P]$ is given by (20.6) with ω_i and μ_i replaced by $(y_i - \hat{\mu}_i)^2$ and $\hat{\mu}_i$.

When doubt exists about the form of the variance function, the use of the PML estimator is recommended. Computationally this is essentially the same as Poisson ML, with the qualification that the variance matrix must be recomputed. The calculation of robust variances is often an option in standard packages.

These results for Poisson PML estimation are qualitatively similar to those for PML estimation in the linear model under normality. They extend more generally to PML estimation based on densities in the linear exponential family. In all cases consistency requires only correct specification of the conditional mean (Nelder and Wedderburn, 1972; Gouriéroux et al., 1984a). This has led to a vast statistical literature on generalized linear models (see McCullagh and Nelder, 1989). These permit valid inference providing the conditional mean is correctly specified and nest many types of data as special cases – continuous (normal), count (Poisson), discrete (binomial), and positive (gamma) as detailed in Section 5.7.4. Many methods for complications, such as time-series and panel data models, are presented within the more general GLM framework rather than specifically for count data.

Some econometricians find it more natural to use the GMM framework rather than GLM. Then the starting point is the conditional moment $E[y_i - \exp(\mathbf{x}'_i \beta) | \mathbf{x}_i] = \mathbf{0}$. If data are independent over i and the conditional variance is a multiple of the mean it can be shown that the optimal choice of instrument is \mathbf{x}_i , leading to the estimating equations (20.5); for more detail, see Cameron and Trivedi (1998, pp. 37–44). The GMM framework has been fruitful for panel data on counts (see Section 20.5.3) and for **endogenous regressors**. Fully specified parametric simultaneous equations models for counts are in their infancy, so instrumental variables methods are appealing. Given instruments \mathbf{z}_i , $\dim(\mathbf{z}) \geq \dim(\mathbf{x})$, satisfying $E[y_i - \exp(\mathbf{x}'_i \beta) | \mathbf{z}_i] = \mathbf{0}$, a consistent estimator of β minimizes

$$Q(\beta) = \left[\sum_{i=1}^N (y_i - \exp(\mathbf{x}'_i \beta)) \mathbf{z}_i \right]' \mathbf{W} \left[\sum_{i=1}^N (y_i - \exp(\mathbf{x}'_i \beta)) \mathbf{z}_i \right], \quad (20.20)$$

where \mathbf{W} is a symmetric weighting matrix.

The pros and cons of this approach are as follows. A major advantage is that the approach makes fewer distributional assumptions and hence avoids a possible model misspecification. However, the discreteness in the outcome variable and its natural heteroskedasticity are ignored, leading to a possible loss of efficiency. A suitable choice of \mathbf{W} matrix may mitigate the problem. Further, by emphasizing the first moment of the distribution, when potentially there may be significant additional information in

the higher moments, the IV estimator may be sensitive to the presence of large counts in the data. Table 20.2 illustrates features of some types of data that are awkward to model using a GMM-type estimator.

20.5.2. Least-Squares Estimation

When attention is focused on modeling just the conditional mean, least-squares methods are inferior to the approach of the previous section.

Linear least-squares regression of y on \mathbf{x} leads to consistent parameter estimates if the conditional mean is linear in \mathbf{x} . However, for count data the specification $E[y|\mathbf{x}] = \mathbf{x}'\beta$ is inadequate as it permits negative values of $E[y|\mathbf{x}]$. For similar reasons the linear probability model is inadequate for binary data.

Transformations of y may be considered. In particular, the logarithmic transformation regresses $\ln y$ on \mathbf{x} . This transformation is problematic if the data contain zeros, as is often the case. One standard solution is to add a constant term, such as 0.5, and to model $\ln(y + .5)$ by OLS. This ad hoc method introduces problems of retransformation if we are interested in $E[y|\mathbf{x}]$ rather than $E[\ln y|\mathbf{x}]$; see Mullahy (1998). However, conversion to a linear model has the advantage of convenience if, for example, there is an endogenous right-hand variable that needs to be “instrumented.”

It is instead better to use nonlinear least squares with the exponential mean specification; that is, estimate the nonlinear regression model $y = \exp(\mathbf{x}'\beta) + u$. It is important that statistical inference for the NLS estimator be based on Eicker–White robust standard errors since the error term in this regression will be heteroskedastic.

For counts the NLS estimator is generally less efficient than the Poisson pseudo-MLE. The NLS first-order condition is $\sum_i (y_i - \exp(\mathbf{x}'_i\beta)) \exp(\mathbf{x}'_i\beta) \mathbf{x}_i = \mathbf{0}$. This weights the residuals differently than does the Poisson pseudo-MLE (see (20.5)). The NLS weights are optimal if $V[y_i|\mathbf{x}_i]$ is constant (homoskedastic) whereas the Poisson psedu-MLE weights are optimal if $V[y_i|\mathbf{x}_i]$ is a multiple of $E[y_i|\mathbf{x}_i]$. The latter is a much better model for handling the inherent heteroskedasticity of count data.

20.5.3. Semiparametric Models

By **semiparametric models** we mean partially parametric models that have an infinite-dimensional component, as developed in Section 9.7. The curse of dimensionality motivates us to put some structure on the conditional mean function.

One class of semiparametric models incompletely specifies the conditional mean. Leading examples are single-index models and partially linear models. Single-index models specify $\mu_i = g(\mathbf{x}'_i\beta)$, where the functional form $g(\cdot)$ is left unspecified. Partially linear models specify $\mu_i = \exp(\mathbf{x}'_i\beta + g(\mathbf{z}_i))$, where the functional form $g(\cdot)$ is left unspecified. In both cases \sqrt{N} -consistent asymptotically normal estimators of β can be obtained, without knowledge of $g(\cdot)$.

A second example is optimal estimation of the regression parameters β , when $\mu_i = \exp(\mathbf{x}'_i\beta)$ is assumed but $V[y_i|\mathbf{x}_i] = \omega_i$ is left unspecified. The infinite-dimensional component arises because as $N \rightarrow \infty$ there are infinitely many variance parameters ω_i . An optimal estimator of β , called an adaptive estimator, is one that is as efficient

as that when ω_i is known. Delgado and Knesner (1997) extend results for the linear regression model to count data with exponential conditional mean function, using kernel regression methods to estimate weights to be used in a second-stage nonlinear least-squares regression. In their application the estimator shows little gain over specifying $\omega_i = \mu_i(1 + \alpha\mu_i)$, overdispersion of the NB2 form.

20.6. Multivariate Counts and Endogenous Regressors

In this section we very briefly present extension from cross-section to other types of count data (see Cameron and Trivedi, 1998, for further detail). For multivariate count data many models have been proposed but preferred methods have not yet been established. For related panel data there is more agreement in the econometrics literature on which methods to use, though a wider range of models is considered in the statistics literature; see Section 23.7.

20.6.1. Multivariate Data

In some data sets more than one count is observed. For example, data on the utilization of several different types of health service, such as doctor visits and hospital days, may be available. Joint modeling will improve efficiency and provide richer models of the data if counts are correlated. This section briefly reviews **bivariate count models** related to the main models of this chapter. The reader familiar with multiequation linear models with correlated errors, e.g. the **SUR model** in Section 6.9.3, may think of a generalization to multiequation count models with correlated errors. Assume that we observe several count variables for the same individual (e.g., number of visits to a doctor and number of prescribed medications taken). The source of correlation may lie in unobserved heterogeneity. Joint estimation that takes account of correlated errors will yield more efficient estimates, but at the cost of additional computational complexity.

Semiparametric Methods

A partially parametric approach views this as a seemingly unrelated regressions problem, adapting methods for the linear regression model to count data where the conditional means are nonlinear and the data are heteroskedastic; see Section 6.10.3.

Gouriéroux, Monfort, and Trognon (1984b) propose a moment-based approach to derive the bivariate Poisson-type model. They specify a model by defining first two moments of y_1 and y_2 and estimate it by a quasi-generalized pseudo-maximum likelihood procedure. This model allows for overdispersion and is more general than the bivariate Poisson model, but it does not maintain the integer-valued property of the counts.

Delgado (1992) treats a multivariate count model as a multivariate nonlinear model and suggests a semiparametric generalized least-squares estimator. The covariance matrix of the residuals is estimated using the k -NN method. The approach differs from

that of Gouriéroux, Monfort, and Trognon (1984) in the choice of the estimator for the covariance matrix.

Most parametric studies have used the **bivariate Poisson**. One way this distribution is derived is to suppose that the two counts y_1 and y_2 are generated as $y_1 = z_1 + w$ and $y_2 = z_2 + w$, where all of z_1 , z_2 , and w are independent and Poisson distributed, with positive parameters λ_1 , λ_2 , and λ_{12} , respectively, which may be parameterized as a function of exogenous covariates. This is called the method of **trivariate reduction**.

The marginal distribution of y_j is Poisson $[\lambda_j + \lambda_{12}]$ and, therefore, this model restricts the conditional mean to be equal to the conditional variance for each count variable, so

$$E[y_j | \mathbf{x}_j] = V[y_j | \mathbf{x}_j] \quad (20.21)$$

for $j = 1, 2$, where \mathbf{x}_j is a vector of explanatory variables. The correlation coefficient is given by

$$\text{Cor}[y_1, y_2] = \frac{\lambda_{12}}{\sqrt{(\lambda_1 + \lambda_{12})(\lambda_2 + \lambda_{12})}}, \quad (20.22)$$

which is positive, because $\lambda_{12} > 0$.

Fully Parametric Methods

Several recent studies develop better parametric models by introducing correlated unobserved heterogeneity for each count. The related issues were discussed in Sections 6.10.1 and 19.3.

Marshall and Olkin (1990) consider a model with **multiplicative unobserved heterogeneity** in the marginal distributions of both counts in the following way. Let y_j be $\mathcal{P}[\lambda_j v]$, $j = 1, 2$, where \mathcal{P} denotes Poisson distribution with mean $\lambda_j v$ and v has gamma distribution with density

$$g(v) = \frac{v^{\alpha-1} \exp(-v)}{\Gamma(\alpha)}.$$

The random variable v can be interpreted as common (shared) unobserved heterogeneity. The resulting model is a **one-factor model**. The **bivariate negative binomial** (BVNB) distribution of two counts is defined as

$$\begin{aligned} f(y_1, y_2 | \mathbf{x}_1, \mathbf{x}_2) &= \int_0^\infty f_1(y_1 | \mathbf{x}_1, v) f_2(y_2 | \mathbf{x}_2, v) g(v) dv \\ &= \int \left[\prod_{j=1}^2 \frac{\exp(-\lambda_j v)(\lambda_j v)^{y_j}}{y_j!} \right] \frac{v^{\alpha-1} \exp(-v)}{\Gamma(\alpha)} dv \\ &= \frac{\Gamma(y_1 + y_2 + \alpha)}{y_1! y_2! \Gamma(\alpha)} \left[\frac{\lambda_1}{\lambda_1 + \lambda_2 + 1} \right]^{y_1} \left[\frac{\lambda_2}{\lambda_1 + \lambda_2 + 1} \right]^{y_2} \\ &\quad \times \left[\frac{1}{\lambda_1 + \lambda_2 + 1} \right]^\alpha. \end{aligned} \quad (20.23)$$

This mixture has a closed-form solution, but the model restricts the unobserved heterogeneity to be the identical component for both count variables. The joint likelihood

is built up with terms like (20.23). The marginal distributions are univariate negative binomial and the correlation between the two count variables,

$$\text{Cor}[y_1, y_2] = \frac{\lambda_1 \lambda_2}{\sqrt{(\lambda_1^2 + \alpha \lambda_1)(\lambda_2^2 + \alpha \lambda_2)}}, \quad (20.24)$$

must be positive.

Other models with more **flexible correlation structures**, but that also require computationally advanced methods, have been proposed by Cameron and Johansson (1998), Munkin and Trivedi (1999), and Chib and Winkelmann (2001).

Munkin and Trivedi (1999) consider a generalization of the BVNB model as follows:

$$f(y_1, y_2 | \mathbf{x}_1, \mathbf{x}_2) = \int_0^\infty \int_0^\infty f_1(y_1 | \mathbf{x}_1, v_1) f_2(y_2 | \mathbf{x}_2, v_2) g(v_1, v_2) dv_1 dv_2, \quad (20.25)$$

where the joint distribution is built up from the two marginal models, each conditioned on a separate unobserved heterogeneity variable, v_1 and v_2 , respectively, that are specified to give a bivariate normal distribution. Conditional on $(\mathbf{x}_1, \mathbf{x}_2, v_1, v_2)$ each marginal distribution is Poisson, with multiplicative unobserved normal heterogeneity. The model is therefore a **bivariate Poisson–log-normal mixture**. The likelihood function is the product over the sample of terms like (20.25). The authors interpret this as a “**two-factor model**.” This specification is more flexible as it does not restrict the sign or size of correlation between the two unobserved components. However this additional flexibility introduces computational complexity because the bivariate integral in (20.25) does not have an analytical solution and hence must be handled using a simulation-based approach (discussed in Chapter 12). 2.4 and in Munkin and Trivedi (1999). If the dimension of the model, the number of y variables, increases, then so does the order of numerical integration involved. This feature combined with a possibly large sample size can make computational burden very significant. Chib and Winkelmann (2001) suggest an alternative Bayesian MCMC approach, which, while retaining the flexibility of the aforementioned specification, can handle a high-dimensional outcome vector. They demonstrate the feasibility of their approach with a six-dimensional mixed Poisson–log-normal model.

Another recently developed approach to modeling correlated counts is the **copula approach** described in Section 19.3. Here one begins with the specification of marginal distributions; the joint distribution is obtained by combining the marginals using a copula. Examples for dependent durations were given in Section 19.3. See also Cameron, Li, Trivedi, and Zimmer (2004).

20.6.2. Count Models with Endogenous Regressors

Simultaneous models for count variables arise in a number of contexts. For example, in Cameron et al. (1988) the focus is on a count variable (medical utilization), but one of the covariates, the health insurance status of the subject, is an endogenous choice. Mullahy (1997) in a cross-section context, and Crépon and Duguet (1997b) in a panel data context, apply the GMM approach to count models with endogenous regressors. A

very well known example from health economics involves models of counts of health services, such as doctor visits, and one of the regressors would be the health insurance status of the individual. The assumption that the choice of health insurance and the error on the outcome equation are uncorrelated is unrealistic, and hence the insurance regressor is likely to be endogenous. Chapter 22 provides more examples and details of panel count models with endogenous regressors.

Currently the econometric literature provides two approaches to the estimation of models with endogenous regressors: one based on the GMM/IV approach and the other based on stronger assumptions of maximum likelihood. We consider each in turn.

The first approach (Mullahy, 1997) begins with a moment condition. Consider the exponential mean model with additive zero-mean error term,

$$y_i = E[y_i | \mathbf{x}_i] + \nu_i = \exp(\mathbf{x}'_i \boldsymbol{\beta}) + \nu_i, \quad (20.26)$$

$$E[\nu_i | \mathbf{x}_i] \neq 0. \quad (20.27)$$

Suppose that we have available instrumental variables \mathbf{z}_i that satisfy the moment conditions

$$E[\nu_i | \mathbf{z}_i] = 0, \quad (20.28)$$

$$E[y_i - \exp(\mathbf{x}'_i \boldsymbol{\beta}) | \mathbf{z}_i] = 0.$$

Then the GMM or nonlinear IV estimation is feasible, assuming that there are enough moment conditions available. This approach has already been discussed in Section 6.5.3. The reader is referred to this section for details and related discussion. However, note that in implementing this approach the count nature of the variable is ignored and the model is treated like any other nonlinear model with an exponential mean. Also, note that heteroskedasticity is highly likely with counted data and hence the GMM/IV procedure should accommodate this complication.

Mullahy has pointed out that a multiplicative error term specification has certain advantages. This, however, leads to a different moment condition. Let

$$E[y_i | \mathbf{x}_i, \nu_i] = \exp(\mathbf{x}'_i \boldsymbol{\beta})\nu_i. \quad (20.29)$$

This leads to the moment condition

$$E\left[\frac{y_i}{\exp(\mathbf{x}'_i \boldsymbol{\beta})} - 1 | \mathbf{z}_i\right] = 0, \quad (20.30)$$

which is a special case of the nonlinear moment condition $E[r(y_i, \mathbf{x}_i, \boldsymbol{\beta}) | \mathbf{z}_i] = 0$ discussed in Section 6.5. Provided suitable and sufficient moment conditions are available, the GMM approach can be followed. Once again, however, for a counted variable, heteroskedasticity is likely and efficiency loss will occur because the count feature of the variable has been ignored.

Alternative approaches that simultaneously handle the count feature of the dependent variable and the problem of endogenous regressors are more **parametric** (Terza, 1998). Deb and Trivedi (2004) develop a joint model of counts (Y) with insurance plan variable (D) as regressors and a binary choice model for the insurance plan.

Endogeneity in their model arises from the presence of correlated unobserved heterogeneity in the outcome (count) equation and the binary choice equation. Their model has the following structure:

$$\Pr[Y_i = y_i | \mathbf{x}_i, D_i, l_i] = f(\mathbf{x}'_i \boldsymbol{\beta} + \gamma_1 D_i + \lambda l_i), \quad (20.31)$$

$$\Pr[D_i = 1 | \mathbf{z}_i, l_i] = g(\mathbf{z}'_i \boldsymbol{\alpha} + \delta l_i), \quad (20.32)$$

where l_i are **latent factors** reflecting unobserved heterogeneity and δ and λ are associated factor loadings. The joint distribution of selection and outcome variables, conditional on the common latent factors, can be written as

$$\Pr[Y_i = y_i, D_i = 1 | \mathbf{x}_i, \mathbf{z}_i, l_i] = f(\mathbf{x}'_i \boldsymbol{\beta} + \gamma_1 D_i + \lambda l_i)g(\mathbf{z}'_i \boldsymbol{\alpha} + \delta l_i), \quad (20.33)$$

because (Y, D) are assumed to be conditionally independent.

The problem in estimation arises because the l_i are unknown. Although the l_i are unknown, assume that h , the distribution of l_i , is known and can therefore be integrated out of the joint density, that is,

$$\Pr[Y_i = y_i, D_i = 1 | \mathbf{x}_i, \mathbf{z}_i] = \int [f(\mathbf{x}'_i \boldsymbol{\beta} + \gamma_1 D_i + \lambda l_i)g(\mathbf{z}'_i \boldsymbol{\alpha} + \delta l_i)] h(l_i) dl_i. \quad (20.34)$$

Cast in this form, the unknown parameters of the model may be estimated by maximum likelihood.

For simplicity we assume $h(l_i)$ has no unknown parameters. Then the maximum likelihood estimator maximizes the joint likelihood function $L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | y_i, D_i, \mathbf{x}_i, \mathbf{z}_i)$, where $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}, \gamma_1, \lambda)$ and $\boldsymbol{\theta}_2 = (\boldsymbol{\alpha}, \delta)$ refer to parameters in the outcome and plan choice equations, respectively, and L refers to the joint likelihood whose i th component is defined in (20.34). For identification additional normalization restrictions may be needed.

The main practical problem of estimation given suitable specifications for f , g , and h is that the integral does not have, in general, a closed-form solution. The MSL estimator involves replacing the expectation by a simulated sample analogue (average), that is,

$$\tilde{\Pr}[Y_i = y_i, D_i = 1 | \mathbf{x}_i, \mathbf{z}_i] = \frac{1}{S} \sum_{s=1}^S [f(\mathbf{x}'_i \boldsymbol{\beta} + \gamma_1 D_i + \lambda \tilde{l}_{is})g(\mathbf{z}'_i \boldsymbol{\alpha} + \delta \tilde{l}_{is})]. \quad (20.35)$$

where \tilde{l}_{is} is the s th draw (from a total of S draws) of a pseudo-random number from the density h and $\tilde{\Pr}$ denotes the simulated probability. A simulated likelihood function for the data can then be defined. The MSL estimator maximizes the simulated log-likelihood.

This approach, developed for an endogenous dummy regressor in a count regression model, can be extended to multiple dummies, and multiple outcomes, whether discrete or continuous. The limitation comes from the burden of estimation, which is very heavy compared with an IV-type estimator. Further, as in any simultaneous equation model, identifiability is an issue. Applied work typically includes some nontrivial explanatory variables in the \mathbf{z} vector that are excluded from the \mathbf{x} vector.

20.7. Count Example: Further Analysis

We now reconsider the earlier analysis based on the Poisson regression by using more flexible parametric models beginning with the NB2 model.

The results for the NB2 model are given in the last columns of Table 20.5, presented in Section 20.3. Here too we report the robust standard errors and t -ratios. First note that the overdispersion coefficient α is highly significant. The Wald test statistic is 8.926, leading to a decisive rejection of the null of equidispersion ($\alpha = 0$). Consistent with this is the large increase in the log-likelihood, from $-60,087$ to $-42,777$. Clearly, the improvement in the fit of the model is considerable. Because the models are nested it is unnecessary to report AIC and BIC.

Row 3 in Table 20.6 shows the predicted frequencies from the NB2 model. These are very close to the observed frequencies and confirm the improvement in the fit of the model as a result of overdispersion being accounted for.

The coefficients themselves, however, seem fairly stable among alternative estimation methods, and all effects are measured with precision, reflecting the impact of the large sample. These features of the results are encouraging, suggesting that the NB2 model is reasonable. As predicted by basic economic theory, utilization and the coinsurance rate (LC) are negatively correlated. The estimated impact does not seem sensitive to the treatment of overdispersion.

Additional modeling refinements are possible. For example, Deb and Trivedi (2002) compare the performance of the two-part (hurdle) model with a two-component finite mixture model and find the latter to fit better. However, even the hurdle model fits better than the NB2 model. Although such refinements provide additional information, none of the results given here can be regarded as misleading on the essential question of price sensitivity of utilization.

The NB2 model works well for doctor visits. For other count outcomes, however, even more flexible models than NB2 may be necessary.

20.8. Practical Considerations

Those with experience of nonlinear least squares will find it easy to use packaged software for Poisson regression, which is a widely available option in popular econometrics and statistics packages. Care is needed to ensure that robust standard errors are obtained. Many econometrics packages also include negative binomial regression and the basic panel data models. Popular statistics packages include count regression in a generalized linear models module. Standard packages also produce some goodness-of-fit statistics, such as the pseudo- R^2 measures, for the Poisson model see Section 8.7.1.

More recently developed models, such as finite mixture models, most time-series models, and dynamic panel data models, require developing one's own programs. A promising route is to use matrix programming languages in conjunction with software for implementing estimation based on user-defined objective functions. For simple models many computer programs make it possible to implement maximum

likelihood estimation and (highly desirable) robust variance estimation for user-defined functions.

In addition to reporting parameter estimates it is useful to have an indication of the magnitude of the estimated effects, as discussed in Section 20.2.3. As noted in Section 20.2.4, care should be taken to ensure that reported standard errors and t -statistics for the Poisson regression model are based on variance estimates robust to overdispersion.

In addition to estimation it is strongly recommended that specification tests be used to assess the adequacy of the estimated model. For Poisson cross-section regression overdispersion tests are easy to implement. For any parametric model one can compare the actual and fitted frequency distribution of counts, although it is not always easy to understand the respect in which a model fails when the distribution of observed counts is highly dispersed. Formal statistical specification and goodness-of-fit tests based on actual and fitted frequencies are available.

In most practical situations one is likely to face the problem of model selection. For likelihood-based models that are nonnested one can use selection criteria, such as the Akaike information criteria, that are based on the fitted log-likelihood but with degrees-of-freedom penalty for models with many parameters.

20.9. Bibliographic Notes

- 20.2** All the topics dealt with in this chapter are treated at greater length and depth by Cameron and Trivedi (1998), who also provide a comprehensive bibliography. Winkelmann (1997) also provides a treatment of the econometric literature on counts. The statistics literature generally analyzes counts in the context of GLM. The standard reference is McCullagh and Nelder (1989). The econometrics literature generally underemphasizes the contributions of the GLM literature. Fahrmeier and Tutz (1994) provide a recent and more econometric exposition of GLMs. The material in Section 20.2 is standard and appears in many places.
- 20.3** Deb and Trivedi (2002) give a detailed analysis of these RHIE data.
- 20.4** Cameron and Trivedi (1986) provide an early presentation and application of the negative binomial. Hausman et al. (1984) applied the model and its variants to panel data. For the finite mixture approach of Section 20.4.3 see Deb and Trivedi (1997). Applications of the hurdle model in Section 20.4.5 include those by Mullahy (1986), who first proposed the model, Pohlmeier and Ulrich (1995), and Gurmu and Trivedi (1996).
- 20.5** The quasi-MLE of Section 20.5.1 is presented in detail by Gouriéroux et al. (1984a,b) and by Cameron and Trivedi (1986).
- 20.6** Regression models for the types of data discussed in Section 20.6 are in their infancy. The notable exception is that (static) panel data count models are well established, with the standard reference being Hausman et al. (1984). See also Brännäs and Johansson (1996). Developing adequate regression models for multivariate count data and models with endogenous regressors is currently an active area; see Terza (1998), and Deb and Trivedi (2004).

Exercises

20–1 Suppose that Y is Poisson distributed with mean μ .

- (a) Verify that the first four moments are, respectively, μ , μ , μ , and $3\mu^2 + \mu$.
- (b) Show that there is a linear relationship between $\Pr[Y = j]$ and $\Pr[Y = j - 1]$, $j = 1, 2, \dots$.
- (c) Consider the Poisson MLE in the regression case with $\mu_i = \exp(\mathbf{x}'\beta)$. Possible estimates of the variance of the Poisson MLE include $V[\hat{\beta}] = [\sum_i \hat{\mu}_i \mathbf{x}_i \mathbf{x}_i']^{-1}$ and $\tilde{V}[\hat{\beta}] = [\sum_i (y_i - \hat{\mu}_i)^2 \mathbf{x}_i \mathbf{x}_i']^{-1}$. Show that they are asymptotically equivalent (upon scaling by N) if the data density is correctly specified.

20–2 Now consider overdispersion in the Poisson model.

- (a) Suppose $Y|\mu \sim P[\mu]$, where $\mu = \exp(\beta_0 + \beta_1 x)$, $\beta_0 = \gamma_0 + \varepsilon$, and ε is an unobserved random variable with $E[\varepsilon] = 0$, $V[\varepsilon] = \sigma^2 > 0$. Show that $V[Y] > E[Y]$.
- (b) Consider the NB2 model with the variance function $\mu + \alpha\mu^2$ and the probability mass function given in (20.12). Using graphs for four different values of $\alpha \in [0, 3]$, describe the behavior of the probability mass for different realized values of Y ; in your answer concentrate on the behavior of the function near the origin and in the right tail.
- (c) For the NB2 density given in (20.12) in Section 20.4.1, show that as $\alpha \rightarrow 0$ the density goes to the Poisson. [This could be tricky.]

20–3 Consider the Poisson regression model with conditional mean $\mu = \exp(\mathbf{x}'\beta)$. Treat the estimation problem as an unweighted nonlinear squares problem in which $y = E[y|\mathbf{x}] + \varepsilon$, where $E[y|\mathbf{x}] = \exp(\mathbf{x}'\beta)$ and $\varepsilon \sim \text{iid}[0, \sigma^2]$.

- (a) Derive the nonlinear least-squares equations for (β, σ^2) . Compare the least-squares and the maximum likelihood equations for β and explain the difference between them.
- (b) Derive the *weighted* nonlinear least-squares equations for β . Explain your choice of weights. [Weights are used to handle heteroskedasticity].
- (c) Compare the weighted nonlinear least-squares and the maximum likelihood equations and explain the similarities, if any.

20–4 Consider a finite mixture density $f(y|\theta) = \sum_{j=1}^C \pi_j f_j(y|\theta_j)$, an additive mixture of C distinct latent classes, or subpopulations, with unknown mixing proportions π_1, \dots, π_C , where $\sum_{j=1}^C \pi_j = 1$, $\pi_j > 0$. Here y is a count variable, and the j th component density $f_j(y_i|\theta_j)$ for the i th observation is expressed as

$$f_j(y_i) = \frac{\Gamma(y_i + \psi_{ji})}{\Gamma(\psi_{ji}) \Gamma(y_i + 1)} \left(\frac{\psi_{ji}}{\lambda_{ji} + \psi_{ji}} \right)^{\psi_{ji}} \left(\frac{\lambda_{ji}}{\lambda_{ji} + \psi_{ji}} \right)^{y_i},$$

where $\lambda_{ji} = \exp(\mathbf{x}'_i \beta_j)$, $\psi_{ji} = \lambda_{ji}^k / \alpha_j$, $\alpha_j > 0$ and $\theta_j = (\beta_j, \alpha_j)$. Here k is either 0 or 1. This model is the finite mixture negative binomial with C components and specializes to the finite mixture Poisson if $\alpha_j = 0$.

- (a) Show that $E[y_i|\mathbf{x}_i] = \bar{\lambda}_i = \sum_{j=1}^C \pi_j \lambda_{ji}$ and $V(y_i|\mathbf{x}_i) = \sum_{j=1}^C \pi_j \lambda_{ji}^2 [1 + \alpha_j \lambda_{ji}^{-k}] + \bar{\lambda}_i - \bar{\lambda}_i^2$.
- (b) Show that any mixture model based on the first moment alone is not identified.

- (c) Show that the C -component Poisson mixture based on the first two moments is identified.
- 20–5** (Adapted from Baltagi and Li, 1999) A simple test of overdispersion in a Poisson model given in Section 20.2.4 tests the null hypothesis of zero coefficient in the regression of $[(y_i - \hat{\mu}_i)^2 - y_i]/\hat{\mu}_i$ on $\hat{\mu}_i$. An alternative test proposed in the literature (Baltagi and Li, 1999) involves the same test but is based on the regression of $(y_i - \hat{\mu}_i)^2$ on $\hat{\mu}_i$. The latter can be motivated by the idea of tests based on the Gauss–Newton regression, (see Section 10.3.9). Analyze the differences between the tests and the implications of the differences for the manner of implementing the second test.
- 20–6** For this problem use a 50% subsample of the data used in this chapter.
- (a) Estimate Poisson and negative binomial regression with MDU as the dependent variable and the following explanatory variable: LC, IDP, LINC, FEMALE, EDUDEC, XAGE, BLACK, HLTHG, HLTHF, and HLTHP. Carry out a likelihood ratio test of the null hypothesis that the variables LC and IDP have no effect on MDU.
 - (b) Test for overdispersion in the Poisson regression using the variance formulations (20.9) with $g(\mu) = \mu$ and (20.10) with $g(\mu) = \mu^2$ in this chapter. Which version of the variance formulation gets more support from the data? What do you conclude from this exercise?
 - (c) Estimate the negative binomial model (NB2). Compare the estimate of the overdispersion parameter with that in part (b). Explain the similarities and differences.
 - (d) Using the results from the negative binomial estimation, compare the estimated marginal effect of a change in LC for an average individual in excellent health (baseline) and an average individual in poor health ($HLTHP = 1$).
 - (e) For this Poisson specification estimate the “hurdle version” consisting of a zero part (logit or probit) and a positive part (truncated-at-zero Poisson). Compare these results with those from a regular Poisson model. Analyze the similarities and differences between the implications of the two models. Based on your analysis, which model do you regard as a better explanation of the data?

PART FIVE

Models for Panel Data

Cross-section models have certain inherent limitations. They are predominantly equilibrium models that generally do not shed light on intertemporal dependence of events. They also cannot satisfactorily resolve fundamental issues about the sources of persistence in behavior. Such persistence may be behavioral, i.e. arising from true state dependence, or it may be spurious, being an artifact of the inability to control for heterogeneous behavior in the population. Because panel data, also called longitudinal data, contain periodically repeated observations of the same subjects, they have a large potential for resolving issues that cross-section models cannot satisfactorily handle. Chapters 21 through 23 present methods for panel data. We progress systematically from linear models for continuous data in Chapter 21 to nonlinear panel data models for limited dependent variables in Chapter 23. Both fixed effects and random effects models are considered. A persistent theme through these three chapters is the importance of using panel-robust methods of inference.

Chapter 21, which reviews the key general results for linear panel data regression models, can be read easily by those with a good grasp of linear regression; it does not require the material covered in Parts 2 to 4. We recommend that even those who are interested in more advanced material should quickly peruse through the contents of this chapter first to gain familiarity with key concepts and definitions.

Chapter 22 covers important extensions of Chapter 21, especially to dynamic panels which allow for Markovian dependence structure of current variables. The analysis is in the GMM framework that is currently favored by many practitioners in this area. The analysis here is at times intricate, involving many issues of detail. A strong grasp of GMM will be helpful in absorbing the main results of this chapter.

The results of Chapters 21 and 22 do not extend to nonlinear panel models of Chapter 23 in a general and unified fashion. There are relatively fewer general results for limited dependent variable panel models. Despite this, in Chapter 23 we begin by

PART FIVE: MODELS FOR PANEL DATA

presenting an analysis of some general issues and approaches. Later sections of this chapter present panel data extensions of the counterpart cross-section models studied in Part 4. These sections analyze four categories of models for binary, count, censored, and duration data, respectively, and should be accessible to a suitably prepared reader familiar with the parallel cross-section models.

Linear Panel Models: Basics

21.1. Introduction

Panel data are repeated observations on the same cross section, typically of individuals or firms in microeconomics applications, observed for several time periods. Other terms used for such data include **longitudinal data** and **repeated measures**. The focus is on data from a **short panel**, meaning a large cross section of individuals observed for a few time periods, rather than a long panel such as a small cross section of countries observed for many time periods.

A major advantage of panel data is increased precision in estimation. This is the result of an increase in the number of observations owing to combining or **pooling** several time periods of data for each individual. However, for valid statistical inference one needs to control for likely correlation of regression model errors over time for a given individual. In particular, the usual formula for OLS standard errors in a pooled OLS regression typically overstates the precision gains, leading to underestimated standard errors and t -statistics that can be greatly inflated.

A second attraction of panel data is the possibility of consistent estimation of the **fixed effects** model, which allows for unobserved individual heterogeneity that may be correlated with regressors. Such unobserved heterogeneity leads to **omitted variables bias** that could in principle be corrected by instrumental variables methods using only a single cross section, but in practice it can be difficult to obtain a valid instrument. Data from a short panel, with as few as two periods, offers an alternative way to proceed if the unobserved individual-specific effects are assumed to be additive and time-invariant.

Most disciplines in applied statistics other than microeometrics treat any unobserved individual heterogeneity as being distributed independently of the regressors. Then the effects are called **random effects**, though a better term is *purely* random effects. Compared to fixed effects models this stronger assumption has the advantage of permitting consistent estimation of all parameters, including coefficients of time-invariant regressors. However, random effects and pooled estimators are inconsistent

if the true model is one with fixed effects. Economists often view the assumptions for the random effects model as being unsupported by the data.

A third attraction of panel data is the possibility of learning more about the **dynamics** of individual behavior than is possible from a single cross section. Thus a cross section may yield a poverty rate of 20% but we need panel data to determine whether the same 20% are in poverty each year. As a related example, panel data may determine whether high serial correlation of individual earnings or unemployment spell length is due to an individual specific tendency to have high earnings or a long unemployment spell, or whether it is a consequence of having past high earnings or unemployment. This topic is deferred to Chapter 22.

The linear panel data models and associated estimators are conceptually simple, aside from the fundamental issue of whether or not fixed effects are necessary. The considerable algebra used to derive the properties of panel data estimators should not distract one from an understanding of the basics: The statistical properties of panel data estimators vary with the assumed model and its treatment of unobserved effects. Furthermore, much of the algebra does not generalize to nonlinear panel models.

The current chapter presents the basic estimators for various linear panel data models. A lengthy introduction in Sections 21.2 and 21.3 provides, respectively, the commonly used models and estimators and an application to the relationship between annual hours worked and wages. The important distinction between fixed and random effects models is studied in Section 21.4. Sections 21.5–21.7 present additional detail on estimation for, respectively, pooled models, individual-specific fixed effects models, and individual-specific random effects models. Section 21.8 considers other basic aspects such as inference and prediction in linear panel data models.

21.2. Overview of Models and Estimators

Panel data provide information on individual behavior both across time and across individuals.

Even for linear regression, standard panel data analysis uses a much wider range of models and estimators than is the case with cross-section data. Several standard models are presented in Section 21.2.1, followed by several estimators presented in Section 21.2.2. Table 21.1 gives a summary that also indicates that several of the estimators are inconsistent if the dgp is the **individual-specific fixed effects model**.

Obtaining correct standard errors of estimators is also more complicated than in the cross-section case. One needs to control for correlation over time in errors for a given individual, in addition to possible heteroskedasticity. This topic is covered in Section 21.2.3.

21.2.1. Panel Data Models

A very general linear model for panel data permits the intercept and slope coefficients to vary over both individual and time, with

$$y_{it} = \alpha_{it} + \mathbf{x}'_{it}\boldsymbol{\beta}_{it} + u_{it}, \quad i = 1, \dots, N, \quad t = 1, \dots, T,$$

Table 21.1. Linear Panel Model: Common Estimators and Models^a

Estimator of β	Assumed Model		
	Pooled (21.1)	Random Effects (21.3) and (21.5)	Fixed Effects (21.3) Only
Pooled OLS (21.1)	Consistent	Consistent	Inconsistent
Between (21.7)	Consistent	Consistent	Inconsistent
Within (or Fixed Effects) (21.8)	Consistent	Consistent	Consistent
First Differences (21.9)	Consistent	Consistent	Consistent
Random Effects (21.10)	Consistent	Consistent	Inconsistent

^a This table considers only consistency of estimators of β . For correct computation of standard errors see Section 21.2.3.

where y_{it} is a scalar dependent variable, \mathbf{x}_{it} is a $K \times 1$ vector of independent variables, u_{it} is a scalar disturbance term, i indexes individual (or firm or country) in a cross section, and t indexes time.

This model is too general and is not estimable as there are more parameters to estimate than observations. Further restrictions need to be placed on the extent to which α_{it} and β_{it} vary with i and t , and on the behavior of the error u_{it} .

Pooled Model

The most restrictive model is a **pooled model** that specifies **constant coefficients**, the usual assumption for cross-section analysis, so that

$$y_{it} = \alpha + \mathbf{x}'_{it}\beta + u_{it}. \quad (21.1)$$

If this model is correctly specified and regressors are uncorrelated with the error then it can be consistently estimated using **pooled OLS**. The error term is likely to be correlated over time for a given individual, however, in which case the usual reported standard errors should not be used as they can be greatly downward biased. Furthermore, the pooled OLS estimator is inconsistent if the fixed effects model, defined in the following, is appropriate.

Individual and Time Dummies

A simple variant of the model (21.1) permits intercepts to vary across individuals and over time while slope parameters do not. Then $y_{it} = \alpha_i + \gamma_t + \mathbf{x}'_{it}\beta + u_{it}$, or

$$y_{it} = \sum_{j=1}^N \alpha_j d_{j,it} + \sum_{s=2}^T \gamma_s d_{s,it} + \mathbf{x}'_{it}\beta, \quad (21.2)$$

where the N **individual dummies** $d_{j,it}$ equal one if $i = j$ and equal zero otherwise, the $(T - 1)$ **time dummies** $d_{s,it}$ equal one if $t = s$ and equal zero otherwise, and it is assumed that \mathbf{x}_{it} does not include an intercept. (If an intercept is included then one of the N individual dummies must be dropped).

This model has $N + (T - 1) + \dim[\mathbf{x}]$ parameters that can be consistently estimated if both $N \rightarrow \infty$ and $T \rightarrow \infty$. We focus on **short panels** where $N \rightarrow \infty$ but T does not. Then the γ_s can be consistently estimated, so the $(T - 1)$ time dummies are simply incorporated into the regressors \mathbf{x}_{it} . The challenge then lies in estimating the parameters β controlling for the N individual intercepts α_i . One possibility is to instead have dummies for groups of observations, such as grouping by region, in which case the clustering methods of Chapter 24 are relevant. Here instead we specify a full set of N individual intercepts, which causes problems as $N \rightarrow \infty$.

Fixed Effects and Random Effects Models

The **individual-specific effects model** allows each cross-sectional unit to have a different intercept term though all slopes are the same, so that

$$y_{it} = \alpha_i + \mathbf{x}'_{it}\beta + \varepsilon_{it}, \quad (21.3)$$

where ε_{it} is iid over i and t . This is a more parsimonious way to express model (21.2), with any time dummies included in the regressors \mathbf{x}_{it} . The α_i are random variables that capture **unobserved heterogeneity**, already studied in Sections 18.2–18.5 and 20.4.

Throughout this chapter we make the assumption of **strong exogeneity** or **strict exogeneity**

$$\text{E}[\varepsilon_{it} | \alpha_i, \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}] = 0, \quad t = 1, \dots, T, \quad (21.4)$$

so that the error term is assumed to have mean zero conditional on past, current, and future values of the regressors. Chamberlain (1980) gives a detailed discussion of exogeneity assumptions and tests for exogeneity for panel data. Strong exogeneity rules out models with lagged dependent variables or with endogenous variables as regressors; these models are deferred to Chapter 22.

One variant of the model (21.3) treats α_i as an unobserved random variable that is potentially correlated with the observed regressors \mathbf{x}_{it} . This variant is called the **fixed effects (FE) model** as early treatments modeled these effects as parameters $\alpha_1, \dots, \alpha_N$ to be estimated. If fixed effects are present and correlated with \mathbf{x}_{it} then many estimators such as pooled OLS are inconsistent. Instead, alternative estimation methods that eliminate the α_i are needed to ensure consistent estimation of β in a short panel.

The other variant of the model (21.3) assumes that the unobservable individual effects α_i are random variables that are distributed independently of the regressors. This model is called the **random effects (RE) model**, which usually makes the additional assumptions that

$$\begin{aligned} \alpha_i &\sim [\alpha, \sigma_\alpha^2], \\ \varepsilon_{it} &\sim [0, \sigma_\varepsilon^2], \end{aligned} \quad (21.5)$$

so that both the random effects and the error term in (21.3) are assumed to be iid. Note that no specific distributions have been specified in (21.5). A more precise term for this model is the **one-way individual-specific random effects model**, or more simply the **random intercept model**, to distinguish the model with more general random effects

models such as the **mixed linear models** presented in Section 22.8. Yet another name is the **random components model**.

The term fixed effect is potentially misleading and the term random effect is more precisely a purely random effect. To avoid such confusion, M-J. Lee (2002) calls a fixed effect a “related effect” and a random effect an “unrelated effect.” We use the traditional notation and terminology, but it should be clear that α_i is a random variable in both fixed and random effects models.

Equicorrelated Model

The RE model can be viewed as a specialization of the pooled model, as the α_i can be subsumed into the error term. Then (21.3) can be viewed as regression of y_{it} on \mathbf{x}_{it} with composite error term $u_{it} = \alpha_i + \varepsilon_{it}$, and (21.5) implies that

$$\text{Cov}[(\alpha_i + \varepsilon_{it}), (\alpha_s + \varepsilon_{is})] = \begin{cases} \sigma_\alpha^2, & t \neq s, \\ \sigma_\alpha^2 + \sigma_\varepsilon^2, & t = s. \end{cases} \quad (21.6)$$

The RE model therefore imposes the constraint that the composite error u_{it} is **equicorrelated**, since $\text{Cor}[u_{it}, u_{is}] = \sigma_\alpha^2 / [\sigma_\alpha^2 + \sigma_\varepsilon^2]$ for $t \neq s$ does not vary with the time difference $t - s$. Clearly, pooled OLS will be consistent but inefficient under the RE model. The random effects model is also called the **equicorrelated model** or **exchangeable errors model**.

Fixed versus Random Effects Models

The fundamental distinction is between models with and without fixed effects. The modern econometrics literature emphasizes fixed effects, but we also provide details for the random effects model.

Some authors, including Chamberlain (1980, 1984) and Wooldridge (2002), use the notation

$$y_{it} = c_i + \mathbf{x}'_{it}\beta + \varepsilon_{it}$$

in (21.3) to make it very clear that the individual effect is a random variable in both fixed and random effects models. Both models assume that

$$E[y_{it}|c_i, \mathbf{x}_{it}] = c_i + \mathbf{x}'_{it}\beta.$$

The individual-specific effect c_i is unknown and in short panels cannot be consistently estimated, so we cannot estimate $E[y_{it}|c_i, \mathbf{x}_{it}]$. Instead, we can eliminate c_i by taking the expectation with respect to c_i , leading to

$$E[y_{it}|\mathbf{x}_{it}] = E[c_i|\mathbf{x}_{it}] + \mathbf{x}'_{it}\beta.$$

For the RE model it is assumed that $E[c_i|\mathbf{x}_{it}] = \alpha$, so $E[y_{it}|\mathbf{x}_{it}] = \alpha + \mathbf{x}'_{it}\beta$ and hence it is possible to identify $E[y_{it}|\mathbf{x}_{it}]$. In the FE model, however, $E[c_i|\mathbf{x}_{it}]$ varies with \mathbf{x}_{it} and it is not known how it varies, so we cannot identify $E[y_{it}|\mathbf{x}_{it}]$. It is nonetheless possible to consistently estimate β in the FE model with short panels (as will be

discussed in the following). Thus it is possible in the FE model to identify the marginal effect

$$\boldsymbol{\beta} = \partial E[y_{it}|c_i, \mathbf{x}_{it}]/\partial \mathbf{x}_{it},$$

even though the conditional mean is not identified. For example, it is possible to identify the effect on earnings of an additional year of schooling, controlling for individual effects, even though the individual effects and the conditional mean are not identified.

In short panels the FE model permits only **identification** of the **marginal effect** $\partial E[y_{it}|c_i, \mathbf{x}_{it}]/\partial \mathbf{x}_{it}$, and even then **only for time-varying regressors**, so the marginal effect of race or gender, for example, is not identified. The RE model permits identification of all components of $\boldsymbol{\beta}$ and of $E[y_{it}|\mathbf{x}_{it}]$, but the key RE assumption that $E[c_i|\mathbf{x}_{it}]$ is constant is viewed as untenable in many microeconomics applications.

21.2.2. Panel Data Estimators

We now introduce several commonly used panel data estimators of $\boldsymbol{\beta}$, with further detail provided in Sections 21.5–21.7. The estimators differ in the extent to which cross-section and time-series variation in the data are used, and their properties vary according to whether or not the fixed effects model is the appropriate model.

A regressor x_{it} may be either **time-invariant**, with $x_{it} = x_i$ for $t = 1, \dots, T$, or **time-varying**. For some estimators, notably the within and first differences estimators defined in the following, only the coefficients of time-varying regressors are identified.

Pooled OLS

The **pooled OLS estimator** is obtained by stacking the data over i and t into one long regression with NT observations, and estimating by OLS

$$y_{it} = \alpha + \mathbf{x}'_{it}\boldsymbol{\beta} + u_{it}, \quad i = 1, \dots, N, \quad t = 1, \dots, T.$$

If $\text{Cov}[u_{it}, \mathbf{x}_{it}] = \mathbf{0}$ then either $N \rightarrow \infty$ or $T \rightarrow \infty$ is sufficient for consistency.

The pooled OLS estimator is clearly consistent if the pooled model (21.1) is appropriate and regressors are uncorrelated with the error term. The usual OLS variance matrix based on iid errors, however, is not appropriate here as the errors for a given individual i are almost certainly positively correlated over t . The NT correlated observations have less information than NT independent observations.

To understand this correlation, note that for a given individual we expect considerable correlation in y over time, so that $\text{Cor}[y_{it}, y_{is}]$ is high. Even after inclusion of regressors $\text{Cor}[u_{it}, u_{is}]$ may remain nonzero, and it often can still be quite high. For example, if a model overpredicts individual earnings in one year it may also overpredict earnings for the same individual in other years. The RE model accommodates this correlation, with $\text{Cor}[u_{it}, u_{is}] = \sigma_\alpha^2/[\sigma_\alpha^2 + \sigma_\varepsilon^2]$ for $t \neq s$ from (21.6).

The usual OLS output treats each of the T years as independent pieces of information, but the information content is less than this given the positive error correlation. This leads to overstatement of estimator precision that can be very large, as illustrated

in Section 21.3.2 and formally demonstrated in Section 21.5.4. One therefore needs to use panel-corrected standard errors (see Section 21.2.3) whenever OLS is applied in a panel setting. Many corrections are possible, depending on the correlation and heteroskedasticity structure assumed for the errors and whether the panel is short or long (see Section 21.5).

The pooled OLS estimator is inconsistent if the true model is the fixed effects model. To see this, rewrite the model (21.3) as

$$y_{it} = \alpha + \mathbf{x}'_{it}\beta + (\alpha_i - \alpha + \varepsilon_{it}).$$

Then pooled OLS regression of y_{it} on \mathbf{x}_{it} and an intercept leads to an inconsistent estimator of β if the individual effect α_i is correlated with the regressors \mathbf{x}_{it} , since such correlation implies that the combined error term $(\alpha_i - \alpha + \varepsilon_{it})$ is correlated with the regressors.

In summary, pooled OLS is appropriate if the constant-coefficients or random effects models are appropriate, but panel-corrected standard errors and t -statistics must be used for statistical inference. Pooled OLS is inconsistent if the fixed effects model is appropriate.

Between Estimator

The pooled OLS estimator uses variation over both time and cross-sectional units to estimate β .

The between estimator in short panels instead uses just the cross-sectional variation. Begin with the individual-specific effects model (21.3). Averaging over all years yields $\bar{y}_i = \alpha_i + \bar{\mathbf{x}}'_i\beta + \bar{\varepsilon}_i$, which can be rewritten as the **between model**

$$\bar{y}_i = \alpha + \bar{\mathbf{x}}'_i\beta + (\alpha_i - \alpha + \bar{\varepsilon}_i), \quad i = 1, \dots, N, \quad (21.7)$$

where $\bar{y}_i = T^{-1} \sum_t y_{it}$, $\bar{\varepsilon}_i = T^{-1} \sum_t \varepsilon_{it}$, and $\bar{\mathbf{x}}_i = T^{-1} \sum_t \mathbf{x}_{it}$.

The **between estimator** is the OLS estimator from regression of \bar{y}_i on an intercept and $\bar{\mathbf{x}}_i$. It uses variation between different individuals and is the analogue of cross-section regression, which is the special case $T = 1$.

The between estimator is consistent if the regressors $\bar{\mathbf{x}}_i$ are independent of the composite error $(\alpha_i - \alpha + \bar{\varepsilon}_i)$ in (21.7). This will be the case for the constant-coefficients model and the random effects model. In contrast, for the fixed effects model the between estimator is inconsistent as α_i is then assumed to be correlated with \mathbf{x}_{it} and hence $\bar{\mathbf{x}}_i$.

Within Estimator or Fixed Effects Estimator

The within estimator is an estimator that, unlike the pooled OLS or between estimators, exploits the special features of panel data. In a short panel it measures the association between individual-specific deviations of regressors from their time-averaged values and individual-specific deviations of the dependent variable from its time-averaged value. This is done using the variation in the data over time.

Specifically, begin with the individual-specific effects model (21.3), which nests (21.1) as the special case $\alpha_i = \alpha$. Then taking the average over time yields $\bar{y}_i = \alpha_i + \bar{\mathbf{x}}_i'\beta + \bar{\varepsilon}_i$. Subtracting this from y_{it} in (21.3) yields the **within model**

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)' \beta + (\varepsilon_{it} - \bar{\varepsilon}_i), \quad i = 1, \dots, N, \quad t = 1, \dots, T, \quad (21.8)$$

as the α_i terms cancel.

The **within estimator** is the OLS estimator in (21.8). A special feature of this estimator is that it yields consistent estimates of β in the fixed effects model, whereas the pooled OLS and between estimators do not.

From Section 21.6 the within estimator has several interpretations. It is called the **fixed effects estimator** as it is the efficient estimator of β in the model (21.3) if α_i are fixed effects and the error ε_{it} is iid. This chapter focuses on a literature that treats fixed effects as **nuisance parameters** that can be ignored since interest lies solely in estimation of β . If instead the fixed effects are of interest they can also be estimated. In short panels these estimates of the individual α_i are inconsistent, though their distribution or their variation with a key variable may be informative. If N is not too large an alternative and simpler way to compute the within estimator is by **least-squares dummy variable estimation**. This directly estimates (21.2) by OLS regression of y_{it} on \mathbf{x}_{it} and the N individual dummy variables and yields the within estimator for β , along with estimates of the N fixed effects (see Section 21.6.4). Yet another interpretation of the within estimator is the covariance estimator. Finally, taking deviations from individual-specific averages is equivalent to taking residuals from auxiliary regression of y_{it} and \mathbf{x}_{it} on individual dummies and then working with the residuals.

A major limitation of within estimation is that the coefficients of time-invariant regressors are not identified in the within model, since if $x_{it} = x_i$ then $\bar{x}_i = x_i$ so $(x_{it} - \bar{x}_i) = 0$. Many studies seek to estimate the effect of time-invariant regressors. For example, in panel wage regressions we may be interested in the effect of gender or race. For this reason many practitioners prefer not to use the within estimator. Pooled OLS or random effects estimators permit estimation of coefficients of time-invariant regressors, but these estimators are inconsistent if the fixed effects model is the correct model.

First-Differences Estimator

The first-differences estimator also exploits the special features of panel data. In a short panel it measures the association between individual-specific one-period changes in regressors and individual-specific one-period changes in the dependent variable.

Specifically, begin with the individual-specific effects model (21.3). Then lagging one period yields $y_{i,t-1} = \alpha_i + \mathbf{x}'_{i,t-1}\beta + \varepsilon_{i,t-1}$. Subtracting this from y_{it} in (21.3) yields the **first-differences model**

$$y_{it} - y_{i,t-1} = (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})' \beta + (\varepsilon_{it} - \varepsilon_{i,t-1}), \quad i = 1, \dots, N, \quad t = 2, \dots, T, \quad (21.9)$$

as the α_i terms cancel.

The **first-differences estimator** is the OLS estimator in (21.9). Like the within estimator, this estimator yields consistent estimates of β in the fixed effects model, though

the coefficients of time-invariant regressors are not identified. The first-differences estimator is less efficient than the within estimator for $T > 2$ if ε_{it} is iid.

Random Effects Estimator

The random effects estimator is an estimator that also exploits the special features of panel data.

Begin with the individual-specific effects model (21.3), but assume a random effects model where α_i and ε_{it} are iid as in (21.5). Pooled OLS is consistent but pooled GLS will be more efficient. The **feasible GLS estimator** (see Section 4.5.1) of the RE model, called the **random effects estimator**, can be calculated from OLS estimation of the transformed model

$$y_{it} - \widehat{\lambda} \bar{y}_i = (1 - \widehat{\lambda})\mu + (\mathbf{x}_{it} - \widehat{\lambda} \bar{\mathbf{x}}_i)' \boldsymbol{\beta} + v_{it}, \quad (21.10)$$

where $v_{it} = (1 - \widehat{\lambda})\alpha_i + (\varepsilon_{it} - \widehat{\lambda} \bar{\varepsilon}_i)$ is asymptotically iid, and $\widehat{\lambda}$ is consistent for

$$\lambda = 1 - \frac{\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2 + T\sigma_\alpha^2}}. \quad (21.11)$$

Section 21.7 provides a derivation of (21.10) and ways to estimate σ_α^2 and σ_ε^2 and hence to estimate λ . Note that $\widehat{\lambda} = 0$ corresponds to pooled OLS, $\widehat{\lambda} = 1$ corresponds to within estimation, and $\widehat{\lambda} \rightarrow 1$ as $T \rightarrow \infty$. This is a two-step estimator of $\boldsymbol{\beta}$.

The RE estimator is fully efficient under the RE model, though the efficiency gain compared to pooled OLS need not be great. It is inconsistent, however, if the fixed effects model is the correct model.

21.2.3. Panel-Robust Statistical Inference

The various panel models include error terms denoted u_{it} , ε_{it} , and α_i . In many microeconometrics applications it is reasonable to assume independence over i . However, the errors are potentially (1) **serially correlated** (i.e., correlated over t for given i) and/or (2) **heteroskedastic**. Valid statistical inference requires controlling for both of these factors.

The White heteroskedastic consistent estimator of Section 4.4.5 is easily extended to short panels since for the i th observation the error variance matrix is of finite dimension $T \times T$ while $N \rightarrow \infty$. Thus panel-robust standard errors can be obtained without assuming specific functional forms for either within-individual error correlation or heteroskedasticity. More efficient estimators using GMM are deferred to Section 22.2.3.

It is crucial to note that frequently the panel commands in many computer packages calculate default standard errors assuming iid model errors, leading to erroneous inference. In particular, for pooled OLS regression of y_{it} on \mathbf{x}_{it} without any control for individual effects it is very likely that $\text{Cov}[u_{it}, u_{is}] > 0$ for $t \neq s$. Ignoring this serial correlation can lead to greatly underestimated standard errors and over-estimated t -statistics, as demonstrated in the Section 21.3 data example and shown algebraically in Section 21.5.4. Once fixed or random individual-specific effects are included the serial correlation in errors can be greatly reduced, but it may not be completely eliminated.

Additionally, one may need to control for potential heteroskedasticity as is routinely done for cross-section data.

Panel-Robust Sandwich Standard Errors

The panel estimators of Section 21.2.2 can be obtained by OLS estimation of θ in the pooled regression

$$\tilde{y}_{it} = \tilde{\mathbf{w}}'_{it} \theta + \tilde{u}_{it}, \quad (21.12)$$

where different panel estimators correspond to different transformations \tilde{y}_{it} , $\tilde{\mathbf{w}}_{it}$, and \tilde{u}_{it} of y_{it} , $\mathbf{w}'_{it} = [1 \mathbf{x}'_{it}]$, and u_{it} . The key is that \tilde{y}_{it} is a known function of only y_{i1}, \dots, y_{iT} , and similarly for $\tilde{\mathbf{w}}_{it}$ and \tilde{u}_{it} .

In the simplest case of pooled OLS, no transformation is necessary and $\theta = [\alpha \beta']'$. For the within estimator $\tilde{y}_{it} = y_{it} - \bar{y}_i$, $\tilde{\mathbf{w}}_{it} = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)$, where only time-varying regressors appear, and θ equals the coefficients of the time-varying regressors. For first-differences estimation $\tilde{y}_{it} = y_{it} - y_{i,t-1}$, $\tilde{\mathbf{w}}_{it} = (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})$ and again only coefficients of time-varying regressors are identified. For random effects $\tilde{y}_{it} = y_{it} - \hat{\lambda}\bar{y}_i$ and $\tilde{\mathbf{w}}'_{it} = (\mathbf{w}_{it} - \hat{\lambda}\bar{\mathbf{w}}_i)$ and $\theta = [\alpha \beta']'$. Such transformations can induce serial correlation even if underlying errors are uncorrelated.

It is convenient to stack observations over time periods for a given individual, leading to

$$\tilde{\mathbf{y}}_i = \tilde{\mathbf{W}}_i \theta + \tilde{\mathbf{u}}_i,$$

where $\tilde{\mathbf{y}}_i$ is a $T \times 1$ vector in the preceding examples, except for the first-differences model where it is $(T - 1) \times 1$, and $\tilde{\mathbf{W}}_i$ is a $T \times q$ matrix or, for the first-differences model, a $(T - 1) \times q$ matrix. Further stacking over the N individuals yields

$$\tilde{\mathbf{y}} = \tilde{\mathbf{W}}\theta + \tilde{\mathbf{u}}.$$

Three representations of the OLS estimator are therefore

$$\begin{aligned} \hat{\theta}_{OLS} &= [\tilde{\mathbf{W}}'\tilde{\mathbf{W}}]^{-1}\tilde{\mathbf{W}}'\tilde{\mathbf{y}} \\ &= \left[\sum_{i=1}^N \tilde{\mathbf{W}}_i'\tilde{\mathbf{W}}_i \right]^{-1} \sum_i \tilde{\mathbf{W}}_i'\tilde{\mathbf{y}}_i \\ &= \left[\sum_{i=1}^N \sum_{t=1}^T \tilde{\mathbf{w}}_{it}\tilde{\mathbf{w}}'_{it} \right]^{-1} \sum_{i=1}^N \sum_{t=1}^T \tilde{\mathbf{w}}_{it}\tilde{\mathbf{y}}_{it}, \end{aligned}$$

where in the third expression the sum is from $t = 2$ to T in the case of the first-differences estimator. The most convenient representation to use varies with the context.

To consider consistency, note that if the model is correctly specified then the usual algebra yields $\hat{\theta}_{OLS} = \theta + [\tilde{\mathbf{W}}'\tilde{\mathbf{W}}]^{-1}\tilde{\mathbf{W}}'\tilde{\mathbf{u}}$ or

$$\hat{\theta}_{OLS} = \theta + \left[\sum_{i=1}^N \tilde{\mathbf{W}}_i'\tilde{\mathbf{W}}_i \right]^{-1} \sum_{i=1}^N \tilde{\mathbf{W}}_i'\tilde{\mathbf{u}}_i.$$

Given independence over i the essential condition for consistency is $E[\tilde{\mathbf{W}}_i' \tilde{\mathbf{u}}_i] = \mathbf{0}$. This generally requires a stronger assumption than $E[u_{it} | \mathbf{w}_{it}] = 0$. A sufficient assumption is that of strong exogeneity given in (21.4). See Chapter 22 for estimation under assumptions weaker than strong exogeneity that permit, for example, lagged dependent variables as regressors.

The asymptotic variance of $\hat{\theta}_{OLS}$ is then

$$V[\hat{\theta}_{OLS}] = \left[\sum_{i=1}^N \tilde{\mathbf{W}}_i' \tilde{\mathbf{W}}_i \right]^{-1} \sum_{i=1}^N \tilde{\mathbf{W}}_i' E[\tilde{\mathbf{u}}_i \tilde{\mathbf{u}}_i' | \tilde{\mathbf{W}}_i] \tilde{\mathbf{W}}_i \left[\sum_{i=1}^N \tilde{\mathbf{W}}_i' \tilde{\mathbf{W}}_i \right]^{-1},$$

given independence of errors over i . Consistent estimation of $V[\hat{\theta}_{OLS}]$ in this panel setting is analogous to the cross-section problem of obtaining a consistent estimate of $V[\hat{\theta}_{OLS}]$ that is robust to heteroskedasticity of unknown form. The only complication is the appearance of a vector \mathbf{u}_i rather than a scalar u_i , which poses no problem if the panel is short as then the dimension of \mathbf{u}_i is finite.

This leads to a **panel-robust estimate** of the asymptotic variance matrix of the pooled OLS estimator, one that controls for both serial correlation and heteroskedasticity, given by

$$\hat{V}[\hat{\theta}_{OLS}] = \left[\sum_{i=1}^N \tilde{\mathbf{W}}_i' \tilde{\mathbf{W}}_i \right]^{-1} \sum_{i=1}^N \tilde{\mathbf{W}}_i' \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i' \tilde{\mathbf{W}}_i \left[\sum_{i=1}^N \tilde{\mathbf{W}}_i' \tilde{\mathbf{W}}_i \right]^{-1}, \quad (21.13)$$

where $\hat{\mathbf{u}}_i = \tilde{\mathbf{u}}_i = \tilde{\mathbf{y}}_i - \tilde{\mathbf{W}}_i \hat{\theta}$. The estimator in (21.13) assumes independence over i and $N \rightarrow \infty$, the case for short panels, but otherwise permits $V[u_{it}]$ and $Cov[u_{it}, u_{is}]$ to vary with i , t , and s . An equivalent expression is

$$\hat{V}[\hat{\theta}_{OLS}] = \left[\sum_{i=1}^N \sum_{t=1}^T \tilde{\mathbf{w}}_{it} \tilde{\mathbf{w}}_{it}' \right]^{-1} \sum_{i=1}^N \sum_{t=1}^T \sum_{s=1}^T \tilde{\mathbf{w}}_{it} \tilde{\mathbf{w}}_{is}' \hat{u}_{it} \hat{u}_{is} \left[\sum_{i=1}^N \sum_{t=1}^T \tilde{\mathbf{w}}_{it} \tilde{\mathbf{w}}_{it}' \right]^{-1},$$

where $\hat{u}_{it} = \tilde{y}_{it} - \tilde{\mathbf{w}}_{it}' \hat{\theta}$. This estimator was proposed by Arellano (1987) for the fixed effects estimator.

Panel-robust standard errors based on (21.13) can be computed by use of a regular OLS command, if the command has a **cluster-robust** standard error option (see Section 24.5.2). Since the clustering here is on the individual one selects the identifier for individual i as the **cluster variable**. This method was used to obtain the panel-robust standard errors given in Table 24.1.

The term “robust” standard error can cause confusion. A common error made in pooled regression is to estimate the OLS regression (21.12) using the standard robust standard error option (see Section 4.4.5). However, this only adjusts for heteroskedasticity, and in practice in a panel setting it is much more important to correct for the correlation in individual errors. Another common error, though one that has smaller impact, is to use cluster-robust standard errors that assume homoskedasticity so that $E[\mathbf{u}_i \mathbf{u}_i']$ is constant over i .

Panel Bootstrap Standard Errors

The **bootstrap method** provides an alternative way to obtain panel-robust standard errors. The key assumption is that observations are independent over i , so one does a bootstrap pairs procedure that resamples **with replacement over i** and uses all observed time periods for a given individual. For data $\{(\mathbf{y}_i, \mathbf{X}_i), i = 1, \dots, N\}$ this yields B pseudo-samples and for each **pseudo-sample** one performs OLS regression of $\tilde{\mathbf{y}}_{it}$ on $\tilde{\mathbf{w}}_{it}$, yielding B estimates $\hat{\theta}_b, b = 1, \dots, B$.

The **panel bootstrap estimate** of the variance matrix is then

$$\widehat{\mathbf{V}}_{\text{Boot}}[\hat{\theta}] = \frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}_b - \bar{\hat{\theta}}) (\hat{\theta}_b - \bar{\hat{\theta}})', \quad (21.14)$$

where $\bar{\hat{\theta}} = B^{-1} \sum_b \hat{\theta}_b$. This bootstrap provides **no asymptotic refinement** (see Section 11.2.2). Given independence over i the estimate is consistent as $N \rightarrow \infty$. It is asymptotically equivalent to the estimate (21.13), just as in the cross-section case bootstrap pairs are asymptotically equivalent to White's heteroskedastic consistent estimate. This bootstrap does not offer an asymptotic refinement though bootstrap with asymptotic refinement is possible (see Section 11.6.2).

This bootstrap method can be applied to any panel estimator that relies on independence over i and $N \rightarrow \infty$, including the pooled feasible GLS estimators of Section 21.5.2 for short panels. The key is to resample over i only, and not over both i and t .

Discussion

The importance of correcting standard errors for serial correlation in errors at the individual level cannot be overemphasized. Computer packages currently do not automatically do this. Bertrand, Duflo, and Mullainathan (2004) illustrate the resulting downward bias in standard error computation, in the context of difference-in-differences estimation (see Section 22.6). They find that the panel-robust and panel bootstrap methods work well, even though in their application with state-year data N (the number of states) is relatively small whereas the asymptotic theory uses $N \rightarrow \infty$.

The following example (see Table 21.2) also shows the importance of correcting standard errors for any error serial correlation and autocorrelation.

21.3. Linear Panel Example: Hours and Wages

An important issue in labor economics is the responsiveness of labor supply to wages. The standard textbook model of labor supply suggests that for people already working the effect of a wage increase on labor supply is ambiguous, with an income effect pushing in the direction of less work offsetting a substitution effect in the direction of more work.

Cross-section analysis for adult males finds a relatively small positive response to hours worked. However, it is possible that this association is spurious, merely reflecting a greater unobserved desire to work being positively associated with higher wages.

Panel data analysis can control for this, under the assumption that the unobserved desire to work is time-invariant. For example, the within estimator does so by measuring the extent to which an individual works above-average (or below-average) hours in periods with above-average (or below-average) wages.

The data on 532 males for each of the 10 years from 1979 to 1988 come from Ziliak (1997). The variable of interest is $\ln\text{hrs}$, the natural logarithm of annual hours worked. The single explanatory variable is $\ln\text{wg}$, the natural logarithm of hourly wage. We consider the regression model

$$\ln\text{hrs}_{it} = \alpha_i + \beta \ln\text{wg}_{it} + \varepsilon_{it},$$

where the individual-specific effect α_i is simplified to α in some models and β measures the wage elasticity of labor supply. The error term ε_{it} is assumed to be independent over i , but it may be correlated over t for given i . As noted we expect β , the labor supply elasticity, to be small and positive.

Ziliak (1997) additionally included a quadratic in age, number of children, and an indicator variable for bad health. These regressors and year dummies make relatively small difference to the estimate of β and its standard error, and for simplicity they are omitted here. In Chapter 22 we consider more general models that permit $\ln\text{wg}$ to be endogenous and permit lags of $\ln\text{hrs}$ to appear as a regressor.

21.3.1. Data Summary

For the 5,320 observations, the sample means of $\ln\text{hrs}$ and $\ln\text{wg}$ are respectively 7.66 and 2.61, implying geometric means of 2,120 hours and \$13.60 per hour. The sample standard deviations are respectively 0.29 and 0.43, indicating considerably greater variability in percentage terms in wages rather than hours.

For panel data it is useful to know whether variability is mostly across individuals or across time. The total variation of a series x_{it} around its grand mean \bar{x} can be decomposed as

$$\begin{aligned} \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x})^2 &= \sum_{i=1}^N \sum_{t=1}^T [(x_{it} - \bar{x}_i) + (\bar{x}_i - \bar{x})]^2 \\ &= \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i)^2 + \sum_{i=1}^N \sum_{t=1}^T (\bar{x}_i - \bar{x})^2, \end{aligned}$$

as the cross-product term sums to zero. In words, the total sum of squares equals the **within sum of squares** plus the **between sum of squares**. This leads to **within standard deviation** s_W and **between standard deviation** s_B , where

$$s_W^2 = \frac{1}{NT - N} \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i)^2$$

and

$$s_B^2 = \frac{1}{N-1} \sum_{i=1}^N (\bar{x}_i - \bar{x})^2.$$

Table 21.2. Hours and Wages: Standard Linear Panel Model Estimators^a

	POLS	Between	Within	First Diff	RE-GLS	RE-MLE
α	7.442	7.483	7.220	.001	7.346	7.346
β	.083	.067	.168	.109	.119	.120
Robust se ^b	(.030)	(.024)	(.085)	(.084)	(.051)	(.052)
Boot se	[.030]	[.019]	[.084]	[.083]	[.056]	[.058]
Default se	{.009}	{.020}	{.019}	{.021}	{.014}	{.014}
R^2	.015	.021	.016	.008	.014	.014
RMSE	.283	.177	.233	.296	.233	.233
RSS	427.225	0.363	259.398	417.944	288.860	288.612
TSS	433.831	17.015	263.677	420.223	293.023	292.773
σ_α	.000		.181		.161	.162
σ_ε	.283		.232		.233	.233
λ	0.000	—	1.000	—	.585	.586
N	5320	532	5320	4788	5320	5320

^a Shown are pooled OLS (POLS), between, within, first-differences, random effects (RE) GLS and MLE linear panel regression of lnhrs on lnwg. Standard errors for the slope coefficients are panel robust in parentheses, panel bootstrap in square brackets, and default estimates that assume iid errors in curly braces. The R^2 , root mean square error (RMSE), residual sum of squares (RSS), total sum of squares (TSS), and sample size come from the appropriate regression given in Section 21.2. The parameter λ is defined after (21.11).

^b se, standard error.

The within and between sample standard deviations are, respectively, 0.22 and 0.18 for lnhrs and 0.19 and 0.39 for lnwg. The larger total variation in wages compared to hours is therefore due to between individual variation being much higher for wages. Within individuals the variation is actually somewhat smaller for wages than it is for hours.

21.3.2. Comparison of Panel Data Estimators

Table 21.2 summarizes results from application of the standard panel estimators defined in Section 21.2.2 to these data, along with three different estimates of the standard errors. As detailed in the following, statistical inference should use either the panel-robust standard error or the panel bootstrap standard error.

Slope Parameter Estimates

The estimate of the slope parameter β differs across the different estimation methods. The between estimate that uses only cross-section variation is less than the pooled OLS estimate. The within or fixed effects estimate of 0.168 is much higher than the pooled OLS estimate of 0.083 and is borderline statistically significant using a two-tailed test at 5% and standard error estimate of 0.084 or 0.085. The first-differences estimate of 0.109 is also higher than that of pooled OLS but is considerably less than the within estimate, which also uses only time-series variation. The RE estimates of 0.119 or 0.120 lie between the between and within estimates. This is expected, as RE estimates

can be shown to be a **weighted average of between and within estimates**. The two RE estimates are very close to each other as here the estimates of the variances σ_α^2 and σ_ε^2 are similar, leading to similar values $\hat{\lambda} = 0.585$ and $\hat{\lambda} = 0.586$ in the regression (21.10). The RE estimates are surprisingly less efficient than the pooled OLS estimates, a sign that the RE model fails to model the error correlation well.

Which estimates are preferred? The within and first-difference estimators are consistent under all models (pooled, RE, and FE) whereas the other estimators are inconsistent under the fixed effects model. The most robust estimates are therefore the within or first-differences estimates of 0.168 or 0.109.

There is, however, an efficiency loss in using these more robust estimators, with standard errors of 0.83 to 0.85 that are much larger than those from pooled OLS and RE estimates. A formal Hausman test (see Section 21.4.3 for details and discussion) can be used to test whether or not the individual effects are fixed. Given the relative imprecision of estimation in this example, the Hausman test does not reject the null hypothesis of random effects, despite the large difference between FE and RE estimates. So the more efficient random effects estimates could be used here. Another advantage of random effects estimation is that it permits estimation of the coefficients of time-invariant estimators.

Standard Error Estimation

We now turn to comparison of the standard error estimates. From Section 21.2.3, inference should be based on panel-robust standard errors that permit errors to be correlated over time for a given individual and to have variances and covariances that differ across individuals. Also, as detailed in later sections, the standard errors for estimators based on deviations from means, such as (21.8) and (21.10), need to account for loss of $N + K$ rather than K degrees of freedom.

The first standard error estimate is computed by the panel-robust method given in (21.13), and the second is computed by the panel bootstrap given in (21.14) with 500 replications. For brevity these estimates are called panel robust, though they are additionally robust to heteroskedasticity. The two estimates are very close, aside from the random effects models where the panel-robust standard errors are underestimated because they are computed for the regression (21.10), which ignores estimation error in $\hat{\lambda}$.

The third standard error estimate is the standard default computer output that is based on the assumption of iid errors. In this example the correctly estimated standard errors are a remarkable three to four times as large as the default standard errors. The one exception is the between estimator, an estimator with standard errors that need only correction for heteroskedasticity since it uses only cross-section variation.

For example, for the pooled OLS estimator of β the default standard error is 0.09, leading to incorrect t -statistic of 9.07. The panel-robust standard error is a much larger 0.30, leading to correct t -statistic of a much smaller 2.83. Default standard errors assume independence of model errors over t for given i when in practice they are likely to be positively correlated. This erroneous assumption overestimates the benefit of additional time periods, leading to downward bias in standard errors (see Section 21.5.4). Additionally, ignoring heteroskedasticity in errors also leads to bias,

though this bias could be in either direction. For these data a failure to control for heteroskedasticity also imparts a large downward bias: The standard error of $\hat{\beta}_{\text{POLS}}$ controlling for heteroskedasticity, but not for correlation over t for given i , is 0.020. For other data, correction for heteroskedasticity is usually much less important than the correction for panel correlation.

For the within and between estimators inclusion of the term α_i should control for some of the correlation in the error across time for a given individual. For these data, however, the differences between panel-robust and nonrobust standard errors remain large, in part because of failure to additionally control for heteroskedasticity.

Clearly panel-robust standard errors should be used.

21.3.3. Graphical Analysis

It is insightful to perform a graphical comparison of overall, between, and fixed effects (within or first-differences) regressions. Such plots are rarely performed in panel data regression, but they are easily applied here as there is only one regressor.

All plots include a nonparametric regression curve using the Lowess smoother (see Section 9.6.2) and a linear regression curve that corresponds to the estimates given in Table 21.2.

Figure 21.1 plots $\ln \text{hrs}$ against $\ln \text{wg}$ for all firms in all years (5,320 observations). The plot suggests a positive relationship, roughly linear except at the extreme ends, and from Table 21.2 the line has slope 0.083 with a low $R^2 = 0.015$.

The between estimator (21.7) regresses \bar{y}_i on \bar{x}_i . The corresponding plot for the $\ln \text{hrs}$ - $\ln \text{wg}$ data is given in Figure 21.2 and again shows a positive relationship.

The within or fixed effects estimator (21.8) regresses $(y_{it} - \bar{y}_i)$ on $(x_{it} - \bar{x}_i)$. Figure 21.3 gives the related plot of $(y_{it} - \bar{y}_i + \bar{y})$ on $(x_{it} - \bar{x}_i + \bar{x})$, where $\bar{y} = N^{-1} \sum_i \bar{y}_i$ and $\bar{x} = N^{-1} \sum_i \bar{x}_i$ are the grand means of y and x . Comparison with Figure 21.1 shows that differencing the individual mean leads to a considerable decrease in the range of variability in $\ln \text{wg}$, with less of a decrease in the variability of $\ln \text{hrs}$.

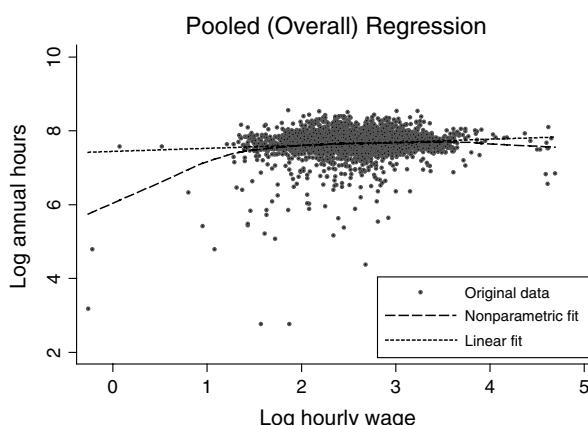


Figure 21.1: Hours and wages: pooled (overall) regression. Natural logarithm of annual hours worked plotted against natural logarithm of hourly wage. Data for 532 U.S. males for each of the ten years 1979–88.

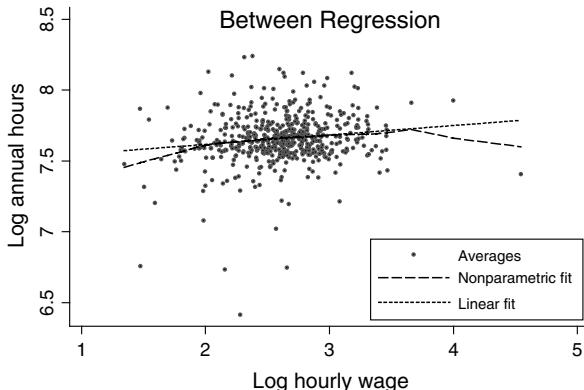


Figure 21.2: Hours and wages: between regression. Ten-year average of log hours plotted against ten-year average of log wage for 532 men. Same sample as Figure 21.1.

The slope does appear steeper than that for pooled OLS, and from Table 21.2 the slope increased from 0.083 to 0.168.

The first-differences estimator (21.9) regresses $(y_{it} - \bar{y}_{i,t-1})$ on $(x_{it} - \bar{x}_{i,t-1})$. The corresponding plot for the lnhrs – lnwg data is given in Figure 21.4. The figure is qualitatively similar to Figure 21.3.

The conclusion of the preceding analysis is that there is greater response to wage changes using time-series variation than using cross-section variation.

21.3.4. Residual Analysis

It is instructive to consider the autocorrelation patterns of the data and of residuals. For example, for residuals $\hat{u}_{it} = y_{it} - \hat{y}_{it}$ the autocorrelation between period s and period t is calculated as $\hat{\rho}_{st} = c_{st}/\sqrt{c_{ss}c_{tt}}$, $s, t = 1, \dots, T$, where the covariance estimate $c_{st} = (N - 1)^{-1} \sum_i (\hat{u}_{it} - \bar{\hat{u}}_t)(\hat{u}_{is} - \bar{\hat{u}}_s)$ and $\bar{\hat{u}}_t = N^{-1} \sum_i \hat{u}_{it}$.

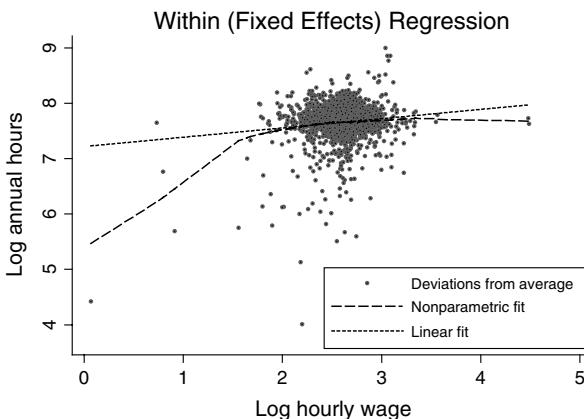


Figure 21.3: Hours and wages: within (fixed effects) regression. Deviation of log hours from ten-year average plotted against deviation of log wage from ten-year average using ten years of data for 532 men. Same sample as Figure 21.1.



Figure 21.4: Hours and wages: first differences regression. First difference of log hours plotted against first difference of log wage using ten years of data for 532 men. Same sample as Figure 21.1.

Table 21.3 gives the residual autocorrelations after pooled OLS regression of $\ln\text{hrs}$ on $\ln\text{wg}$. The autocorrelations generally lie between 0.2 and 0.4 for data two to nine periods apart. The decay rate is very slow, and the autocorrelations appear closer to a random effects model that assumes that $\text{Cor}[u_{it}, u_{is}]$ is constant for $t \neq s$ than to an AR(1) model that has exponential decay.

The autocorrelations for $\ln\text{hrs}$ before regression are very similar to those given in Table 21.3, since $\hat{u}_{it} \simeq y_{it}$ as evident from the poor explanatory power of pooled OLS with $R^2 = 0.015$. The autocorrelations for the regressor $\ln\text{wg}$, not tabulated here, are much higher, ranging from approximately 0.9 at one lag, to 0.7 at nine lags.

The correlations of the residuals from the within regression are given in Table 21.4. If the original errors ε_{it} in (21.3) are iid then it can be shown that the transformed errors $\varepsilon_{it} - \bar{\varepsilon}_i$ have autocorrelations at all lags equal to $-1/(T-1) = -0.11$. There is some departure from this here, particularly for the first lag, which is always positive.

Table 21.3. Hours and Wages: Autocorrelations of Pooled OLS Residuals^a

	u79	u80	u81	u82	u83	u84	u85	u86	u87	u88
upols79	1.00									
upols80	.33	1.00								
upols81	.44	.40	1.00							
upols82	.30	.31	.57	1.00						
upols83	.21	.23	.37	.47	1.00					
upols84	.20	.23	.32	.34	.64	1.00				
upols85	.24	.32	.41	.35	.39	.58	1.00			
upols86	.20	.19	.28	.25	.31	.35	.40	1.00		
upols87	.20	.32	.33	.29	.31	.34	.39	.35	1.00	
upols88	.16	.25	.30	.26	.21	.25	.34	.55	.53	1.00

^a Note: Autocorrelations of residuals are from pooled OLS regression of $\ln\text{hrs}$ on $\ln\text{wg}$ for 532 men in 10 years. The autocorrelations die slowly.

Table 21.4. Hours and Wages: Autocorrelations of Within Regression Residuals^a

	u79	u80	u81	u82	u83	u84	u85	u86	u87	u88
ufe79	1.00									
ufe80	.10	1.00								
ufe81	.21	.08	1.00							
ufe82	.00	-.04	.26	1.00						
ufe83	-.26	-.27	-.21	.01	1.00					
ufe84	-.26	-.27	-.30	-.20	.32	1.00				
ufe85	-.18	-.10	-.11	-.17	-.16	.17	1.00			
ufe86	-.19	-.25	-.26	-.27	-.17	-.14	-.08	1.00		
ufe87	-.15	-.05	-.16	-.20	-.24	-.21	-.09	-.09	1.00	
ufe88	-.17	-.11	-.14	-.18	-.38	-.31	.13	.24	.24	1.00

^a Autocorrelations of residuals are from within (fixed effects) regression of lnhrs on lnwg for 532 men in 10 years.

The correlations of the residuals from random effects regression are quite similar to those for fixed effects given in Table 21.4. The correlations of residuals from first-differences regression are qualitatively similar to the theoretical result that if the original errors ε_{it} in (21.3) are iid then the transformed errors $\varepsilon_{it} - \varepsilon_{it-1}$ have autocorrelations of 0.5 at lag one and 0 at other lags.

21.4. Fixed Effects versus Random Effects Models

The fixed effects model has the attraction of allowing one to use panel data to establish causation under weaker assumptions (presented in Section 21.4.1) than those needed to establish causation with cross-section data or with panel data models without fixed effects, such as pooled models and random effects models.

In some studies causation is clear, so random effects may be appropriate. For example, in a controlled experiment such as crop yield from different amounts of fertilizers applied to different fields the causation is clear. In other cases it may be sufficient to use a random effects analysis to measure the extent of correlation, with determination of causation left to further research taking other approaches. The effect of smoking on lung cancer is an example. Economists are unusual in instead preferring a fixed effects approach, however, because of a desire to measure **causation** in spite of reliance on observational data.

The fixed effects model has several practical weaknesses. Estimation of the coefficient of any time-invariant regressor, such as an indicator variable for gender, is not possible as it is absorbed into the individual-specific effect. Coefficients of time-varying regressors are estimable, but these estimates may be very imprecise if most of the variation in a regressor is cross sectional rather than over time. Prediction of the conditional mean is not possible. Instead, only changes in the conditional mean caused by changes in time-varying regressors can be predicted. Even coefficients of time-varying regressors may be difficult or theoretically impossible to identify in nonlinear

models with fixed effects (see Chapter 23). For these reasons economists also use random effects models, even if causal interpretation may then be unwarranted.

21.4.1. Fixed Effects Example

Consider the effect of computer use on wage. Several cross-sectional studies, most notably those by Krueger (1993) and DiNardo and Pischke (1997), find that computer use in a job is associated with substantially higher wages, even after controlling for many determinants of the wage such as education, age, gender, industry, and occupation. As emphasized by DiNardo and Pischke (1997) this does not necessarily imply causation, if regressors are correlated with the error term owing to endogeneity or omitted variables.

Specifically, we suppose that in the cross section

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \alpha_i + \varepsilon_i,$$

where y is the natural logarithm of wage, \mathbf{x} is a vector of individual characteristics including an indicator variable for computer use at work, and ε is an error that is assumed to be independent of \mathbf{x} . The complication is the addition of the unobserved variable α , which is assumed to be correlated with computer use at work, and hence with the observed regressors \mathbf{x} , even though the components of \mathbf{x} other than computer use, such as occupation and education, may partly control for computer use at work. Regression of y on \mathbf{x} leads to **omitted variables bias** leading to inconsistent estimates of $\boldsymbol{\beta}$ as the combined error ($\alpha + \varepsilon$) is correlated with \mathbf{x} .

Panel data offer a way around this problem, if we assume that the unobserved variable α_i is time-invariant. Then

$$y_{it} = \mathbf{x}'_{it} \boldsymbol{\beta} + \alpha_i + \varepsilon_{it},$$

where again ε is uncorrelated with \mathbf{x} and α is correlated with \mathbf{x} . Differencing eliminates α_i (see Section 21.2.2), permitting consistent estimation of $\boldsymbol{\beta}$. For the computer use example, the causative effect of computer use on wages is then measured by the association between individual changes in wages and individual movements to or from a job with a computer. Haisken-DeNew and Schmidt (1999) found no effect using German panel data.

This fixed effects panel approach permits determination of causation under weaker assumptions than those of cross-section analysis, but it still requires assumptions. The key assumption is that the unobservables α_i are time-invariant, rather than being of more general form α_{it} . In the computer use example it is being assumed that an individual's propensity to have a job with a computer may be endogenous, but the unobserved component of the effect of this propensity α_i on wage is constant over time once we control for observables \mathbf{x}_{it} . Essentially the particular time periods in which an individual's job does or does not involve use of a computer are assumed to be purely random, once we control for time-invariant unobservable α_i and observable \mathbf{x}_{it} .

A random effects or pooled panel approach does not have similar properties. It instead assumes away the original concern that α is correlated with \mathbf{x} , since it assumes that α is iid $[0, \sigma^2]$ and hence is uncorrelated with \mathbf{x} . This leads to inconsistent

parameter estimates if in fact α is correlated with \mathbf{x} , whereas the fixed effects estimator is consistent if α is correlated with \mathbf{x} , provided α is time-invariant.

21.4.2. Conditional versus Marginal Analysis

Fixed effects estimation is a **conditional analysis**, measuring the effect of \mathbf{x}_{it} on y_{it} controlling for the individual effect α_i . Prediction is possible only for individuals in the particular sample being used, and even then it is only possible if the panel is long enough so that α_i can be consistently estimated. Random effects estimation is instead an example of **marginal analysis** or **population-averaged analysis**, as the individual effects are integrated out as iid random variables. The random effects estimators can be applied outside the sample.

If the true model is a random effects model, then whether to perform a conditional or marginal analysis will vary with the application. If analysis is for a random sample of countries then one uses random effects, but if one is intrinsically interested in the particular countries in the sample then one does fixed effects estimation even though this can entail a loss of efficiency.

If the true model instead has individual-specific effects correlated with regressors, however, then a random effects analysis is no longer meaningful as the random effects estimator is inconsistent. Instead, alternative estimators such as the fixed effects and first-differences estimators are necessary. Because of the desire to determine causation microeconomic applications emphasize these latter estimators.

21.4.3. Hausman Test

If individual effects are fixed the within estimator $\hat{\beta}_W$ is consistent whereas the random effects estimator $\tilde{\beta}_{RE}$ is inconsistent. Here β refers to the vector of coefficients of just the time-varying regressors. One can therefore test whether fixed effects are present by using a Hausman test of whether there is a statistically significant difference between these estimators. Alternatively, any other pair of estimators with similar properties, such as first differences versus pooled OLS, can be used.

A large value of the Hausman test statistic leads to rejection of the null hypothesis that the individual-specific effects are uncorrelated with regressors and to the conclusion that fixed effects are present. It may still be possible to avoid using a fixed effects model. If regressors are correlated with individual-specific effects caused by omitted variables, then one can add further regressors, either time varying or time-invariant, and again perform a Hausman test in this larger model to see whether fixed effects are still necessary. Even if such correlation persists it may be possible to estimate a random effects model using instrumental variables methods (see Sections 22.4.3–22.4.4).

Computation When RE Is Fully Efficient

We begin by assuming that the true model is the random effects model (21.3) with α_i iid $[0, \sigma_\alpha^2]$ uncorrelated with regressors and error ε_{it} iid $[0, \sigma_\varepsilon^2]$.

Then the estimator $\tilde{\beta}_{\text{RE}}$ is fully efficient, so from Section 8.3 the **Hausman test** statistic simplifies to

$$H = (\tilde{\beta}_{1,\text{RE}} - \hat{\beta}_{1,W})' [\hat{V}[\tilde{\beta}_{1,W}] - \hat{V}[\tilde{\beta}_{1,\text{RE}}]]^{-1} (\tilde{\beta}_{1,\text{RE}} - \hat{\beta}_{1,W}),$$

where β_1 denotes the subcomponent of β corresponding to time-varying regressors since only that component can be estimated by the within estimator. This test statistic is asymptotically $\chi^2(\dim[\beta_1])$ distributed under the null hypothesis.

Hausman (1978) showed that an asymptotically equivalent version of this test is to perform a Wald test of $\gamma = \mathbf{0}$ in the auxiliary OLS regression,

$$y_{it} - \hat{\lambda}\bar{y}_i = (1 - \hat{\lambda})\mu + (\mathbf{x}_{1it} - \hat{\mathbf{x}}_{1i})'\beta_1 + (\mathbf{x}_{1it} - \bar{\mathbf{x}}_{1i})'\gamma + v_{it}, \quad (21.15)$$

where \mathbf{x}_{1it} denotes the time-varying regressors and $\hat{\lambda}$ is defined in (21.11) and only the time-varying regressors are used. This algebraic result can be interpreted as follows. The individual-specific effects model (21.10) implies that $v_{it} = (1 - \hat{\lambda})\alpha_i + (\varepsilon_{it} - \hat{\lambda}\bar{\varepsilon}_i)$. The random effects estimator is actually obtained by OLS estimation of (21.15) with $\gamma = \mathbf{0}$ (see (21.10)). If instead the fixed effects specification is valid then the error v_{it} will be correlated with the regressors, via correlation of α_i with regressors. This correlation leads to additional functions of the regressors, such as $(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)$, being statistically significant variables in (21.15).

Computation When RE Is Not Fully Efficient

The simple form of the Hausman test is invalid if α_i or ε_{it} are not iid, which is more than likely given heteroskedasticity inherent in much microeconomics data. Then the RE estimator is not fully efficient under the null hypothesis so the expression $\hat{V}[\tilde{\beta}_W] - \hat{V}[\tilde{\beta}_{\text{RE}}]$ in the formula for H needs to be replaced by the more general $\hat{V}[\tilde{\beta}_{\text{RE}} - \tilde{\beta}_W]$ (see Section 8.3).

For short panels this variance matrix can be consistently estimated by bootstrap resampling over i (see Section 21.2.3). Then a panel-robust Hausman test statistic is

$$H_{\text{Robust}} = (\tilde{\beta}_{1,\text{RE}} - \hat{\beta}_{1,W})' [\hat{V}_{\text{Boot}}[\tilde{\beta}_{1,\text{RE}} - \hat{\beta}_{1,W}]]^{-1} (\tilde{\beta}_{1,\text{RE}} - \hat{\beta}_{1,W}), \quad (21.16)$$

where

$$\hat{V}_{\text{Boot}}[\tilde{\beta}_{1,\text{RE}} - \hat{\beta}_{1,W}] = \frac{1}{B-1} \sum_{b=1}^B (\hat{\delta}_b - \bar{\delta})(\hat{\delta}_b - \bar{\delta})',$$

b denotes the b th of B bootstrap replications (see Section 21.2.3), and $\hat{\delta} = \tilde{\beta}_{1,\text{RE}} - \hat{\beta}_{1,W}$. This test statistic can be applied to subcomponents of β_1 and can use alternative estimators such as $\tilde{\beta}_{1,\text{POLS}}$ in place of $\tilde{\beta}_{1,\text{RE}}$ and $\hat{\beta}_{1,\text{FD}}$ in place of $\hat{\beta}_{1,W}$.

Alternatively, Wooldridge (2002) suggests estimating the auxiliary OLS regression (21.15) and testing $\gamma = \mathbf{0}$ using panel-robust standard errors. If the effects are random, though not necessarily such that α_i and ε_{it} are iid, then $v_{it} = (1 - \hat{\lambda})\alpha_i + (\varepsilon_{it} - \hat{\lambda}\bar{\varepsilon}_i)$ is still uncorrelated with regressors though v_{it} is no longer asymptotically iid, so **cluster-robust standard errors** need to be used. If the effects are fixed then the error v_{it} is correlated with the regressors, leading to significance of additional functions of the

regressors such as $(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)$. This robust version of the auxiliary regression for the Hausman test is preferred to one that assumes v_{it} is asymptotically iid, on the usual grounds of minimizing distributional assumptions. However, it is not clear whether this test actually coincides with the Hausman test when RE is inefficient.

Hausman Test Example

For the Inhrs–Inwg example estimates given in Table 21.2, a comparison of FE and RE estimates using the default standard errors yields $H \simeq (0.168 - 0.119)^2 / (0.019^2 - 0.014^2)$. This leads to $H = 14 > \chi^2_{.05}(1) = 3.84$, so the random effects model is rejected.

This test is not appropriate, however. The statistic H is inflated because the usual standard errors in this example are greatly downward biased (see Section 21.3.2). Furthermore, this bias is a signal that the RE estimator is not fully efficient under H_0 , so that the more general form of the Hausman test needs to be used.

The auxiliary regression (21.15) yields a panel-robust t -statistic for $\hat{\gamma}$ of 1.28 and hence $H^* = 1.28^2 = 1.65$, leading to nonrejection of the random effects model at 5%. Even though the wage elasticity estimates differ by 0.049, the estimates are sufficiently imprecise that the difference is not statistically significant. Note that if the nonrobust t -statistic for $\hat{\gamma}$ is used instead, then $t^2 = 13.69$, close to the previous incorrect Hausman test statistic.

21.4.4. Richer Models for Random Effects

The random effects model specifies that the random effect α_i is distributed independently of regressors. Richer models, closer in spirit to fixed effects models, relax this assumption.

Mundlak (1978) allowed individual effects in the panel model (21.3) to be determined by **time averages** of the regressors, so that $\alpha_i = \bar{\mathbf{x}}'_i \boldsymbol{\pi} + w_i$, where w_i is iid. Then efficient GLS estimation of β and $\boldsymbol{\pi}$ in this expanded model leads to an estimator of β that equals the fixed effects estimator in model (21.3). By contrast the usual random effects estimator of β in model (21.3) that erroneously specifies iid random effects will be inconsistent.

Chamberlain (1982, 1984) considered an even richer model for the random effects, with $\alpha_i = \mathbf{x}'_{1i} \boldsymbol{\pi}_1 + \cdots + \mathbf{x}'_{Ti} \boldsymbol{\pi}_T + w_i$, a **weighted sum** of the regressors. He proposed estimation by minimum distance methods (see Section 22.2.7 for details), leading to an estimator of β that equals the fixed effects estimator.

More generally, mixed linear models and hierarchical linear models of Section 24.6 permit quite general models for random intercepts and also random slope parameters. Bayesian analysis of panel data also uses this framework. See Section 22.8 for details.

In linear models the fixed effects approach is used if the unobserved individual effect is correlated with regressors. In more complicated models, such as nonlinear models, fixed effects models are not always estimable and richer random effects models provide an alternative approach.

21.5. Pooled Models

The **pooled cross-section time-series model** or **constant-coefficients model** is

$$y_{it} = \alpha + \mathbf{x}_{it}'\boldsymbol{\beta} + u_{it}. \quad (21.17)$$

In the statistics literature the model is called a **population-averaged model**, as there is no explicit model of y_{it} conditional on individual effects. Instead, any individual effects have implicitly been averaged out. The random effects model is a special case where the error u_{it} is equicorrelated over t for given i (see Section 21.2.1).

The main complication for statistical inference, assuming no fixed effects, is that the distribution of least-squares estimators of this model varies with the assumed distribution of u_{it} . In short panels, panel-robust standard errors can be obtained using (21.13).

Here we instead focus on GLS estimation using many of the different specifications, including equicorrelation, for the covariance structure of u_{it} over time and individuals that have been proposed in the literature.

Although we focus on pooled GLS estimation of (21.17), a model without individual-specific fixed effects, the methods of this section can be applied more generally to pooled GLS estimation of the transformed model (21.12) of Section 21.2.3.

21.5.1. Pooled OLS, FGLS, and WLS Estimators

It is convenient to use matrix notation. Combining observations over time for a given individual, define

$$\mathbf{y}_i = \mathbf{W}_i \boldsymbol{\delta} + \mathbf{u}_i, \quad (21.18)$$

where $\boldsymbol{\delta} = [\alpha \ \boldsymbol{\beta}']'$ is a $(K + 1) \times 1$ parameter vector, \mathbf{y}_i and \mathbf{u}_i are $T \times 1$ vectors with t th entries y_{it} and u_{it} , respectively, and \mathbf{W}_i is a $T \times (K + 1)$ matrix with t th row $\mathbf{w}'_{it} = [1 \ \mathbf{x}_{it}]'$. Stacking all individuals yields

$$\mathbf{y} = \mathbf{W}\boldsymbol{\delta} + \mathbf{u}, \quad (21.19)$$

where \mathbf{y} and \mathbf{u} are $NT \times 1$ vectors, for example $\mathbf{y} = [\mathbf{y}'_1 \dots \mathbf{y}'_N]'$, and \mathbf{W} is an $NT \times (K + 1)$ regressor matrix whose first column is a vector of ones. We assume that $E[\mathbf{u}|\mathbf{W}] = \mathbf{0}$, so errors are strictly exogenous, and define $\Omega = E[\mathbf{u}\mathbf{u}'|\mathbf{W}]$.

There are several possible least-squares estimators of this model, summarized in Table 21.5.

First, **pooled OLS** is consistent and asymptotically normal. However, in a panel setting it is unlikely that $\Omega = \sigma^2 \mathbf{I}_{NT}$, so OLS is inefficient except in some special cases such as when all regressors are time-invariant. More importantly, the usual OLS variance estimate of $\sigma^2(\mathbf{W}'\mathbf{W})^{-1}$ should not be used and a panel-robust estimate such as that in (21.13) needs to be used.

Second, **pooled feasible GLS** (PFGLS) is consistent and fully efficient if Ω is correctly specified and $\hat{\Omega}$ is consistent for Ω . Some of the very large range of structures on u_{it} and hence Ω that have been proposed in the panel literature and incorporated

Table 21.5. Pooled Least-Squares Estimators and Their Asymptotic Variances

Estimator	Formula ^a	Variance Matrix ^b
Pooled OLS: $\hat{\delta}_{\text{POLS}}$	$(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{y}$	$(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\widehat{\Omega}\mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}$
Pooled FGLS: $\hat{\delta}_{\text{PFGLS}}$	$(\mathbf{W}'\widehat{\Omega}^{-1}\mathbf{W})^{-1}\mathbf{W}'\widehat{\Omega}^{-1}\mathbf{y}$	$(\mathbf{W}'\widehat{\Omega}^{-1}\mathbf{W})^{-1}$
Pooled WLS: $\hat{\delta}_{\text{PWLS}}$	$(\mathbf{W}'\widehat{\Sigma}^{-1}\mathbf{W})^{-1}\mathbf{W}'\widehat{\Sigma}^{-1}\mathbf{y}$	$(\mathbf{W}'\widehat{\Sigma}^{-1}\mathbf{W})^{-1}\mathbf{W}'\widehat{\Sigma}^{-1}\widehat{\Omega}\widehat{\Sigma}^{-1}\mathbf{W}$ $\times(\mathbf{W}'\widehat{\Sigma}^{-1}\mathbf{W})^{-1}$

^a The formulas are for the model $\mathbf{y} = \mathbf{W}\delta + \mathbf{u}$ defined in (21.19) and error matrix Ω .

^b For computation of $\widehat{\Omega}$ for the variance matrices of POLS and PWLS see the text; in those cases $\widehat{\Omega}$ need not be consistent for Ω . For pooled FGLS it is assumed that $\widehat{\Omega}$ is consistent for Ω .

into regression packages are given in Sections 21.5.2 and 21.5.3 for, respectively, short and long panels.

Third, the **pooled weighted LS** (PWLS) estimator guards against misspecification of Ω . It posits a **working matrix** Σ for the error variance matrix Ω but then performs inference that is valid even if $\Sigma \neq \Omega$. Ordinary least squares is an example, with $\Sigma = \sigma^2 \mathbf{I}_{NT}$, but other choices of Σ may improve efficiency.

Estimation of the variance matrix of the pooled OLS estimator requires an $\widehat{\Omega}$ such that $(NT)^{-1}\mathbf{W}'\widehat{\Omega}\mathbf{W}$ consistently estimates $(NT)^{-1}\mathbf{W}'\Omega\mathbf{W}$.

For short panels this is possible by direct application of the results of Section 21.2.3. Estimation of the variance matrix of the pooled WLS estimator requires an $\widehat{\Omega}$ such that $(NT)^{-1}\mathbf{W}'\widehat{\Sigma}^{-1}\widehat{\Omega}\widehat{\Sigma}^{-1}\mathbf{W}$ consistently estimates $(NT)^{-1}\mathbf{W}'\Sigma^{-1}\Omega\Sigma^{-1}\mathbf{W}$. The panel-robust estimate for OLS given in (21.13) can be adapted to pooled WLS by replacing $\mathbf{W}'\Sigma^{-1}\Omega\Sigma^{-1}\mathbf{W}$, or equivalently $\sum_i \mathbf{W}_i'\Sigma_i^{-1}\mathbf{E}[\mathbf{u}_i\mathbf{u}_i'|\mathbf{W}_i]\Sigma_i^{-1}\mathbf{W}_i$ given independence over i , by the quantity $\sum_i \mathbf{W}_i'\widehat{\Sigma}_i^{-1}\widehat{\mathbf{u}}_i\widehat{\mathbf{u}}_i'\widehat{\Sigma}_i^{-1}\mathbf{W}_i$, where $\widehat{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{W}_i\widehat{\delta}$. Alternatively, a panel bootstrap can be used.

21.5.2. Error Variance Matrix for Short Panels

In short panels there are few time periods but many individuals, usually people or firms. It is assumed that errors are independent over individuals, so that $\text{Cov}[u_{it}, u_{js}] = 0$, $i \neq j$. In such cases it is convenient to revert to summation notation. For example, the PFGLS estimator given in Table 21.5 becomes

$$\widehat{\beta}_{\text{PFGLS}} = \left[\sum_{i=1}^N \mathbf{W}_i'\widehat{\Omega}_i^{-1}\mathbf{W}_i \right]^{-1} \sum_{i=1}^N \mathbf{W}_i'\widehat{\Omega}_i^{-1}\mathbf{y}_i, \quad (21.20)$$

where $\widehat{\Omega}_i$ is consistent for

$$\Omega_i = \mathbf{E}[\mathbf{u}_i\mathbf{u}_i'|\mathbf{W}_i], \quad (21.21)$$

and Ω_i is nondiagonal as errors for a given individual are likely to be correlated over time. Note that $\widehat{\Omega}_i$ needs to come from estimation of a specified model for Ω_i , and we cannot use $\widehat{\Omega}_i = \widehat{\mathbf{u}}_i\widehat{\mathbf{u}}_i'$ (see the related discussion after equation (5.88)).

Equicorrelated Errors

The most commonly used error structure is the random effects model presented in Section 21.2.1. Then from (21.6) Ω_i has common diagonal entries $\sigma_\alpha^2 + \sigma_\varepsilon^2$ and common off-diagonal entries σ_α^2 . Equivalently, the errors are **equicorrelated**, with Ω_i having common diagonal entries σ^2 and common off-diagonal entries $\rho\sigma^2$. Implementation of FGLS requires only estimation of σ_α^2 and σ_ε^2 , or of σ^2 and ρ (see Sections 21.2.2 and 21.7).

ARMA Errors

An alternative error structure is to assume an ARMA error model. For example, an AR(1) error model specifies that $u_{it} = \rho u_{i,t-1} + \varepsilon_{it}$, where ε_{it} are iid. Then $\text{Cov}[u_{it}, u_{is}] = \rho^{|t-s|}\sigma^2$. In this case the covariance between errors falls as the number of time periods between the errors increases. The RE model and an AR(1) error model are compared in Section 21.5.4.

Baltagi and Li (1991) combine the two error models to consider a random effects model with AR(1) errors. This can be easily generalized to the AR(p) case, and methods for moving average and **ARMA errors** (see Section 5.8.7) in random effects models have also been developed more recently. A summary is given in Baltagi (2001, Chapter 5).

Homoskedastic Errors with Unstructured Autocorrelation

For FGLS estimation in short panels there is actually no need to impose as much structure as that imposed by an RE model or an AR(1) error model, if the assumption is made that the $T \times T$ matrix Ω_i is constant over i . Then there are “only” $T(T+1)/2$ covariance parameters to estimate. A consistent estimate of Ω_i is then $\widehat{\Omega}_i$ with (t, s) th entry $\widehat{\sigma}_{ts} = N^{-1} \sum_{i=1}^N \widehat{u}_{it} \widehat{u}_{is}$. The preceding models also assume homoskedasticity, but place additional structure on Ω_i .

Robust Inference

All of the preceding specifications assume that error covariances are the same across individuals, which rules out heteroskedasticity. Provided the panel is short one can nonetheless use the preceding restrictive error variance matrix models as the basis for pooled WLS estimation, but then obtain robust standard errors as discussed after Table 21.5. Alternatively, richer mixed models, presented in Chapter 22, can be estimated.

The assumption of independence over i is maintained throughout Chapters 21–23, though it can be relaxed even for small T provided structure can be placed on the correlation. An example is an explicit model for spatial correlation for panel data on regions such as states or countries, with correlations declining as physical distance between individual observations increases.

21.5.3. Error Variance Matrix for Long Panels

In **long panels** there are many time periods but relatively few individuals. Such data can arise in microeconomics analysis if the individual observational unit is one of only a few regions, such as a state or country, or firms, but these are observed over enough time periods to base inference on the assumption that $T \rightarrow \infty$.

Correlation across time for a given individual can be introduced using an ARMA model for the errors, with the parameters of the ARMA model permitted to differ across individuals as now N is fixed and $T \rightarrow \infty$. For example, consider an AR(1) error with $u_{it} = \rho_i u_{i,t-1} + \varepsilon_{it}$, where $\varepsilon_{it} \sim [0, \sigma_i^2]$ is heteroskedastic and ρ_i also differs across individuals. Separate regressions of y_{it} on w_{it} with AR(1) errors for each individual using T time periods yields consistent estimates $\hat{\rho}_i$ and $\hat{\sigma}_i^2$, since $T \rightarrow \infty$. These can then be used for feasible GLS estimation of δ using all NT observations. For details see Kmenta (1986). This model permits both heteroskedasticity across individuals and correlation over time for a given individual. Pesaran (2004) proposes a considerably richer model that is estimated by GLS.

For long panels it is possible to introduce correlation across individuals, so that $\text{Cov}[u_{it}, u_{jt}] \neq 0$ for $i \neq j$, since N is fixed and asymptotic results rely on $T \rightarrow \infty$. In particular, one can perform pooled GLS estimation as done earlier, with the assumption of independence across individuals, but then calculate standard errors using the method of Newey and West (1987b), mentioned briefly in Section 6.4.4, that permits arbitrary cross-sectional dependence and serial dependence, provided the serial dependence dies away sufficiently fast. For details see Arellano (2003, p. 19).

Time-series considerations for panel data are discussed in more detail in Section 22.5 for models with lagged dependent variables as regressors.

21.5.4. The Impact of Autocorrelated Errors

Panel data regression models have errors that are usually autocorrelated over time for a given individual. If fixed effects are absent then pooled OLS regression gives consistent parameter estimates. However, the **error correlation** can lead to **large bias** in standard errors for pooled OLS if autocorrelation is ignored and to relatively small efficiency gains as the length of a panel is increased.

The analysis is particularly simple for estimation of the mean of y based on T observations for one individual (so $N = 1$) with equicorrelation. Then $y_t = \beta + u_t$, and the OLS estimator is the sample mean, so $\hat{\beta} = \bar{y} = T^{-1} \sum_t y_t$. The OLS estimator has true variance $V[\hat{\beta}] = V[\bar{y}] = T^{-2} \sum_t \sum_s \text{Cov}[u_t, u_s]$. Assuming equicorrelation the double sum has T variances equal to σ^2 and $T(T - 1)$ covariances all equal to $\rho\sigma^2$. Hence $V[\bar{y}] = T^{-1}\sigma^2(1 + (T - 1)\rho)$. Thus the iid result that $V[\bar{y}] = T^{-1}\sigma^2$ needs to be modified by inflation by a multiple $(1 + \rho(T - 1))$. In particular $V[\bar{y}]$ approaches σ^2 as $\rho \rightarrow 1$.

Table 21.6 presents the impact of correlation on the variance of \bar{y} for different values of T and ρ , where for simplicity we normalize $\sigma^2 = 1$. The precision of estimation falls considerably as ρ increases, and the estimate of $V[\bar{y}]$ under the assumption of independence given in the first column (assuming σ^2 is known for simplicity) can

Table 21.6. Variances of Pooled OLS Estimator with Equicorrelated Errors^a

T	$\rho = 0.0$	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$	$\rho = 1.0$
1	1.00	1.00	1.00	1.00	1.00	1.00
2	0.50	0.60	0.70	0.80	0.90	1.00
5	0.20	0.36	0.52	0.68	0.84	1.00
10	0.10	0.28	0.46	0.64	0.82	1.00

^a Given are the variances of the pooled OLS estimator as the correlation ρ of equicorrelated errors increases, for an intercept-only model with error variance normalized to one assuming errors are correlated though homoskedastic.

greatly underestimate the true variance. Furthermore, for $\rho > 0$ the gain in precision due to increase in the number of time periods is much smaller than with independent data where a doubling of the number of time periods will halve estimator variance. For example, if $\rho = 0.4$ then with five time periods the estimator variance is only 0.52 times that with one period, instead of the much lower multiple of 0.2 with independent data. Moreover, a doubling from 5 to 10 time periods leads to only a small reduction in estimator variance from 0.52 to 0.46.

This result holds more generally for balanced panel regression with equicorrelated errors and regressors that are time-invariant, where the true variance of the OLS estimator is $(1 + \rho(T - 1))$ times that assuming independent errors (see Kloek, 1981). In practice time-varying regressors are also included and clear analytical results are more difficult to obtain. For regression with intercept and single time-varying regressor, Scott and Holt (1982) show that the variance of the slope coefficient is inflated by the multiple $(1 + \hat{\rho}_x \rho(T - 1))$, where $\hat{\rho}_x$ can be viewed as an estimate of the individual-specific autocorrelation in x . For panel data $\hat{\rho}_x$ is often high so that there is still considerable inflation. These results also apply to other forms of clustered data and are presented in more detail in Section 24.5.2.

The preceding analysis assumes equicorrelated errors, a property of the RE model. If instead errors are AR(1) there is greater benefit from increasing panel length. Then $\text{Cov}[u_t, u_s] = \rho^{|t-s|}\sigma^2$, so $V[\bar{y}] = T^{-2}\sigma^2[T + 2\sum_{s=1}^{T-1}(T-s)\rho^s]$. For example, if $\rho = 0.8$ then $V[\bar{y}] = 0.72\sigma^2$ for $T = 5$ and $0.54\sigma^2$ for $T = 10$, lower than the corresponding values from Table 21.6 of $0.84\sigma^2$ and $0.82\sigma^2$ for equicorrelation with $\rho = 0.8$, but still much higher than values of $0.2\sigma^2$ and $0.1\sigma^2$ for $\rho = 0.0$.

Microeconomists gravitate to the RE model or equicorrelated error models for short panels as an outgrowth of the literature on clustered data presented in Chapter 24. For example, consider data on different siblings in a family for many families. Then it is natural to assume that correlations of unobservables across siblings in the same family are the same for different sibling pairs. For example, the correlation between the first and second siblings equals that between the first and third siblings. Those using long panel data instead often have a time-series background and naturally assume that correlation declines over time, leading to models such as an AR(1) error.

Determining which model of time-series correlation is more reasonable really depends on the data. Many short panels used in microeconomics applications yield

pooled OLS residual autocorrelations that are qualitatively similar to those given in Table 21.3. These are closer to an RE model than an AR(1) model, though an ARMA(1,1) model may do well. Better still may be an RE model with AR(1) error. In all cases error correlation leads to a loss of information and the usual OLS standard errors underestimate the true standard errors. For short panels one can base inference on panel robust standard errors (see Section 21.2.3) that do not require specifying a model for the error correlation.

21.5.5. Hours and Wages Pooled GLS Example

A variety of pooled GLS estimates and associated default and robust standard errors of the model $y_{it} = \alpha_i + \beta x_{it} + u_{it}$ for the lnhrs on lnwg regression are given in Table 21.7. All assume the error u_{it} is independent over i and identically distributed over i , and then have different assumptions on correlation in u_{it} over t .

The first column of Table 21.7, for the pooled OLS estimator, repeats the first column of Table 21.2.

Pooled GLS estimates assuming equicorrelated errors are given in the second column of Table 21.7. These coincide with the RE-GLS column in Table 21.2, since the random effects model implies equicorrelated errors (see (21.6)).

Pooled GLS estimates assuming AR(1) errors, so that $u_{it} = \rho u_{it-1} + \varepsilon_{it}$ where ε_{it} is iid, are given in the third column of Table 21.7. The slope coefficient estimate is close to the pooled OLS estimate.

Pooled GLS estimates with no structure placed on error correlation aside from homoskedasticity, so that $\text{Cov}[u_{it}, u_{is}] = \sigma_{ts}$, are given in the fourth column of Table 21.7. Then σ_{ts} is consistently estimated given small T by $\widehat{\sigma}_{ts} = N^{-1} \sum_{i=1}^N \widehat{u}_{it} \widehat{u}_{is}$ for all t and s . These are again close to the pooled OLS estimate.

It is clear from Table 21.7 that panel-robust standard errors should be used rather than the default standard errors, which here assume homoskedasticity and correctly-specified model for serial correlation.

Table 21.7. Hours and Wages: Pooled OLS and GLS Estimates^a

Estimator	POLS		PGGLS	
	None	Equi	AR1	General
α	7.442	7.346	7.440	7.426
β	.083	.120	.084	.091
Robust se	(.029)	(.052)	(.037)	(.050)
Boot se	[.032]	[.060]	[.050]	[-]
Default se	{.009}	{.014}	{.012}	{.014}

^a Pooled OLS and GLS linear panel regression of lnhrs on lnwg for a short panel assuming independence and identical distribution over i and no fixed effects. Pooled GLS estimators assume equicorrelated or random effects errors (equi), AR(1) errors (AR1), or no structure on the correlations (general). Standard errors for the slope coefficients are panel robust in parentheses, panel bootstrap in square brackets, and usual default estimates that assume iid errors in curly braces.

21.6. Fixed Effects Model

The **fixed effects model** specifies

$$y_{it} = \alpha_i + \mathbf{x}'_{it}\beta + \varepsilon_{it}, \quad (21.22)$$

where the individual-specific effects $\alpha_1, \dots, \alpha_N$ measure unobserved heterogeneity that is possibly correlated with the regressors, \mathbf{x}_{it} and β are $K \times 1$ vectors, and to begin with the errors ε_{it} are iid $[0, \sigma^2]$.

The challenge for estimation is the presence of the N individual-specific effects that increase in number as $N \rightarrow \infty$. For practical purposes we are most interested in the K slope parameters β , which give the marginal effect of change in regressors since $\partial E[y_{it}] / \partial \mathbf{x}_{it} = \beta$. The N parameters $\alpha_1, \dots, \alpha_N$ are **nuisance parameters** or **incidental parameters** that are not of intrinsic interest. Nevertheless, their presence potentially prevents estimation of the parameters β that are of interest.

Remarkably, for the linear model there are several ways to consistently estimate β despite the presence of these nuisance parameters. These include (1) OLS in the within model (21.8); (2) direct OLS estimation of the model (21.2) with indicator variables for each of the N fixed effects; (3) GLS in the within model (21.8); (4) ML estimation conditional on the individual means \bar{y}_i , $i = 1, \dots, N$; and (5) OLS in the first-differences model (21.9).

The first two methods always lead to the same estimator for β . So too does the third if additionally the ε_{it} in (21.22) are iid and the fourth if $\varepsilon_{it} \sim \mathcal{N}[0, \sigma^2]$. The last estimator differs from the others for $T > 2$. Such equivalences generally do not hold in nonlinear models, which are considered in Chapter 23.

The essential results for the within estimator are given in the next Section. The first-differences estimator, presented in Section 21.6.2, is extensively used in Chapter 22 when regressors are no longer strongly exogenous. The other estimators are presented in the remainder of Section 21.6, which some readers may wish to skip.

21.6.1. Within or Fixed Effects Estimator

The within model is obtained by subtraction of the time-averaged model $\bar{y}_i = \alpha_i + \bar{\mathbf{x}}'_i\beta + \bar{\varepsilon}_i$ from the original model. Then

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)' \beta + (\varepsilon_{it} - \bar{\varepsilon}_i), \quad (21.23)$$

so the fixed effect α_i is eliminated, along with time-invariant regressors since $\mathbf{x}_{it} - \bar{\mathbf{x}}_i = \mathbf{0}$ if $\mathbf{x}_{it} = \mathbf{x}_i$ for all t .

Using OLS estimation yields the **within estimator** or **fixed effects estimator** $\hat{\beta}_W$, where

$$\hat{\beta}_W = \left[\sum_{i=1}^N \sum_{t=1}^T (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)' \right]^{-1} \sum_{i=1}^N \sum_{t=1}^T (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(y_{it} - \bar{y}_i). \quad (21.24)$$

The individual fixed effects α_i can then be estimated by

$$\hat{\alpha}_i = \bar{y}_i - \bar{\mathbf{x}}'_i \hat{\beta}_W, \quad i = 1, \dots, N. \quad (21.25)$$

The estimate $\hat{\alpha}_i$ is unbiased for α_i , and it is consistent provided $T \rightarrow \infty$ since $\hat{\alpha}_i$ averages T observations. In short panels the estimates $\hat{\alpha}_i$ are inconsistent, but $\hat{\beta}_W$ is nonetheless consistent for β . The α_i are viewed as **nuisance parameters or ancillary parameters** that fortunately do not need to be consistently estimated to obtain consistent estimates of the more important slope parameters β . This remarkable result need not carry over to more complicated fixed effects models such as nonlinear models.

Consistency of the Within Estimator

The within estimator of β is consistent if $\text{plim}(NT)^{-1} \sum_i \sum_t (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\varepsilon_{it} - \bar{\varepsilon}_i) = \mathbf{0}$. This should happen if either $N \rightarrow \infty$ or $T \rightarrow \infty$ and

$$\mathbb{E}[\varepsilon_{it} - \bar{\varepsilon}_i | \mathbf{x}_{it} - \bar{\mathbf{x}}_i] = 0. \quad (21.26)$$

Owing to the presence of the averages $\bar{\mathbf{x}}_i = T^{-1} \sum_t \mathbf{x}_{it}$ and $\bar{\varepsilon}_i$ this condition is stronger than $\mathbb{E}[\varepsilon_{it} | \mathbf{x}_{it}] = 0$. A sufficient condition for (21.26) is the strong exogeneity condition that $\mathbb{E}[\varepsilon_{it} | \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}] = 0$. This precludes within estimation with lagged endogenous variables as regressors (see Section 22.5).

Asymptotic Distribution of the Within Estimator

The distribution of $\hat{\beta}_W$ appears potentially complicated because the error $(\varepsilon_{it} - \bar{\varepsilon}_i)$ in the within model (21.8) is correlated over t for given i . It is shown in the following that the usual OLS results nonetheless apply. Under the strong assumption that ε_{it} is iid,

$$\text{V}[\hat{\beta}_W] = \sigma_\varepsilon^2 \left[\sum_{i=1}^N \sum_{t=1}^T \ddot{\mathbf{x}}_{it} \ddot{\mathbf{x}}'_{it} \right]^{-1}, \quad (21.27)$$

where $\ddot{\mathbf{x}}_{it} = \mathbf{x}_{it} - \bar{\mathbf{x}}_i$. A consistent and unbiased estimate of σ_ε^2 is $\hat{\sigma}_\varepsilon^2 = [N(T-1) - K]^{-1} \sum_i \sum_t \hat{\varepsilon}_{it}^2$, where the degrees of freedom equal the sample size NT less the number of model parameters K and the N individual effects. Note that if the regression (21.23) is estimated using a standard least-squares package then we need to inflate the reported variances by $[N(T-1) - K]^{-1}[NT - K]$.

For short panels (21.13) yields the robust estimate of the asymptotic variance

$$\text{V}[\hat{\beta}_W] = \left[\sum_{i=1}^N \sum_{t=1}^T \ddot{\mathbf{x}}_{it} \ddot{\mathbf{x}}'_{it} \right]^{-1} \sum_{i=1}^N \sum_{t=1}^T \sum_{s=1}^T \ddot{\mathbf{x}}_{it} \ddot{\mathbf{x}}'_{is} \hat{\varepsilon}_{it} \hat{\varepsilon}'_{is} \left[\sum_{i=1}^N \sum_{t=1}^T \ddot{\mathbf{x}}_{it} \ddot{\mathbf{x}}'_{it} \right]^{-1}, \quad (21.28)$$

where $\hat{\varepsilon}_{it} = \varepsilon_{it} - \bar{\varepsilon}_i$. This preferred estimate permits arbitrary autocorrelations for the ε_{it} and arbitrary heteroskedasticity.

Derivation of the Variance of the Within Estimator

We now derive the estimates of the variance of the within estimator given in (21.27) and (21.28), using matrix algebra. We begin with the model for the i th observation

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

where \mathbf{x}_{it} and β are $K \times 1$ vectors. For the i th individual, stack all T observations, so

$$\begin{bmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \alpha_i + \begin{bmatrix} \mathbf{x}'_{i1} \\ \vdots \\ \mathbf{x}'_{iT} \end{bmatrix} \beta + \begin{bmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{iT} \end{bmatrix}, \quad i = 1, \dots, N,$$

or

$$\mathbf{y}_i = \mathbf{e}\alpha_i + \mathbf{X}_i\beta + \varepsilon_i, \quad i = 1, \dots, N, \quad (21.29)$$

where $\mathbf{e} = (1, 1, \dots, 1)'$ is a $T \times 1$ vector of ones, \mathbf{X}_i is a $T \times K$ matrix, and \mathbf{y}_i and ε_i are $T \times 1$ vectors.

To transform model (21.29) to the within model, which subtracts the individual-specific mean, introduce the $T \times T$ matrix

$$\mathbf{Q} = \mathbf{I}_T - T^{-1}\mathbf{e}\mathbf{e}'. \quad (21.30)$$

Premultiplication by the matrix \mathbf{Q} creates deviations from the mean, since

$$\mathbf{Q}\mathbf{W}_i = \mathbf{W}_i - \mathbf{e}\bar{\mathbf{w}}_i', \quad (21.31)$$

where \mathbf{W}_i is a $T \times m$ matrix with t th row \mathbf{w}'_{it} and $\bar{\mathbf{w}}_i = T^{-1} \sum_{t=1}^T \mathbf{w}_{it}$ is a $m \times 1$ vector of averages. The result (21.31) is obtained using $\mathbf{e}'\mathbf{W}_i = T\bar{\mathbf{w}}_i'$. Note also that $\mathbf{Q}\mathbf{Q}' = \mathbf{Q}$, using $\mathbf{e}'\mathbf{e} = T$ and $\mathbf{Q}\mathbf{e} = \mathbf{0}$, so \mathbf{Q} is idempotent.

Premultiplying the fixed effects model (21.29) for the i th individual by \mathbf{Q} yields

$$\mathbf{Q}\mathbf{y}_i = \mathbf{Q}\mathbf{X}_i\beta + \mathbf{Q}\varepsilon_i, \quad i = 1, \dots, N, \quad (21.32)$$

using $\mathbf{Q}\mathbf{e} = \mathbf{0}$. This is the within model (21.23), since equivalently $\mathbf{y}_i - \mathbf{e}\bar{y}'_i = (\mathbf{X}_i - \mathbf{e}\bar{\mathbf{x}}'_i)\beta + (\varepsilon_i - \mathbf{e}\bar{\varepsilon}_i)$ using (21.31). Thus premultiplication by \mathbf{Q} yields the within model. An OLS estimation of (21.32) yields $\hat{\beta}_W$ with variance matrix, assuming independence over i , equal to

$$\text{V}[\hat{\beta}_W] = \left[\sum_{i=1}^N \mathbf{X}'_i \mathbf{Q}' \mathbf{Q} \mathbf{X}_i \right]^{-1} \sum_{i=1}^N \mathbf{X}'_i \mathbf{Q}' \mathbf{V}[\mathbf{Q}\varepsilon_i | \mathbf{X}_i] \mathbf{Q} \mathbf{X}_i \left[\sum_{i=1}^N \mathbf{X}'_i \mathbf{Q}' \mathbf{Q} \mathbf{X}_i \right]^{-1}. \quad (21.33)$$

Begin with the strong assumption that ε_{it} are iid $[0, \sigma_\varepsilon^2]$, so that ε_i are iid $[\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}]$. The $T \times 1$ error $\mathbf{Q}\varepsilon_i$ is then independent over i with mean zero and variance $\text{V}[\mathbf{Q}\varepsilon_i] = \mathbf{Q}\text{V}[\varepsilon_i]\mathbf{Q}' = \sigma_\varepsilon^2 \mathbf{Q}\mathbf{Q}' = \sigma_\varepsilon^2 \mathbf{Q}$. Then

$$\begin{aligned} \sum_{i=1}^N \mathbf{X}'_i \mathbf{Q}' \mathbf{V}[\mathbf{Q}\varepsilon_i | \mathbf{X}_i] \mathbf{Q} \mathbf{X}_i &= \sum_{i=1}^N \mathbf{X}'_i \mathbf{Q}' \sigma_\varepsilon^2 \mathbf{Q} \mathbf{Q} \mathbf{X}_i \\ &= \sigma_\varepsilon^2 \sum_{i=1}^N \mathbf{X}'_i \mathbf{Q}' \mathbf{Q} \mathbf{X}_i, \end{aligned}$$

so that (21.33) simplifies to the estimate given in (21.27), using

$$(\mathbf{Q}\mathbf{X}_i)'(\mathbf{Q}\mathbf{X}_i) = \sum_{t=1}^T (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)',$$

At the time of writing many packages use (21.27) but alternative estimators may be better. In particular, the assumption of serially uncorrelated error ε_{it} is easily relaxed. If ε_i are iid $[\mathbf{0}, \Sigma_i]$ we use the more general form of the variance matrix (21.33) with $\text{Cov}[\mathbf{Q}\varepsilon_i, \mathbf{Q}\varepsilon_j] = \mathbf{0}$, for $i \neq j$, and $V[\mathbf{Q}\varepsilon_i]$ replaced by $(\mathbf{Q}\hat{\varepsilon}_i)(\mathbf{Q}\hat{\varepsilon}_i)'$, where $\hat{\varepsilon}_i = \mathbf{y}_i - \mathbf{X}_i\hat{\beta}_W$. This yields the estimate given in (21.28).

From the derivation it should be clear that $\hat{\beta}_W$ is also consistent in the random effects model, though as shown in Section 21.7 it is less efficient than the random effects estimator if the random effects model is appropriate.

GLS Estimation of the Within Model

The within model (21.32) can also be estimated by feasible GLS.

If in fact ε_{it} are iid $[0, \sigma_\varepsilon^2]$, however, then there are no gains to doing GLS. To see this, note that then $\mathbf{Q}\varepsilon_i$ is independent of $\mathbf{Q}\varepsilon_j$, $i \neq j$, with $V[\mathbf{Q}\varepsilon_i] = \sigma_\varepsilon^2 \mathbf{Q}$, so the **GLS estimator** is

$$\hat{\beta}_{W,GLS} = \left[\sum_{i=1}^N \mathbf{X}'_i \mathbf{Q}' \mathbf{Q}^{-1} \mathbf{Q} \mathbf{X}_i \right]^{-1} \sum_{i=1}^N \mathbf{X}'_i \mathbf{Q}' \mathbf{Q}^{-1} \mathbf{Q} \mathbf{y}_i,$$

where the generalized inverse \mathbf{Q}^{-1} is used as \mathbf{Q} is not of full rank. However, $\mathbf{Q}' \mathbf{Q}^{-1} \mathbf{Q} = \mathbf{Q}' \mathbf{Q}$ since $\mathbf{Q}' \mathbf{Q}^{-1} \mathbf{Q} = \mathbf{Q}$, for a generalized inverse, and $\mathbf{Q} = \mathbf{Q} \mathbf{Q}'$ as \mathbf{Q} here is idempotent. Replacing $\mathbf{Q}' \mathbf{Q}^{-1} \mathbf{Q}$ by $\mathbf{Q}' \mathbf{Q}$ in the formula for $\hat{\beta}_{W,GLS}$ yields the OLS estimator in (21.32).

There can be gains to GLS if other models for ε_{it} are assumed. The approach is essentially the same as that in Section 21.5.2 for pooled GLS without fixed effects, except that first the fixed effect must be eliminated. This leads to error $\mathbf{Q}\varepsilon_i$ that is less than full rank, so we first drop one time period and apply pooled GLS to only $(T - 1)$ time periods. It is easier, and often not much less efficient, to instead just use the usual within FE estimator and then obtain panel-robust standard errors using (21.28).

MacCurdy (1982b) gives a Box–Jenkins-type analysis for identification and estimation of ARMA processes for ε_{it} in a fixed effects model for a short panel. For short panels it is not necessary to assume an ARMA process for ε_{it} or even stationarity, since for $N \rightarrow \infty$ we can always consistently estimate $\text{Cov}[u_{it}, u_{is}]$ by $N^{-1} \sum_i \hat{u}_{it} \hat{u}_{is}$. Nonetheless, there may be interest in determining the ARMA process for the errors.

21.6.2. First-Differences Estimator

The within model is obtained by subtraction of the time-averaged model $\bar{y}_i = \alpha_i + \bar{\mathbf{x}}_i' \beta + \bar{\varepsilon}_i$ from the original model. Alternatively, one can subtract the model lagged

one period, $y_{i,t-1} = \alpha_i + \mathbf{x}_{i,t-1}'\beta + \varepsilon_{i,t-1}$. Then

$$(y_{it} - y_{i,t-1}) = (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\beta + (\varepsilon_{it} - \varepsilon_{i,t-1}), \quad t = 2, \dots, T, \quad (21.34)$$

so the fixed effect α_i is eliminated. An OLS estimation yields the **first-differences estimator**

$$\hat{\beta}_{FD} = \left[\sum_{i=1}^N \sum_{t=2}^T (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})(\mathbf{x}_{it} - \mathbf{x}_{i,t-1})' \right]^{-1} \sum_{i=1}^N \sum_{t=2}^T (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})(y_{it} - y_{i,t-1}). \quad (21.35)$$

Note that there are only $N(T - 1)$ observations in this regression. An easy error to make in implementation is to stack all NT observations and then subtract the first lag. Then only the $(1, 1)$ observation is dropped, whereas all T first-period observations $(i, 1)$, $i = 1, \dots, N$, must be dropped after differencing.

Consistency of the First-Differences Estimator

Consistency of the first differences estimator requires that $E[\varepsilon_{it} - \varepsilon_{i,t-1} | \mathbf{x}_{it} - \mathbf{x}_{i,t-1}] = 0$. This is a stronger condition than $E[\varepsilon_{it} | \mathbf{x}_{it}] = 0$ but a weaker condition than the strong exogeneity condition needed for consistency of the within estimator.

Asymptotic Distribution of the First-Differences Estimator

Statistical inference requires adjusting the usual OLS standard errors to account for the correlation over time in the error term $\varepsilon_{it} - \varepsilon_{i,t-1}$. To obtain the asymptotic variance of $\hat{\beta}_{FD}$, stack the model for the i th individual as

$$\Delta \mathbf{y}_i = \Delta \mathbf{X}'_i \beta + \Delta \varepsilon_i,$$

where $\Delta \mathbf{y}_i$ is a $(T - 1) \times 1$ vector with entries $(y_{i2} - y_{i1}), \dots, (y_{iT} - y_{i,T-1})$, and $\Delta \mathbf{X}_i$ is a $(T - 1) \times K$ vector with rows $(\mathbf{x}_{i2} - \mathbf{x}_{i1})', \dots, (\mathbf{x}_{iT} - \mathbf{x}_{i,T-1})'$. Then

$$\hat{\beta}_{FD} = \left[\sum_{i=1}^N (\Delta \mathbf{X}_i)'(\Delta \mathbf{X}_i) \right]^{-1} \sum_{i=1}^N (\Delta \mathbf{X}_i)'(\Delta \mathbf{y}_i) \quad (21.36)$$

has variance matrix, assuming independence over i , of

$$V[\hat{\beta}_{FD}] = \left[\sum_{i=1}^N (\Delta \mathbf{X}_i)'(\Delta \mathbf{X}_i) \right]^{-1} \left[\sum_{i=1}^N (\Delta \mathbf{X}_i)' V[\Delta \varepsilon_i | \Delta \mathbf{X}_i](\Delta \mathbf{X}_i) \right] \left[\sum_{i=1}^N (\Delta \mathbf{X}_i)'(\Delta \mathbf{X}_i) \right]^{-1}. \quad (21.37)$$

The simplest assumption is that ε_{it} are iid $[0, \sigma_\varepsilon^2]$. Then the error $(\varepsilon_{it} - \varepsilon_{i,t-1})$ is now an MA(1) error, with variance $2\sigma_\varepsilon^2$ and one-period apart autocovariance σ_ε^2 for individual i . It follows that $V[\Delta \varepsilon_i]$ equals σ_ε^2 times a $(T - 1) \times (T - 1)$ matrix with entries of 2 on the diagonal, entries of 1 on the immediate off-diagonals, and 0s elsewhere.

A more realistic assumption is that ε_{it} is correlated over time for given i , so that $Cov[\varepsilon_{it}, \varepsilon_{is}] \neq 0$ for $t \neq s$, but is still independent over i . From (21.13), for short panels an estimator that is robust to general forms of autocorrelation and

heteroskedasticity is (21.37) with $V[\Delta\varepsilon_i]$ replaced by $(\widehat{\Delta\varepsilon}_i)'(\widehat{\Delta\varepsilon}_i)$. One should never use the usual OLS standard errors from OLS regression of the first-differences model (21.37), as these are only correct in the unlikely event that ε_{it} is a random walk, so that $(\varepsilon_{it} - \varepsilon_{i,t-1})$ are iid.

For $T = 2$ the first-differences and within estimators are equal since $\bar{y} = (y_1 + y_2)/2$ so $(y_1 - \bar{y}) = (y_1 - y_2)/2$ and $(y_2 - \bar{y}) = -(y_1 - y_2)/2$, and similarly for \mathbf{x} . For $T > 2$ the two estimators differ. Under the simplest assumption that ε_{it} are iid, it can be shown that the GLS estimator of the first-difference model (21.34) equals the within estimator. The estimator $\widehat{\beta}_{FD}$ instead estimates (21.34) by OLS and is less efficient than $\widehat{\beta}_W$. For this reason the first-difference estimator is not mentioned much in introductory courses. However, it is used extensively once lagged dependent variables are introduced (see Chapter 22). Then the within estimator is inconsistent. The first-differences estimator is also inconsistent, but relies on weaker exogeneity assumptions that permit consistent IV estimation.

21.6.3. Conditional ML Estimator

The conditional MLE maximizes the joint likelihood of y_{11}, \dots, y_{NT} conditional on the individual averages $\bar{y}_1, \dots, \bar{y}_T$. This method has the attraction that, for the linear panel model under normality, the fixed effects α_i are eliminated, so maximization is with respect to β alone.

Assume that y_{it} conditional on regressors \mathbf{x}_{it} and parameters α_i, β , and σ^2 are iid with normal distribution $\mathcal{N}[\alpha_i + \mathbf{x}'_{it}\beta, \sigma^2]$. Then the **conditional likelihood function** is

$$\begin{aligned} L_{COND}(\beta, \sigma^2, \alpha) &= \prod_{i=1}^N f(y_{i1}, \dots, y_{iT} | \bar{y}_i) \\ &= \prod_{i=1}^N \frac{f(y_{i1}, \dots, y_{iT}, \bar{y}_i)}{f(\bar{y}_i)} \\ &= \prod_{i=1}^N \frac{(2\pi\sigma^2)^{-T/2}}{(2\pi\sigma^2/T)^{-1/2}} \exp \left\{ \sum_{t=1}^T -[(y_{it} - \mathbf{x}'_{it}\beta)^2 + (\bar{y}_i - \bar{\mathbf{x}}'_i\beta)^2]/2\sigma^2 \right\}. \end{aligned} \quad (21.38)$$

The first equality defines the conditional likelihood assuming independence over i . The second equality always holds since, suppressing subscript i , $f(y_1, \dots, y_T | \bar{y}) = f(y_1, \dots, y_T, \bar{y})/f(\bar{y})$ and $f(y_1, \dots, y_T, \bar{y}) = f(y_1, \dots, y_T)$ as knowledge of $\bar{y} = T^{-1} \sum_i y_i$ adds nothing given knowledge of y_1, \dots, y_T . The third equality under normality comes after considerable algebra that is left as an exercise.

The key result is that the fixed effects α do not appear in the final equality in (21.38), so $L_{COND}(\beta, \sigma^2, \alpha)$ is in fact $L_{COND}(\beta, \sigma^2)$, and we need to maximize the conditional log-likelihood function (21.38) with respect to β and σ^2 only. The resulting **conditional ML estimator** $\widehat{\beta}_{CML}$ solves the first-order conditions

$$\frac{1}{\sigma^2} \sum_{t=1}^T \sum_{i=1}^N [(y_{it} - \mathbf{x}'_{it}\beta)\mathbf{x}_{it} - (\bar{y}_i - \bar{\mathbf{x}}'_i\beta)\bar{\mathbf{x}}_i] = \mathbf{0},$$

or equivalently

$$\sum_{t=1}^T \sum_{i=1}^N [(y_{it} - \bar{y}_i) - (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)' \boldsymbol{\beta}] (\mathbf{x}_{it} - \bar{\mathbf{x}}_i) = \mathbf{0}.$$

However, these are just the first-order conditions from OLS regression of $(y_{it} - \bar{y}_i)$ on $(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)$.

The conditional MLE $\widehat{\boldsymbol{\beta}}_{CML}$ therefore equals the within estimator $\widehat{\boldsymbol{\beta}}_W$.

Intuitively, the method yields a consistent estimator because conditioning on \bar{y}_i in (21.38) eliminated the fixed effects. More formally, \bar{y}_i is a sufficient statistic for α_i and conditioning on the sufficient statistic enables consistent estimation of $\boldsymbol{\beta}$ (see Section 23.2.2).

21.6.4. Least-Squares Dummy Variable Estimator

Consider the original fixed effects model (21.22) before any differencing. An OLS analysis can be applied directly to the model, simultaneously estimating $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

In principle no special software is needed. One simply estimates the OLS regression of y_{it} on \mathbf{x}_{it} and a set of N indicator variables $d_{1,it}, \dots, d_{N,it}$, where $d_{j,it}$ equals one if $j = i$ and equals zero otherwise. However, as N gets large there are too many regressors to permit inversion of the $(N + K) \times (N + K)$ regressor matrix. Some matrix algebra, however, reduces the problem to inversion of a $K \times K$ matrix.

The resulting estimator of $\boldsymbol{\beta}$ turns out to equal the within estimator. This is a special case of the so-called Frisch-Waugh Theorem for a subset regression. If dummy variables are partialled out by regression of all the variables on the dummies, and if the residuals from these regressions are used in a second stage regression, then we get the same estimates as in the full regression. But these residuals here are simply deviations from their respective means, i.e. the within regression. For completeness we now present the relevant matrix algebra.

Stack the $T \times 1$ vectors in (21.29) over all N individuals to yield the **fixed effects dummy variable model**

$$\begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \end{bmatrix} = \begin{bmatrix} \mathbf{e} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{e} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix} + \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_N \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{bmatrix},$$

or

$$\mathbf{y} = [(\mathbf{I}_N \otimes \mathbf{e}) \quad \mathbf{X}] \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} + \boldsymbol{\varepsilon}, \quad (21.39)$$

where \mathbf{y} is an $NT \times 1$ vector, the Kronecker product $(\mathbf{I}_N \otimes \mathbf{e})$ is an $NT \times N$ block-diagonal matrix, and \mathbf{X} is the $NT \times K$ matrix of nonconstant regressors.

An OLS estimation of this model yields the **least-squares dummy variable (LSDV) estimator**

$$\begin{aligned} \begin{bmatrix} \widehat{\alpha}_{\text{LSDV}} \\ \widehat{\beta}_{\text{LSDV}} \end{bmatrix} &= \begin{bmatrix} (\mathbf{I}_N \otimes \mathbf{e})'(\mathbf{I}_N \otimes \mathbf{e}) & (\mathbf{I}_N \otimes \mathbf{e})'\mathbf{X} \\ \mathbf{X}'(\mathbf{I}_N \otimes \mathbf{e}) & \mathbf{X}'\mathbf{X} \end{bmatrix}^{-1} \times \begin{bmatrix} (\mathbf{I}_N \otimes \mathbf{e})'\mathbf{y} \\ \mathbf{X}'\mathbf{y} \end{bmatrix} \\ &= \begin{bmatrix} T\mathbf{I}_N & T\bar{\mathbf{X}} \\ T\bar{\mathbf{X}}' & \mathbf{X}'\mathbf{X} \end{bmatrix}^{-1} \times \begin{bmatrix} \bar{\mathbf{y}} \\ \mathbf{X}'\mathbf{y} \end{bmatrix}, \end{aligned}$$

where the matrix of sample means $\bar{\mathbf{X}} = [\bar{\mathbf{x}}'_1 \cdots \bar{\mathbf{x}}'_N]'$, $\bar{\mathbf{x}}_i = T^{-1} \sum_{t=1}^T \mathbf{x}_{it}$, $\bar{\mathbf{y}} = [\bar{y}_1 \cdots \bar{y}_N]'$, and $\bar{y}_i = T^{-1} \sum_{t=1}^T y_{it}$. Using the formula for partitioned inverse and performing further algebra leads to

$$\begin{bmatrix} \widehat{\alpha}_{\text{LSDV}} \\ \widehat{\beta}_{\text{LSDV}} \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{y}} - \bar{\mathbf{X}}\widehat{\beta}_w \\ [\mathbf{X}'\mathbf{X} - \bar{\mathbf{X}}'\bar{\mathbf{X}}]^{-1}(\mathbf{X}'\mathbf{y} - \bar{\mathbf{X}}'\bar{\mathbf{y}}) \end{bmatrix}. \quad (21.40)$$

Reexpressing this in summation notation, we have $\widehat{\beta}_{\text{LSDV}} = \widehat{\beta}_w$ defined in (21.24) and $\widehat{\alpha}_{\text{LSDV}} = \widehat{\alpha}_{\text{FE}}$ defined in (21.25), so the LSDV estimators equal the within or fixed effects estimator

For short panels an obvious potential problem is that consistent estimation of β and α is not guaranteed as there are $N + K$ parameters to estimate and $N \rightarrow \infty$. Remarkably, consistent estimation of β is possible, even though α is inconsistently estimated, unless additionally $T \rightarrow \infty$.

This estimator is second-moment efficient if ε_{it} are iid $[0, \sigma^2]$. It follows that the within estimator of β is more efficient than alternative differencing estimators that also eliminate α_i , such as subtracting the first observation or the previous period's observation. If additionally the errors are normally distributed, the LSDV estimator equals the MLE by the usual equivalence of OLS and MLE in the linear model with spherical normal errors.

21.6.5. Covariance Estimator

Suppose data belong to one of N classes, with y_{it} denoting the t th observation in the i th class. The **analysis of variance** decomposes the total variation of y_{it} around the grand mean \bar{y} , $\sum_i \sum_t (y_{it} - \bar{y})^2$, into **within-group** variation $\sum_i \sum_t (y_{it} - \bar{y}_i + \bar{y})^2$ and **between-group** variation $\sum_i (\bar{y}_i - \bar{y})^2$, where \bar{y}_i is the mean in the i th group. Group membership becomes more important as between-group variation increases. The **analysis of covariance** extends this approach to introduce regressors, in which case the residual sum of squares is similarly decomposed. This framework is widely used in applied statistics.

For short panels each individual is viewed as a class, observed for several time periods. The model (21.3) is called the **analysis-of-covariance model**, as it permits the mean residual in the i th class to differ over classes. The estimator of this model, the within estimator, is accordingly also called the **covariance estimator**.

21.7. Random Effects Model

The **random effects model** (21.3) can be rewritten as

$$y_{it} = \mu + \mathbf{x}'_{it}\beta + \alpha_i + \varepsilon_{it}, \quad i = 1, \dots, N, \quad t = 1, \dots, T, \quad (21.41)$$

or

$$y_{it} = \mathbf{w}'_{it}\delta + \alpha_i + \varepsilon_{it}, \quad (21.42)$$

where $\mathbf{w}_{it} = [1 \ \mathbf{x}_{it}]$ and $\delta = [\mu \ \beta']'$. The individual-specific effects α_i are assumed to be realizations of iid random variables with distribution $[0, \sigma_\alpha^2]$ and the error ε_{it} is iid $[0, \sigma_\varepsilon^2]$. The nonrandom scalar intercept μ is added so that, unlike in (21.5), the random effects can be normalized to have zero mean.

The model can alternatively be viewed as a special case of a **random coefficient** or **varying coefficient model**, where only the intercept coefficient is random. The model can be re-expressed as $y_{it} = \mu + \mathbf{x}'_{it}\beta + u_{it}$, where the error term u_{it} has two components $u_{it} = \alpha_i + \varepsilon_{it}$. For this reason the random effects model is also called the **error components model**. Even clearer terminology may be the **random intercept model**. Richer mixed models also permit random slopes, see Chapter 22.

There are many consistent estimators of the random effects model, including (1) GLS estimation in the model (21.42); (2) ML estimation in the model (21.42) assuming α_i and ε_{it} are normally distributed; (3) OLS estimation in the model (21.42); and (4) fixed effects model estimators such as the within and first-differences estimators, though these only estimate the coefficients of time-varying regressors. The first two estimators are asymptotically equivalent but can vary in finite samples depending on the specific estimates used for σ_α^2 and σ_ε^2 . The remaining estimators are consistent, though they are inefficient if in fact α_i and ε_{it} are iid.

21.7.1. GLS Estimator

The **random effects estimator** of μ and β is the feasible GLS estimator of the model (21.42), and it is shown later in this section that it can be implemented by OLS regression of the transformed equation

$$y_{it} - \hat{\lambda}\bar{y}_i = (1 - \hat{\lambda})\mu + (\mathbf{x}_{it} - \hat{\lambda}\bar{\mathbf{x}}_i)'\beta + v_{it}, \quad (21.43)$$

where $v_{it} = (1 - \hat{\lambda})\alpha_i + (\varepsilon_{it} - \hat{\lambda}\bar{\varepsilon}_i)$ and $\hat{\lambda}$ is consistent for

$$\lambda = 1 - \sigma_\varepsilon^2 / (T\sigma_\alpha^2 + \sigma_\varepsilon^2)^{1/2}. \quad (21.44)$$

Equivalently,

$$\hat{\delta}_{RE} = \begin{bmatrix} \hat{\mu}_{RE} \\ \hat{\beta}_{RE} \end{bmatrix} = \left[\sum_{i=1}^N \sum_{t=1}^T (\mathbf{w}_{it} - \hat{\lambda}\bar{\mathbf{w}}_i)(\mathbf{w}_{it} - \hat{\lambda}\bar{\mathbf{w}}_i)' \right]^{-1} \sum_{i=1}^N \sum_{t=1}^T (\mathbf{w}_{it} - \hat{\lambda}\bar{\mathbf{w}}_i)(y_{it} - \hat{\lambda}\bar{y}_i), \quad (21.45)$$

where $\mathbf{w}_{it} = [1 \ \mathbf{x}_{it}]$ and $\bar{\mathbf{w}}_i = [1 \ \bar{\mathbf{x}}_i]$. Consistency requires $NT \rightarrow \infty$, through either $N \rightarrow \infty$ or $T \rightarrow \infty$ or both.

Assuming that ε_{it} and α_i are iid, the usual OLS output from OLS regression of (21.43) can be used to obtain the variance matrix estimate, so that

$$\text{V} \begin{bmatrix} \widehat{\mu}_{\text{RE}} \\ \widehat{\beta}_{\text{RE}} \end{bmatrix} = \sigma_{\varepsilon}^2 \left[\sum_{i=1}^N \sum_{t=1}^T (\mathbf{w}_{it} - \widehat{\lambda} \bar{\mathbf{w}}_i)(\mathbf{w}_{it} - \widehat{\lambda} \bar{\mathbf{w}}_i)' \right]^{-1}. \quad (21.46)$$

Alternatively, for short panels a robust variance estimate that permits quite general behavior for $\alpha_i + \varepsilon_{it}$ can be obtained using (21.13). This yields

$$\text{V} \begin{bmatrix} \widehat{\mu}_{\text{RE}} \\ \widehat{\beta}_{\text{RE}} \end{bmatrix} = \left[\sum_{i=1}^N \sum_{t=1}^T \widetilde{\mathbf{w}}_{it} \widetilde{\mathbf{w}}'_{it} \right]^{-1} \sum_{i=1}^N \sum_{t=1}^T \sum_{s=1}^T \widetilde{\mathbf{w}}_{it} \widetilde{\mathbf{w}}'_{is} \widehat{\varepsilon}_{it} \widehat{\varepsilon}'_{is} \left[\sum_{i=1}^N \sum_{t=1}^T \widetilde{\mathbf{w}}_{it} \widetilde{\mathbf{w}}'_{it} \right]^{-1}, \quad (21.47)$$

where $\widetilde{\mathbf{w}}_{it} = \mathbf{w}_{it} - \widehat{\lambda} \bar{\mathbf{w}}_i$ and $\widetilde{\varepsilon}_{it} = \widehat{\varepsilon}_{it} - \widehat{\lambda} \widehat{\varepsilon}_i$ where $\widehat{\varepsilon}_{it}$ is the RE residual. This estimate permits arbitrary autocorrelations for the ε_{it} and arbitrary heteroskedasticity.

Equation (21.46) requires consistent estimates of the **variance components** σ_{ε}^2 and σ_{α}^2 . From the within or fixed effects regression of $(y_{it} - \bar{y}_i)$ on $(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)$ we obtain

$$\widehat{\sigma}_{\varepsilon}^2 = \frac{1}{N(T-1)-K} \sum_i \sum_t ((y_{it} - \bar{y}_i) - (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)' \widehat{\beta}_W)^2. \quad (21.48)$$

From the between regression of \bar{y}_i on an intercept and $\bar{\mathbf{x}}_i$, an equation that has error term with variance $\sigma_{\alpha}^2 + \sigma_{\varepsilon}^2/T$, we obtain

$$\widehat{\sigma}_{\alpha}^2 = \frac{1}{N-(K+1)} \sum_i (\bar{y}_i - \widehat{\mu}_B - \bar{\mathbf{x}}_i' \widehat{\beta}_B)^2 - \frac{1}{T} \widehat{\sigma}_{\varepsilon}^2. \quad (21.49)$$

More efficient estimators of the variance components σ_{ε}^2 and σ_{α}^2 are possible (see, for example, Amemiya, 1985), but these will not necessarily increase the efficiency of $\widehat{\beta}_{\text{RE}}$. A wide range of estimators are possible. The variance estimator (21.49) can be negative, in which case programs often set $\widehat{\sigma}_{\alpha}^2 = 0$, so $\widehat{\lambda} = 0$ and estimation is then by pooled OLS.

To verify that the feasible GLS estimator simplifies to OLS estimation of (21.43), stack (21.42) by observations from all T time periods for given i in the same way as for the fixed effects model. Then

$$\mathbf{y}_i = \mathbf{W}_i \boldsymbol{\delta} + (\mathbf{e} \boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i), \quad (21.50)$$

where \mathbf{y}_i , \mathbf{e} , $\boldsymbol{\varepsilon}_i$, and \mathbf{X}_i are defined after (21.29), and $\mathbf{W}'_i = [\mathbf{e} \quad \mathbf{X}'_i]$. To estimate by GLS we need to obtain the variance matrix $\boldsymbol{\Omega}$ of the $T \times 1$ vector error $(\mathbf{e} \boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i)$. Given independence of $\boldsymbol{\alpha}_i$ and $\boldsymbol{\varepsilon}_{it}$ we have $E[(\mathbf{e} \boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i)(\mathbf{e} \boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i)'] = E[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}'_i] + E[\boldsymbol{\alpha}_i^2] \mathbf{e} \mathbf{e}'$. Since $\boldsymbol{\varepsilon}_{it}$ are iid $[0, \sigma_{\varepsilon}^2]$ and $\boldsymbol{\alpha}_i$ are iid $[0, \sigma_{\alpha}^2]$ we obtain

$$\boldsymbol{\Omega} = \sigma_{\varepsilon}^2 \mathbf{I}_T + \sigma_{\alpha}^2 \mathbf{e} \mathbf{e}' = \sigma_{\varepsilon}^2 \left[\mathbf{Q} + \frac{1}{\psi^2} (\mathbf{I}_T - \mathbf{Q}) \right],$$

where $\mathbf{Q} = \mathbf{I}_T - T^{-1} \mathbf{e} \mathbf{e}'$ was introduced in (21.30) and $\psi^2 = \sigma_{\varepsilon}^2 / [\sigma_{\varepsilon}^2 + T \sigma_{\alpha}^2]$. Using $\mathbf{Q} \mathbf{Q}' = \mathbf{Q}$ we can easily verify that $\boldsymbol{\Omega}^{-1} = \sigma_{\varepsilon}^{-2} [\mathbf{Q} + \psi^2 (\mathbf{I}_T - \mathbf{Q})]$ and

$$\boldsymbol{\Omega}^{-1/2} = \frac{1}{\sigma_{\varepsilon}} [\mathbf{Q} + \psi (\mathbf{I}_T - \mathbf{Q})]. \quad (21.51)$$

The GLS estimator is obtained by premultiplication of (21.50) by any scalar multiple of $\Omega^{-1/2}$. Now

$$\begin{aligned} [\mathbf{Q} + \psi(\mathbf{I}_T - \mathbf{Q})] \mathbf{y}_i &= \mathbf{y}_i - \mathbf{e}\bar{\mathbf{y}}_i' + \psi(\mathbf{y}_i - (\mathbf{y}_i - \mathbf{e}\bar{\mathbf{y}}_i')) \\ &= \mathbf{y}_i - \lambda\mathbf{e}\bar{\mathbf{y}}_i', \end{aligned}$$

where $\lambda = (1 - \psi)$. Performing similar algebra for \mathbf{W}_i , $\mathbf{e}\alpha_i$, and ε_i in (21.50) yields the following model:

$$\mathbf{y}_i - \lambda\mathbf{e}\bar{\mathbf{y}}_i' = (\mathbf{W}_i - \lambda\bar{\mathbf{W}})' \boldsymbol{\delta} + (1 - \lambda)\alpha_i + (\varepsilon_i - \lambda\mathbf{e}\bar{\varepsilon}_i'), \quad (21.52)$$

where the transformed error in (21.52) has variance matrix $\sigma_\varepsilon^2 \mathbf{I}_T$. The GLS estimator is the OLS estimator of (21.52), but (21.52) is just a stacked version of (21.43) with the scalar λ replaced by a consistent estimate.

The random effects estimator $\widehat{\beta}_{RE}$ of the slope parameters converges to the within estimator as $T \rightarrow \infty$ since then $\lambda \rightarrow 1$. Otherwise, $\widehat{\beta}_{RE}$ can be shown, after some algebra, to equal a **matrix-weighted combination** of the within estimator and the between estimator. If the random effects model is appropriate, this weighted average works better than using the within estimator alone. However, if the fixed effects model is appropriate then this weighted average is inconsistent, as the between estimator is then inconsistent. The estimator of the intercept can be shown to simplify to $\widehat{\mu}_{RE} = \bar{y} - \bar{\mathbf{X}}\widehat{\beta}_{RE}$. For more details see, for example, Hsiao (2003, p. 36) or Greene (2003).

21.7.2. ML Estimator

In the derivation in the previous section, normality of the errors is not assumed. If they are in fact **normal**, we can maximize the log-likelihood function with respect to $\boldsymbol{\beta}$, μ , σ_ε^2 , and σ_α^2 . For given σ_ε^2 and σ_α^2 the MLE for $\boldsymbol{\beta}$ and μ is the same as the GLS estimator, but the MLE gives estimators $\widetilde{\sigma}_\varepsilon^2$ and $\widetilde{\sigma}_\alpha^2$ that differ from those given in (21.48) and (21.49).

Thus the MLE for $\boldsymbol{\beta}$ and μ is given by (21.45) with $\widehat{\lambda}$ replaced by the alternative consistent estimate $\widetilde{\lambda} = 1 - \widetilde{\sigma}_\varepsilon^2 / (T\widetilde{\sigma}_\alpha^2 + \widetilde{\sigma}_\varepsilon^2)^{1/2}$. Asymptotically, the MLE and GLS estimators of the random effects model are equivalent, but the two will differ in finite samples.

For the MLE there may be two local maxima rather than one of the likelihood for $0 < \psi^2 \leq 1$, so care is needed to ensure a global maximum.

21.7.3. Other Estimators

Many different estimators of $\boldsymbol{\beta}$ are consistent if the random effects model is the correct model. In particular, the pooled OLS, within, first-differences, and between estimators are all consistent. However they are inefficient if α_i and ε_{it} are iid, and the within and first-differences estimators can only estimate the coefficients of time-varying regressors.

21.8. Modeling Issues

In this section we consider some practical issues that arise in linear panel data models, even in the absence of complications such as endogeneity and lagged dependent variables, topics that are deferred to Chapter 22.

21.8.1. Tests for Pooling

The random effects model restricts all regression parameters to be the same in different cross sections and time periods, whereas the fixed effects models imposes parameter constancy except for the intercept, which may vary across individuals. **Tests of poolability** test the appropriateness of these constraints.

These tests are usually done using a Chow test (see Greene, 2003, p. 130) based on the tests for equality of regressors in two linear regressions assuming a common variance. Depending on the assumptions about errors, the Chow test may be applied to models estimated by OLS or by GLS. Baltagi (2001, Chapter 4) and Hsiao (2003, Chapter 2) provide detailed coverage.

For short panels it is not possible to allow the slope parameters to differ across individuals, as then the number of parameters goes to infinity. However, parameters can be permitted to vary over time. The model $y_{it} = \gamma + \mathbf{x}'_{it}\beta + u_{it}$ is then tested against $y_{it} = \gamma_t + \mathbf{x}'_{it}\beta_t + u_{it}$. The most obvious method is to assume random effects with $u_{it} = \varepsilon_{it} + \alpha_i$, estimate the restricted model ($\gamma_t = \gamma$ and $\beta_t = \beta$) using the random effects GLS estimator, and compare the restricted and unrestricted residual sums of squares in the transformed models. If more robust inference is preferred then panel-robust standard errors should be obtained and a Wald test performed. For short panels it is common to specify models with slope parameters β constant, though the intercept γ_t may be permitted to vary over time by inclusion of time dummies as additional regressors.

21.8.2. Tests for Individual-Specific Effects

Breusch and Pagan (1980) derived Lagrange-multiplier tests for the presence of individual-specific random effects against the null hypothesis assumption of iid errors. These have the advantage of being easily implemented by an auxiliary regression that requires only residuals from pooled OLS estimates. Alternatively, one can assume normality and do a likelihood ratio test of the random effects MLE against the MLE of the constant-coefficients model, or a Wald test of $\sigma_\alpha = 0$ in the random effects model.

In practice one often rejects the null hypothesis that the errors in the constant-coefficients model are iid. It is easiest to immediately estimate by pooled OLS with panel-robust standard errors or by random effects GLS.

For a short panel formal tests for the presence of individual-specific fixed effects are not possible because of the incidental parameters problem. It is not possible to test whether N parameters are zero when there are only NT observations and T is small. Instead, the Hausman test of Section 21.4.3 is used to test the null hypothesis of random effects against the alternative of fixed effects.

21.8.3. Prediction

Prediction in models without individual effects is straightforward: Use $\hat{y}_{js} = \mathbf{x}'_{js}\hat{\beta}$. This is a prediction of the population average $E[y_{js}|\mathbf{x}_{js}]$.

Prediction for a given individual conditional on the individual-specific effect is more difficult. This is prediction of $E[y_{js}|\mathbf{x}_{js}, \alpha_i]$. We consider out-of-sample forecasts for the i th individual using the random effects model (21.42). Then $y_{i,t+s} = \mathbf{w}'_{it}\delta + u_{i,t+s}$, where $u_{i,t+s} = \alpha_i + \varepsilon_{i,t+s}$. The obvious predictor replaces δ by $\hat{\delta}_{RE}$ and $u_{i,t+s}$ by either 0 or \bar{u}_i , where $\bar{u}_i = \bar{y}_i - \bar{\mathbf{w}}'_i\hat{\delta}_{RE}$ is the average within-sample residual for the i th individual. However, this is inefficient as it ignores the correlation between $u_{i,t+s}$ and in-sample errors induced by the individual-specific random effect α_i . The problem is an example of the more general problem of prediction within a GLS rather than an OLS framework. For this special case the best linear unbiased predictor (see Section 22.8.3) is $\hat{y}_{i,t+s} = \mathbf{x}'_{it}\hat{\delta}_{RE} + (T\sigma_\alpha^2/(T\sigma_\alpha^2 + \sigma_\varepsilon^2))\bar{u}_i$. For the fixed effects model the obvious predictor is $\hat{y}_{i,t+s} = \mathbf{x}'_{it}\hat{\beta}_W + \hat{\alpha}_{i,FE}$, but again this is inconsistent in short panels.

21.8.4. Two-Way Effects Models

The analysis to date has focused on the one-way model, which is (21.1) with $u_{it} = \alpha_i + \varepsilon_{it}$. A more general model is the **two-way effects model**, with $u_{it} = \alpha_i + \gamma_t + \varepsilon_{it}$, which additionally allows for time-specific effects. Then

$$y_{it} = \alpha_i + \gamma_t + \mathbf{x}'_{it}\beta + \varepsilon_{it}, \quad i = 1, \dots, N, \quad t = 1, \dots, T. \quad (21.53)$$

This model was presented originally in (21.2).

As already noted, for short panels the usual approach is to treat the time-specific effects as fixed and estimate them as the coefficients of time dummies that are included in the regressors, with analysis then differing according to whether the individual-specific effects are treated as fixed or random.

If both α_i and γ_t are fixed then the OLS estimator of β in (21.53) is equivalent to regression of $y_{it} - \bar{y}_i - \bar{y}_t + \bar{\bar{y}}$ on $\mathbf{x}_{it} - \bar{\mathbf{x}}_i - \bar{\mathbf{x}}_t + \bar{\bar{\mathbf{x}}}$, where $\bar{y}_i = T^{-1} \sum_{t=1}^T y_{it}$, $\bar{y}_t = N^{-1} \sum_{i=1}^N y_{it}$, and $\bar{\bar{y}} = (NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T y_{it}$, with similar definitions for $\bar{\mathbf{x}}_i$, $\bar{\mathbf{x}}_t$, and $\bar{\bar{\mathbf{x}}}$. This method of estimation is convenient if T is large.

If instead both α_i and γ_t are random then the error term will have a component γ_t that induces error correlation across individuals, whereas we have focused on independence over i . It can be shown that the GLS estimator can be computed by OLS estimation of y_{it}^* on a constant and \mathbf{x}_{it}^* ,

$$y_{it}^* = y_{it} - \lambda_1 \bar{y}_i - \lambda_2 \bar{y}_t + \lambda_3 \bar{\bar{y}},$$

where \bar{y}_i , \bar{y}_t , and $\bar{\bar{y}}$ have already been defined and \mathbf{x}_{it}^* is defined analogously to y_{it}^* . For this and other results for the two-way effects model see Hsiao (2003) or Baltagi (2001).

21.8.5. Unbalanced Panel Data

The discussion thus far has assumed the panel is **balanced**, meaning that data are available for every individual in every year. For panel data on different regions this is often the case. In contrast, for panel surveys of individuals there is usually a drop off or **attrition** over time in the proportion of individuals still answering the survey. Moreover, some individuals may miss one or more periods but return later, in some cases by design as in the case of **rotating panels** such as the CPS, where households are surveyed for four consecutive months, not surveyed for eight months, and then surveyed for another four months. Such panels where different individuals appear in different years are called **unbalanced panels** or **incomplete panels**.

Let d_{it} be an indicator variable equal to one if the it th observation is observed and equal to zero otherwise. Then for the individual-specific effects model (21.3) the FE estimator is consistent if the strong exogeneity assumption (21.4) becomes

$$E[u_{it}|\alpha_i, \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}, d_{i1}, \dots, d_{iT}] = 0, \quad (21.54)$$

and the RE estimator is consistent if additionally α_i is independent of the other conditioning variables. The fixed and random effects estimators can then be applied to unbalanced data with relatively little adjustment. This should be clear from the initial presentation of the estimators as OLS estimators in various models given in Section 21.2.2. For example, for the random effects model replace $\widehat{\lambda}$ in (21.10) by $\widehat{\lambda}_i = 1 - \sigma_\varepsilon^2 / (T_i \sigma_\alpha^2 + \sigma_\varepsilon^2)^{1/2}$, where T_i is the number of observations for individual i (see Baltagi, 1985, and Wansbeek and Kapteyn, 1989). Davis (2002) considers multi-way random effects models. For the fixed effects model an individual observation must be observed at least twice in the sample and degrees of freedom must be appropriately adjusted. Baltagi (2001) gives a lengthy treatment of unbalanced panels. Econometrics packages that estimate the more standard of the panel models presented in Chapters 21–23 usually automatically handle missing observations.

At times it may be convenient to convert an unbalanced panel into a balanced panel, by including in the sample only those individuals with data available in all years. This obviously can greatly reduce efficiency because of the loss of many observations. Furthermore, if data are not randomly missing this can exacerbate potential problems of a nonrepresentative sample.

One reason for missing data can be that although most variables are observed, at least one variable is not. For example, the **nonresponse rate** to income questions can be quite high. Rather than drop an entire observation because data for one regressor, income, is missing there may be efficiency gains to using the imputation methods presented in Chapter 27.

Unbalanced panels require special methods if the reason for individuals dropping out of the sample is correlated with the error term, so that (21.54) does not hold. For example, those individuals with unusually low wages (after controlling for observed characteristics) may be more likely to drop out of a panel sample. The result is an unrepresentative panel that will lead to **attrition bias** if wage is the dependent variable. Consistent estimation requires use of sample selection methods extended to panel data (see Section 23.5.2).

21.8.6. Measurement Error

Measurement error in regressors leads to inconsistent parameter estimates in cross-section regression models. If panel data methods are used that involve differencing of the data, the result may be a large increase in the inconsistency caused by measurement error depending on the assumptions made about the dgp. This is pursued in Chapter 26.

21.9. Practical Considerations

The various estimators presented in this chapter are easily implemented. The most foolproof method is to use the panel commands available in econometric packages such as LIMDEP, STATA, and TSP, all of which have the added advantage of usually handling unbalanced panels. Most estimators can alternatively be estimated using an appropriate pooled OLS regression on transformed data that requires only a cross-section package, though standard errors may then differ from panel package standard errors because the latter may ignore autocorrelation induced by transformation and may use different degrees of freedom.

A weakness of panel commands in packages is that they currently compute standard errors based on restrictive distributional assumptions such as iid errors in the fixed effects models, and iid individual effect and iid errors in the random effects model. To compute the more robust standard error estimates presented in this chapter may require panel estimation with a panel bootstrap or estimation of an appropriate pooled OLS regression using an option to compute cluster-robust standard errors.

In microeconometric analysis there is a fundamental distinction between models with and models without fixed effects. If a model without fixed effects is preferred it should be justified by passing a Hausman test. If this test rejects the random effects model then it may still be possible to consistently estimate coefficients of time-invariant regressors using the instrumental variables methods presented in the next chapter.

21.10. Bibliographic Notes

Most textbooks, such as Greene's (2003), include at least a chapter on panel data models. Wooldridge (2002) has several chapters that cover both linear and nonlinear panel models. Econometrics monographs on panel data include those by Hsiao (1986, 2003), Baltagi (1995, 2001), Matyas and Sevestre (1995), M-J. Lee (2002), and Arellano (2003). The last three books place greater emphasis on the methods presented in Chapter 22 and 23. Diggle, Liang, and Zeger (1994, 2002) is a standard statistics reference.

- 21.4** Mundlak (1978) wrote a classic article on fixed versus random effects models. Hausman (1978) used tests between these two models to illustrate his testing approach.
- 21.6** Kuh (1959) and Hoch (1962) provide two early panel data applications to estimation of investment functions and of Cobb–Douglas production functions. These studies contrast use of within estimates using time-series variation and between estimates using cross-section variation.

Exercises

21–1 (Adapted from Baltagi, 1999) Consider the panel model $y_{it} = \alpha + \beta x_{it} + u_{it}$, where α and β are scalars.

- (a) Show by appropriate subtraction that this model implies

$$y_{it} - \bar{y} = \beta(x_{it} - \bar{x}_i) + \beta(\bar{x}_i - \bar{x}) + (u_{it} - \bar{u}),$$

where $\bar{y} = (NT)^{-1} \sum_{i,t} y_{it}$, $\bar{x} = (NT)^{-1} \sum_{i,t} x_{it}$ and $\bar{x}_i = T^{-1} \sum_t x_{it}$.

- (b) For the corresponding unrestricted least-squares regression

$$y_{it} - \bar{y} = \beta_1(x_{it} - \bar{x}_i) + \beta_2(\bar{x}_i - \bar{x}) + (u_{it} - \bar{u}),$$

show that the least-squares estimator of β_1 is the within estimator and that of β_2 is the between estimator.

- (c) Show that if $u_{it} = \mu_i + v_{it}$, where $\mu_i \sim \text{iid}[0, \sigma_\mu^2]$ and $v_{it} \sim \text{iid}[0, \sigma_v^2]$, and the two are mutually independent across both i and t , the OLS and the GLS estimators are equivalent.

21–2 Consider estimation of the fixed effects linear regression model $y_{it} = \alpha_i + \mathbf{x}'_{it}\beta + \varepsilon_{it}$, where α_i are fixed effects possibly correlated with \mathbf{x}_{it} . Stacking all T observations for individual i yields $\mathbf{y}_i = \alpha_i \mathbf{e} + \mathbf{X}_i \beta + \varepsilon_i$ (see (21.29) for definitions). Consider the estimator $\hat{\beta} = [\sum_{i=1}^N \mathbf{X}'_i \mathbf{J} \mathbf{X}_i]^ {-1} \times \sum_{i=1}^N \mathbf{X}'_i \mathbf{J} \mathbf{y}_i$, where \mathbf{J} is a $T \times T$ matrix of known constants such that $\mathbf{J}\mathbf{e} = \mathbf{0}$. [Note that an example of \mathbf{J} is $\mathbf{Q} = \mathbf{I}_T - T^{-1} \mathbf{e} \mathbf{e}'$.]

- (a) Provide a motivation for the estimator $\hat{\beta}$.
 (b) Find $E[\hat{\beta}]$. For simplicity assume that \mathbf{X}_i are fixed regressors and that ε_{it} are iid $[0, \sigma^2]$. Is $\hat{\beta}$ unbiased for β ?
 (c) Find $V[\hat{\beta}]$. For simplicity assume that \mathbf{X}_i are fixed regressors and that ε_{it} are iid $[0, \sigma^2]$.
 (d) Now suppose ε_{it} are independent over i but correlated over t with $V[\varepsilon_i] = \Omega_i$. Give $V[\hat{\beta}]$.
 (e) Suppose that the effects α_i are random $(0, \sigma_\alpha^2)$ rather than fixed. Would the estimator in this exercise be consistent?

21–3 (Adapted from Baltagi, 1998) Consider the fixed effects, two-way error component panel data model

$$y_{it} = \alpha + \mathbf{x}'_{it}\beta + \mu_i + \lambda_t + \varepsilon_{it},$$

where α is a scalar, \mathbf{x}_{it} is a $k \times 1$ vector of exogenous regressors, β is a $K \times 1$ vector, μ and λ denote fixed individual and time effects, respectively, and $\varepsilon_{it} \sim \text{iid}[0, \sigma^2]$.

- (a) Show that the within estimator of β , which is best linear unbiased, can be obtained by applying two within (one-way) transformations on this model. The first is the within transformation ignoring the time effects followed by the within transformation ignoring the individual effects.
 (b) Show that the order of these two within (one-way) transformations is unimportant. Give an intuitive explanation for this result.

21–4 Use a 50% random subsample of the wage-hours data in Section 21.3

- (a) Can β be directly interpreted as a labor supply elasticity? Explain.

- (b) For the following estimators: (1) pooled OLS, (2) between, (3) within, (4) first differences, (5) random effects GLS, (6) random effects MLE give (i) $\hat{\beta}$ (estimated coefficient of $\ln w_g$), (ii) default standard error, and (iii) panel bootstrap standard error with 200 replications.
- (c) Are the estimates of β similar?
- (d) Is there a systematic difference between default standard errors and panel-robust standard errors?
- (e) Will the pooled OLS estimator in part (b) be consistent for β in a fixed effects model? Will the pooled OLS estimator be consistent for β in a random effects model?
- (f) Perform a Hausman test of the difference between the fixed and random effects (GLS) estimates of β in this model. Do this manually using the earlier regression output with the default standard errors. What do you conclude and which model is favored?
- (g) Given the preceding evidence, do you believe that the labor supply curve is upward sloping? Explain.

Linear Panel Models: Extensions

22.1. Introduction

The previous chapter presented variants of the linear panel data model with a fixed or random intercept and regressors that are strongly exogenous. Now we move on to various extensions for linear models, with focus on relaxation of the strong exogeneity assumption to permit consistent estimation of models with endogenous variables and/or lagged dependent variables as regressors.

The use of instrumental variables is a standard method to handle endogenous regressors. It is much easier to obtain instruments with panel data than with cross-section data, since exogenous regressors in other time periods can be used as instruments for endogenous regressors in the current time period. The only complication is to first control for any fixed or random effects.

Panel data permit regressors to additionally include lagged dependent variables, data unavailable with a single cross section. This permits estimation of dynamic models that distinguish between persistence of earnings, for example, as the result of variation around an unobserved individual-specific effect, as in Chapter 21, and persistence caused by the outcomes of previous periods directly determining the outcome of the current period. The estimators of Chapter 21 that control for individual-specific effects become inconsistent, however, if lagged dependent variables are regressors. Instrumental variables estimation using longer lags as instruments leads to consistent estimation.

Panel data provide an excess of moment conditions available for estimation, owing to an abundance of instruments, and panel model errors are usually not iid. The natural estimation framework is that of panel GMM, presented in detail in Section 22.2 and illustrated with an application to estimation of the labor supply elasticity in Section 22.3. Further details on estimation with individual-specific effects and regressors that are endogenous or lagged dependent variables are presented in Sections 22.4 and 22.5. The discussion is quite extensive due to the many possible variations that are covered. These include the presence of individual specific effects that may be fixed or

random, different exogeneity assumptions, and models that may be just-identified or over-identified.

The remainder of this chapter considers other stand-alone topics that generally do not require reading of Sections 22.2–22.5. Models closely related to panel data models are presented in Sections 22.6–22.8, namely repeated cross-section data, differences in differences, and hierarchical models.

22.2. GMM Estimation of Linear Panel Models

The panel regression models in Chapter 21 restricted the scalar dependent variable y_{it} to depend on just the contemporaneous value of regressors \mathbf{x}_{it} , even though potentially all of $\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}$ could be regressors under the Chapter 21 assumption of strong exogeneity. This introduces the possibility of more efficient estimation using excluded regressors from other periods as instruments in the current period.

Furthermore, regressors in other periods may be valid instruments for current-period regressors that are either endogenous or lags of the dependent variable. So instruments are readily available to permit consistent IV estimation in situations where failure of the strong exogeneity assumption leads to inconsistency of the Chapter 21 estimators.

This section provides a general presentation of panel GMM estimation, a very useful framework for **panel IV estimation** that is used extensively throughout Sections 22.2–22.5. Then we introduce the use of exogenous variables (regressors or instruments) in periods other than the current period as an instrument. Once this groundwork is laid it is a relatively minor adaptation to incorporate fixed or random effects, typically included in panel models. This is deferred to subsequent sections.

22.2.1. Panel GMM

Consider the **linear panel model**

$$y_{it} = \mathbf{x}'_{it}\beta + u_{it}, \quad (22.1)$$

where the regressors \mathbf{x}_{it} may have both time-varying and time-invariant components and may include an intercept. Here there is **no individual-specific effect** α_i , an assumption relaxed from Section 22.3 on, and \mathbf{x}_{it} is assumed to include only current-period variables, an assumption relaxed in Section 22.5. Observations are assumed to be independent over i and a **short panel** with T fixed and $N \rightarrow \infty$ is assumed.

Begin by stacking all T observations for the i th individual,

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{u}_i, \quad (22.2)$$

where \mathbf{y}_i and \mathbf{u}_i are $T \times 1$ vectors and \mathbf{X}_i is a $T \times K$ matrix with t th row \mathbf{x}'_{it} , so

$$\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{bmatrix}; \quad \mathbf{X}_i = \begin{bmatrix} \mathbf{x}'_{i1} \\ \vdots \\ \mathbf{x}'_{iT} \end{bmatrix}; \quad \mathbf{u}_i = \begin{bmatrix} u_{i1} \\ \vdots \\ u_{iT} \end{bmatrix}.$$

The model (22.2) defines a linear system of equations, so the results of Section 6.9.5 for systems IV estimation with data independent over i are directly applicable.

Assume the existence of a $T \times r$ matrix of **instruments** \mathbf{Z}_i , where $r \geq K$ is the number of **instruments**, that satisfy the r moment conditions

$$E[\mathbf{Z}'_i \mathbf{u}_i] = \mathbf{0}. \quad (22.3)$$

The GMM estimator based on these moment conditions minimizes the associated quadratic form

$$Q_N(\boldsymbol{\beta}) = \left[\sum_{i=1}^N \mathbf{Z}'_i \mathbf{u}_i \right]' \mathbf{W}_N \left[\sum_{i=1}^N \mathbf{Z}'_i \mathbf{u}_i \right],$$

where \mathbf{W}_N denotes an $r \times r$ weighting matrix. Given $\mathbf{u}_i = \mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}$, some algebra yields the **panel GMM estimator**

$$\hat{\boldsymbol{\beta}}_{\text{PGMM}} = \left[\left(\sum_{i=1}^N \mathbf{X}'_i \mathbf{Z}_i \right) \mathbf{W}_N \left(\sum_{i=1}^N \mathbf{Z}'_i \mathbf{X}_i \right) \right]^{-1} \left(\sum_{i=1}^N \mathbf{X}'_i \mathbf{Z}_i \right) \mathbf{W}_N \left(\sum_{i=1}^N \mathbf{Z}'_i \mathbf{y}_i \right).$$

The essential condition for consistency of this estimator is assumption (22.3).

In many applications \mathbf{Z}_i is composed of current and lagged values of exogenous regressors. For example, suppose all regressors are contemporaneously exogenous. Then $E[\mathbf{x}_{it} \mathbf{u}_{it}] = \mathbf{0}$ implies (22.3) with $\mathbf{Z}_i = [\mathbf{x}'_{i1} \dots \mathbf{x}'_{iT}]$. In this case the model is **just identified** and, since $\mathbf{Z}_i = \mathbf{X}_i$, $\hat{\boldsymbol{\beta}}_{\text{PGMM}}$ simplifies to the pooled OLS estimator of Chapter 21. If it is additionally assumed that $E[\mathbf{x}_{it-1} \mathbf{u}_{it}] = \mathbf{0}$, then \mathbf{x}_{it-1} is available as additional instruments for the it th observation, the model is **over-identified**, and more efficient estimation is possible using the PGMM estimator.

The use of **various exogeneity assumptions** to form the instrument matrix \mathbf{Z}_i is detailed in Section 22.2.4. The analysis requires adaptation in panel data models with individual-specific effects α_i . This is illustrated in an empirical application in Section 22.3 and is dealt with explicitly in Sections 22.4 and 22.5.

22.2.2. Panel-Robust Statistical Inference

To express the distribution of the panel GMM estimator it is convenient to use more compact notation. Rewrite

$$\hat{\boldsymbol{\beta}}_{\text{PGMM}} = [\mathbf{X}' \mathbf{Z} \mathbf{W}_N \mathbf{Z}' \mathbf{X}]^{-1} \mathbf{X}' \mathbf{Z} \mathbf{W}_N \mathbf{Z}' \mathbf{y}, \quad (22.4)$$

where $\mathbf{X}' = [\mathbf{X}'_1 \cdots \mathbf{X}'_N]$, $\mathbf{Z}' = [\mathbf{Z}'_1 \cdots \mathbf{Z}'_N]$, and $\mathbf{y}' = [\mathbf{y}'_1 \cdots \mathbf{y}'_N]$. Then $\hat{\boldsymbol{\beta}}_{\text{PGMM}}$ is **asymptotically normal** with estimated asymptotic variance matrix

$$\widehat{V}[\hat{\boldsymbol{\beta}}_{\text{PGMM}}] = [\mathbf{X}' \mathbf{Z} \mathbf{W}_N \mathbf{Z}' \mathbf{X}]^{-1} \mathbf{X}' \mathbf{Z} \mathbf{W}_N (N \widehat{\mathbf{S}}) \mathbf{W}'_N \mathbf{Z}' \mathbf{X} [\mathbf{X}' \mathbf{Z} \mathbf{W}_N \mathbf{Z}' \mathbf{X}]^{-1}, \quad (22.5)$$

see Equation (6.97), where $\widehat{\mathbf{S}}$ is a consistent estimate of the $r \times r$ matrix

$$\mathbf{S} = \text{plim} \frac{1}{N} \sum_{i=1}^N \mathbf{Z}'_i \mathbf{u}_i \mathbf{u}'_i \mathbf{Z}_i, \quad (22.6)$$

and independence over i has been assumed. The essential assumption for this result is that $N^{-1/2}\mathbf{Z}'\mathbf{u} = N^{-1/2} \sum_i \mathbf{Z}'_i \mathbf{u}_i \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{S}]$. A White-type robust estimate of \mathbf{S} is

$$\widehat{\mathbf{S}} = \frac{1}{N} \sum_{i=1}^N \mathbf{Z}'_i \widehat{\mathbf{u}}_i \widehat{\mathbf{u}}'_i \mathbf{Z}_i, \quad (22.7)$$

where the $T \times 1$ estimated residual $\widehat{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{X}_i \widehat{\beta}$.

The estimate (22.5) yields **panel-robust standard** errors allowing for both heteroskedasticity and correlation over time. Alternatively, the **panel bootstrap** could be used. For further discussion see Section 21.2.3 where the same issues apply.

22.2.3. One-Step and Two-Step Panel GMM

Different full-rank weighting matrices \mathbf{W}_N in (22.4) lead to different systems GMM estimators, except in the just-identified case of $r = K$ when the PGMM estimator simplifies to the IV estimator $[\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{Z}'\mathbf{y}$ for any \mathbf{W}_N . The discussion mirrors that in Section 6.4.2. The two leading choices of \mathbf{W}_N are given here.

One-Step GMM

The **one-step GMM** or **two-stage least-squares estimator** uses weighting matrix $\mathbf{W}_N = [\sum_i \mathbf{Z}'_i \mathbf{Z}_i]^{-1} = [\mathbf{Z}'\mathbf{Z}]^{-1}$, leading to

$$\widehat{\beta}_{2SLS} = [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}. \quad (22.8)$$

The motivation for this estimator is that it can be shown to be the optimal PGMM estimator based on (22.3) if $\mathbf{u}_i | \mathbf{Z}_i$ is iid $[\mathbf{0}, \sigma^2 \mathbf{I}_T]$.

This estimator is called one-step GMM because given the data it can be directly calculated using Equation (22.8). It is called 2SLS as it can instead be obtained in two stages by (1) OLS of \mathbf{X}_i on \mathbf{Z}_i , yielding prediction $\widehat{\mathbf{X}}_i$, and (2) OLS of \mathbf{y}_i on $\widehat{\mathbf{X}}_i$. An estimate of the variance matrix of $\widehat{\beta}_{2SLS}$ that is both panel and heteroskedasticity robust is that given in (22.5) with $\mathbf{W}_N = [\mathbf{Z}'\mathbf{Z}]^{-1}$.

Two-Step GMM

The most efficient GMM estimator based on the unconditional moment condition (22.3) uses weighting matrix $\mathbf{W}_N = \widehat{\mathbf{S}}^{-1}$, where $\widehat{\mathbf{S}}$ is consistent for \mathbf{S} defined in (22.6); see Section 6.4.2 for the general result. Using $\widehat{\mathbf{S}}$ in (22.7) yields the **two-step GMM estimator**

$$\widehat{\beta}_{2SGMM} = [\mathbf{X}'\widehat{\mathbf{S}}^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\widehat{\mathbf{S}}^{-1}\mathbf{Z}'\mathbf{y}. \quad (22.9)$$

Then (22.5) simplifies and $\widehat{\mathbf{V}}[\widehat{\beta}_{2SGMM}] = [\mathbf{X}'\mathbf{Z}(N\widehat{\mathbf{S}})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}$.

This is called **two-step GMM** since a first-step consistent estimator of β such as $\widehat{\beta}_{2SLS}$ is needed to form the residuals $\widehat{\mathbf{u}}_i$ used to compute $\widehat{\mathbf{S}}$.

Efficiency Gains

In this chapter the focus is on situations where \mathbf{Z} cannot contain all of \mathbf{X} , because of endogeneity of some components of \mathbf{X} . Then panel GMM provides consistent estimates when OLS does not. Two-step GMM provides the most efficient estimator based on the moment condition $E[\mathbf{Z}'_i \mathbf{u}_i] = \mathbf{0}$.

Even if regressors are strongly exogenous, two-step GMM has the attraction of being **more efficient** than pooled OLS. To see this, suppose that \mathbf{X} is strongly exogenous. Setting $\mathbf{Z} = \mathbf{X}$, the two-step GMM estimator simplifies to $[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{y}$ and there is no benefit to panel GMM. However, if instead \mathbf{Z} equals \mathbf{X} as well as some additional variables, such as powers of the regressors or regressor values in periods other than the current period, then the two-step GMM method is at least as efficient as OLS, with equality applying if the errors u_{it} are iid.

Even more efficient estimators than $\widehat{\beta}_{2SGMM}$ are possible, by widening the definition of \mathbf{Z}_i , by using the optimal moment condition based on $E[\mathbf{u}_i | \mathbf{Z}_i] = \mathbf{0}$, which need not be $E[\mathbf{Z}'_i \mathbf{u}_i] = \mathbf{0}$ (see Section 22.4.3), and by using additional moment restrictions. We shy away from calling two-step GGM the **optimum GMM** estimator, as in Section 6.3, because it is only optimal given (22.3).

Tests of Overidentifying Restrictions

If there are r instruments and only K parameters to estimate, then panel GMM estimations leaves $(r - K)$ overidentifying restrictions. From Section 6.3.8 this permits a **test of overidentifying restrictions**

$$\text{OIR} = \left[\sum_{i=1}^N \widehat{\mathbf{u}}'_i \mathbf{Z}_i \right] (N\widehat{\mathbf{S}})^{-1} \left[\sum_{i=1}^N \mathbf{Z}'_i \widehat{\mathbf{u}}_i \right], \quad (22.10)$$

where $\widehat{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{Z}'_i \widehat{\beta}_{2SGMM}$, $\widehat{\mathbf{S}}$ is given in (22.7), and independence over i is assumed but heteroskedasticity and correlation over t for given i is permitted. Note that $\widehat{\beta}_{2SGMM}$ must be used, not $\widehat{\beta}_{2SLS}$.

This test statistic is distributed as $\chi^2(r - K)$ under the null hypothesis that the overidentifying restrictions are valid. If OIR is large then the overidentifying moment conditions are rejected and we conclude that some of the instruments in \mathbf{Z}_i are correlated with the error and hence are endogenous.

22.2.4. Selection of Instruments

The discussion so far has assumed the existence of a $T \times r$ matrix of instruments \mathbf{Z}_i that satisfies (22.3). Now we provide a lengthy discussion of how to obtain instruments in a panel setting.

In cross-section models, endogenous variables are instrumented by variables that do not appear as regressors in the equation of interest. Such variables can also be used as instruments in the panel case. With panel models, however, the additional periods of data provide additional moment conditions and additional instruments that can easily lead to identification or overidentification of β .

The number of moment conditions and instruments available expands as progressively stronger assumptions are made about the correlation between u_{it} and \mathbf{z}_{is} , $s, t = 1, \dots, T$. We consider the effect of progressively stronger **exogeneity assumptions**, see Section 2.3, following M.-J. Lee (2002). The emphasis is on using exogenous components of the regressors as instruments more than once, but the technique also applies to more traditional instruments that are variables excluded from the regression (22.1).

Summation Assumption

An obvious procedure is to define \mathbf{Z}_i similarly to \mathbf{X}_i . Then

$$\mathbf{Z}_i = \begin{bmatrix} \mathbf{z}'_{i1} \\ \mathbf{z}'_{i2} \\ \vdots \\ \mathbf{z}'_{iT} \end{bmatrix}, \quad \mathbf{u}_i = \begin{bmatrix} u_{i1} \\ u_{i2} \\ \vdots \\ u_{iT} \end{bmatrix}, \quad (22.11)$$

where \mathbf{z}_{it} is $r \times 1$ and $E[\mathbf{Z}'_i \mathbf{u}_i] = \mathbf{0}$ if the **summation assumption**

$$E \left[\sum_{t=1}^T \mathbf{z}_{it} u_{it} \right] = \mathbf{0} \quad (22.12)$$

is satisfied.

This assumption corresponds to that used in pooled OLS regression of y_{it} on \mathbf{x}_{it} , since if $\mathbf{z}_{it} = \mathbf{x}_{it}$ in (22.12) then the PGMM estimator defined in (22.4) simplifies to $(\sum_i \mathbf{Z}'_i \mathbf{X}_i)^{-1} \sum_i \mathbf{Z}'_i \mathbf{y}_i$.

For this estimator to be feasible requires at least that the order condition be met, so that $r \geq K$. Under the summation assumption it is just as difficult to find instruments with panel data as it is with cross-section data.

Contemporaneous Exogeneity Assumption

A stronger and more natural assumption is the **contemporaneous exogeneity assumption** that

$$E[\mathbf{z}_{it} u_{it}] = \mathbf{0}, \quad t = 1, \dots, T, \quad (22.13)$$

so that the instruments are assumed to be contemporaneously uncorrelated with the error term.

This presents many more moment conditions, as in principle there are as many as Tr moment conditions, where $r = \dim[\mathbf{z}_{it}]$. To use these we define

$$\mathbf{Z}_i = \begin{bmatrix} \mathbf{z}'_{i1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{z}'_{i2} & & \vdots \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{z}'_{iT} \end{bmatrix}, \quad \mathbf{u}_i = \begin{bmatrix} u_{i1} \\ u_{i2} \\ \vdots \\ u_{iT} \end{bmatrix}, \quad (22.14)$$

where \mathbf{Z}_i is now $Tr \times T$. The moment condition (22.3) holds, since $E[\mathbf{Z}'_i \mathbf{u}_i] = \mathbf{0}$ by (22.13), but now (22.3) defines Tr moment conditions that can be used to estimate the K components of β .

This remarkable result of an apparent surfeit of moment restrictions comes about because of the implicit assumption that β is time-invariant, so that each additional time period offers additional moment restrictions.

The number of additional moment restrictions is reduced to the extent that β is time varying. In particular, the intercept is often permitted to vary over time by inclusion in \mathbf{x}_{it} of $(T - 1)$ time dummies $d_{s,it} = 1$ if $t = s$ and 0 otherwise, for $s = 2, \dots, T$. Then the condition $E[d_{s,it} u_{it}] = 0$ cannot be used as it duplicates the condition $E[1 \times u_{it}] = 0$ implied by inclusion of an intercept in \mathbf{x}_{it} . In the preceding example, if \mathbf{x}_{1it} includes time dummies then there are only $TK - (T - 1)$ moment conditions available. Any time-invariant regressors can be used only once as an instrument.

Weak Exogeneity Assumption

Moment condition (22.13) considers only contemporaneous correlation between instruments and regressors. A stronger assumption is the **weak exogeneity assumption** or **predetermined instruments assumption** that additionally lagged values of the instruments are uncorrelated with the current-period error, so that

$$E[\mathbf{z}_{is} u_{it}] = \mathbf{0}, \quad s \leq t, \quad t = 1, \dots, T. \quad (22.15)$$

Condition (22.15) permits $\mathbf{z}_{i1}, \dots, \mathbf{z}_{it}$ to be instruments for u_{it} , though future values of \mathbf{z}_{is} cannot be so used. The instrument \mathbf{Z}_i is structured similarly to (22.14), except that \mathbf{z}'_{it} is replaced by the expanded instrument vector $[\mathbf{z}'_{i1}, \dots, \mathbf{z}'_{it}]$ that increases in size as t increases.

Conditions of this sort arise in rational expectations models and in models of intertemporal decision making under uncertainty that lead to **Euler conditions** of the form $E[u_{it} | \mathcal{I}_{it}] = 0$, where \mathcal{I}_{it} is the information set available at time t and an example of u_{it} is given in Section 6.2.7. If the information set includes current and past values of \mathbf{z}_{it} then $E[u_{it} | \mathbf{z}_{is}] = 0$, $s \leq t$, leading to (22.15).

More generally these conditions become relevant in dynamic models with lagged dependent variables as regressors (see Section 22.5). In some instances contemporaneous correlation is not ruled out, so that the inequality $s \leq t$ in (22.15) is replaced by $s < t$.

Note that time-invariant instruments can only be used once. Thus if $\mathbf{z}_{it} = [\mathbf{z}_{1i} \ \mathbf{z}_{2it}]$, then \mathbf{z}_{1i} and $\mathbf{z}_{2i1}, \dots, \mathbf{z}_{2it}$ are available as instruments.

Strong Exogeneity Assumption

A stronger assumption than weak exogeneity is the **strong exogeneity assumption** that future values of instruments are also uncorrelated with the current period error, so that

$$E[\mathbf{z}_{is} u_{it}] = \mathbf{0}, \quad s, t = 1, \dots, T. \quad (22.16)$$

Then current, past, and future values of \mathbf{z}_{is} are valid instruments for u_{it} .

This assumption was maintained for the regressors \mathbf{x}_{it} throughout Chapter 21, since $E[u_{it}|\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}] = 0$ implies $E[u_{it}|\mathbf{x}_{is}] = 0$, $1 \leq s \leq T$, and hence $E[\mathbf{x}_{is}u_{it}] = \mathbf{0}$. It may be appropriate for static models, but for dynamic models at most weak exogeneity of instruments can be assumed.

Condition (22.16) permits $\mathbf{z}_{i1}, \dots, \mathbf{z}_{iT}$ to be instruments for u_{it} . The instrument \mathbf{Z}_i is structured similarly to (22.14), except that \mathbf{z}'_{it} in (22.14) is replaced by the expanded instrument vector $[\mathbf{z}'_{i1}, \dots, \mathbf{z}'_{iT}]$.

As for the weak exogeneity case, time-invariant instruments can be used only once. If $\mathbf{z}_{it} = [\mathbf{z}_{1i} \ \mathbf{z}_{2ii}]$ then $T(r_{\text{TI}} + Tr_{\text{TV}})$ moment conditions are available, where r_{TI} and r_{TV} denote the numbers of time-invariant and time-varying instruments.

The extraordinary number of moment conditions, as many as rT^2 , is due to exclusion restrictions implicitly made in the panel model (22.1). For simplicity suppose all components of \mathbf{x}_{it} are strongly exogenous and we wish to use these as instruments whenever possible. In general y_{it} could depend on the regressors in all time periods, $\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}$. In contrast, the panel model $y_{it} = \mathbf{x}'_{it}\beta + u_{it}$ with $E[\mathbf{x}_{it}u_{it}] = \mathbf{0}$ excludes all but \mathbf{x}_{it} from the model for y_{it} . The strong exogeneity assumption that $E[\mathbf{x}_{is}u_{it}] = \mathbf{0}$ then permits the excluded regressors \mathbf{x}_{is} , $s \neq t$, to be used as instruments in addition to \mathbf{x}_{it} .

Redundant Instruments

If \mathbf{z}_{it} is varying over both i and t then its lags and leads can also be used as an instrument, depending on the exogeneity assumptions made. For the it th observation the available instruments are \mathbf{z}_{it} under contemporaneous exogeneity, $\mathbf{z}_{i1}, \dots, \mathbf{z}_{it}$ under weak exogeneity, and $\mathbf{z}_{i1}, \dots, \mathbf{z}_{iT}$ under strong exogeneity. This makes identification possible using only exogenous regressors as instruments. Only under the summation assumption are the difficulties of finding valid instruments comparable to those in the cross-section case.

In practice, however, there are not as many available instruments as the preceding discussion suggests. **Time-invariant instruments** $\mathbf{z}_{it} = \mathbf{z}_i$ can be used only once, since then $\mathbf{z}_{it} = \mathbf{z}_{is}$ for all s and t . For example, this is the case for an intercept or for a race or gender indicator. If the instrument is a regressor and lagged values of the regressor appear in the model then the number of available instruments is reduced. Time-varying instruments that vary in a systematic way may also not be available in all periods. Thus instruments that are the product of time dummies and a time-invariant regressor should be included only once if a complete set of time dummies is used. Examples include time dummies and time dummies interacted with race or gender indicators. Instruments that are a linear function of time should be used only once. For example, if year is an instrument then lagged years should not also be used. This comment does not apply to age, which increases linearly for each individual but varies across individuals.

It is clearly easy to inadvertently use **redundant instruments**. The panel GMM estimators are still feasible and the usual results are valid if there are still sufficient nonredundant instruments. For example, if r instruments are used and two of these are redundant the model is still estimable provided $r \geq K + 2$ as $\mathbf{Z}'\mathbf{X}$ is still of full

rank K . Singularity problems in GMM estimation may arise if too many redundant instruments are used, leading to an underidentified model. Even if the model remains overidentified, the degrees of freedom in a test of overidentifying restrictions will be reduced if some instruments are redundant.

Weak Instruments

Weak instruments, not to be confused with weak exogeneity, were introduced in Section 4.9. There is no well-established formal test of **weak instruments**. Standard R^2 and F -statistic diagnostics are given in Section 4.9. It is the incremental explanatory power of the instruments that matters. So a partial R^2 that controls for exogenous regressors that are also in the instrument set should be used. Moreover, whereas the endogenous regressor is regressed on all instruments, the F -statistic should be one of the overall significance of the subset of the instruments that are not exogenous regressors.

Since the errors here are not iid, the F -statistic should be based on panel robust standard errors. It can be calculated as W/r^* , where W is the Wald chi-square test statistic for exclusion restrictions given in Section 7.2.7 and r^* is the number of instruments that are not regressors in the original model.

22.2.5. Computation of Panel GMM Estimators

The moment conditions discussed in the preceding section provide the instrument matrix \mathbf{Z}_i . Then, given \mathbf{Z}_i , one can estimate β by $\hat{\beta}_{2SLS}$ defined in (22.8) or by $\hat{\beta}_{2SGMM}$ defined in (22.9).

The 2SLS estimator is easier to implement than the two-step GMM. Consider estimation under the summation assumption, in which case \mathbf{Z}_i is defined in (22.11). Then $\hat{\beta}_{2SLS}$ is given in (22.8), where $\mathbf{Z}'\mathbf{X} = \sum_i \mathbf{Z}'_i \mathbf{X}_i = \sum_i \sum_t \mathbf{z}'_{it} \mathbf{x}'_{it}$ and similar algebra applies for the other cross-products. This yields the standard textbook formula for 2SLS, except that summation is over both i and t . Thus $\hat{\beta}_{2SLS}$ can be obtained by 2SLS regression of y_{it} on \mathbf{x}_{it} using a cross-section 2SLS package. **Panel-robust** standard errors can then be obtained using a cluster-robust option that permits clustering on i , or by a **panel bootstrap** that resamples over i rather than both i and t . The approaches are similar to those for pooled LS given in Section 21.2.3, which provides additional detail.

For assumptions other than the summation assumption one can still use a cross-section 2SLS package by appropriately defining the instrument matrix \mathbf{Z}_i , which then has a more complicated form. For the contemporaneous exogeneity assumption, \mathbf{Z}_i is defined in (22.14). This is in the same form as (22.11) if the t th row in (22.11), \mathbf{z}'_{it} , is replaced by

$$[\mathbf{0}'_{r_1} \cdots \mathbf{0}'_{r_{t-1}} \mathbf{z}'_{it} \mathbf{0}'_{r_{t+1}} \cdots \mathbf{0}'_{r_T}], \quad (22.17)$$

where $r_s = \dim[\mathbf{z}_{is}]$ and $\mathbf{0}_{r_s}$ denotes an $r_s \times 1$ vector of zeros. Similarly, for the weak exogeneity assumption, \mathbf{Z}_i is as in (22.11) with the t th row in (22.11), \mathbf{z}'_{it} , replaced by

$$[\mathbf{0}'_{r_1} \cdots \mathbf{0}'_{r_{t-1}} (\mathbf{z}'_{it})' \mathbf{0}'_{r_{t+1}} \cdots \mathbf{0}'_{r_T}], \quad (22.18)$$

where $(\mathbf{z}'_{it})' = [\mathbf{z}'_{i1} \dots \mathbf{z}'_{it}]$ and $r_s = \dim[\mathbf{z}'_{is}]$, and for the strong exogeneity assumption, \mathbf{Z}_i is as in (22.11) with the t th row in (22.11), \mathbf{z}'_{it} , replaced by

$$[\mathbf{0}'_{r_1} \dots \mathbf{0}'_{r_{t-1}} (\mathbf{z}'_{it})' \mathbf{0}'_{r_{t+1}} \dots \mathbf{0}'_{r_T}], \quad (22.19)$$

where $(\mathbf{z}'_{it})' = [\mathbf{z}'_{i1} \dots \mathbf{z}'_{iT}]$ and $r_s = \dim[\mathbf{z}'_{is}]$. A practical example of generating the instruments is given in Section 22.3.

In practice there can be too many moment conditions. For example, with 10 periods of data and 5 time-varying regressors the strong exogeneity assumption yields as many as $5 \times 10^2 = 500$ moment conditions (and the preceding row vector has 500 entries) with only 5 parameters to estimate. The marginal value of an instrument may be very slight, because of increasing multicollinearity among the instruments, leading to a situation of weak instruments. Good practice is to treat time-varying instruments that vary little over time as time-invariant. For example, use only the data for the first period as an instrument. Even instruments that vary considerably over time might be used for only a few periods rather than in all possible periods.

Computation of the more efficient $\widehat{\beta}_{2SGMM}$ is not possible using only a 2SLS package. Instead, either more specialized software is needed or the estimator needs to be programmed using a matrix language algorithm.

Table 22.1 provides a summary of the four exogeneity assumptions and the resulting valid instruments.

22.2.6. Variations on GMM Estimation

Although $\widehat{\theta}_{2SGMM}$ is more efficient than $\widehat{\theta}_{2SLS}$, several studies find it to have greater finite-sample bias than $\widehat{\theta}_{2SLS}$, especially when r is much greater than K . For explanation see the discussion of finite-sample bias of optimal GMM in Section 6.3.5.

One approach is to be judicious in the use of instruments, though then potential efficiency gains due to additional instruments are lost.

Several authors have proposed alternative GMM estimators that may be less likely to be biased in finite samples. Many of these are presented in Section 6.4.4 and are used in the panel study by Ziliak (1997).

Table 22.1. Panel Exogeneity Assumptions and Resulting Instruments

Exogeneity Assumption	Moment Condition	Instrument Vector ^a
Summation	$E[\sum_t \mathbf{z}_{it} u_{it}] = \mathbf{0}$	$[\mathbf{z}_{it}]$
Contemporaneous	$E[\mathbf{z}_{it} u_{it}] = \mathbf{0}$, all t	$[\mathbf{0}'_{r_1} \dots \mathbf{0}'_{r_{t-1}} \mathbf{z}'_{it} \mathbf{0}'_{r_{t+1}} \dots \mathbf{0}'_{r_T}]$
Weak	$E[\mathbf{z}_{is} u_{it}] = \mathbf{0}$, $s \leq t$, all t	$[\mathbf{0}'_{r_1} \dots \mathbf{0}'_{r_{t-1}} (\mathbf{z}'_{it})' \mathbf{0}'_{r_{t+1}} \dots \mathbf{0}'_{r_T}]$
Strong	$E[\mathbf{z}_{is} u_{it}] = \mathbf{0}$, all s and t	$[\mathbf{0}'_{r_1} \dots \mathbf{0}'_{r_{t-1}} (\mathbf{z}'_{it})' \mathbf{0}'_{r_{t+1}} \dots \mathbf{0}'_{r_T}]$

^a The instrument vector is the t th row of \mathbf{Z}_i in (22.11); $(\mathbf{z}'_{it})' = [\mathbf{z}'_{i1} \dots \mathbf{z}'_{it}]$, $(\mathbf{z}'_{it})' = [\mathbf{z}'_{i1} \dots \mathbf{z}'_{iT}]$; and $r_s = \dim[\mathbf{z}'_{is}]$ or $\dim[\mathbf{z}'_{is}]$ or $\dim[\mathbf{z}'_{is}]$.

22.2.7. Chamberlain's Optimal Distance Estimator

Consider estimation of the individual-specific effects model

$$y_{it} = \alpha_i + \mathbf{x}'_{it}\beta + u_{it}, \quad (22.20)$$

when regressors are strongly exogenous as in Chapter 21. In Sections 21.2.3 and 21.6.1 methods to obtain panel-robust standard errors for the within estimator were presented.

If panel-robust inference is warranted, because ε_{it} are not iid, then the estimators detailed in Chapter 21 are actually inefficient. More efficient estimation is possible using optimal GMM applied to an overidentified model. Here \mathbf{x}_{is} , $s \neq t$, are available as additional instruments and GMM can be applied to a transformed model if elimination of α_i is necessary (see Section 22.4.2). The efficiency improvement is analogous to that for cross-section data with heteroskedasticity (see Section 6.3.5).

Chamberlain (1982, 1984) proposed the following more efficient estimator. The model (22.20) can be stacked to yield

$$\mathbf{y}_i = \mathbf{e}\alpha_i + (\mathbf{I}_T \otimes \beta')\mathbf{x}_i + \mathbf{u}_i, \quad (22.21)$$

where $\mathbf{e} = (1, 1, \dots, 1)'$ is a $T \times 1$ vector of ones, $\mathbf{x}_i = [\mathbf{x}'_{i1} \dots \mathbf{x}'_{iT}]$ is a $TK \times 1$ vector, and \mathbf{y}_i and \mathbf{u}_i are $T \times 1$ vectors. Equation (22.21) makes clear the restrictions that are implicitly made in static models that specify that y_{it} depends only on contemporaneous \mathbf{x}_{it} . Chamberlain used linear projection arguments that rely on weaker assumptions than those of conditional expectation. Let

$$E^*[\alpha_i | \mathbf{x}_i] = \mu + \sum_t \lambda'_t \mathbf{x}_{it} = \mu + \lambda' \mathbf{x}_i,$$

where E^* denotes linear projection. Given $E[\mathbf{u}_i | \alpha_i, \mathbf{x}_i] = \mathbf{0}$, (22.21) implies

$$E^*[\mathbf{y}_i | \mathbf{x}_i] = \mathbf{e}\mu + (\mathbf{I}_T \otimes \beta' + \mathbf{e}\lambda')\mathbf{x}_i.$$

This imposes restrictions on the unrestricted linear projection $E^*[\mathbf{y}_i | \mathbf{x}_i] = \pi_0 + \boldsymbol{\pi}' \mathbf{x}_i$, specifically that $\boldsymbol{\pi} - \mathbf{I}_T \otimes \beta' + \mathbf{e}\lambda' = \mathbf{0}$.

Rather than use GMM, Chamberlain proposed the following two-step procedure. First, obtain $\hat{\boldsymbol{\pi}}$ by multivariate OLS regression of \mathbf{y}_i on intercepts and \mathbf{x}_i . Second, obtain the **optimal MD estimator** (see Section 6.7) that minimizes

$$Q_N(\beta, \lambda) = (\text{Vec}[\hat{\boldsymbol{\pi}} - \mathbf{I}_T \otimes \beta' - \mathbf{e}\lambda'])' \mathbf{W}_N (\text{Vec}[\hat{\boldsymbol{\pi}} - \mathbf{I}_T \otimes \beta' - \mathbf{e}\lambda']),$$

where the optimal weighting matrix $\mathbf{W}_N = (\hat{\mathbf{V}}[\text{Vec}[\hat{\boldsymbol{\pi}}]])^{-1}$. This yields estimator $\hat{\beta}$ that is more efficient than OLS estimation of (22.20) if u_{it} is heteroskedastic.

Minimun distance estimation has been supplanted by GMM; see Arellano (2003, pp. 22–23) and Crépon and Mairesse (1995) for comparison of Chamberlain's MD estimator with GMM. However, Chamberlain's approach of obtaining moment restrictions via exogeneity assumptions and assumptions on the individual effects has had a big impact on the panel literature. His MD estimator is also used for estimation of covariance structures (see Section 22.5.4).

22.3. Panel GMM Example: Hours and Wages

We return to the hours–wages example of Section 21.3. Unlike as in Chapter 21 regressors are now permitted to be endogenous, and unlike as in Section 22.2 an individual-specific fixed effect is included. Estimation is by the IU methods of Section 22.2, after first-differencing to eliminate the fixed effects.

The regression model is

$$\lnhrs_{it} = \alpha_i + \beta_1 \lnwg_{it} + \beta_2 \text{kids}_{it} + \beta_3 \text{age}_{it} + \beta_4 \text{agesq}_{it} + \beta_5 \text{disab}_{it} + u_{it},$$

where interest lies in the intertemporal substitution wage elasticity of labor supply, β_1 , the coefficient of \lnwg , and the additional regressors are number of children, age, age squared, and an indicator for disability.

MacCurdy (1981) derived this relationship using a life-cycle labor supply model under uncertainty. The model is then a “ λ -constant” model where α_i here equals λ_i , a multiple of the marginal utility of initial wealth that is time-invariant but will differ across individuals. Since λ_i depends on variables and constraints it needs to be treated as a fixed rather than random effect. The labor supply literature presents several methods for controlling for this fixed effect.

One method, discussed further in Section 22.4.2, is to first difference the regression model, yielding

$$\Delta \lnhrs_{it} = \beta_1 \Delta \lnwg_{it} + \beta_2 \Delta \text{kids}_{it} + \beta_3 \Delta \text{age}_{it} + \beta_4 \Delta \text{agesq}_{it} + \beta_5 \Delta \text{disab}_{it} + \Delta u_{it}. \quad (22.22)$$

Estimation by OLS is then consistent for β if all regressors are exogenous. Note that this differencing induces serial correlation in the error even if u_{it} are iid, so panel-robust standard errors should be used.

Ziliak (1997) instead permitted \lnwg_{it} to be contemporaneously correlated with u_{it} , because of measurement error in wage or because of kink points in the budget constraint. Then the OLS estimator of (22.22) is inconsistent.

Ziliak proposed IV estimation using suitably lagged regressors as instruments. Assume that past wages are uncorrelated with the error, so that \lnwg is weakly exogenous aside from being contemporaneously correlated with the error. Then $E[\lnwg_{is} u_{it}] = 0$ for $s \leq t - 1$ implies that for the differenced model error $E[\lnwg_{is} \Delta u_{it}] = 0$ for $s \leq t - 2$, so \lnwg lagged two or more periods may be used as an instrument in the first-differences model. Note that this means that at least three periods of the original data are needed to identify β .

Ziliak's study focused on the properties of panel GMM estimators with endogenous regressors, so he treated all the regressors in (22.22) as endogenous and used as instruments lags of one or more periods in the levels of the other four regressors. For simplicity an intercept and time dummies, individual-invariant instruments that can be only used once, were not included. Results here change little with inclusion of an intercept as the dependent variable is in differenced form. Since $\lnwg_{i,t-2}$ is always used as an instrument the first two years are dropped and only the eight years 1981–1988 are used to estimate (22.22).

Table 22.2. Hours and Wages: GMM-IV Linear Panel Model Estimators^a

	Base Case			Stacked	
	OLS	2SLS	2SGMM	2SLS	2SGMM
β_1	0.112	0.209	0.547	0.543	0.330
Panel se	(.096)	(.374)	(.327)	(.209)	(.110)
Het se	[.079]	[.423]	[-]	[.226]	[-]
Default se	{.023}	{.389}	{ - }	{.169}	{ - }
RMSE	.283	.296	.307	.307	.298
Instruments	5	9	9	72	72
OIR Test	—	—	5.45	—	69.51
dof	—	—	4	—	67
p-value	—	—	.244	—	.393
N	4256	4256	4256	4256	4256

^a Differenced regression uses annual data from 1981–1988 for 532 men. Reported are β_1 , the coefficient of $\Delta \lnwg$, and three estimated standard errors: panel robust in parentheses, heteroskedastic robust in square brackets, and usual default estimates that assume iid errors in curly braces. All regressions additionally include Δkids , Δage , Δagesq , and Δdisab as regressors but their coefficient estimates are not reported. The instruments are \lnwg lagged twice and kids , age , agesq , and disab lagged both once and twice. For the base case there are 9 instruments and for stacked instruments there are $8 \times 9 = 72$ instruments. RMSE is the root mean square error of the residual. OIR is the over identifying restrictions test statistic, dof is the degrees of freedom, and p-value is the p-value for this test.

Table 22.2 presents a small subset of the many results given in tables 1 and 2 of Ziliak (1997). For completeness various standard error estimates are given but the panel-robust standard errors should be used.

OLS: The column OLS reports OLS estimation of (22.22). The labor supply elasticity of 0.112 differs a little from the estimate of 0.109 in the First-Diff column of Table 21.2 as here the four demographic variables are also included as regressors and an additional year of data has been dropped. Because first differences are modeled the model fit is poor, and the R^2 with additional inclusion of an intercept is 0.006.

2SLS with Base-Case Instruments: The base-case instruments use \mathbf{Z}_i defined in (22.11), where \mathbf{z}_{it} has nine entries: $\lnwg_{i,t-2}$, $\text{kids}_{i,t-1}$, $\text{age}_{i,t-1}$, $\text{agesq}_{i,t-1}$, $\text{disab}_{i,t-1}$, $\text{kids}_{i,t-2}$, $\text{age}_{i,t-2}$, $\text{agesq}_{i,t-2}$, and $\text{disab}_{i,t-2}$. The model is then over-identified with nine instruments and five parameters to estimate. The 2SLS estimate of β_1 is much less precise than the OLS estimate, with standard error increasing fourfold from 0.096 to 0.374. For the other regressors, not reported, the efficiency loss is much less.

2SLS with Stacked Instruments: The base case is GMM based on the nine moment conditions $E[\sum_{t=3}^{10} \mathbf{z}_{it} u_{it}] = \mathbf{0}$. The stacked instruments instead use 72 ($= 8 \times 9$) moment conditions $E[\mathbf{z}_{it} u_{it}] = \mathbf{0}$, $t = 3, \dots, 10$, where \mathbf{z}_{it} is as in the base case. Then use \mathbf{Z}_i defined in (22.14), where here \mathbf{Z}_i is 8 years by 72 instruments. The t th row of \mathbf{Z}_i is given in (22.17), where \mathbf{z}_{it} here is the 9×1 column vector of instruments for the base case. To construct the instruments first generate 72 variables ztj equal to zero for all i and t , where t denotes the year and j denotes the j th

instrument. Then replace zs_{jt} by $z_{it,j}$ if $t = s$ but leave $zs_{jt} = 0$ if $t \neq s$. For example, if $t = 3$ (the third year) set $z35$ equal to $disab_{i,2}$ if the fifth instrument is $disab_{i,t-1}$ and keep $zt5$ equal to zero for $t \neq 3$. The 2SLS estimates can then be obtained by standard 2SLS regression of $\Delta\lnhrs_{it}$ on the five regressors in (22.22) with these 72 constructed variables as instruments. Using the expanded instruments we have that the standard error of the 2SLS estimate falls from 0.374 to 0.209 and is only twice that of the original OLS estimate.

Two-step GMM: The two-step GMM estimates in Table 22.2 differ from those in table 1 of Ziliak (1997) as a panel-robust estimate of $\widehat{\mathbf{S}}$ defined in (22.7) is used here to form the weighting matrix, whereas Ziliak used the heteroskedastic-robust $\widehat{\mathbf{S}} = N^{-1} \sum_i \widehat{\mathbf{u}}_{it}^2 \mathbf{z}_{it} \mathbf{z}_{it}'$. As expected, the two-step GMM estimator is more efficient than 2SLS, with standard error falling from 0.374 to 0.327 with base-case instruments and from 0.209 to 0.110 with stacked instruments. This last standard error is not much larger than that for OLS.

Test of Overidentifying Restrictions: The test statistic for overidentifying restrictions is given in (22.10). From Table 22.2 for both base case and stacked instruments the test statistic has p-value much higher than 0.05, so the restrictions are not rejected and we conclude that the overidentifying instruments are valid instruments.

Test of Weak Instruments: Diagnostics for weak instruments were presented in Section 22.2.4 and Section 5.9. Since none of the regressors appear in the instrument set the overall F -statistic from the first-stage regression is used rather than a subset of regressors F -statistic. For the base-case instruments, regression of $\Delta\lnwg$ on the nine instruments and a constant term yields panel-robust $F = 2.80$, and similar regression for the 72 stacked instruments yields $F = 1.90$, indicating finite-sample bias is very likely. Similar regressions for Δkids , Δage , Δagesq , and Δdisab , regressors in (22.22) that are also being treated here as endogenous, yield $F > 8.5$ in all cases. Shea's partial R^2 (see Section 4.9.1) is 0.0036 for $\Delta\lnwg$ and exceeds 0.075 for the other four endogenous regressors. The weak instruments problem is therefore due to the problems of finding a good instrument for $\Delta\lnwg$.

Efficiency Gains: In this example panel GMM estimators were used to control for endogeneity. However, even if all the regressors are assumed to be strongly exogenous, panel GMM is still attractive as it is more efficient than OLS unless the errors u_{it} are iid; see the discussion after (22.20). As an example, the panel two-step GMM estimator with instrument set the base-case instruments plus the five original regressors in (22.22) yields $\widehat{\beta}_1 = 0.016$ with a standard error of 0.076, lower than the OLS standard error of 0.096.

22.4. Random and Fixed Effects Panel GMM

We now augment the panel data model (22.1) by including a time-invariant additive individual-specific effect α_i , so

$$y_{it} = \alpha_i + \mathbf{x}'_{it}\beta + \varepsilon_{it}. \quad (22.23)$$

Then the error term in (22.1) is now modeled as $u_{it} = \alpha_i + \varepsilon_{it}$. For simplicity the same notation is used for both fixed and random effects models, so in the case of random effects model the common intercept μ in Section 21.7 is subsumed into $\mathbf{x}'_{it}\beta$.

Some components of the regressors \mathbf{x}_{it} are assumed to be **endogenous**, with $E[\mathbf{x}_{it}(\alpha_i + \varepsilon_{it})] \neq \mathbf{0}$, so that the OLS estimator of β is inconsistent. In this section we propose IV estimators that yield consistent estimates of β in a variety of settings, including fixed effects, random effects, a hybrid of the two, and systems of equations.

22.4.1. Random Effects or Fixed Effects?

Recall from Chapter 21 that the individual-specific effect α_i can be viewed as random in both the FE and RE models. This random variable α_i was independent of \mathbf{x}_{it} in the RE model but correlated with \mathbf{x}_{it} in the FE model. For the RE model all coefficients are estimable, whereas in the FE model coefficients of time-invariant regressors are not estimable as consistent estimation requires elimination of α_i and the time-invariant regressors by differencing.

In this chapter with endogenous regressors we view a model to be a **random effects** model if instruments \mathbf{Z}_i exist that satisfy $E[\mathbf{Z}'_i(\alpha_i + \varepsilon_{it})] = \mathbf{0}$. Then the methods of Section 22.2 will permit consistent estimation of all regression parameters. If instead it is possible only to find instruments such that $E[\mathbf{Z}'_i\varepsilon_{it}] = \mathbf{0}$, but $E[\mathbf{Z}'_i\alpha_i] \neq \mathbf{0}$, we view the model to be a **fixed effects** model. Then α_i must be eliminated by differencing, in which case only the coefficients of time-varying regressors will be identified.

22.4.2. IV for Fixed Effects Models

The various differencing operations given in Section 21.2 applied to (22.23) lead to a **transformed model** of the form

$$\tilde{y}_{it} = \tilde{\mathbf{x}}'_{it}\beta + \tilde{\varepsilon}_{it},$$

where the tilda denotes a differencing transformation that eliminates α_i , and leading examples are given in the following. Upon stacking we get

$$\tilde{\mathbf{y}}_i = \tilde{\mathbf{X}}'_i\beta + \tilde{\varepsilon}_i. \quad (22.24)$$

If $E[\mathbf{x}_{it}\varepsilon_{it}] \neq \mathbf{0}$ then $E[\tilde{\mathbf{x}}_{it}\tilde{\varepsilon}_{it}] \neq \mathbf{0}$ and LS estimation of (22.24) leads to inconsistent estimates.

We now consider IV estimation, assuming existence of instruments \mathbf{Z}_i that satisfy $E[\mathbf{Z}'_i\tilde{\varepsilon}_i] = \mathbf{0}$. Then panel GMM estimation (IV, 2SLS, or 2SGMM) of (22.24) with instruments \mathbf{Z}_i yields consistent estimates of the coefficients of time-varying regressors. Panel-robust standard errors can be computed as discussed in Section 22.2.2.

One way that instruments may be obtained is through logic similar to that in the cross-section case. A valid instrument is a variable correlated with the regressor but not the error, yet is also one that can be excluded from the right-hand side of (22.23). Another way to obtain instruments, emphasized here, is through use of exogenous regressors in periods other than the current period, using the exogeneity assumptions detailed in Section 22.2.4.

The primitive assumptions for instrument availability are those on correlation between \mathbf{z}_{is} and ε_{it} . However, here it is correlation between \mathbf{z}_{is} and the differenced error $\tilde{\varepsilon}_{it}$ that matters. In general, differencing, necessary to eliminate the fixed effect, reduces the number of available instruments. Some differencing operations lead to greater loss than others and can even lead to inconsistent IV estimation. We consider three differencing operations with focus on **weakly exogenous instruments**. This can be a more realistic assumption in practice, especially for application to dynamic models.

IV for the First-Differences Model

The **first-differences IV estimator** is the IV or 2SLS or panel GMM estimator of the **first-differences model**

$$y_{it} - y_{i,t-1} = (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\beta + (\varepsilon_{it} - \varepsilon_{i,t-1}), \quad t = 2, \dots, T. \quad (22.25)$$

The weak exogeneity assumption that $E[\mathbf{z}_{is}\varepsilon_{it}] = \mathbf{0}$ for $s \leq t$ implies $E[\mathbf{z}_{is}(\varepsilon_{it} - \varepsilon_{i,t-1})] = \mathbf{0}$ for $s \leq t-1$. First differencing therefore shortens the time series on the available instrument set by one period, so that only $\mathbf{z}_{i,t-1}, \mathbf{z}_{i,t-2}, \dots$ are available as instruments. Assuming weak exogeneity, these yield a consistent IV estimator of β .

The use of lagged regressors as instruments was first proposed by Anderson and Hsiao (1981) in the context of dynamic panel models and was expanded upon by Holtz-Eakin, Newey, and Rosen (1988) and Arellano and Bond (1991) (see Section 22.5.3). Section 22.3 provided a detailed empirical example of this approach.

Note that one can instead use transformed instruments $\tilde{\mathbf{z}}_{is} = \Delta\mathbf{z}_{is} = \mathbf{z}_{is} - \mathbf{z}_{i,s-1}$, $s \leq t-1$. However, there is no gain, since using $\Delta\mathbf{z}_{i,t-1}, \dots, \Delta\mathbf{z}_{i2}, \mathbf{z}_{i1}$ is equivalent to using $\mathbf{z}_{i,t-1}, \dots, \mathbf{z}_{i2}, \mathbf{z}_{i1}$ as instruments, and only \mathbf{z}_{i1} and not $\Delta\mathbf{z}_{i1}$ can be computed if data begin in period 1.

IV for the Within or Mean-Differenced Model

The **within IV estimator** is the IV or 2SLS or panel GMM estimator of the **within model** or **mean-differenced model**

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\beta + (\varepsilon_{it} - \bar{\varepsilon}_i). \quad (22.26)$$

Then $E[\mathbf{z}_{is}\varepsilon_{it}] = \mathbf{0}$ for $s \leq t$ no longer implies $E[\mathbf{z}_{is}(\varepsilon_{it} - \bar{\varepsilon}_i)] = \mathbf{0}$ even for s much less than t . To see this suppose that $E[\mathbf{z}_{is}\varepsilon_{it}] \neq \mathbf{0}$ for $s > t$. Then $E[\mathbf{z}_{is}\bar{\varepsilon}_i] \neq \mathbf{0}$ for all s since $\bar{\varepsilon}_i = T^{-1} \sum \varepsilon_{it}$ includes past ε_{it} , which are correlated with \mathbf{z}_{is} .

Thus IV estimation of the within model leads to inconsistent estimation of β if the instruments are weakly exogenous or if they satisfy the even weaker assumptions of contemporaneous exogeneity or the summation condition. The within transformation can only be used if the instruments are actually strongly exogenous.

IV for the Forward Orthogonal Deviations Model

An alternative method to first differences, one that also requires that instruments be only weakly exogenous rather than strongly exogenous, was proposed by Arellano and Bover (1995). We also present this method, even though first differences are used much more.

For the stacked model (22.2) for the i th observation, the first-difference transformation yields model $\mathbf{D}\mathbf{y}_i = \mathbf{D}\mathbf{X}_i\beta + \mathbf{D}\varepsilon_i$, where \mathbf{D} is a $(T - 1) \times T$ matrix with entry \mathbf{D}_{ts} , $t = 1, \dots, T - 1$, $s = 1, \dots, T$, equal to minus one if $s = t$, equal to one if $s = t + 1$, and equal to zero otherwise. If ε_{it} are iid the transformed error is MA(1) and $V[\mathbf{D}\mathbf{u}_i] = \sigma^2 \mathbf{D}\mathbf{D}'$. The GLS estimator then premultiplies $\mathbf{D}\varepsilon_i$ by $(\mathbf{D}\mathbf{D}')^{-1/2}$, or premultiplies ε_i by $(\mathbf{D}\mathbf{D}')^{-1/2}\mathbf{D}$. This yields a transformed model of the form (22.24) where the tilde denotes premultiplication by $(\mathbf{D}\mathbf{D}')^{-1/2}\mathbf{D}$.

If the upper triangular Cholesky factorization is used to obtain $(\mathbf{D}\mathbf{D}')^{-1/2}$, then this yields the **forward orthogonal deviation model**

$$c_t(y_{it} - \bar{y}_{it}^F) = c_t(\mathbf{x}_{it} - \bar{\mathbf{x}}_{it}^F)' \beta + c_t(\varepsilon_{it} - \bar{\varepsilon}_{it}^F) \quad (22.27)$$

(see Arellano, 2003, p. 17), where $c_t^2 = (T - t)/(T - t + 1)$ and the superscript F denotes that only future values are used to form the average. For example, $\bar{y}_{it}^F = (T - t)^{-1} \sum_{s=t+1}^T y_{is}$.

The transformation is called **orthogonal deviations** because the transformed errors $c_t(\varepsilon_{it} - \bar{\varepsilon}_i^F)$ have unit variance and are uncorrelated. The adjective **forward** is added as the transformed error depends only on current and future values of the original error. An OLS estimation of (22.27) yields the within estimator of Chapter 21, so the orthogonal deviations transformation is optimal if indeed ε_{it} are iid.

The **forward orthogonal deviations IV estimator** is the IV or 2SLS or panel GMM estimator of the model (22.27). For weakly exogenous instruments, $E[\mathbf{z}_{is}\varepsilon_{it}] = \mathbf{0}$ for $s \leq t$ implies $E[\mathbf{z}_{is}(\varepsilon_{it} - \bar{\varepsilon}_i^F)] = \mathbf{0}$ for $s \leq t$. Forward orthogonal deviations therefore lead to no loss in the number of available instruments. The transformation is usually not applied to the instruments as $(\mathbf{z}_{it} - \bar{\mathbf{z}}_i^F)$ involves future values of \mathbf{z}_{it} that in many applications are correlated with ε_{it} .

22.4.3. IV for Random Effects Models

The model stacked for the i th observation is

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{e}\alpha_i + \varepsilon_i,$$

where \mathbf{e} is a $T \times 1$ vector of ones. Consistent but inefficient estimates can be obtained by directly applying the panel GMM estimators of Section 22.2 given instruments \mathbf{Z}_i , obtained through exclusion restrictions or through appropriate exogeneity restrictions, such that $E[\mathbf{Z}_i'(\mathbf{e}\alpha_i + \varepsilon_i)] = \mathbf{0}$. Here we go further and consider more efficient estimation that, as in Chapter 21, controls for error correlation over time given the error components model $u_{it} = \alpha_i + \varepsilon_{it}$.

IV Estimation of Transformed Model

Assume that the instruments \mathbf{Z}_i satisfy $E[\mathbf{u}_i | \mathbf{Z}_i] = \mathbf{0}$ and $V[\mathbf{u}_i | \mathbf{Z}_i] = \Omega_i$, where Ω_i has the same form as the standard RE model with diagonal entries $\sigma_\alpha^2 + \sigma_\varepsilon^2$ and off-diagonal entries σ_α^2 . Note that this is a stronger assumption than $E[\mathbf{Z}'_i \mathbf{u}_i] = \mathbf{0}$ and will therefore place restrictions on available instruments.

Given the conditional moment condition $E[\mathbf{u}_i | \mathbf{Z}_i] = \mathbf{0}$, from Section 6.3.7 the optimal unconditional moment condition is

$$E[\mathbf{Z}'_i \Omega_i^{-1} \mathbf{u}_i] = E[(\Omega_i^{-1/2} \mathbf{Z}_i)' (\Omega_i^{-1/2} \mathbf{u}_i)] = \mathbf{0}.$$

This leads to GMM estimation in the transformed system $\mathbf{y}_i^* = \mathbf{X}_i^* \boldsymbol{\beta} + \mathbf{u}_i^*$ with transformed instruments \mathbf{Z}_i^* , where the asterisk denotes premultiplication by the $T \times T$ matrix $\Omega_i^{-1/2}$ or a consistent estimate $\widehat{\Omega}_i^{-1/2}$.

From Section 21.7.1 premultiplication by $\widehat{\Omega}_i^{-1/2}$ leads to the model

$$y_{it} - \widehat{\lambda} \bar{y}_i = (\mathbf{x}_{it} - \widehat{\lambda} \bar{\mathbf{x}}_i)' \boldsymbol{\beta} + \{(1 - \widehat{\lambda})\alpha_i + (\varepsilon_{it} - \widehat{\lambda} \bar{\varepsilon}_i)\}, \quad (22.28)$$

where $\widehat{\lambda}$ is a consistent estimate of $\lambda = 1 - \sigma_\varepsilon / \sqrt{\sigma_\varepsilon^2 + T\sigma_\alpha^2}$. The **random effects IV estimator** is the IV or 2SLS estimator of this model with transformed instruments $\tilde{\mathbf{z}}_{it} = (\mathbf{z}_{it} - \widehat{\lambda} \bar{\mathbf{z}}_i)$, or equivalently with instruments $\mathbf{z}_{it} - \bar{\mathbf{z}}_i$ and $\bar{\mathbf{z}}_i$.

This method requires a consistent estimate $\widehat{\lambda}$ of λ . For σ_ε^2 we use $\widehat{\sigma}_\varepsilon^2 = \sum_i \tilde{\varepsilon}_{it}^2 / N(T-1)$, where $\tilde{\varepsilon}_{it}$ is the residual from within IV regression of $y_{it} - \bar{y}_i$ on $(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)$ with instruments $(\mathbf{z}_{it} - \bar{\mathbf{z}}_i)$ (see (22.26)). Also, $(\sigma_\varepsilon^2 + T\sigma_\alpha^2)$ can be estimated by $\sum_i \bar{u}_i^2 / N$, where \bar{u}_i is the residual from the between IV regression of \bar{y}_i on $\bar{\mathbf{x}}_i$ with instruments $\bar{\mathbf{z}}_i$. The resulting IV estimator of $\boldsymbol{\beta}$ is called the **error components 2SLS (EC2SLS) estimator** by Baltagi (1981).

These results are dependent on specification of a particular functional form for Ω_i . The results in Section 22.2.2 permit inference that is robust to misspecification of Ω_i , using (22.5) where \mathbf{y} , \mathbf{X} , \mathbf{Z} , and $\mathbf{W}_N = [\mathbf{Z}' \mathbf{Z}]^{-1}$ are replaced by the transformed variables in (22.28).

A more important restriction is that this method can only be used if the original instruments are strongly exogenous. Here consistency requires that $E[\mathbf{Z}'_i \Omega_i^{-1} \mathbf{u}_i] = \mathbf{0}$, a much stronger assumption than $E[\mathbf{Z}'_i \mathbf{u}_i] = \mathbf{0}$, which essentially requires that $E[\mathbf{u}_i | \mathbf{Z}_i] = \mathbf{0}$. For example, suppose $E[\mathbf{z}_{it} \alpha_i] = \mathbf{0}$ for all t whereas $E[\mathbf{z}_{is} \varepsilon_{it}] = \mathbf{0}$ for $s \leq t$ but $E[\mathbf{z}_{it} \varepsilon_{it}] \neq \mathbf{0}$ for $s > t$. Then $E[\mathbf{z}_{it} \bar{\varepsilon}_i] \neq \mathbf{0}$, leading to correlation of instruments with the error term in (22.28).

22.4.4. IV for the Hausman–Taylor Hybrid Model

A leading example of endogeneity involves regressors correlated with the individual-specific effect α_i . This leads to inconsistency of the RE estimator of Chapter 21. An obvious solution is to instead use the within (or fixed effects) estimator, which is consistent. However, then the coefficients of time-invariant individual regressors cannot be identified. This defeats the purpose of many panel studies – estimation of the effect of time-invariant regressors, such as the effect of the level of schooling in a postschooling earnings regression.

Hausman and Taylor (1981) considered the following variant of (22.23):

$$y_{it} = \mathbf{x}'_{1it}\beta_1 + \mathbf{x}'_{2it}\beta_2 + \mathbf{w}'_{1i}\gamma_1 + \mathbf{w}'_{2i}\gamma_2 + \alpha_i + \varepsilon_{it}, \quad (22.29)$$

where some regressors are assumed to be correlated with α_i whereas others are not, and \mathbf{w} is introduced to denote time-invariant regressors. Specifically, \mathbf{x}_{1it} and \mathbf{w}_{1i} are uncorrelated with α_i but \mathbf{x}_{2it} and \mathbf{w}_{2i} are correlated with α_i . All regressors are assumed to be uncorrelated with ε_{it} . In this model the α_i can be viewed as a **hybrid** of random and fixed effects.

Hausman and Taylor (1981) proposed making use of the time-varying exogenous regressor \mathbf{x}_{1it} in two ways: to estimate β_1 and as an instrument for \mathbf{w}_{2i} , permitting estimation of γ . Then γ is identified if the number of time-varying exogenous regressors equals or exceeds the number of time-invariant endogenous regressors. Amemiya and MacCurdy (1986) proposed a more efficient estimator that uses \mathbf{x}_{1it} in $(T + 1)$ ways: to estimate β_1 and as T instruments for \mathbf{w}_{2i} , permitting identification if $\dim[\mathbf{w}_{2i}] \geq T\dim[\mathbf{x}_{1it}]$. This approach to obtaining instruments from exogenous regressors in periods other than the current period has already been discussed in detail in Section 22.2.4.

Various **projections**, some equivalent, can be used to generate suitable instruments. Breusch, Mizon, and Schmidt (1989) provided a simpler presentation and projection that permits estimation using a 2SLS package.

First consider consistent but inefficient estimation that ignores the panel correlation structure of $(\alpha_i + \varepsilon_{it})$. The within transformation eliminates correlation with α_i , so $\tilde{\mathbf{x}}_{2it} = \mathbf{x}_{2it} - \bar{\mathbf{x}}_2$ can be used as instrument for endogenous \mathbf{x}_{2it} . The instrument for \mathbf{x}_{1it} is similarly $\tilde{\mathbf{x}}_{1it}$, rather than the more obvious \mathbf{x}_{1it} . Then $\bar{\mathbf{x}}_{1i}$ is used as an instrument for endogenous \mathbf{w}_{2i} , whereas the exogenous \mathbf{w}_{1i} is used as an instrument for itself.

Now consider efficient estimation under the random effects assumption that the components α_i and ε_{it} are homoskedastic. Then from (22.27) the **random effects differencing transformation** (see 22.28) leads to

$$\tilde{y}_{it} = \tilde{\mathbf{x}}'_{1it}\beta_1 + \tilde{\mathbf{x}}'_{2it}\beta_2 + \tilde{\mathbf{w}}'_{1i}\gamma_1 + \tilde{\mathbf{w}}'_{2i}\gamma_2 + v_{it}, \quad (22.30)$$

where, for example, $\tilde{\mathbf{x}}_{1it} = \mathbf{x}_{1it} - \hat{\lambda}\bar{\mathbf{x}}_{1i}$, where an estimator for the scalar $\hat{\lambda}$ has been presented at the end of the preceding section. The Hausman–Taylor estimator is equivalent to IV estimation of (22.30) using as instruments $\tilde{\mathbf{x}}_{1it}$, $\tilde{\mathbf{x}}_{2it}$, \mathbf{w}_{1i} , and $\bar{\mathbf{x}}_{1i}$. The exogenous time-varying regressors $\mathbf{x}_{1it} = \tilde{\mathbf{x}}_{1it} + \bar{\mathbf{x}}_{1i}$ are used as instrument twice, with the within difference $\tilde{\mathbf{x}}_{1it}$ used as an instrument for \mathbf{x}_{1it} and the time average $\bar{\mathbf{x}}_{1i}$ used as an instrument for \mathbf{w}_{2i} . The estimator of Amemiya and MacCurdy (1986) instead uses as instruments $\tilde{\mathbf{x}}_{1it}$, $\tilde{\mathbf{x}}_{2it}$, \mathbf{w}_{1i} and $\mathbf{x}_{1it}, \dots, \mathbf{x}_{1iT}$, so that the entire history of \mathbf{x}_{1i} rather than just the time average is used as an instrument. This requires that $E[\mathbf{x}_{1i}\alpha_i] = \mathbf{0}$ for $t = 1, \dots, T$, a stronger assumption than $E[\bar{\mathbf{x}}_{1i}\alpha_i] = \mathbf{0}$ (see Section 22.2.4). Breusch et al. (1989) proposed an even more efficient estimator using $\tilde{\mathbf{x}}_{2is}$, $s \neq t$, as additional instruments.

The major limitation of this approach is that it requires specification of which regressors are either correlated or not correlated with α_i . In a post schooling log-wage regression, Hausman and Taylor begin by assuming that all three time-varying regressors (experience, bad health, and unemployment last year) are exogenous, two

time-invariant regressors (race and union status) are exogenous, and the time-invariant regressor of interest (schooling) is endogenous. In this specification there are two overidentifying restrictions. A model specification test is possible using a Hausman test based on the difference between $\hat{\beta}_{HT}$ and $\hat{\beta}_W$, since the within estimator for β is consistent regardless of which components of \mathbf{x}_{it} and \mathbf{w}_i are correlated with α_i . Cornwall and Rupert (1988) provide an empirical study that contrasts the various estimators.

22.4.5. SUR and Simultaneous Equations Estimation

The preceding panel data analysis has focused exclusively on estimation of a single equation in isolation. In some cases it may be desired to estimate a system of equations, such as a system of demand equations, where dependent variables and regressors are observed for many individuals at several points in time. If there are no cross-equation restrictions on the parameters then single-equation estimation can yield consistent estimates, but more efficient estimation is possible using joint equation estimation that exploits error correlation across equations.

In the Chapter 21 framework of strongly exogenous regressors, the more efficient estimator is an extension of seemingly unrelated regressions from cross-section to panel data. The **error components SUR model** specifies the g th of G equations to be given by

$$y_{git} = \mathbf{x}'_{git}\beta + \alpha_{gi} + \varepsilon_{git}, \quad g = 1, \dots, G, \quad (22.31)$$

where, as in the cross-section case, α_{gi} is independent over i , ε_{git} is independent over i and t , and α_{gi} and ε_{git} are independent of each other. However, the error components are allowed to be correlated across components, so that $\text{Cov}[\alpha_{gi}, \alpha_{hi}] \neq 0$ and $\text{Cov}[\varepsilon_{git}, \varepsilon_{hit}] \neq 0$ for $g \neq h$. Then the Chapter 21 methods yield consistent estimates. The obvious single-equation estimator is the random effects estimator that is feasible GLS controlling for the correlation within each equation. More efficient GLS estimators that additionally control for cross-equation correlation in the errors are detailed in Avery (1977) and Baltagi (1980).

Similar efficiency gains can be found when the system is one of **simultaneous equations**, where now in (22.31) the regressor \mathbf{x}_{git} may include one or more endogenous regressors y_{hit} from other equations. Then IV or GMM estimation of each single equation yields consistent estimates, with the obvious estimator given the error components structure being the random effects IV or EC2SLS estimator of Section 22.4.3. More efficient estimates are obtained by systems estimation, using the **error components three-stage least-squares (EC3SLS) estimator** proposed by Baltagi (1981).

The systems estimators are more difficult to implement and separate estimation of each equation may be adequate. Even if this simpler approach is taken, however, much can be gained in specifying a system of simultaneous equations as it permits identification of the coefficients of endogenous regressors using as instruments exogenous regressors excluded from the equation of interest. This provides a more traditional approach to obtaining instruments than using as instruments exogenous regressors from time periods other than the current one.

22.5. Dynamic Models

In this section we consider the usual individual-specific effects panel data model, with the complication that the regressors include the dependent variable lagged once. Then the model is a **dynamic model** with

$$y_{it} = \gamma y_{i,t-1} + \mathbf{x}'_{it}\beta + \alpha_i + \varepsilon_{it}, \quad i = 1, \dots, N, \quad t = 1, \dots, T. \quad (22.32)$$

As usual the panel is short with data independent over i . It is assumed that $|\gamma| < 1$, an assumption relaxed in Section 22.5.4.

An important result is that even if α_i is a random effect, OLS estimation of (22.32) leads to inconsistent estimation of γ and β . This is because the regressor $y_{i,t-1}$ is correlated with α_i and hence with the composite error term ($\alpha_i + \varepsilon_{it}$). Alternative estimators are needed even with random effects.

We consider estimation when α_i is a fixed effect, $|\gamma| < 1$, the error ε_{it} is serially uncorrelated, and the panel is short (see Section 22.5.3). Although this is the base case for microeconomics applications there exists a vast literature that changes one or more of these assumptions. More generally the individual-specific effect may be purely random, errors may be serially correlated, data may be nonstationary, and the panel may be a long panel, but we barely touch on this literature.

22.5.1. True State Dependence and Unobserved Heterogeneity

Before considering estimation, we note that time-series correlation in y_{it} is now induced directly by $y_{i,t-1}$ in addition to the indirect effect via α_i already considered in Chapter 21. These two causes lead to quite different interpretations of **correlation over time** in, for example, individual earnings or welfare recipiency.

For simplicity let $\beta = \mathbf{0}$ so that $y_{it} = \gamma y_{i,t-1} + \alpha_i + \varepsilon_{it}$. Then $E[y_{it}|y_{i,t-1}, \alpha_i] = \gamma y_{i,t-1} + \alpha_i$ and $\text{Cor}[y_{it}, y_{i,t-1}|\alpha_i] = \gamma$. Conditional on α_i , the standard time-series results for an AR(1) model apply with dependence over time in y_{it} determined solely by the autoregressive parameter γ . However, α_i is unknown and we actually observe $E[y_{it}|y_{i,t-1}] = \gamma y_{i,t-1} + E[\alpha_i|y_{i,t-1}]$ and $\text{Cor}[y_{it}, y_{i,t-1}] \neq \gamma$. Specifically, from (22.32) with $\beta = \mathbf{0}$

$$\begin{aligned} \text{Cor}[y_{it}, y_{i,t-1}] &= \text{Cor}[\gamma y_{i,t-1} + \alpha_i + \varepsilon_{it}, y_{i,t-1}] \\ &= \gamma + \text{Cor}[\alpha_i, y_{i,t-1}] \\ &= \gamma + \frac{(1 - \gamma)}{1 + (1 - \gamma)\sigma_\varepsilon^2/(1 + \gamma)\sigma_\alpha^2}, \end{aligned} \quad (22.33)$$

where the second equality assumes $\text{Cor}[\varepsilon_{it}, y_{i,t-1}] = 0$ and the third equality is obtained after some algebra for the special case of random effects with ε_{it} iid $[0, \sigma_\varepsilon^2]$ and α_i iid $[0, \sigma_\alpha^2]$.

Result (22.33) makes it clear that there are two possible reasons for correlation between y_{it} and $y_{i,t-1}$.

True state dependence occurs when correlation over time is due to the causal mechanism that $y_{i,t-1}$ last period determines y_{it} this period. This dependence is

relatively large if the individual effect $\alpha_i \simeq 0$ as then $\text{Cor}[y_{it}, y_{i,t-1}] \simeq \gamma$. More generally, this happens when σ_α^2 is very small relative to σ_ε^2 .

Correlation due to **unobserved heterogeneity** arises even if there is no causal mechanism, so $\gamma = 0$, but nonetheless there is correlation since $\text{Cor}[y_{it}, y_{i,t-1}]$ simplifies to $\sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_\varepsilon^2)$ if $\gamma = 0$, as in Chapter 21.

Both extremes permit this correlation to be arbitrarily close to one because either $\gamma \rightarrow 1$ or $\sigma_\varepsilon^2/\sigma_\alpha^2 \rightarrow 0$. However, these give two quite different explanations with quite different policy implications. A true state dependence explanation for earnings y_{it} being continuously high over time even after controlling for regressors x_{it} is that future earnings are determined by past earnings and γ is large. An unobserved heterogeneity explanation is that actually γ is small, but important variables have been omitted from x_{it} , leading to a high α_i in each time period. For duration data the distinction between true state dependence and unobserved heterogeneity was explored in Chapter 18. The static linear panel models of Chapter 21 considered only unobserved heterogeneity.

22.5.2. Inconsistency of Standard Panel Estimators

The estimators from the previous chapter are all **inconsistent** if the regressors include lagged dependent variables, even in the case of the random effects model. We consider estimation of the model given in (22.32), where the literature usually assumes that ε_{it} are serially uncorrelated.

First consider **OLS estimation** of y_{it} on $y_{i,t-1}$ and \mathbf{x}_{it} . The error term is then $(\alpha_i + \varepsilon_{it})$, which is correlated with the regressor $y_{i,t-1}$ since lagging the equation gives $y_{i,t-1} = \gamma y_{i,t-2} + \mathbf{x}'_{i,t-1}\beta + \alpha_i + \varepsilon_{i,t-1}$, so that $y_{i,t-1}$ is correlated with α_i . Note that this is a departure from earlier results for OLS estimation of the random effects model without lagged dependent variable, as then OLS of y_{it} on \mathbf{x}_{it} yields a consistent, albeit inefficient, estimator. This is also a departure from the usual OLS result that regression of y_{it} on $y_{i,t-1}$ yields a consistent estimate (though one biased in small samples) if the error is serially uncorrelated.

Second, consider the **within estimator**, which regresses $(y_{it} - \bar{y}_i)$ on $(y_{i,t-1} - \bar{y}_{i,-1})$ and $(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)$. This regression has error term $(\varepsilon_{it} - \bar{\varepsilon}_i)$. Now by (22.32), y_{it} is correlated with ε_{it} , so $y_{i,t-1}$ is correlated with $\varepsilon_{i,t-1}$ and hence $\bar{\varepsilon}_i$. However, this implies that the regressor $(y_{i,t-1} - \bar{y}_i)$ is correlated with the error $(\varepsilon_{it} - \bar{\varepsilon}_i)$. Thus OLS estimation of the within model leads to inconsistent parameter estimates, because the regressor is correlated with the error term. Consistency requires that $\bar{\varepsilon}_i$ becomes very small relative to ε_{it} , which requires $T \rightarrow \infty$, which occurs in long panels but not in short panels. A leading reference is Nickell (1981).

Inconsistency also arises for the **random effects estimator** given in Chapter 21, since this is a linear combination of the within and between estimators. For random effects models Anderson and Hsiao (1981) instead considered ML estimation when $\varepsilon_{it} \sim \mathcal{N}[0, \sigma^2]$; see also Bhargava and Sargan (1983). In short panels the distribution of the MLE depends on the assumptions made on y_{i0} , the initial value of the dependent variable. Anderson and Hsiao (1981) distinguish among the following **initial condition** assumptions: (1) fixed initial observations, (2) random initial observations with a

common mean, (3) random initial observations with different means, and (4) random initial observations with a stationary distributions.

The **first differences OLS estimator** is also inconsistent, but an IV variant leads to consistent estimates. We now present this estimator.

22.5.3. Arellano–Bond Estimator

Model (22.32) leads to the first-differences model

$$y_{it} - y_{i,t-1} = \gamma(y_{i,t-1} - y_{i,t-2}) + (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\beta + (\varepsilon_{it} - \varepsilon_{i,t-1}), \quad t = 2, \dots, T. \quad (22.34)$$

The OLS estimator is inconsistent because $y_{i,t-1}$ is correlated with $\varepsilon_{i,t-1}$ from (22.32), so the regressor $(y_{i,t-1} - y_{i,t-2})$ is correlated with the error $(\varepsilon_{it} - \varepsilon_{i,t-1})$ in (22.34).

Anderson and Hsiao (1981) proposed estimating (22.34) using the **instrumental variables estimator** with $y_{i,t-2}$ as an instrument for $(y_{i,t-1} - y_{i,t-2})$. This is a valid instrument, since $y_{i,t-2}$ is not correlated with $(\varepsilon_{it} - \varepsilon_{i,t-1})$ assuming the errors ε_{it} are serially uncorrelated. Furthermore, $y_{i,t-2}$ is a good instrument since it is correlated with $(y_{i,t-1} - y_{i,t-2})$. The method requires availability of three periods of data for each individual. An alternative is to use $\Delta y_{i,t-2}$ as an instrument for $\Delta y_{i,t-1}$, which will require four periods of data. Anderson and Hsiao (1981) present results suggesting that the IV estimator is more efficient using $\Delta y_{i,t-2}$ rather than $y_{i,t-2}$ as the instrument in the usual case that $\gamma > 0$. In either case $(\mathbf{x}_{it} - \mathbf{x}_{i,t-1})$ is used as an instrument for itself.

More efficient estimation is possible by using additional lags of the dependent variable as instruments. For example, both $y_{i,t-2}$ and $y_{i,t-3}$ might be used as instruments. The model is then overidentified, so estimation should be by 2SLS or panel GMM. Furthermore, the number of instruments available is highest for the dependent variable observed at time t closest to the final time period T . In period 3 only y_{i1} is available as an instrument, in period 4 both y_{i1} and y_{i2} are available, in period 5 y_{i1} , y_{i2} , and y_{i3} are available, and so on. Holtz-Eakin et al. (1988) and Arellano and Bond (1991) proposed panel GMM estimators using these wider unbalanced instrument sets.

The microeconomics literature refers to the resulting panel GMM estimator as the **Arellano–Bond estimator**. The general procedure has already been presented in Section 22.4.2, where dynamics were not explicitly introduced. The estimator is

$$\hat{\beta}_{AB} = \left[\left(\sum_{i=1}^N \tilde{\mathbf{X}}_i' \mathbf{Z}_i \right) \mathbf{W}_N \left(\sum_{i=1}^N \mathbf{Z}_i' \tilde{\mathbf{X}}_i \right) \right]^{-1} \left(\sum_{i=1}^N \tilde{\mathbf{X}}_i' \mathbf{Z}_i \right) \mathbf{W}_N \left(\sum_{i=1}^N \mathbf{Z}_i' \tilde{\mathbf{y}}_i \right), \quad (22.35)$$

where $\tilde{\mathbf{X}}_i$ is a $(T-2) \times (K+1)$ matrix with t th row $(\Delta y_{i,t-1}, \Delta \mathbf{x}'_{it})$, $t = 3, \dots, T$, $\tilde{\mathbf{y}}_i$ is a $(T-2) \times 1$ vector with t th row Δy_{it} , and \mathbf{Z}_i is a $(T-2) \times r$ matrix of instruments

$$\mathbf{Z}_i = \begin{bmatrix} \mathbf{z}'_{i3} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{z}'_{i4} & & \vdots \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{z}'_{iT} \end{bmatrix}, \quad (22.36)$$

where often $\mathbf{z}'_{it} = [y_{i,t-2}, y_{i,t-3}, \dots, y_{i1}, \Delta x'_{it}]$. Lags of \mathbf{x}_{it} or $\Delta \mathbf{x}_{it}$ can additionally be used as instruments, and for moderate or large T there may be a maximum lag of y_{it} that is used as an instrument, such as not more than $y_{i,t-4}$. Two-stage LS and two-step GMM correspond to different weighting matrices \mathbf{W}_N (see Section 22.2.3).

The method is easily adapted to an AR(p) model, with $\gamma y_{i,t-1}$ in (22.32) replaced by $\gamma_1 y_{i,t-1} + \gamma_2 y_{i,t-2} + \dots + \gamma_p y_{i,t-p}$, though more than three periods of data will be needed to permit consistent estimation.

The empirical example in Section 22.3 is essentially an Arellano–Bond estimation example, since a first differences model is estimated by IV with lagged regressors used as instruments.

Ahn and Schmidt (1995) noted that more efficient estimation is possible using additional moment conditions. Consider the pure time-series version of (22.32) where $\beta = \mathbf{0}$, and make the standard assumption that ε_{it} is uncorrelated with α_i , ε_{is} for $s \neq t$ and the initial observation y_{i1} . The Arellano–Bond estimator uses the moment conditions $E[y_{is} \Delta u_{it}] = 0$ for $s \leq t - 2$, where $u_{it} = \varepsilon_{it} + \alpha_i$. Ahn and Schmidt (1995) obtain a more efficient estimator by additionally using the moment conditions $E[u_{iT} \Delta u_{it}] = 0$. They show that this estimator, which makes efficient use of the second moment assumptions, is asymptotically equivalent to the optimal minimum distance estimator of Chamberlain (1982, 1984).

Additional assumptions lead to additional moment conditions and hence more efficient estimation. If $V[\varepsilon_{it}] = V[\varepsilon_{is}]$ then $E[\bar{u}_i \Delta u_{it}] = 0$ (see Ahn and Schmidt, 1995), assuming homoskedasticity of ε_{it} . Arellano and Bover (1995) propose using the condition $E[u_{it} \Delta y_{is}] = 0$ for $s \leq t - 1$. Blundell and Bond (1998) consider these and additional assumptions and show that the benefit can be considerable, especially when γ is high and T is small. Arellano and Honore (2001) present many assumptions that might be made and the corresponding moment conditions that can be used in estimation.

Hsiao, Pesaran, and Tahmisioglu (2002) propose a **transformed ML estimator**. Assume that ε_{it} are iid $\mathcal{N}[0, \sigma^2]$, an assumption that can be relaxed. Rather than form the likelihood based on $\varepsilon_{i1}, \dots, \varepsilon_{iT}$, they form the likelihood based on the error differences $\Delta \varepsilon_{i1}, \dots, \Delta \varepsilon_{iT}$. For the pure time series AR(1) model $\Delta \varepsilon_{it} = \Delta y_{it} - \gamma \Delta y_{i,t-1}$ for $t > 1$. The density of $\Delta \varepsilon_{i1}$ depends on the assumptions made about initial conditions: either $\Delta \varepsilon_{i1} = \Delta y_{i1}$ or $\Delta \varepsilon_{i1} = \Delta y_{i1} - b$, where $b = E[\Delta y_{i1}]$ is an additional parameter to be estimated. The resulting estimator is a quasi-MLE that retains consistency even if ε_{it} are nonnormal. If ε_{it} are iid $[0, \sigma^2]$ then the transformed MLE is more efficient than the preceding GMM estimators.

22.5.4. Estimation of Covariance Structures

Covariance structures are models that specify a structure for the covariance matrix of the regression error. Applications include structures for error dynamics and for measurement error. The goal is to estimate the parameters of the structure.

As an example, suppose that y_{it} is generated by a random effects model with MA(1) error, so that

$$y_{it} = \alpha_i + \varepsilon_{it} + \phi \varepsilon_{i,t-1},$$

where $\alpha_i \sim [0, \sigma_\alpha^2]$ and $\varepsilon_{it} \sim [0, \sigma_\varepsilon^2]$ and $|\phi| < 1$. Then the autocovariances $\gamma_j = \text{Cov}[y_{it}, y_{it-j}]$ satisfy $\gamma_0 = \sigma_\alpha^2 + (1 + \phi^2)\sigma_\varepsilon^2$, $\gamma_1 = \sigma_\alpha^2 + \phi\sigma_\varepsilon^2$, and $\gamma_j = \sigma_\alpha^2$ for $j \geq 2$. If $T = 3$ these equations yield estimates $\hat{\sigma}_\alpha^2$, $\hat{\sigma}_\varepsilon^2$, and $\hat{\phi}$ given autocovariance estimates $\hat{\gamma}_0$, $\hat{\gamma}_1$, and $\hat{\gamma}_2$. If $T > 3$ the model is overidentified as there are only three variance parameters to estimate but more than three autocovariance estimates. An obvious estimator is the minimum distance estimator.

In general let $\boldsymbol{\theta}$ denote the q structural parameters and suppose $\mathbf{g}(\boldsymbol{\theta}) = \boldsymbol{\gamma}$, where $\boldsymbol{\gamma} = [\gamma_0, \dots, \gamma_{T-1}]'$ is the vector of $T \geq q$ autocovariances. Then the **minimum distance estimator** $\hat{\boldsymbol{\theta}}_{MD}$ minimizes

$$Q_N(\boldsymbol{\theta}) = (\hat{\boldsymbol{\gamma}} - \mathbf{g}(\boldsymbol{\theta}))' \mathbf{W}_N (\hat{\boldsymbol{\gamma}} - \mathbf{g}(\boldsymbol{\theta})), \quad (22.37)$$

where $\hat{\boldsymbol{\gamma}} = [\hat{\gamma}_1, \dots, \hat{\gamma}_{T-1}]'$,

$$\hat{\gamma}_j = [N(T-j)]^{-1} \sum_{t=j+1}^T \sum_{i=1}^N (y_{it} - \bar{y}_t)(y_{i,t-j} - \bar{y}_{t-j}), \quad (22.38)$$

and $\bar{y}_{t-j} = N^{-1} \sum_i y_{i,t-j}$. The weighting matrix \mathbf{W}_N and further details on MD estimation are provided in Section 6.7. The restrictions of the model can be tested by use of the chi-squared test statistic given in Section 6.7. The discussion thus far has already imposed the restriction of covariance stationarity. One can more generally permit $\gamma_{tj} \neq \gamma_{sj}$ for $t \neq s$, where $\gamma_{tj} = \text{Cov}[y_{it}, y_{i,t-j}]$. Then $\boldsymbol{\gamma}$ has $T(T+1)/2$ entries γ_{tj} , $t = j+1, \dots, T$ and $j = 0, \dots, T-1$. The stationarity assumption is itself a testable assumption. Moreover, regressors can be incorporated by replacing y_{it} by the residual $y_{it} - \mathbf{x}'_{it}\beta$.

Abowd and Card (1989) provided an early application of this approach to joint modeling of earnings and hours. Altonji and Segal (1996) demonstrated that the optimal MD estimator can be quite biased in finite samples (see Section 6.3.5). Many of the applications are to models of earnings; see Baker and Solon (2003) for a recent example.

The MD approach is well suited to estimation of covariance structures. The panel data sets can be large, but by first estimating the autocovariances the estimation is reduced to minimizing (22.37). Other estimation approaches are possible. In particular, see MacCurdy (1982b), who presents Box–Jenkins type models for panel data.

22.5.5. Nonstationary Panels

The panel literature on unit roots and nonstationarity emphasizes panels where both N and T are large. For **unit root tests** a key early paper is that by Levin and Lin (1992), ultimately published as Levin, Lin, and Chu (2002); Pesaran and Smith (1995) wrote an early paper that considered **cointegration**. Phillips and Moon (1999) and Pedroni (2004) provide general theory for inference with nonstationary panel data. Analysis is simplest using a **sequential limit theory** where, say, first N is fixed and $T \rightarrow \infty$ and subsequently $N \rightarrow \infty$. A more robust approach uses **joint limits** where $T \rightarrow \infty$ and $N \rightarrow \infty$ simultaneously. Recent reviews of the literature include those by Phillips and Moon (2000) and Baltagi (2001, Chapter 12).

Less consideration has been given to nonstationary data in **short panels**. Harris and Tzavalis (1999) consider the unit root tests of Levin and Lin (1992) in short panels. Let $\hat{\gamma}$ denote the within estimate of γ in the AR(1) fixed effects model $y_{it} = \alpha_i + \gamma y_{i,t-1} + \varepsilon_{it}$, where $\varepsilon_{it} \sim \mathcal{N}[0, \sigma^2]$. We consider the null hypothesis of a unit root, so $\gamma = 1$, and no intercept $\alpha_i = 0$, which corresponds to the pure time series case 2 in Hamilton (1994, p. 490). Under this null hypothesis the unit root test statistic

$$\frac{\sqrt{N}(\hat{\gamma} - 1 + 3/(T+1))}{[3(17T^2 - 20T + 17)]/[5(T-1)(T+1)^3]} \xrightarrow{d} \mathcal{N}[0, 1]$$

as $N \rightarrow \infty$ for fixed T . Large negative values of this statistic lead to rejection of the unit root hypothesis. Levin and Lin (1992) provide additional tests, such as for models with individual time trends.

Binder, Hsiao, and Pesaran (2003) consider short panel estimation of fixed effect dynamic panel models with unit roots and cointegration. With unit roots the Arellano–Bond estimator is inconsistent, though the extensions due to Ahn and Schmidt (1995) and others discussed at the end of Section 22.5.3 yield consistent estimates. Binder et al. (2003) propose quasi-ML estimators that perform better in finite samples when unit roots are present.

22.6. Difference-in-Differences Estimator

The evaluation literature presented in Chapter 25 focuses on measuring the **treatment effect**, in the simplest case the impact or marginal effect of a single binary regressor that equals one if treatment occurs and equals zero if treatment does not occur. For example, interest may lie in measuring the effect on earnings of a policy change (the binary treatment) that alters tax rates or welfare eligibility or access to training for some individuals but not for others.

In this section we relate one of the methods of Chapter 25 to panel methods. Specifically the treatment effect can be measured using standard panel data methods if panel data are available before and after the treatment and if not all individuals receive the treatment. Then the first-differences estimator for the fixed effects model reduces to a simple estimator called the differences-in-differences estimator, introduced in Section 3.4.2 and also studied in Section 25.5. The latter estimator has the advantage that it can also be used when repeated cross-section data rather than panel data are available. However, it does rely on model assumptions that are often not made explicit. The treatment here follows Blundell and MaCurdy (2000).

22.6.1. Fixed Effects with Binary Treatment

Let the binary regressor of interest be

$$D_{it} = \begin{cases} 1 & \text{if individual } i \text{ receives treatment in period } t, \\ 0 & \text{otherwise.} \end{cases} \quad (22.39)$$

Assume a fixed effects model for y_{it} with

$$y_{it} = \phi D_{it} + \delta_t + \alpha_i + \varepsilon_{it}, \quad (22.40)$$

where δ_t is a time-specific fixed effect and α_i is an individual-specific fixed effect. As noted in Section 21.2.1 this is equivalent to regression of y_{it} on D_{it} and a full set of time dummies with the complication of individual-specific fixed effects. For simplicity there are no other regressors.

The individual effects α_i can be eliminated by first differencing. Then

$$\Delta y_{it} = \phi \Delta D_{it} + (\delta_t - \delta_{t-1}) + \Delta \varepsilon_{it}. \quad (22.41)$$

The treatment effect ϕ can be consistently estimated by pooled OLS regression of Δy_{it} on ΔD_{it} and a full set of time dummies.

22.6.2. Differences in Differences

Now consider specialization to only two time periods. Furthermore, suppose treatment occurs only in period 2, so that in period 1 $D_{i1} = 0$ for all individuals and in period 2 $D_{i2} = 1$ for the treated and $D_{i2} = 0$ for the nontreated. Then the subscript t can be dropped from (22.41) and

$$\Delta y_i = \phi D_i + \delta + v_i, \quad (22.42)$$

where D_i is a binary treatment variable indicating whether or not the individual received treatment.

The treatment effect can be estimated by OLS regression of Δy on an intercept and the binary regressor D . Define $\bar{\Delta y}^{\text{tr}}$ to denote the sample average of Δy_i for the treated ($D_i = 1$) and $\bar{\Delta y}^{\text{nt}}$ to denote the sample average of Δy_i for the nontreated ($D_i = 0$). Then the OLS estimator reduces to

$$\hat{\phi} = \bar{\Delta y}^{\text{tr}} - \bar{\Delta y}^{\text{nt}}. \quad (22.43)$$

This estimator is called the **differences-in-differences (DID) estimator**, since one estimates the time difference for the treated and untreated groups and then takes the difference in the time differences.

The estimator is appealing for its intuitive simplicity. Additionally, it can be extended from panel data to the case where separate cross sections are available in the two periods. In the second period compute the averages \bar{y}_2^{tr} and \bar{y}_2^{nt} for the treated and untreated groups. Compute similar averages \bar{y}_1^{tr} and \bar{y}_1^{nt} in the first pretreatment period. This assumes that it is possible to identify in the first period whether or not an individual is eligible for treatment. This is easy if, for example, the treatment applies only to women and data on gender are available. Then compute

$$\hat{\phi} = (\bar{y}_2^{\text{tr}} - \bar{y}_1^{\text{tr}}) - (\bar{y}_2^{\text{nt}} - \bar{y}_1^{\text{nt}}). \quad (22.44)$$

As an example, if average annual earnings for the group eligible for treatment equals 10,000 before treatment and 13,000 after treatment then $\bar{y}_2^{\text{tr}} - \bar{y}_1^{\text{tr}} = 3,000$. Similarly, if average annual earnings for the group not eligible for treatment equals 15,000 before treatment and 17,000 after treatment then $\bar{y}_2^{\text{nt}} - \bar{y}_1^{\text{nt}} = 2,000$. The DID estimate of the treatment effect $\hat{\phi}$ is then $3,000 - 2,000 = 1,000$.

22.6.3. Assumptions Underlying Differences in Differences

The preceding formulation of the DID estimator makes explicit the underlying assumptions for consistent estimation of ϕ .

First, it is assumed that the time effects δ_t are common across treated and untreated individuals. For example, time trends may differ by gender, in which case identifying ϕ is problematic if treatment depends on gender. The common trends assumption is needed if either panel or cross-section data are used.

Second, if cross-section data are used then the composition of the treated and untreated groups is assumed to be stable before and after the change. With panel data differencing eliminates the fixed effects α_i . With repeated cross-section data the original model (22.40) implies that $\bar{y}_t^{\text{tr}} = \phi + \delta_t + \bar{\alpha}_t^{\text{tr}} + \bar{\varepsilon}_t^{\text{tr}}$ and $\bar{y}_t^{\text{nt}} = \delta_t + \bar{\alpha}_t^{\text{nt}} + \bar{\varepsilon}_t^{\text{nt}}$. Given that treatment only occurs in the second period it follows that

$$\phi = (\bar{y}_2^{\text{tr}} - \bar{y}_1^{\text{tr}}) - (\bar{y}_2^{\text{nt}} - \bar{y}_1^{\text{nt}}) + (\bar{\alpha}_2^{\text{tr}} - \bar{\alpha}_1^{\text{tr}}) - (\bar{\alpha}_2^{\text{nt}} - \bar{\alpha}_1^{\text{nt}}) + v,$$

where $v = (\bar{\varepsilon}_2^{\text{nt}} - \bar{\varepsilon}_1^{\text{nt}}) - (\bar{\varepsilon}_2^{\text{tr}} - \bar{\varepsilon}_1^{\text{tr}})$. Consistency of $\hat{\phi}$ in (22.44) occurs if $\text{plim}(\bar{\alpha}_2^{\text{tr}} - \bar{\alpha}_1^{\text{tr}}) = 0$ and $\text{plim}(\bar{\alpha}_2^{\text{nt}} - \bar{\alpha}_1^{\text{nt}}) = 0$. This will happen if assignment to treatment is random. However, often this is not the case.

22.6.4. Richer Models

In practice richer models are used. An obvious extension is to include regressors other than the treatment indicator and time dummies. By grouping data the individual-specific effects can at least be permitted to differ on average across groups. The general procedure is to estimate

$$y_{igt} = \phi D_{igt} + \delta_t + \alpha_i + \varepsilon_{it},$$

where g denotes the g th group.

In a classic example of DID estimation, Card (1990) studied the effect on unemployment of low-wage workers in Miami of a sudden influx of immigrants from Cuba. This example is also reviewed in Angrist and Krueger (1999). Athey and Imbens (2002) present extension to nonlinear models.

22.7. Repeated Cross Sections and Pseudo Panels

The key potential advantages of panel data arise from being able to observe subjects over time. This makes it possible to control for unobserved individual heterogeneity, differences in initial conditions, and dynamic dependence of outcomes. In many cases, however, genuine panel data are unavailable.

22.7.1. Repeated Cross Sections

We consider analysis when data are for several **repeated cross sections**, derived from responses to a series of independent sample surveys, where independence means that

each subject appears in only one survey. An example is the U.K. Family Expenditure Survey, which collects a large annual sample of household expenditure data but each year surveys different families. Also, if only a very short panel is available (e.g., $T = 2$) then data from repeated cross sections are appealing if they can generate a larger and richer sample.

For a **random effects** model repeated cross-section data pose no challenges. One simply performs a pooled regression of y_{it} on \mathbf{x}_{it} (see Section 21.5) and statistical inference is actually simplified as correction is needed only for heteroskedasticity since here errors are independent over both i and t .

With **fixed effects**, however, pooled regression leads to inconsistent parameter estimates. Furthermore, alternative methods such as the within or first-differences estimation are infeasible if individuals are observed at only one point in time. In this section repeated cross-section data are used to construct **pseudo panels** or **synthetic panel data** that have some of the advantages of genuine panel data, most notably the ability to control for fixed effects. A special case is the DID estimator presented in Section 22.6.

22.7.2. Pseudo Panels

Browning, Deaton, and Irish (1985) and Deaton (1985), in their empirical studies based on the U.K. Family Expenditure Survey, considered methods for analyzing repeated cross-section data. Their suggestion was to convert the individual-level data into **cohort-level** data. Although individual household expenditures cannot be tracked through time, it is possible to do so for cohorts of individuals.

A **cohort** is defined as “a group with fixed membership, individuals of which can be identified as they show up in the surveys” (Deaton, 1985, p. 109). An example is an age cohort such as males born between 1965 and 1970. For large samples, successive surveys will generate random samples of members of each cohort.

Time series of sample averages of cohorts can form the basis of regression models. Whether synthetic panels based on cohort data can substitute for genuine panel data is a key issue. The topic of repeated cross section deals with inference procedures for such models. Here we focus on static pseudo panel models. Collado (1997) and Girma (2000) also consider the dynamic case.

The starting point is the static linear regression with individual fixed effects α_i , based on T successive cross sections,

$$y_{it} = \alpha_i + \mathbf{x}'_{it}\beta + u_{it}, \quad t = 1, \dots, T. \quad (22.45)$$

The explanatory variables are assumed to be strongly exogenous with respect to parameters of interest, β , so $E[\mathbf{x}'_{it}u_{is}] = \mathbf{0}, \forall t, s$. For simplicity, we assume that N observations are available for each cross section. Each individual is observed in only one time period, so the individual-specific effects α_i cannot be swept out by differencing the individual-level data.

Let g be a random variable that determines cohort membership for each i , such that i belongs to cluster c if and only if g_i belongs to the set I_c . Assume that there are C

cohorts, and c is the cohort subscript, $c = 1, \dots, C$. Taking expectations conditional on g_i yields

$$E[y_{it}|g_i \in I_c] = E[\alpha_i|g_i \in I_c] + E[\mathbf{x}'_{it}|g_i \in I_c]\beta + E[u_{it}|g_i \in I_c]. \quad (22.46)$$

This generates a **cohort population** version of the model (22.45) given by

$$y_{ct}^* = \alpha_c^* + \mathbf{x}'_{ct}\beta + u_{ct}^*, \quad (22.47)$$

where the asterisks denote unobservable population cohort averages. For example, $y_{ct}^* = E[y_{it}|g_i \in I_c]$.

The parameter $\alpha_c^* = E[\alpha_i|g_i \in I_c]$ is the **cohort fixed effect**. An important assumption made in the case of fixed effects is that the population is stationary so that α_c^* can be assumed to be constant over time. This is qualitatively similar to the assumption needed for consistency of the DID estimator made at the end of Section 22.6.3. Under the usual weak exogeneity assumptions $E[u_{ct}^*|\mathbf{x}_{ct}^*] = 0$. However, the unobserved fixed effect α_c^* will be correlated with \mathbf{x}_{ct}^* if α_i is correlated with \mathbf{x}_{it} in the original model (22.45). Estimation needs to control for the fixed effect.

In practice the population cohort means are unobservable and we instead work with **cohort-time averages** \bar{y}_{ct} and $\bar{\mathbf{x}}_c$. The regression is then

$$\bar{y}_{ct} = \bar{\alpha}_c + \bar{\mathbf{x}}'_c\beta + \bar{u}_{ct}, \quad c = 1, \dots, C, \quad t = 1, \dots, T. \quad (22.48)$$

This step introduces an additional source of error, since \bar{y}_{ct} and $\bar{\mathbf{x}}_c$ are error-contaminated estimates of the population cohort averages, that is,

$$\begin{aligned} \bar{y}_{ct} &= y_{ct}^* + \xi_{ct}, \\ \bar{\mathbf{x}}_c &= \mathbf{x}_{ct}^* + \mathbf{v}_{ct}. \end{aligned} \quad (22.49)$$

If the **measurement error** is very small, owing to the number of observations per cohort per time period (N_{ct}) being very large, then $\bar{y}_{ct} \approx y_{ct}^*$ and $\bar{\mathbf{x}}_c = \mathbf{x}_{ct}^*$ and the measurement error can be ignored. A consistent estimate of β can be obtained by within estimation of (22.48), that is, OLS regression of $(\bar{y}_{ct} - \bar{y}_c)$ on $(\bar{\mathbf{x}}_c - \bar{\mathbf{x}}_c)$, where $\bar{y}_c = T^{-1} \sum_t \bar{y}_{ct}$ and $\bar{\mathbf{x}}_c = T^{-1} \sum_t \bar{\mathbf{x}}_c$.

Unfortunately, the measurement error is often too large to ignore. Then within estimation of (22.48), or even OLS estimation of (22.48) when $\bar{\alpha}_c$ is a random effect, leads to inconsistent estimation of β . Instead, errors-in-variables estimators need to be used. These can be implemented here since the individual-level data yield necessary estimates of the moments of the measurement error, see Section 26.3.3.

22.7.3. Measurement Error Estimators for Pseudo Panels

A classic solution to measurement errors is to use replicated observations to estimate the covariance matrix of the measurement error, and to then use these estimates to “correct” the sample moments of the error-contaminated variables before applying the least-squares procedure (see Section 26.3.4). Deaton (1985) proposed using this method in the current setting.

Assume that individual observations satisfy the equations

$$\begin{aligned} y_{it} &= y_{ct}^* + \varepsilon_{it} \\ \mathbf{x}_{it} &= \mathbf{x}_{ct}^* + \eta_{it}, \end{aligned}$$

a setup similar to that in Section 26.2.1, except that there is also measurement error in the dependent variable, and assume that for any individual in a given cohort c ,

$$\begin{bmatrix} \varepsilon_{it} \\ \eta_{it} \end{bmatrix} \sim \text{iid} \left[\begin{bmatrix} 0 \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01}' \\ \sigma_{01} & \Sigma \end{bmatrix} \right].$$

Sample estimates of (Σ, σ_{01}) , denoted $(\widehat{\Sigma}, \widehat{\sigma}_{01})$, can be obtained given $(\bar{y}_{ct}, \bar{\mathbf{x}}_{ct})$ from using all individual-level data. Define \mathbf{d}_c to be the $C \times 1$ column vector of dummy variables corresponding to the fixed effects α_c^* (see Section 21.2.1), which is a regressor vector that is clearly not subject to estimation error. Then provided T is sufficiently large and the relevant inverses exist, the regression

$$\begin{bmatrix} \widehat{\alpha}_{ct} \\ \widehat{\beta}_{ct} \end{bmatrix} = \left(\sum_{c=1}^C \sum_{t=1}^T \begin{bmatrix} \mathbf{d}_c' \mathbf{d}_c & \mathbf{d}_c' \bar{\mathbf{x}}_{ct} \\ \bar{\mathbf{x}}_{ct}' \mathbf{d}_c & \bar{\mathbf{x}}_{ct}' \bar{\mathbf{x}}_{ct} - \widehat{\Sigma} \end{bmatrix} \right)^{-1} \left[\sum_{c=1}^C \sum_{t=1}^T \left(\bar{\mathbf{x}}_{ct}' \mathbf{d}_c' \bar{y}_{ct} - \widehat{\sigma}_{01} \right) \right] \quad (22.50)$$

will provide consistent estimates of the cohort regression as $CT \rightarrow \infty$. This estimator is the same as that given in Section 26.3.4, with adaptation here because \bar{y}_{ct} is also measured with error and with simplification because only a subset of the regressors, $\bar{\mathbf{x}}_{ct}$, is measured with error. Verbeek and Nijman (1992) provide a more detailed discussion of the sampling properties, and Deaton (1985) presents variance estimation. See also Verbeek (1995).

The preceding estimator essentially controls for the cohort fixed effects by estimating the least-squares dummy variable model, adjusting for measurement error by use of replicated data using the estimator given in Section 26.3.4.

Collado (1997) considered an alternative approach of eliminating the cohort effects by first differencing, and then controlling for measurement error through instrumental variables estimation, an alternative identification strategy for measurement error given in Section 26.3.2.

Substituting (22.49) into (22.47) gives

$$\begin{aligned} \bar{y}_{ct} - \xi_{ct} &= \alpha_c^* + (\bar{\mathbf{x}}_{ct}' - \mathbf{v}_{ct}') \beta + u_{ct}^*, \\ \bar{y}_{ct} &= \alpha_c^* + \bar{\mathbf{x}}_{ct}' \beta + w_{ct}, \end{aligned}$$

where the error $w_{ct} = u_{ct}^* - \mathbf{v}_{ct}' \beta + \xi_{ct}$. First differencing eliminates α_c^* , leading to

$$\Delta \bar{y}_{ct} = \Delta \bar{\mathbf{x}}_{ct}' \beta + \Delta w_{ct}, \quad t = 2, \dots, T. \quad (22.51)$$

Now because of the measurement error terms the explanatory variables $\Delta \bar{\mathbf{x}}_{ct}'$ will be correlated with Δw_{ct} , and hence applying least squares will lead to inconsistent estimation. Consistent estimates can be obtained by IV estimation based on lagged levels of exogenous variables, that is, $\bar{\mathbf{x}}_{c,t-1}'$. This approach has the advantage of ready extension to models with lagged dependent variables. For details see Collado (1997).

22.8. Mixed Linear Models

The model called the random effects model by econometricians specifies only the intercept coefficient to be random. Richer random effects models, widely used in other areas of applied statistics, additionally permit the slope parameters to be random. In this section we present mixed linear models – also called mixed effects models, hierarchical, or multilevel linear models (see Chapter 24), random coefficients models, and variance components models.

These models are applied in a setting where the pooled OLS estimator is still consistent. In particular, there are no fixed effects. Because the mixed linear models framework provides enough structure to permit estimation by feasible GLS, its estimates are more efficient.

22.8.1. Mixed Linear Models

The **mixed linear model** specifies

$$y_{it} = \mathbf{z}'_{it}\boldsymbol{\beta} + \mathbf{w}'_{it}\boldsymbol{\alpha}_i + \varepsilon_{it}, \quad (22.52)$$

where the regressors \mathbf{z}_{it} include an intercept, \mathbf{w}_{it} is a vector of observable characteristics, $\boldsymbol{\alpha}_i$ is a random zero-mean vector, and ε_{it} is an error term. This model is called a **mixed model** as it has both **fixed parameters** $\boldsymbol{\beta}$ and zero-mean **random parameters** or **random effects** $\boldsymbol{\alpha}_i$.

The random intercept model $y_{it} = \mathbf{z}'_{it}\boldsymbol{\beta} + \alpha_i + \varepsilon_{it}$ is a special case of (22.52) with $\mathbf{w}'_{it}\boldsymbol{\alpha}_i = \alpha_i$.

Another special case of (22.52) is the **random coefficients model** or **random parameters model**. In the regression setting we suppose that

$$y_{it} = \mathbf{z}'_{it}\boldsymbol{\beta}_i + \varepsilon_{it},$$

a regular linear regression, except that the regression parameter vector now differs across individuals according to

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\alpha}_i,$$

where $\boldsymbol{\alpha}_i$ is a zero-mean random vector. Substitution yields $y_{it} = \mathbf{z}'_{it}\boldsymbol{\beta} + \mathbf{z}'_{it}\boldsymbol{\alpha}_i + \varepsilon_{it}$, which is (22.52) with $\mathbf{w}_{it} = \mathbf{z}_{it}$.

Many applications lie between random intercept and random coefficients models, with \mathbf{w}_{it} often a subset of \mathbf{z}_{it} . In particular, standard mixed and random **ANOVA** models are also a special case, where the k th component of the vector \mathbf{w}_{it} is either zero or one, according to various possible models for clustering the data. For example, one of the components in \mathbf{z}_{it} may be a race or gender indicator variable. Then the conditional mean of y_{it} varies with gender or race. It may also be felt that the conditional variance of y_{it} also varies with gender or race, which can be captured by inclusion in \mathbf{w}_{it} . The mixed model is an outgrowth of ANOVA models. The **hierarchical linear model** or **multi-level linear model** (see Section 24.6.2) can also be expressed as a special case of (22.52).

22.8.2. Estimation

The goal is to estimate the fixed regression parameters β and the variances and covariance parameters of the distributions for α_i and ε_{it} . One of the early treatments of this model was in a Bayesian context by Lindley and Smith (1972). A simple example of their general treatment was the random coefficients model with $y_{it} \sim \mathcal{N}[\mathbf{z}'_{it}\beta_i, \sigma^2]$, where $\beta_i \sim \mathcal{N}[\gamma, \Gamma]$. See Koop (2003), for example, for **Bayesian analysis** of the linear panel data model.

Here we follow the **classical approach**, based on the work of Harville (1977), who gives references to the earlier literature. The mixed model (22.52) can be split into a deterministic component $\mathbf{x}'_{it}\beta$ and a random component $\mathbf{w}'_{it}\alpha_i + \varepsilon_{it}$. The stochastic assumptions include the assumption that the regressors \mathbf{x}_{it} are independent of the zero-mean random components α_i and ε_{it} . So pooled OLS regression of y_{it} on \mathbf{x}_{it} provides consistent estimates of β . We are essentially in the world of Section 21.5, with feasible GLS estimation possible as structure has been placed on the variance matrix of the error term $\mathbf{w}'_{it}\alpha_i + \varepsilon_{it}$. In this section we present the feasible GLS estimator along with two different methods to estimate the variances and covariances of α_i and ε_{it} and consider prediction of the random components α_i .

Combine observations over time for a given individual in the usual way, so that (22.52) becomes

$$\mathbf{y}_i = \mathbf{Z}_i\beta + (\mathbf{W}_i\alpha_i + \varepsilon_i). \quad (22.53)$$

The usual assumptions are that α_i and ε_i are independent over i and independent of each other with $\alpha_i \sim [\mathbf{0}, \Sigma_\alpha]$ and $\varepsilon_i \sim [\mathbf{0}, \Sigma_\varepsilon]$, so that the error term

$$\mathbf{W}_i\alpha_i + \varepsilon_i \sim [\mathbf{0}, \Omega_i = \mathbf{W}_i\Sigma_\alpha\mathbf{W}'_i + \Sigma_\varepsilon].$$

Then the **feasible GLS estimator** is

$$\hat{\beta}_{\text{FGLS}} = \left[\sum_{i=1}^N \mathbf{Z}_i' \hat{\Omega}_i^{-1} \mathbf{Z}_i \right]^{-1} \sum_{i=1}^N \mathbf{Z}_i' \hat{\Omega}_i^{-1} \mathbf{y}_i, \quad (22.54)$$

where $\hat{\Omega}_i$ is consistent for Ω_i .

Implementation requires consistent estimation of Ω_i . This has already been discussed in Section 21.7 for the simpler case of a random intercept, in which case there were several different ways to consistently estimate the variance components σ_α^2 and σ_ε^2 , with complications such as bias and the possibility of negative estimates. Similar issues arise here in estimation of Σ_α and Σ_ε .

We present two estimators based on the additional assumption of normal distribution for the random components. The presentation is for the more general model

$$\mathbf{y} = \mathbf{Z}\beta + (\mathbf{W}\alpha + \varepsilon), \quad (22.55)$$

which can be obtained, for example, by appropriate stacking of (22.53). It is assumed that $\alpha \sim \mathcal{N}[\mathbf{0}, \mathbf{G}]$ and $\varepsilon \sim \mathcal{N}[\mathbf{0}, \mathbf{R}]$, where in the current application \mathbf{G} and \mathbf{R} are functions of Σ_α and Σ_ε . The **feasible GLS estimator** for the mixed model is

$$\hat{\beta}_{\text{FGLS}} = [\mathbf{Z}'\hat{\mathbf{V}}^{-1}\mathbf{Z}]^{-1} \mathbf{Z}'\hat{\mathbf{V}}^{-1}\mathbf{y},$$

where $\hat{\mathbf{V}}$ is consistent for $\mathbf{V} = \mathbf{V}[\mathbf{W}\alpha + \varepsilon] = \mathbf{W}\mathbf{G}\mathbf{W}' + \mathbf{R}$. See Swamy (1970).

The obvious method for obtaining $\widehat{\mathbf{V}}$ is maximum likelihood. The log-likelihood function based on the multivariate normal, after concentrating out β which is equal to the GLS estimator $[\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{y}$, is

$$\ln L(\mathbf{G}, \mathbf{R}) = -\frac{1}{2} \ln |\mathbf{V}| - \frac{NT}{2} \ln \mathbf{r}'\mathbf{V}^{-1}\mathbf{r} - \frac{NT}{2} \left[1 + \ln \left(\frac{2\pi}{NT} \right) \right],$$

where $\mathbf{r} = \mathbf{y} - \mathbf{Z}[\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{y}$ and $|\mathbf{V}|$ denotes the determinant of \mathbf{V} . Maximization with respect to the parameters in \mathbf{G} and \mathbf{R} yields $\widehat{\mathbf{V}} = \mathbf{W}\widehat{\mathbf{G}}\mathbf{W}' + \widehat{\mathbf{R}}$.

A weakness of ML estimates of variance components are that they are biased in small samples. For example, for cross-section linear regression with homoskedastic errors the MLE $\widehat{\sigma}^2 = N^{-1} \sum_i \widehat{u}_i^2$ is biased and it is better to instead divide by $(N - K)$. For the model (22.53), degree-of-freedom corrections are provided by the **restricted maximum likelihood** estimator that instead maximizes

$$\begin{aligned} \ln L_R(\mathbf{G}, \mathbf{R}) &= -\frac{1}{2} \ln |\mathbf{V}| - \frac{NT - p}{2} \ln \mathbf{r}'\mathbf{V}^{-1}\mathbf{r} - \frac{NT - p}{2} \left[1 + \ln \left(\frac{2\pi}{NT - p} \right) \right] \\ &\quad - \frac{1}{2} \ln |\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}|, \end{aligned}$$

where p is the rank of \mathbf{Z} . For motivation of $\ln L_R(\mathbf{G}, \mathbf{R})$ see Harville (1977).

As an empirical example of a mixed linear model, consider the ln(hours)-ln(wage) regression example of Section 21.3 with both the intercept and slope parameters permitted to be random. Then the random coefficients model yields $\ln\text{hrs} = 7.734 - 0.021\ln\text{wg}$ with slope coefficient standard error of 0.046 (default) or 0.020 (panel bootstrap). The slope coefficient is quite different from the estimates given in Table 21.2.

22.8.3. Prediction

We may wish to **predict** the random parameters α in addition to the fixed parameters β and the covariance parameters.

The joint normal equations for $\widehat{\beta}$ and $\widehat{\alpha}$, given consistent estimates of $\widehat{\beta}$ and $\widehat{\alpha}$, can be written as

$$\begin{bmatrix} \mathbf{Z}'\widehat{\mathbf{R}}^{-1}\mathbf{Z} & \mathbf{Z}'\widehat{\mathbf{R}}^{-1}\mathbf{W} \\ \mathbf{W}'\widehat{\mathbf{R}}^{-1}\mathbf{Z} & \mathbf{W}'\widehat{\mathbf{R}}^{-1}\mathbf{W} + \widehat{\mathbf{G}}^{-1} \end{bmatrix} \begin{bmatrix} \widehat{\beta} \\ \widehat{\alpha} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}'\widehat{\mathbf{R}}^{-1}\mathbf{y} \\ \mathbf{W}'\widehat{\mathbf{R}}^{-1}\mathbf{y} \end{bmatrix}.$$

Solving for $\widehat{\beta}$ gives $\widehat{\beta}_{\text{FGLS}}$ given earlier, whereas

$$\widehat{\alpha} = \widehat{\mathbf{G}}\mathbf{W}'\widehat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{Z}'\widehat{\beta}).$$

In the case of independence over i , this yields $\widehat{\alpha}_i = \widehat{\Sigma}_\alpha \mathbf{W}'_i \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{Z}'_i \widehat{\beta})$. This is the **best linear unbiased predictor** if the variance matrices are known.

22.9. Practical Considerations

The panel 2SLS estimators can actually be estimated using just a 2SLS program for cross-section data (see Section 22.2.5) though computed standard errors need to be

panel robust. Optimal GMM estimators can be implemented using matrix commands in a statistical package or in a programming language such as GAUSS. Statistical packages are increasingly offering panel commands that automatically implement the estimators of this chapter, most notably the Arellano–Bond estimator.

22.10. Bibliographic Notes

This chapter covers an active area of research that appears in several recent texts devoted to panel data, notably those by Baltagi (1995, 2001), Hsiao (1986, 2003), M–J. Lee (2002), and Arellano (2003). More advanced methods are given in Matyas and Sevestre (1995) and in Arellano and Honore (2001).

- 22.2** Chamberlain (1982, 1984) emphasized the use of exogeneity assumptions. He used minimum distance estimation. The subsequent literature has used GMM methods. M–J. Lee (2002) and Arellano (2003) especially emphasize GMM estimation. See also the survey by Ahn and Schmidt (1999).
- 22.4** The model of Hausman and Taylor (1981) is attractive. By assuming that some regressors are uncorrelated with the individual-specific effect it permits identification of the coefficients of time-invariant regressors.
- 22.5** The coverage of linear dynamic models is very brief compared to the size of the literature that began with Balestra and Nerlove (1966). More complete discussions are given in Baltagi (2001, Chapter 8), Hsiao (2003, Chapter 4), and Arellano (2003, Chapter 5–8). The Arellano–Bond (1991) estimator is especially popular as it accommodates dynamic models with fixed effects.
- 22.6** The difference-in-differences approach is very popular because of its simplicity. Although it can be used with repeated cross-section rather than panel data, a panel data interpretation helps make explicit the underlying assumptions. Bertrand et al. (2004) demonstrate the importance of correcting for time series correlation at the individual level using the methods of Section 22.2.3.
- 22.8** Mixed linear models are especially popular in the statistics literature. They are less used in the econometrics literature, because of the reluctance to impose structure on the time-invariant individual-specific fixed effect.

Exercises

- 22–1** Consider the panel GMM estimator of Section 22.2.1.

- (a) Show that minimization with respect to β of the quadratic function $Q_N(\beta)$ given after (22.3) yields the panel GMM estimator given after $Q_N(\beta)$ that is expressed using summation notation.
- (b) Show that this estimator is equivalent to the estimator defined in (22.4).
- (c) For simplicity suppose that the matrices Z and X in (22.4) are nonstochastic and that $y = X\beta + u$ where u has mean 0 and variance Ω . Obtain the finite sample variance matrix of the estimator in (22.4) and compare this to the asymptotic results in (22.5).
- (d) Simplify the panel GMM estimator in the case that $r = K$.

- 22–2** Consider the panel data model $y_{it} = \alpha + \beta x_{it} + \gamma w_{it} + u_{it}$, $i = 1, \dots, N$, $t = 1, \dots, T$, where for simplicity there is no individual-specific effect. Suppose the

scalar regressor x_{it} is correlated with u_{is} for all t and s . For each of the following parts state whether consistent IV estimation of β and γ is possible, and if so give all the suitable instruments, based on the discussion in Section 22.2. Assume that three periods of data are available, so $T = 3$, and note that a variable may not be available as an instrument in all years, and that in different years different instruments may be available.

- (a) The regressor w_{it} satisfies the summation assumption $E[\sum_t w_{it}u_{it}] = 0$.
- (b) The regressor w_{it} satisfies the contemporaneous exogeneity assumption $E[w_{it}u_{it}] = 0, t = 1, \dots, 3$.
- (c) The regressor w_{it} satisfies the weak exogeneity assumption $E[w_{is}u_{it}] = 0, s \leq t, t = 1, \dots, 3$.
- (d) The regressor w_{it} satisfies the strong exogeneity assumption $E[w_{it}u_{it}] = 0, s, t = 1, \dots, 3$.

22–3 Repeat question 3, again with three periods of data, but now consider the panel data model $y_{it} = \alpha_i + \beta x_{it} + \gamma w_{it} + u_{it}$, where α_i is a fixed effect, and consider IV estimation based on the first differences model, $y_{it} - y_{i,t-1} = \beta(x_{it} - x_{i,t-1}) + \gamma(w_{it} - w_{i,t-1}) + (u_{it} - u_{i,t-1})$.

22–4 Consider the differences in differences (DID) estimator presented in Section 22.6. Suppose the time trend term ($\delta_t - \delta_{t-1}$) differs across the treated and untreated groups.

- (a) Will the DID estimator of ϕ based on repeated cross-section data be consistent? Explain your answer.
- (b) Is consistent estimation of ϕ possible if panel data are available? Explain your answer.

22–5 Using the hours and wages data of Ziliak (1997) reproduce as much of Table 22.2 as you can, with appropriate discussion, when the instrument set is expanded to include the third lags of lnwg, kids, age, agesq, and disab and the seven years 1982–88 are used to estimate (22.22).

Nonlinear Panel Models

23.1. Introduction

This chapter extends the linear model panel data methods of Chapters 21 and 22 to the nonlinear regression models presented in Chapters 14–20. We focus on short panels and models with a time-invariant individual-specific effect that may be fixed or may be random. Both static and dynamic models are considered.

There is no one-size-fits-all prescription for nonlinear models with individual specific effects. If individual-specific effects are fixed and the panel is short then consistent estimation of the slope parameters is possible for only a subset of nonlinear models. If individual-specific effects are instead purely random then consistent estimation is possible for a wide range of models.

Section 23.2 presents general approaches that may or may not be implementable for particular models. Section 23.3 provides an application to a nonlinear model with multiplicative individual-specific effects. Specializations to the leading classes of nonlinear models – discrete data, selection models, transition data, and count data – are presented in Sections 23.4–23.7. Semiparametric estimation is surveyed in Section 23.8.

23.2. General Results

General approaches to extending the methods for linear models are presented in this section. We first present the various models – fixed effects, random effects, and pooled models, distinguishing parametric from conditional mean models. Methods to estimate these models and obtain panel-robust standard errors are then presented. Further details for specific nonlinear panel models are provided in subsequent sections.

23.2.1. Individual-Specific Effects Models

The linear individual-specific effects model (see Section 21.2.1) specifies that the dependent variable y_{it} depends on a time-invariant individual-specific effect α_i , as

well as the usual regressors \mathbf{x}_{it} and regression parameters β . The model is written as $y_{it} = \alpha_i + \mathbf{x}'_{it}\beta + u_{it}$, where u_{it} is an error term.

For nonlinear models such as logit and Poisson models there is less motivation for introducing an additive error u_{it} . Instead, it is more natural to directly model the conditional density, or the conditional mean, which in the linear case can be expressed as $E[y_{it}|\alpha_i, \mathbf{x}_{it}] = \alpha_i + \mathbf{x}'_{it}\beta$.

Parametric Models

A fully parametric approach is common for many nonlinear models, most notably models for binary, multinomial, and censored outcomes given in Chapters 14–16.

The standard cross-section models are single-index models, or single-index models with additional scale parameter(s). The **parametric individual-specific effects models** presented in subsequent sections specify conditional density

$$f(y_{it}|\alpha_i, \mathbf{x}_{it}) = f(y_{it}, \alpha_i + \mathbf{x}'_{it}\beta, \gamma), \quad (23.1)$$

where γ denotes additional parameters such as variance parameters. The model is a single-index model in the regressors \mathbf{x}_{it} and the individual effects α_i .

The usual assumption is that $y_{it}|\mathbf{x}_{it}, \alpha_i$ is independent over both i and t . This can be relaxed to permit dependence over t for given i (see Section 23.2.6).

Conditional Mean Models

A quite general nonlinear model for the conditional mean, with unobserved time-invariant individual-specific effect, is

$$E[y_{it}|\alpha_i, \mathbf{x}_{it}] = g(\alpha_i, \mathbf{x}_{it}, \beta), \quad i = 1, \dots, N, \quad t = 1, \dots, T, \quad (23.2)$$

for given function $g(\cdot)$. Three common specifications are an **additive individual-specific effects model**

$$g(\alpha_i, \mathbf{x}_{it}, \beta) = \alpha_i + g(\mathbf{x}_{it}, \beta), \quad (23.3)$$

a **multiplicative individual-specific effects model**,

$$g(\alpha_i, \mathbf{x}_{it}, \beta) = \alpha_i g(\mathbf{x}_{it}, \beta), \quad (23.4)$$

and a **single-index individual-specific effects model**

$$g(\alpha_i, \mathbf{x}_{it}, \beta) = g(\alpha_i + \mathbf{x}'_{it}\beta). \quad (23.5)$$

In each case the function $g(\cdot)$ is specified. The regressors \mathbf{x}_{it} may be time varying or time-invariant and may include a time dummy.

The additive effects model is suited to applications where the range of y_{it} is unbounded, as implicitly assumed with linear regression. The multiplicative effects model is suited to applications where y_{it} is nonnegative unbounded, such as count data, in which case $\alpha_i > 0$ and $g(\cdot) > 0$. The single-index model is a natural starting point for the probit model, for example, with $g(\alpha_i + \mathbf{x}'_{it}\beta) = \Phi(\alpha_i + \mathbf{x}'_{it}\beta)$, where $\Phi(\cdot)$ is the standard normal cdf. The single-index model reduces to the additive model if $g(\cdot)$ is

the identity function. It reduces to the multiplicative model if $g(\cdot)$ is the exponential function, since then $\exp(\alpha_i + \mathbf{x}'_{it}\beta) = \exp(\alpha_i)\exp(\mathbf{x}'_{it}\beta)$.

The moment condition (23.2) conditions only on current period \mathbf{x}_{it} and assumes that regressors are **contemporaneously exogenous** (see Section 22.2.4). Elimination of the individual-specific effects α_i can require stronger exogeneity assumptions. Regressors are **weakly exogenous** if

$$E[y_{it}|\alpha_i, \mathbf{x}_{i1}, \dots, \mathbf{x}_{it}] = g(\alpha_i, \mathbf{x}_{it}, \beta) \quad (23.6)$$

and **strongly exogenous** or **strictly exogenous** if

$$E[y_{it}|\alpha_i, \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}] = g(\alpha_i, \mathbf{x}_{it}, \beta). \quad (23.7)$$

A nonlinear model with additive effects adds relatively few complications. In particular, if the panel model is $y_{it} = \alpha_i + g(\mathbf{x}_{it}, \beta) + u_{it}$, then the approaches of Chapters 21 and 22 will carry through with some modification, including estimation by nonlinear LS and IV rather than linear LS and IV.

This chapter focuses on models with nonadditive individual-specific effects, such as in (23.4) and (23.5). These effects can be treated as fixed effects or as random effects.

23.2.2. Fixed Effects Models

A **fixed effects model** treats the individual-specific effect α_i as an unobserved random variable that may be correlated with the regressors \mathbf{x}_{it} . In short panels joint estimation of the fixed effects $\alpha_1, \dots, \alpha_N$ and the other model parameters, β and possibly γ , generally leads to inconsistent estimation of all parameters. Instead, a variety of methods have been proposed that eliminate the fixed effects in some special settings, permitting consistent estimation of the other model parameters.

The Incidental Parameters Problem

Neyman and Scott (1948) considered inference when some parameters are common to all observations but there are additionally an infinity of parameters, each of which depends on only a finite number of observations. The **common parameters** are of intrinsic interest, whereas the latter parameters are called **incidental parameters**.

Here β and γ are common parameters, but $\alpha_1, \dots, \alpha_N$ are incidental parameters if the panel is short as then each α_i depends on fixed T observations and there are infinitely many α_i since $N \rightarrow \infty$. The incidental parameters are inconsistently estimated as $N \rightarrow \infty$, since only T observations are used to estimate each parameter. The **incidental parameters problem** is that this contaminates the estimation of the common parameters. In general the common parameters are also inconsistently estimated, even though they are finite in number and are estimated using $NT \rightarrow \infty$ observations.

A simple illustration of contamination by incidental parameters is to suppose that $y_{it} \sim \mathcal{N}[\alpha_i, \sigma^2]$. Maximum likelihood estimation yields $\hat{\alpha}_i = \bar{y}_i$, $i = 1, \dots, N$, and $\hat{\sigma}^2 = (NT)^{-1} \sum_i \sum_t (y_{it} - \bar{y}_i)^2$. Then $E[\hat{\sigma}^2] = \sigma^2(T-1)/T$, so $\hat{\sigma}^2$ is inconsistent for σ^2 as $N \rightarrow \infty$ in the short panel setting of fixed T . This inconsistency can be very large, with $\hat{\sigma}^2 \xrightarrow{P} 0.5\sigma^2$ when $T = 2$.

In general if there is an incidental parameters problem, alternative estimation methods are needed that first eliminate the incidental parameters. For some popular models, most notably the panel probit model, there is no solution to the incidental parameters problem. Even where methods exist to consistently estimate β these methods tend to be model specific, as emphasized by Lancaster (2000). No unified solution to the incidental parameters problem exists.

Conditional Likelihood

A statistic t is called **sufficient** for a parameter θ if the distribution of the sample given t does not depend on θ . For individual-specific effects panel models, if a sufficient statistic exists for the nuisance parameter α_i then by conditioning on this sufficient statistic the nuisance parameter is eliminated. The resulting conditional density depends only on the common parameters, permitting consistent estimation.

Let $\mathbf{y}_i = [y_{i1}, \dots, y_{iT}]'$ be a $T \times 1$ vector dependent variable for individual i over all T time periods, and let $\mathbf{X}_i = [\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}]'$ denote the corresponding $T \times K$ matrix of regressors. For a static model \mathbf{y}_i has density

$$f(\mathbf{y}_i | \mathbf{X}_i, \alpha_i, \beta, \gamma) = \prod_{t=1}^T f(y_{it} | \mathbf{x}_{it}, \alpha_i, \beta, \gamma). \quad (23.8)$$

Maximum likelihood estimation based on this density generally leads to inconsistent estimation of β in short panels owing to the incidental parameters problem.

Suppose there exists a **sufficient statistic** \mathbf{s}_i for α_i . Then conditioning on the sufficient statistic \mathbf{s}_i , in addition to the usual conditioning on regressors, leads to **conditional density**

$$f(\mathbf{y}_i | \mathbf{X}_i, \alpha_i, \beta, \gamma, \mathbf{s}_i) = f(\mathbf{y}_i | \mathbf{X}_i, \beta, \gamma, \mathbf{s}_i), \quad (23.9)$$

so that α_i has dropped out. For example, for the linear regression model under normality $\mathbf{s}_i = \bar{\mathbf{y}}_i$ (see Section 21.6.3). Then the **conditional MLE** maximizes the **conditional log-likelihood**

$$\ln L_{\text{COND}}(\beta, \gamma) = \sum_{i=1}^N \ln f(\mathbf{y}_i | \mathbf{X}_i, \beta, \gamma, \mathbf{s}_i). \quad (23.10)$$

The adjective conditional is added here to indicate conditioning on \mathbf{s}_i and not just \mathbf{X}_i .

Andersen (1970) provided a detailed analysis of the conditional MLE. He showed that the conditional MLE is consistent if the density $f(\mathbf{y}_i | \mathbf{X}_i, \alpha_i, \beta)$ is correctly specified, that the information matrix equality holds for the conditional log-likelihood, but in general there is a loss of efficiency as the conditional MLE need not attain the Cramer–Rao lower bound. For normal and Poisson distributions, however, there is no efficiency loss.

The approach requires that a suitable sufficient statistic exists. This is the case for only a few models, essentially those of the linear exponential family. Andersen focused on models without regressors and gave as examples the normal, Poisson, binomial, and gamma. Once regressors are introduced it becomes even more difficult to find

a suitable sufficient statistic. McCullagh and Nelder (1989) provide a quite general discussion and Diggle et al. (2002) restrict their attention to specialized GLMs with canonical link functions.

The leading examples when a sufficient statistic is available are linear models under normality (see Section 21.6.2), logit models (though not probit models) for binary data (see Section 23.4.3), one-parameter gamma (including exponential), and particular parameterizations of the Poisson and negative binomial models for count data (see Section 23.7.3).

Mean-Differenced Transformation

For some models of the conditional mean with additive or multiplicative effects, the individual effects α_i can instead be eliminated by use of an appropriate differencing transformation. This leads to moment conditions that can be used for method of moments or GMM estimation as detailed in Section 23.2.6.

The **mean-differenced transformation** generalizes the within transformation for the linear model given in Section 21.2.2 that eliminates α_i by subtracting individual-specific **means**. It requires strongly exogenous regressors, see (23.7).

For the additive effects model defined in (23.3) with strongly exogenous regressors

$$E[(y_{it} - \bar{y}_i) - (g(\mathbf{x}'_{it}\beta) - \bar{g}_i(\beta)) | \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}] = 0, \quad (23.11)$$

where $\bar{g}_i(\beta) = T^{-1} \sum_{t=1}^T g(\mathbf{x}'_{it}\beta)$ and the result uses $E[\bar{y}_i | \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}] = \alpha_i + \bar{g}_i(\beta)$. For linear models (23.11) simplifies considerably as then $g(\mathbf{x}'_{it}\beta) - \bar{g}_i(\beta) = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)' \beta$.

For the multiplicative effects model defined in (23.4), some algebra leads to

$$E\left[y_{it} - \frac{g(\mathbf{x}'_{it}\beta)}{\bar{g}_i(\beta)} \times \bar{y}_i | \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}\right] = 0, \quad (23.12)$$

using $E[\bar{y}_i | \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}] = \alpha_i \bar{g}_i(\beta)$. For simplicity we call this a **mean-differenced transformation**, though strictly speaking it is a **quasi-difference**. It is also called a (conditional) **mean-scaling transformation**, as equivalently

$$E\left[y_{it} - \frac{\bar{y}_i}{\bar{g}_i(\beta)} g(\mathbf{x}'_{it}\beta) | \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}\right] = 0.$$

First-Differences Transformation

The **first-differences transformation** generalizes the first-difference transformation for the linear model given in Section 21.2.2 that eliminates α_i by subtracting the model lagged one period. We assume regressors are weakly exogenous (see (23.6)).

For the additive effects model,

$$E[(y_{it} - y_{i,t-1}) - (g(\mathbf{x}'_{it}\beta) - g(\mathbf{x}'_{i,t-1}\beta)) | \mathbf{x}_{i1}, \dots, \mathbf{x}_{i,t-1}] = 0, \quad (23.13)$$

where we have used $E[y_{i,t-1} | \mathbf{x}_{i1}, \dots, \mathbf{x}_{i,t-1}] = \alpha_i + g(\mathbf{x}'_{i,t-1}\beta)$.

For the multiplicative effects model defined in (23.4),

$$E \left[y_{it} - \frac{g(\mathbf{x}'_{it}\beta)}{g(\mathbf{x}'_{i,t-1}\beta)} \times y_{i,t-1} | \mathbf{x}_{i1}, \dots, \mathbf{x}_{i,t-1} \right] = 0, \quad (23.14)$$

where we have used $E[y_{i,t-1} | \mathbf{x}_{i1}, \dots, \mathbf{x}_{i,t-1}] = \alpha_i g(\mathbf{x}'_{i,t-1}\beta)$. For simplicity we call it a **first-differences** transformation, though strictly speaking it is a **quasi-difference**.

The first-differences transformation relies on weaker assumptions, conditioning only up to period t . It permits estimation of dynamic models, extending Section 22.5 to nonlinear models. For dynamic multiplicative effects Wooldridge (1997) and Chamberlain (1992) actually proposed use of a variant of (23.14),

$$E \left[\frac{g(\mathbf{x}'_{i,t-1}\beta)}{g(\mathbf{x}'_{it}\beta)} y_{it} - y_{i,t-1} | \mathbf{x}_{i1}, \dots, \mathbf{x}_{i,t-1} \right] = 0. \quad (23.15)$$

Dummy Variable Model Estimation

If the incidental parameters problem is ignored, one can attempt to estimate all parameters, including the individual-specific effects. Introduce a set of N dummy variables $d_{j,it}$ equal to 1 if $i = j$ and equal to 0 otherwise, and then jointly estimate the individual-specific parameters $\alpha_1, \dots, \alpha_N$ along with the other model parameters.

This estimator is computationally feasible, despite the very large number of parameters owing to large N , but the resulting estimates of β and possibly γ are in general inconsistent. Here we consider just parametric models, though similar points hold for conditional mean models.

Thus consider the parametric individual-specific effects model defined in (23.1). Then the method is to obtain ML estimates of β , γ , and $\alpha = [\alpha_1 \dots \alpha_N]'$ that maximize the full log-likelihood

$$\ln L_{FE}(\beta, \gamma, \alpha) = \sum_{i=1}^N \sum_{t=1}^T \ln f(y_{it}, \mathbf{d}'_{it}\alpha + \mathbf{x}'_{it}\beta, \gamma), \quad (23.16)$$

where $\mathbf{d}_{it} = [d_{1,it} \dots d_{N,it}]'$. The first-order conditions with respect to $\delta = [\beta' \gamma']'$ and α are

$$\begin{aligned} \sum_{i=1}^N \sum_{t=1}^T \partial \ln f(y_{it}, \mathbf{d}'_{it}\alpha + \mathbf{x}'_{it}\beta, \gamma) / \partial \delta &= \mathbf{0}, \\ \sum_{t=1}^T \partial \ln f(y_{it}, \alpha_i + \mathbf{x}'_{it}\beta, \gamma) / \partial \alpha_i &= 0, \quad i = 1, \dots, N. \end{aligned}$$

This estimator can be simple to compute despite the large number of parameters, N plus the dimension of δ . As detailed in Greene (2004b), the inverse of the Hessian is easily obtained by partitioning into δ and α and applying the standard partitioned inverse formula, using the simplification that $\partial \ln L(\delta, \alpha) / \partial \alpha_i \partial \alpha_j = 0$ for $j \neq i$ so that the inverse of the large $N \times N$ block corresponding to (α, α) is trivially obtained.

In two special cases there is no incidental parameters problem. First, if $y_{it} \sim \mathcal{N}[\alpha_i + \mathbf{x}'_{it}\beta, \sigma^2]$ then, from Section 21.6.4, the MLE for β is the within estimator,

which is consistent for β even for finite T . Here the incidental parameters problem arises for estimation of σ^2 but not for β . Second, for $y_{it} \sim \mathcal{P}[\exp(\alpha_i + \mathbf{x}'_{it}\beta)]$ there is similarly no incidental parameters problem in estimating β (see Section 23.7.3).

In general, however, there is an incidental parameters problem. The derivative with respect to α_i involves only T observations, rather than all NT observations. This usually spills over to inconsistency of $\hat{\beta}_{\text{ML}}$ and $\hat{\gamma}_{\text{ML}}$ in short panels. It is possible that this inconsistency is moderate in panels that are not too short, such as $T = 10$ or $T = 20$. The simulation study of Greene (2004a) indicates that the nature and extent of bias vary considerably with the particular nonlinear model being studied. The development of methods that are reasonably robust in the presence of fixed effects, though still inconsistent in short panels, is an active area of research.

23.2.3. Random Effects Models

A **random effects model** treats the individual-specific effect α_i as a random variable with specified distribution and eliminates α_i by integrating over this distribution. Random effects are usually applied to parametric models.

Parametric Models

Suppose the i th observation \mathbf{y}_i has unconditional joint density $f(\mathbf{y}_i | \mathbf{X}_i, \alpha_i, \beta, \gamma)$ given in (23.8), and the random effect has density

$$\alpha_i \sim g(\alpha_i | \eta), \quad (23.17)$$

where $g(\alpha_i | \eta)$ does not depend on observables. Then the unconditional joint density for the i th observation is

$$f(\mathbf{y}_i | \mathbf{X}_i, \beta, \gamma, \eta) = \int \left[\prod_{t=1}^T f(y_{it} | \mathbf{x}_{it}, \alpha_i, \beta, \gamma) \right] g(\alpha_i | \eta) d\alpha_i, \quad (23.18)$$

where by unconditional we mean we no longer condition on α_i . The **random effects MLE** of β , γ , and η maximizes the log-likelihood

$$\ln L_{\text{RE}}(\beta, \gamma, \eta) = \sum_{i=1}^N \ln \left(\int \left[\prod_{t=1}^T f(y_{it} | \mathbf{x}_{it}, \alpha_i, \beta) \right] g(\alpha_i | \gamma) d\alpha_i \right). \quad (23.19)$$

In some cases an analytical expression for this integral is possible, basically if $\prod_t f(y_{it} | \alpha_i)$ and $g(\alpha_i)$ are conjugate pairs (see Table 13.2). Examples include normal–normal for linear regression, which yields normal, and Poisson–gamma for count data regression, which yields negative binomial.

In most cases analytical results are not available, but numerical methods or simulation-based methods are likely to work well because the integral is only one dimensional. The usual approach is to choose $f(y_{it})$ to be the density that is thought to best fit the data in the absence of individual effects, and to let $g(\alpha_i)$ be the normal density. The integral is then a univariate integral with respect to a normal random variable. For small T the integral can be well approximated by Gauss–Hermite quadrature

(see Section 12.3.1), which approximates the integral with respect to a normal density by a weighted sum. Butler and Moffitt (1982) provide a detailed exposition for the random effects probit model. Skrondal and Rabe-Hesketh (2004) use quadrature. Alternatively, repeated draws from $g(\alpha_i)$ can be the basis for maximum simulated likelihood estimation (see Section 12.4.2).

The preceding discussion assumed independence over t for given i . If instead y_{it} and y_{is} are correlated over i then it is more efficient to replace $\prod_t f(y_{it} | \mathbf{x}_{it}, \alpha_i, \beta, \gamma)$ by $f(\mathbf{y}_i | \mathbf{X}_i, \alpha_i, \beta, \gamma)$ in (23.18) and (23.19).

Random Coefficients Model

The random effects approach can clearly be generalized to a **random coefficients model**, with random slopes as well as random intercepts, similar to the linear case in Section 22.8.

The natural model is a single-index model with conditional density $f(y_{it}, \mathbf{x}'_{it}(\beta + \alpha_i), \gamma)$ or conditional mean $g(y_{it}, \mathbf{x}'_{it}(\beta + \alpha_i))$ and the univariate integral with respect to scalar α_i will become a multivariate integral with respect to vector α_i , usually assumed to be normally distributed.

Correlated Random Effects Model

The key weakness of the random effects model is that it makes the strong assumption that the random effects are independent of regressors. To overcome this limitation Chamberlain (1980, 1982) proposed a **correlated random effects model**, for background discussion see Section 21.4.4, that specifies

$$\alpha_i = \mathbf{x}'_{li} \boldsymbol{\pi}_1 + \cdots + \mathbf{x}'_{Ti} \boldsymbol{\pi}_T + \xi_i. \quad (23.20)$$

The likelihood above is then maximized with respect to $\beta, \gamma, \boldsymbol{\pi}$, and the parameters of the density of ξ . Unlike linear models this model leads to different estimator than that obtained using the simpler specification of Mundlak (1978) that

$$\alpha_i = \bar{\mathbf{x}}'_i \boldsymbol{\pi} + \xi_i. \quad (23.21)$$

The equation (23.20) can be viewed as an example of a hierarchical model. More general hierarchical models also permit random slopes, with estimation by classical or Bayesian methods. Section 22.8 presented details for the linear model.

Finite Mixture Model

The finite mixture model (see Section 18.5.1) provides an alternative model for the unobserved individual-specific effect. If there are m **latent classes** or types of individuals and for the j th type $\alpha_i = \alpha_j$ then (23.18) becomes

$$f(\mathbf{y}_i | \mathbf{X}_i, \beta, \gamma, \boldsymbol{\pi}) = \sum_{j=1}^m \left[\prod_{t=1}^T f(y_{it} | \mathbf{x}_{it}, \alpha_j, \beta, \gamma) \right] \pi_j.$$

This model is most often used for panel duration models (see Section 18.5.2).

23.2.4. Pooled Models

The pooled model does not explicitly model individual-specific effects. It extends linear pooled regression (see Section 21.5) to nonlinear models.

Conditional Mean Models

For conditional mean models the **pooled model** is

$$E[y_{it} | \mathbf{x}_{it}] = g(\mathbf{x}_{it}, \boldsymbol{\beta}), \quad (23.22)$$

for specified function $g(\cdot)$.

The model (23.22) can be directly estimated by NLS, with inference based on panel-robust standard errors that control for conditional heteroskedasticity and for conditional correlation between y_{it} and y_{is} . More efficient estimation is possible by modeling the heteroskedasticity and correlation. Details are provided in Section 23.2.6.

Pooled versus Random Effects Models

What is the cost of ignoring individual-specific random effects?

The additive effect model $E[y_{it} | \alpha_i, \mathbf{x}_{it}] = \alpha_i + g(\mathbf{x}_{it}, \boldsymbol{\beta})$ leads to (23.22) if $E[\alpha_i | \mathbf{x}_{it}] = 0$. The multiplicative effect model $E[y_{it} | \alpha_i, \mathbf{x}_{it}] = \alpha_i g(\mathbf{x}_{it}, \boldsymbol{\beta})$ implies (23.22) if $E[\alpha_i | \mathbf{x}_{it}] = 1$. So the pooled model will lead to consistent estimation of $\boldsymbol{\beta}$ in a random effects model if the effects are additive or multiplicative and the standard normalizations of the mean of α_i for these models are used.

Otherwise, the pooled model is unlikely to lead to the same parameter estimates as an individual-specific random effects model. For example, consider a probit random effects model with $E[y_{it} | \alpha_i, \mathbf{x}_{it}] = \Phi(\alpha_i + \mathbf{x}_{it}' \boldsymbol{\beta})$, where $\alpha_i \sim N[0, \sigma_\alpha^2]$. Then it can be shown that $E[y_{it} | \mathbf{x}_{it}] = \Phi(\mathbf{x}_{it}' \boldsymbol{\beta} / \sqrt{1 + \sigma_\alpha^2})$, which differs from the natural pooled probit model $E[y_{it} | \mathbf{x}_{it}] = \Phi(\mathbf{x}_{it}' \boldsymbol{\beta})$. Unlike the linear model of Chapter 21, if the true model has individual-specific random effects than ignoring these effects can lead to inconsistent parameter estimates of $\boldsymbol{\beta}$.

The statistics literature uses the pooled model approach extensively for panel versions of **generalized linear models**, such as binary data and count data. The resulting parameter estimates are called **population averaged**, as the random effects are averaged out. The approach is called **marginal analysis**, as $E[y_{it} | \mathbf{x}_{it}]$ is a model that is marginal with respect to the random effects.

Parametric Models

For **pooled parametric models** the starting point is usually

$$f(y_{it} | \mathbf{x}_{it}) = f(y_{it}, \mathbf{x}'_{it} \boldsymbol{\beta}, \gamma) \quad (23.23)$$

for specified density $f(\cdot)$. This model can be directly estimated by ML, with inference based on panel-robust standard errors that control for conditional heteroskedasticity and correlation (see Section 23.2.6).

In general the pooled parametric model estimates of β and γ are unlikely to be consistent with those from a random effects parametric model. The arguments are similar to those for the conditional mean.

23.2.5. Fixed versus Random Effects

The essential result that random effects and pooled model estimators are inconsistent if individual-specific effects are present and are correlated with regressors still holds in nonlinear models. This favors use of fixed effects models on grounds of robustness, though there is a trade-off with loss of efficiency in estimation. A Hausman test can be used (see Section 21.4.4) to test whether a fixed effects model is needed, provided consistent estimation of the fixed effects model is possible.

Other comparisons of fixed versus random effects models for linear models (see Section 21.4) require some adaptation for nonlinear models.

Because of the incidental parameters problem, not all nonlinear models with fixed effects admit consistent parameter estimates. So fixed effects modeling is not always feasible.

If consistent estimation of a nonlinear fixed effects model is possible then, unlike the linear case, the coefficients of time-invariant regressors can be identified. To see this consider the mean-differenced transformation for an additive effects model. For a linear model $E[(y_{it} - \bar{y}_i) - (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)' \boldsymbol{\beta} | \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}] = \mathbf{0}$, with obvious problems for time-invariant regressors as then, considering the j th regressor, $x_{itj} - \bar{x}_{ij} = x_{ij} - x_{ij} = 0$. More generally, from (23.11)

$$E[(y_{it} - \bar{y}_i) - (g(\mathbf{x}'_{it}\boldsymbol{\beta}) - \bar{g}(\boldsymbol{\beta})) | \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}] = \mathbf{0},$$

with no such simplification for nonlinear $g(\cdot)$ unless all K components of \mathbf{x}_{it} are time-invariant.

In fixed effect models with nonadditive effects it is not possible to predict changes in the dependent variable as regressors change. For the general model (23.2), the **marginal effect** $\partial E[y_{it} | \mathbf{x}_{it}, \alpha_i, \beta] / \partial \mathbf{x}_{it} = \partial g(\mathbf{x}_{it}, \alpha_i, \beta) / \partial \mathbf{x}_{it}$ depends on α_i .

The marginal effect can be measured in two special cases. For additive effects (see (23.3)) the marginal effect is $\partial g(\mathbf{x}_{it}, \beta) / \partial \mathbf{x}_{it}$, which does not depend on α_i . For multiplicative effects models (see (23.4)) the marginal effect is $\alpha_i \partial g(\mathbf{x}_{it}, \beta) / \partial \mathbf{x}_{it}$. Then it is possible to measure the relative size of marginal effects for changes in different regressors. In particular, if $E[y_{it} | \mathbf{x}_{it}, \alpha_i, \beta] = \alpha_i \exp(\mathbf{x}'_{it}\boldsymbol{\beta})$, then $(\partial E[y_{it}] / \partial x_{itj}) / (\partial E[y_{it}] / \partial x_{itk}) = \beta_j / \beta_k$.

23.2.6. Estimation and Panel-Robust Statistical Inference

The preceding analysis has concentrated on elimination of the incidental parameter α_i . Now we detail estimation of model parameters, once α_i has been eliminated for models with individual-specific effects.

We assume a short panel and independence of observations over i . The dependent variable y_{it} may be conditionally heteroskedastic and conditionally correlated over t for given i . The situation is similar to that in Section 21.2.3, except that nonlinear

estimators are used instead of simpler linear LS estimators. Standard statistical output that ignores this complication will lead to invalid inference. In the following we present expressions for panel-robust estimates of the variance matrix of parameter estimates. Alternatively, a panel bootstrap can be used (see Section 11.6.2).

GMM Estimation

Panel GMM estimation is appropriate for models based on the conditional mean. The key is specification of the moment condition that is the basis of GMM estimation. Following Section 22.2.1, a natural starting point is

$$E[\mathbf{Z}_i' \mathbf{u}_i(\boldsymbol{\theta})] = \mathbf{0}, i = 1, \dots, N, \quad (23.24)$$

where \mathbf{Z}_i is a $T \times r$ matrix that depends on the regressors, $\mathbf{u}_i(\boldsymbol{\theta})$ is a $T \times 1$ residual vector, and $\boldsymbol{\theta}$ is a $q \times 1$ parameter vector $\boldsymbol{\theta}$. Different panel models lead to different specifications of \mathbf{u}_i and \mathbf{Z}_i . An example is given in the following. A key departure from Chapter 22 is that the residual $\mathbf{u}_i(\boldsymbol{\theta})$ will be nonlinear in $\boldsymbol{\theta}$.

If $r = q$ then there are as many moment conditions as parameters to estimate and we can use the **panel method of moments estimator** $\widehat{\boldsymbol{\theta}}_{\text{MM}}$ that solves

$$\frac{1}{N} \sum_{i=1}^N \mathbf{Z}_i' \mathbf{u}_i(\widehat{\boldsymbol{\theta}}) = \mathbf{0}. \quad (23.25)$$

Using results in Section 6.10.3 on nonlinear systems estimation, we have that this estimator is asymptotically normal with variance matrix consistently estimated by

$$\widehat{\mathbf{V}}[\widehat{\boldsymbol{\theta}}] = \left[\sum_{i=1}^N \widehat{\mathbf{D}}_i' \mathbf{Z}_i \right]^{-1} \sum_{i=1}^N \mathbf{Z}_i' \widehat{\mathbf{u}}_i \widehat{\mathbf{u}}_i' \mathbf{Z}_i \left[\sum_{i=1}^N \mathbf{Z}_i' \widehat{\mathbf{D}}_i \right]^{-1}, \quad (23.26)$$

where $\widehat{\mathbf{D}}_i = \partial \mathbf{u}_i / \partial \boldsymbol{\theta}'|_{\widehat{\boldsymbol{\theta}}}$ and $\widehat{\mathbf{u}}_i = \mathbf{u}_i(\widehat{\boldsymbol{\theta}})$. This yields panel-robust-standard errors in short panels.

If $r > q$ then GMM estimation is necessary, and we use the **panel GMM estimator** $\widehat{\boldsymbol{\theta}}_{\text{GMM}}$ that minimizes

$$\mathcal{Q}_N(\boldsymbol{\theta}) = \left[\frac{1}{N} \sum_{i=1}^N \mathbf{Z}_i' \mathbf{u}_i(\boldsymbol{\theta}) \right]' \mathbf{W}_N \left[\frac{1}{N} \sum_{i=1}^N \mathbf{Z}_i' \mathbf{u}_i(\boldsymbol{\theta}) \right], \quad (23.27)$$

where \mathbf{W}_N is an $r \times r$ weighting matrix. The asymptotic variance matrix for this estimator can be obtained directly from results for the nonlinear systems IV estimator given in Section 6.10.4. Given the moment condition (23.24), the most efficient estimator uses $\mathbf{W}_N = [N^{-1} \sum_i \mathbf{Z}_i' \widehat{\mathbf{u}}_i \widehat{\mathbf{u}}_i' \mathbf{Z}_i]^{-1}$.

More efficient estimators are possible using alternative moment conditions. In particular, if the starting point is a particular conditional moment condition then the optimal unconditional moment condition for GMM estimation is given in Section 6.3.7. The GEE estimator given later follows this approach. A more general treatment is given in Avery, Hansen, and Hotz (1983) and Breitung and Lechner (1999).

GMM Example

As a specific example, consider the first-differences transformation applied to the multiplicative fixed effects model. The starting point is the conditional moment restriction (23.14). This leads to many unconditional moment conditions, one of which is

$$E \left[\mathbf{x}_{it} \left(y_{it} - \frac{g(\mathbf{x}'_{it}\beta)}{g(\mathbf{x}'_{i,t-1}\beta)} \times y_{i,t-1} \right) \right] = \mathbf{0}, \quad t = 1, \dots, T, i = 1, \dots, N.$$

Assume data on $(y_{it}, \mathbf{x}_{it})$ are available for $(T + 1)$ periods, with the initial period then lost because of first differencing. Stacking over T time periods yields (23.24) with $\mathbf{Z}'_i = [\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}]$ and $\mathbf{u}'_i = [u_{i1}, \dots, u_{iT}]$, where $u_{it} = y_{it} - [g(\mathbf{x}'_{it}\beta)/g(\mathbf{x}'_{i,t-1}\beta)]y_{i,t-1}$. Here $\mathbf{Z}'_i \mathbf{u}_i = \sum_t \mathbf{x}_{it} u_{it}$, so the method of moments estimator $\hat{\beta}$ solves

$$\sum_{i=1}^N \sum_{t=1}^T \mathbf{x}_{it} \left[y_{it} - \frac{g(\mathbf{x}'_{it}\beta)}{g(\mathbf{x}'_{i,t-1}\beta)} y_{i,t-1} \right] = \mathbf{0}.$$

Clearly, additional moment conditions can be used, such as $E[\mathbf{x}_{it-1} u_{it}] = \mathbf{0}$, leading to an overidentified model and estimation by GMM. This was discussed extensively for the linear model in Section 22.2.

Generalized Estimating Equations Estimation

The pooled model for the conditional mean specifies $E[y_{it}|\mathbf{x}_{it}] = g(\mathbf{x}_{it}, \beta)$ (see Section 23.2.4). This model can be estimated by GMM methods already given. Here we go further and consider efficient GMM estimation.

Stacking over all T observations gives conditional moment condition

$$E[\mathbf{y}_i - \mathbf{g}_i(\beta)|\mathbf{X}_i] = \mathbf{0}, \quad (23.28)$$

where $\mathbf{g}_i(\beta) = [g(\mathbf{x}_{i1}, \beta), \dots, g(\mathbf{x}_{iT}, \beta)]'$ and $\mathbf{X}_i = [\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}]'$. The optimal unconditional moment condition to use in estimation is then

$$E \left[\frac{\partial \mathbf{g}'_i(\beta)}{\partial \beta} \{V[\mathbf{y}_i|\mathbf{X}_i]\}^{-1} (\mathbf{y}_i - \mathbf{g}_i(\beta)) \right] = \mathbf{0}, \quad (23.29)$$

a result obtained by applying the general result given in Section 6.3.7. This leads to the **generalized estimating equations estimator** $\hat{\beta}_{GEE}$ that solves

$$\sum_{i=1}^N \frac{\partial \mathbf{g}'_i(\beta)}{\partial \beta} \Sigma_i^{-1} (\mathbf{y}_i - \mathbf{g}_i(\beta)) = \mathbf{0}, \quad (23.30)$$

where Σ_i is a working variance matrix for $V[\mathbf{y}_i|\mathbf{X}_i]$. The asymptotic variance matrix of $\hat{\beta}_{GEE}$ is given by (23.26) with $\widehat{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{g}_i(\hat{\beta})$ and $\mathbf{Z}_i = \partial \mathbf{g}'_i(\beta)/\partial \beta|_{\hat{\beta}} \times \widehat{\Sigma}_i$. This variance estimate is panel-robust and is also robust to misspecification of Σ_i .

The GEE estimator, due to Liang and Zeger (1986), is widely used in the statistics literature for panel versions of generalized linear models. Different GLMs correspond to different conditional mean functions $\mathbf{g}_i(\beta)$ and working variance matrices Σ_i .

ML Estimation

For likelihood-based models the starting point is the joint density for all T individuals, $f(\mathbf{y}_i | \mathbf{X}_i, \boldsymbol{\theta})$. For pooled parametric models $\boldsymbol{\theta}' = [\boldsymbol{\beta}', \boldsymbol{\gamma}']$ (see (23.23)), and for random effects parametric models $\boldsymbol{\theta}' = [\boldsymbol{\beta}', \boldsymbol{\gamma}', \boldsymbol{\eta}']$ (see (23.18)).

The standard approach is to let $f(\mathbf{y}_i | \mathbf{X}_i, \boldsymbol{\theta}) = \prod_{t=1}^T f(y_{it} | \mathbf{x}_{it}, \boldsymbol{\theta})$, where $f(y_{it} | \mathbf{x}_{it}, \boldsymbol{\theta})$ is the density for the (i, t) th observation. The implicit assumption of independence over t for given i is usually unwarranted, especially for pooled models that do not include a random effect that permits some correlation over time. Nonetheless, consistent estimates of $\boldsymbol{\theta}$ are obtained even if $f(\mathbf{y}_i | \mathbf{X}_i, \boldsymbol{\theta})$ is misspecified, provided $f(y_{it} | \mathbf{x}_{it}, \boldsymbol{\theta})$ is correctly specified. A sandwich form should then be used for the estimator variance matrix to ensure panel-robust standard errors. The MLE is strictly a quasi-MLE, with detailed discussion given in Section 5.7.5. More generally, this approach is an example of inference with clustered data (see Section 24.5).

More efficient estimation is possible using a richer model for $f(\mathbf{y}_i | \mathbf{X}_i, \boldsymbol{\theta})$ that accommodates correlation over time. However, nonnormal multivariate distributions for \mathbf{y}_i can be restrictive or difficult to work with. For pooled GLMs the GEE estimator can be used instead.

23.2.7. Dynamic Models

Dynamic models with individual-specific effects are of considerable interest as they allow one to distinguish between true state dependence and spurious dependence caused by unobserved heterogeneity (see Section 22.5.1).

For nonlinear models it is not always obvious how to include lagged dependent variables as regressors, since for some types of data there is not always a standard pure time series model. This is illustrated in Section 23.7.4 for the Poisson model. Once an appropriate specification is determined, the standard fixed effects estimators become inconsistent and random effects estimators need to incorporate initial conditions, as was the case for the linear panel model.

Pooled Models

The pooled model ignores random effects and estimates the usual cross-section model where the regressors now include lagged dependent variables. The discussion in Section 23.2.4 is again relevant.

Fixed Effects Models

For fixed effects models the issues are similar to those presented in Section 22.5. The regressors are now weakly exogenous rather than strongly exogenous. The usual fixed effects estimators are inconsistent.

For models with additive effects or multiplicative effects consistent estimation is possible using the first-difference transformation (see Section 23.2.2) and higher lags of the lagged dependent variable as an instrument. For additive effects models this leads to a nonlinear version of the Arellano–Bond estimator given in Section 22.5.3.

For multiplicative effects the first-difference transformation is detailed in Section 23.7.4. For dynamic logit with fixed effects see Section 23.4.3.

Parametric Random Effects Models

For parametric random effects models initial conditions on the lagged dependent variable matter. Usually there is no satisfactory treatment, so the estimates are inconsistent in short panels with inconsistency that declines as T gets larger.

Consider the simplest case where only the first-period lag appears in the model, so the regressors \mathbf{x}_{it} become regressors \mathbf{x}_{it} and y_{it-1} . The random effects density (23.1) becomes $f(y_{it}|y_{it-1}, \mathbf{x}_{it}, \alpha_i, \delta)$ for $t = 2, \dots, T$. However, a similar model for y_{i1} cannot be included because y_{i0} is not observed. One approach treats y_{i1} as exogenous, so that we model the conditional distribution for only $T - 1$ observations y_{i1}, \dots, y_{iT} . An alternative approach presumes a static model for y_{i1} that depends on regressors \mathbf{x}_{i1} and possibly on the marginal effect α_i . Then the joint conditional density of \mathbf{y}_i is

$$\begin{aligned} f(\mathbf{y}_i | \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}, \alpha_i, \delta, \delta_1, \gamma) \\ = \int \left[\prod_{t=2}^T f(y_{it} | y_{it-1}, \mathbf{x}_{it}, \alpha_i, \delta) \right] f_1(y_{i1} | \mathbf{x}_{i1}, \alpha_i, \delta_1) g(\alpha_i | \gamma) d\alpha_i, \end{aligned}$$

rather than (23.18), where $f_1(y_{i1} | \mathbf{x}_{i1}, \alpha_i, \delta_1)$ is the assumed density for the first observation.

In pure time series analysis initial conditions become irrelevant asymptotically, since $T \rightarrow \infty$. In short panels, however, they become very important as T is small and asymptotics instead use $N \rightarrow \infty$.

23.2.8. Endogenous Regressors

The treatment for endogenous variables in nonlinear models is similar to that in the linear case presented in Chapter 22.

Panel GMM is the natural framework. The starting point is a conditional moment restriction $E[\mathbf{u}_i(\theta)|\mathbf{Z}_i] = \mathbf{0}$ for appropriately defined residual $\mathbf{u}_i(\theta)$ and instruments \mathbf{Z}_i . This leads to unconditional moment condition (23.24) that is the basis for GMM estimation. Possible candidates for instruments can include exogenous regressors from periods other than the current one, as detailed in Sections 22.2 and 22.4 for the linear model.

23.3. Nonlinear Panel Example: Patents and R&D

We model the relationship between patents and R&D expenditures, using U.S. data on 346 firms for each of the five years 1975–1979 from Hall, Griliches, and Hausman (1986). The dependent variable y_{it} is Patents, defined as the number of patents applied for during the year that were eventually granted. For simplicity we consider just one explanatory variable x_{it} , real R&D spending during the year (in 1972 dollars).

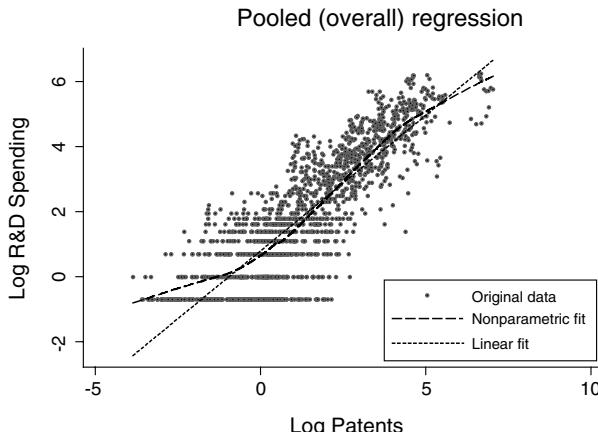


Figure 23.1: Patents and R&D spending: pooled (overall) regression. Natural logarithm of patent applications leading to award plotted against the natural logarithm of R&D spending for 346 firms in each of the five years 1975–79. Zero patents recoded to 0.5 patents.

An obvious starting model is a log–log model, with $E[\ln y_{it} | x_{it}] = \alpha_i + \beta \ln x_{it}$, since then β equals the Patents–R&D elasticity. This model cannot be applied here, as $y_{it} = 0$ for a considerable number of observations and $\ln 0$ is not defined. An ad hoc adjustment is to recode $y_{it} = 0$ as $y_{it} = 0.5$ before taking logs.

Figure 23.1 provides a plot of the adjusted $\ln(\text{Patents})$ against $\ln(\text{R&D})$, along with fitted OLS (with an estimated slope coefficient of 0.834) and nonparametric regression curves, using data for all firms in all years. Patents clearly increase with R&D expenditure. Panel data analysis, particularly fixed effects models, can separate this relationship into cross-section and time-series components. Note that Patents vary greatly across observations, particularly across firms, with a mean of 36.3, a standard deviation of 74.5, and a range of 0 to 608 over all years and firms.

We estimate a multiplicative individual-effects model for the conditional mean with

$$E[y_{it} | x_{it}, \alpha_i] = \alpha_i \exp(\beta \ln x_{it}) = \exp(\gamma_i + \beta \ln x_{it}), \quad (23.31)$$

where $\gamma_i = \ln \alpha_i$. Then β directly estimates the Patents–R&D elasticity, since (23.31) implies $\partial \ln E[y_{it} | x_{it}] / \partial \ln x_{it} = \beta$. Unlike the log–log model, zero values for y_{it} cause no problems.

A richer parametric model recognizes that the dependent variable is a count. A starting point is a Poisson model

$$y_{it} | x_{it}, \gamma_i \sim \mathcal{P}[\exp(\gamma_i + \beta \ln x_{it})]. \quad (23.32)$$

This model, detailed in Section 23.7, has the same conditional mean for y_{it} as that given in (23.31).

Table 23.1 presents a number of estimators for these data. All estimators are consistent under the assumption that the conditional mean is given by (23.31) with α_i a random effect that is independent of x_{it} and has constant mean. All estimators except the last are inconsistent under the assumption that α_i is instead a fixed effect that is correlated with x_{it} . Three standard error estimates are provided: program defaults,

Table 23.1. Patents and R&D Spending: Nonlinear Panel Model Estimators^a

	NLS	Poisson	GEE	Poisson–RE	Poisson–FE
$\gamma = \ln \alpha$	2.529	1.712	2.068	2.313	–
β	.509	.693	.560	.349	–0.038
Panel se ^b	(.055)	(.043)	(.033)	(.033)	(.033)
Boot se	[.054]	[.047]	[.107]	[.119]	{.107}
Usual se	{.011}	{.002}	{.004}	{.033}	{.033}
Sum β	–	.486	.460	.546	.313
N	1730	1730	1730	1730	1620

^a Shown are pooled NLS, pooled Poisson, pooled GEE, Poisson Random Effects (RE), and Poisson Fixed Effects estimates for the nonlinear panel regression (23.31) of ln(Patents) on ln(R&D). Standard errors for the slope coefficients are panel robust in parentheses, bootstrap in square brackets, and usual estimates that assume iid errors in curly braces. The second to last row gives the sum of β coefficients in an expanded model with up to five lags of ln(R&D) as regressors.

^b se, standard error

panel-robust estimates (where available), and bootstrap estimates (without refinement). The details for each column are as follows:

Pooled NLS: The NLS estimates in the first column estimate (23.31) with $\alpha_i = \alpha$ by NLS (see Section 5.8). The default standard error of 0.011 assuming iid errors is much smaller than the correct panel-robust standard error estimate of 0.054.

Pooled Poisson: The Poisson estimates in the second column are for the Poisson model (23.32) with $\alpha_i = \alpha$ estimated by the Poisson MLE assuming independence over i and t . The estimated elasticity is 0.693 compared to the NLS estimate of 0.509. The default standard error of 0.002 imposes the Poisson restriction of variance–mean equality (see Section 20.2.2). Correcting for overdispersion using the sandwich variance matrix estimate (see also Section 20.2.2) increases the standard error estimate to 0.020 and emphasizes the importance of controlling for any overdispersion in count data. Additionally controlling for correlation over t for given i leads to an even higher panel-robust standard error estimate of 0.043.

Pooled GEE: The pooled GEE estimator solves (23.30), where $g(\mathbf{x}_{it}, \boldsymbol{\beta})$ is given by (23.32) with $\alpha_i = \alpha$. The particular specification of the working matrix Σ_i used here is given after (23.55). The estimated elasticity is 0.560 with standard error of 0.033 using the panel-robust estimate discussed after (23.30).

Poisson–RE: The Poisson random effects estimator assumes that $\alpha_i = \ln \gamma_i$ is gamma distributed (see Section 23.7.2). The estimated elasticity is 0.349 with default standard error of 0.033.

Poisson–FE: The Poisson fixed effects estimator assumes that $\alpha_i = \ln \gamma_i$ is a fixed effect, and it is estimated as in Section 23.7.3. The estimated elasticity of –0.038 is now negative, with default standard error of 0.033. For the Poisson fixed effect model, firms with $\sum_t y_{it} = 0$ are dropped, leading here to a loss of $22 \times 5 = 110$ observations.

There is a big difference between fixed and random effects results, favoring fixed effects estimation. The surprising negative estimated elasticity with FE arises because the model is too simple. In particular, R&D expenditure affects patent activity with a lag. Replacing $\beta \ln \mathbf{x}_{it}$ in (23.31) and (23.32) by $\sum_{l=0}^5 \beta_l \ln \mathbf{x}_{i,t-l}$ leads to estimated elasticity $\sum_{l=0}^5 \hat{\beta}_l$ given in the second last row of Table 23.1. The FE estimate of 0.313 is still less than the other estimates, but the difference is now reduced.

23.4. Binary Outcome Data

We consider a binary outcome in which y_{it} takes only the values 0 and 1. For example, data may be available on whether or not an individual is employed in each of several time periods. A key result is that fixed effects estimation is possible for the logit model but not the probit model.

23.4.1. Individual-Specific Effects Binary Models

The natural extension of the binary outcome model from cross-section data (see Section 14.3) to panel data with individual-specific effects is to specify that y_{it} takes only the values 0 and 1, with

$$\Pr[y_{it} = 1 | \mathbf{x}_{it}, \boldsymbol{\beta}, \alpha_i] = \begin{cases} F(\alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta}) & \text{in general,} \\ \Lambda(\alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta}) & \text{for logit model,} \\ \Phi(\alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta}) & \text{for probit model,} \end{cases} \quad (23.33)$$

where $F(\cdot)$ is a cumulative distribution function, $\Lambda(\cdot)$ is the logistic cdf with $\Lambda(z) = e^z/(1 + e^z)$, and $\Phi(\cdot)$ is the standard normal cdf. Given (23.33) and assuming conditional independence, the joint density for the i th observation $\mathbf{y}_i = (y_{i1}, \dots, y_{iT})$ is

$$f(\mathbf{y}_i | \mathbf{X}_i, \alpha_i, \boldsymbol{\beta}) = \prod_{t=1}^T F(\alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta})^{y_{it}} (1 - F(\alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta}))^{1-y_{it}}. \quad (23.34)$$

For binary data the conditional probability is also the conditional mean, so

$$\mathbb{E}[y_{it} | \alpha_i, \mathbf{x}_{it}] = F(\alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta}). \quad (23.35)$$

This is a single-index individual-specific effects model (see (23.5)) that does not simplify to either an additive or multiplicative effects model. Additive and multiplicative effects models are not appropriate as they do not restrict the conditional mean and conditional probability to lie between zero and one.

Binary panel models emphasize the parametric model (23.34), since binary data must be Bernoulli distributed. The conditional mean model (23.35) is rarely used, though it is natural to use this if regressors are endogenous.

23.4.2. Random Effects Binary Models

The random effects MLE assumes that the individual effects are normally distributed, with $\alpha_i \sim \mathcal{N}[0, \sigma_\alpha^2]$. The **random effects MLE** of $\boldsymbol{\beta}$ and σ_α^2 maximizes the

log-likelihood $\sum_{i=1}^N \ln f(\mathbf{y}_i | \mathbf{X}_i, \boldsymbol{\beta}, \sigma_\alpha^2)$, where

$$f(\mathbf{y}_i | \mathbf{X}_i, \boldsymbol{\beta}, \sigma_\alpha^2) = \int f(\mathbf{y}_i | \mathbf{X}_i, \alpha_i, \boldsymbol{\beta}) \frac{1}{\sqrt{2\pi\sigma_\alpha^2}} \exp\left(\frac{-\alpha_i}{2\sigma_\alpha^2}\right)^2 d\alpha_i, \quad (23.36)$$

where $f(\mathbf{y}_i | \mathbf{X}_i, \alpha_i, \boldsymbol{\beta})$ is given in (23.34) with $F = \Lambda$ for the logit model and $F = \Phi$ for the probit model. There is no closed-form solution for the integral (23.36) and it is standard to compute it numerically using **quadrature methods**.

If fixed effects are not present, then an alternative to the random effects model is a pooled binary model that simply specifies that $\Pr[\mathbf{y}_{it} = 1 | \mathbf{x}_{it}] = F(\mathbf{x}'_{it}\boldsymbol{\beta})$. Statistical inference should then be based on panel-robust standard errors (see Section 23.2.6). More efficient estimation is possible using a GMM approach (see Avery et al., 1983) or a GEE approach (see Liang and Zeger, 1986).

23.4.3. Fixed Effects Logit

Fixed effects estimation is possible for the panel logit model, using the conditional MLE, but not for other binary panel models such as panel probit.

For the logit model performing some algebra given in Section 23.4.5 yields that the joint density of $\mathbf{y}_i = (y_{i1}, \dots, y_{iT})$ is

$$f(\mathbf{y}_i | \alpha_i, \mathbf{x}_i, \boldsymbol{\beta}) = \frac{\exp(\alpha_i \sum_t y_{it}) \exp((\sum_t y_{it} \mathbf{x}'_{it}) \boldsymbol{\beta})}{\prod_t [1 + \exp(\alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta})]}. \quad (23.37)$$

This depends on α_i , which we need to eliminate. For observation i there are $\sum_t y_{it}$ outcomes of 1 in the T periods. Define the set $\mathbf{B}_c = \{\mathbf{d}_i | \sum_t d_{it} = \sum_t y_{it} = c\}$ to be the set of all possible sequences of 0s and 1s for which the sum of T binary outcomes $\sum_t y_{it} = c$. Then if we condition on $\sum_t y_{it} = c$ it is shown in Section 23.4.6 that α_i is eliminated and

$$f(\mathbf{y}_i | \sum_t y_{it} = c, \mathbf{x}_i, \boldsymbol{\beta}) = \frac{\exp((\sum_t y_{it} \mathbf{x}'_{it}) \boldsymbol{\beta})}{\sum_{\mathbf{d} \in \mathbf{B}_c} \exp((\sum_t d_{it} \mathbf{x}'_{it}) \boldsymbol{\beta})}, \quad (23.38)$$

a result due to Chamberlain (1980). The density (23.38) is the basis for conditional ML estimation. The only complication is that there are many sets \mathbf{B}_c and sequences within these sets, as we now detail.

First, it is not possible to condition on $\sum_t y_{it} = 0$, since this can only occur if all $y_{it} = 0$, and similarly for $\sum_t y_{it} = T$. This can mean considerable loss of observations if, for example, most people are employed in all periods.

As an example where conditioning works, suppose $T = 2$ and $\sum_t y_{it} = 1$. Then either the sequence $\{0, 1\}$ or $\{1, 0\}$ is possible, and the conditional probability in (23.38) implies that, for example,

$$\begin{aligned} \Pr[y_{i1} = 0, y_{i2} = 1 | y_{i1} + y_{i2} = 1] &= \frac{\exp(\mathbf{x}'_{i1}\boldsymbol{\beta})}{\exp(\mathbf{x}'_{i1}\boldsymbol{\beta}) + \exp(\mathbf{x}'_{i2}\boldsymbol{\beta})} \\ &= \frac{\exp((\mathbf{x}_{i1} - \mathbf{x}_{i0})'\boldsymbol{\beta})}{1 + \exp((\mathbf{x}_{i1} - \mathbf{x}_{i0})'\boldsymbol{\beta})}. \end{aligned}$$

If $T = 3$ then we can condition on $\sum_t y_{it} = 1$, with possible sequences $\{0, 0, 1\}$, $\{0, 1, 0\}$ and $\{1, 0, 0\}$, or on $\sum_t y_{it} = 2$, with possible sequences $\{0, 1, 1\}$, $\{1, 0, 1\}$ and $\{1, 1, 0\}$. Clearly for large T there are many sequences and the conditional density can get complicated.

The conditional density is that of a conditional logit model, where parameters are invariant but regressors vary over alternatives. The number of alternatives varies across individuals, where for individual i each alternative is a specific sequence of 0s and 1s that sum to $\sum_t y_{it}$. It is easiest to use computer code specifically set up for this problem. Even then there can be a large number of alternatives. For example, if $T = 10$ and $\sum_t y_{it} = 5$ then there are 252 alternatives. Consistent but less efficient estimation is possible by dropping some observations, such as for individuals with many alternatives because of a high $\sum_t y_{it}$, or by reducing the number of time periods.

The elimination of the individual-effects α_i makes it impossible to interpret regression coefficients using the original model (23.37). Instead, we use the conditional model (23.38). For example, suppose we have single regressor and $\beta = 0.2$. Then if we consider two time periods and condition on $\sum_t y_{it} = 1$, then

$$\Pr[y_{i1} = 0, y_{i2} = 1 | y_{i1} + y_{i2} = 1] = \frac{\exp(\beta(x_{i1} - x_{i2}))}{1 + \exp(\beta(x_{i1} - x_{i2}))}.$$

It follows that a one-unit difference in x_{i1} versus x_{i2} leads to a conditional probability of this sequence being $\exp(\beta)/[1 + \exp(\beta)]$ compared to a probability of one-half if $x_{i1} = x_{i2}$.

23.4.4. Dynamic Binary Models

Suppose we have a pure time series first-order Markov logit model with no regressors other than the lagged dependent variable:

$$\Pr[y_{it} = 1 | \alpha_i, y_{it-1}] = \frac{\exp(\alpha_i + \gamma y_{it-1})}{1 + \exp(\alpha_i + \gamma y_{it-1})}. \quad (23.39)$$

Then performing some algebra given in Section 23.4.5 gives

$$f(\mathbf{y}_{it} | y_{i1}, y_{iT}, \sum_{t=2}^{T-1} y_{it}, \gamma) = \frac{\exp\left(\gamma \sum_{t=2}^{T-1} y_{it} y_{it-1}\right)}{\sum_{\mathbf{d} \in \mathbf{C}_i} \exp\left(\gamma \sum_{t=2}^{T-1} d_{it} d_{it-1}\right)}, \quad (23.40)$$

where the set $\mathbf{C}_i = \{\mathbf{d}_i | y_{i1}, y_{iT}, \sum_t d_{it} = \sum_t y_{it}\}$ is the set of all possible sequences of 0s and 1s for which the sum of T binary outcomes is $\sum_t y_{it}$, the first outcome is y_{i1} , and the last outcome is y_{iT} .

Conditional ML estimation based on (23.40) leads to a consistent estimate of γ . The minimum number of time periods needed is four. For example, if \mathbf{y}_i is the sequence $\{0, 1, 0, 1\}$ then the set \mathbf{C}_i is composed of the sequences $\{0, 1, 0, 1\}$ and $\{0, 0, 1, 1\}$. The approach is due to Chamberlain (1985), who actually considered a second-order Markov model. Chay, Hoynes, and Hyslop (2001) apply this method to California administrative data on welfare spells and find that, after controlling for

unobserved individual heterogeneity, there remains true state dependence in welfare participation.

The preceding results and discussion apply to pure time series models. Honoré and Kyriazidou (2000) provided a method that allows regressors other than the lagged dependent variable. Thus let (23.39) become

$$\Pr[y_{it} = 1 | \alpha_i, y_{it-1}, \mathbf{x}_{it}] = \frac{\exp(\alpha_i + \mathbf{x}'_{it}\beta + \gamma y_{it-1})}{1 + \exp(\alpha_i + \mathbf{x}'_{it}\beta + \gamma y_{it-1})}. \quad (23.41)$$

Consider four time periods and consider sequences with common binary outcomes in the first and fourth periods, say d_1 and d_4 . Then the probability that the sequence is $\{d_1, 0, 1, d_4\}$, given that it is either $\{d_1, 0, 1, d_4\}$ or $\{d_2, 1, 0, d_4\}$, now depends on α_i . However, the dependence on α_i disappears if $x_{3i} = x_{4i}$. Since few observations have $x_{3i} = x_{4i}$, especially with continuous data, Honoré and Kyriazidou (2000) propose use of kernel smoothing methods with kernel weights that depend on $(x_{3i} - x_{4i})$. Chay and Hyslop (2000) provide an application that implements this method and many other methods for dynamic binary data models.

23.4.5. Multinomial Models

The fixed effects estimator can be generalized to the multinomial logit model, since this model implies a binary logit model for pairwise comparison of alternatives (see Section 15.4.3). For static models Chamberlain (1980) provides a brief exposition and M.-J. Lee (2002) provides more details. Magnac (2000) provides a quite detailed empirical application to individual transitions among six different states in the French labor market using dynamic fixed effects logit models with no regressors other than lagged dependent variables. Honoré and Kyriazidou (2000) consider the multinomial logit model.

For other multinomial models a random effects approach is necessary. These models, such as mixed logit and multinomial probit, are complicated to estimate even in the cross-section case. For details see Train (2003).

23.4.6. Derivations for Fixed Effects Logit

For simplicity suppress the subscript i . For the logit model the joint probability of $\mathbf{y} = (y_1, \dots, y_T)$ given in (23.34) becomes

$$\begin{aligned} f(\mathbf{y}|\alpha) &= \prod_{t=1}^T \left(\frac{\exp(\alpha + \mathbf{x}'_t\beta)}{1 + \exp(\alpha + \mathbf{x}'_t\beta)} \right)^{y_t} \left(\frac{1}{1 + \exp(\alpha + \mathbf{x}'_t\beta)} \right)^{1-y_t} \\ &= \frac{\exp(\sum_t y_t(\alpha + \mathbf{x}'_t\beta))}{\prod_t [1 + \exp(\alpha + \mathbf{x}'_t\beta)]} \\ &= \frac{\exp(\alpha \sum_t y_t) \exp((\sum_t y_t \mathbf{x}'_t)\beta)}{\prod_t [1 + \exp(\alpha + \mathbf{x}'_t\beta)]}, \end{aligned} \quad (23.42)$$

which yields (23.37).

The quantity $\sum_t y_t$ can be shown to be a sufficient statistic for α as follows. Suppose we have an observation for \mathbf{y} such that $\sum_t y_t = c$. Define the set $\mathbf{B}_c = \{\mathbf{d} \mid \sum_t d_t = c\}$ to be the set of all possible sequences of 0s and 1s for which the sum of T binary outcomes is c , and condition on $\sum_t y_t = c$. Then

$$\begin{aligned} f(\mathbf{y} \mid \sum_t y_t = c) &= \frac{\Pr[\mathbf{y}, \sum_t y_t = c]}{\Pr[\sum_t y_t = c]} \\ &= \frac{\Pr[\mathbf{y}]}{\Pr[\sum_t y_t = c]} \\ &= \frac{\Pr[\mathbf{y}]}{\sum_{\mathbf{d} \in \mathbf{B}_c} \Pr[\mathbf{d}]} \\ &= \frac{\exp((\sum_t y_t \mathbf{x}'_t) \boldsymbol{\beta})}{\sum_{\mathbf{d} \in \mathbf{B}_c} \exp((\sum_t d_t \mathbf{x}'_t) \boldsymbol{\beta})}, \end{aligned} \quad (23.43)$$

where the first equality uses Bayes' rule, the second equality uses the fact that knowledge of $\sum_t y_t$ does not add anything given knowledge of \mathbf{y} , the third equality uses the fact that $\Pr[\sum_t y_t = c]$ equals the sum of the probabilities of combinations of 0s and 1s that equal c , and the fourth uses the previous definition of $f(\mathbf{y})$ and considerable simplification that in part relies on $\sum_t y_t = \sum_t d_t$ when we restrict attention to $\mathbf{d} \in \mathbf{B}_c$.

Now consider the dynamic model. Replacing $\mathbf{x}'_t \boldsymbol{\beta}$ in (23.42) by γy_{t-1} yields

$$\begin{aligned} f(\mathbf{y}) &= \frac{\exp(\alpha \sum_{t=2}^T y_t) \exp(\sum_{t=2}^T \gamma y_{t-1} y_t)}{\prod_t [1 + \exp(\alpha + \gamma y_{t-1})]} \\ &= \frac{\exp(\alpha \sum_{t=2}^T y_t) \exp(\sum_{t=2}^T \gamma y_{t-1} y_t)}{[1 + \exp(\alpha)]^{\sum_{t=2}^T (1 - y_{t-1})} [1 + \exp(\alpha + \gamma)]^{\sum_{t=2}^T y_{t-1}}} \\ &= \frac{\exp(\alpha \sum_{t=2}^T y_t) \exp(\sum_{t=2}^T \gamma y_{t-1} y_t)}{[1 + \exp(\alpha)]^{(T-1+y_1-y_T+\sum_{t=2}^T y_t)+} [1 + \exp(\alpha + \gamma)]^{y_1-y_T+\sum_{t=2}^T y_t}}, \end{aligned}$$

where the second equality uses the fact that y_{t-1} is either 0 or 1 and follows after some algebra, and the last equality uses $\sum_{t=2}^T y_{t-1} = y_1 - y_T + \sum_{t=2}^T y_t$. The algebra is then similar to (23.43) except that in addition to conditioning on $\sum_{t=2}^T y_t$ we also need to condition on y_1 and y_T that appear in the denominator. Equivalently, we can condition on $\sum_{t=1}^T y_t$ and y_1 and y_T . This yields

$$f(\mathbf{y}) = \frac{\exp(\sum_{t=2}^T \gamma y_{t-1} y_t)}{\sum_{\mathbf{d} \in \mathbf{C}_c} \exp(\sum_{t=2}^T \gamma d_{t-1} d_t)},$$

where $\mathbf{C} = \{\mathbf{d} \mid d_1 = y_1, d_T = y_T, \sum_{t=1}^T d_t = \sum_{t=1}^T y_t\}$ is the set of all possible sequences of 0s and 1s for which the sum of the T binary outcomes is $\sum_t y_t$, the first outcome is y_1 , and the last outcome is y_T .

23.5. Tobit and Selection Models

We consider censoring, truncation, or selection when panel data are available, rather than data on a single cross-section.

A pooled analysis simply mirrors analysis in the cross-section case, with the adjustment that panel-robust standard errors should be computed (see Section 23.2.8). For example, see Grasdal (2001) who considers selection resulting from panel attrition.

Here we focus instead on panel models with individual-specific effects. Then random effects models can be estimated, if the strong assumption of a purely random effect is warranted, the only complication being that of numerical computation. There are no simple consistent estimators for fixed effects models, however, in the usual microeconometric setting of a short panel. More complicated semiparametric estimators that permit fixed effects in Tobit and generalized Tobit models are given in Section 23.8.

23.5.1. Censored and Truncated Models

For cross-section data the censored Tobit model is given in Section 16.3.1. A panel version with additive individual-specific effect specifies

$$y_{it}^* = \alpha_i + \mathbf{x}'_{it}\beta + \varepsilon_{it}, \quad (23.44)$$

where $\varepsilon_{it} \sim \mathcal{N}[0, \sigma_\varepsilon^2]$, and we observe $y_{it} = y_{it}^*$ if $y_{it}^* > 0$ and $y_{it} = 0$ or is observed to be missing if $y_{it}^* \leq 0$. The joint density for the i th observation $\mathbf{y}_i = (y_{i1}, \dots, y_{iT})$ can be written as

$$f(\mathbf{y}_i | \mathbf{X}_i, \alpha_i, \beta, \sigma_\varepsilon^2) = \prod_{t=1}^T \left[\frac{1}{\sigma_\varepsilon} \phi_{it} \right]^{d_{it}} [1 - \Phi_{it}]^{1-d_{it}}, \quad (23.45)$$

where $\phi_{it} = \phi((y_{it} - \alpha_i - \mathbf{x}'_{it}\beta)/\sigma_\varepsilon)$, $\Phi_{it} = \Phi((\alpha_i + \mathbf{x}'_{it}\beta)/\sigma_\varepsilon)$, and $\phi(\cdot)$ and $\Phi(\cdot)$ denote, respectively, the standard normal pdf and cdf.

The fixed effects MLE maximizes the log-likelihood based on (23.45) with respect to β , σ_ε^2 , and $\alpha_1, \dots, \alpha_N$. In short panels the resulting estimator of β is inconsistent because of the incidental parameters problem, and there is no simple differencing or conditioning method that can provide a consistent estimator. Heckman and MacCurdy (1980) applied the fixed effects MLE to female labor supply. Although recognizing the inconsistency of the estimator, they argued that with $T = 8$ the inconsistency may not be too great. Greene (2004a) provides a recent Monte Carlo study for the fixed effects Tobit MLE.

Random effects estimation is more commonly used because of inconsistency of the fixed effects estimator. Under the assumption that $\alpha_i \sim \mathcal{N}[0, \sigma_\alpha^2]$ the **random effects MLE** of β , σ_ε^2 , and σ_α^2 maximizes the log-likelihood $\sum_{i=1}^N \ln f(\mathbf{y}_i | \mathbf{X}_i, \beta, \sigma_\varepsilon^2, \sigma_\alpha^2)$, where

$$f(\mathbf{y}_i | \mathbf{X}_i, \beta, \sigma_\varepsilon^2, \sigma_\alpha^2) = \int f(\mathbf{y}_i | \mathbf{X}_i, \alpha_i, \beta, \sigma_\varepsilon^2) \frac{1}{\sqrt{2\pi\sigma_\alpha^2}} \exp\left(\frac{-\alpha_i^2}{2\sigma_\alpha^2}\right)^2 d\alpha_i, \quad (23.46)$$

for $f(\mathbf{y}_i | \mathbf{X}_i, \alpha_i, \beta, \sigma_\varepsilon^2)$ given in (23.45). This one-dimensional integral can be computed using **Gaussian quadrature**.

This approach can be extended to other models with censoring or truncation. For example, a right-censored version of the Poisson random effects model in Section 23.7.2 may be used if, say, counts above 10 are recorded only as 10 or more.

There are two weaknesses to the fully parametric approach. First, as in the cross-section case reliance on distributional assumptions becomes much greater when there is censoring or truncation. Second, the assumption of purely random effects independent of regressors may be too strong.

23.5.2. Selection Models

Selection models can arise in panel data for the same reasons as in the cross-section case (see Section 16.5). A generalization of the **Tobit type 2 model** in Section 16.5.1 to a linear panel model with individual specific effects λ_i and δ_i is

$$\begin{aligned} y_{it}^* &= \alpha_i + \mathbf{x}'_{it}\beta + \varepsilon_{it}, \\ d_{it}^* &= \delta_i + \mathbf{z}'_{it}\gamma + v_{it}, \end{aligned} \tag{23.47}$$

where $y_{it} = y_{it}^*$ is observed if $d_{it}^* > 0$ and y_{it} is not observed otherwise.

For the random effects formulation the four unobservables are assumed to be normally distributed. Hausman and Wise (1979) proposed ML estimation, which involves a bivariate integral as α_i may be correlated with δ_i and ε_{it} may be correlated with v_{it} .

The fixed effects estimator is inconsistent in short panels. Note, however, that if $d_{it}^* = \delta_i$, so that selection is due only to time-invariant characteristics of the individual, which may be observed or unobserved, then the fixed effects estimator in the model $y_{it} = \alpha_i + \mathbf{x}'_{it}\beta + \varepsilon_{it}$ is consistent. A fixed effect panel model controls for sample selection, to the extent that it depends on time-invariant characteristics.

Verbeek and Nijman (1992) provide a more detailed discussion of the essential assumptions needed for consistent estimation in these model and propose tests for selectivity bias. Wooldridge (1995) provides a similar analysis under weaker assumptions and presents assumptions that may not be too restrictive in some applications that permit consistent estimation of the fixed effects model. Vella (1998) provides a review and additional references.

The methods for sample selection can be extended to **panel attrition** (see Section 21.8.5) that leads to **attrition bias** if observations on the dependent variable are lost in a nonrandom manner. Then all data for the i th observation are not observed if $d_{it}^* \leq 0$, so \mathbf{z}_{it} in (23.47) needs to be replaced by variables observed in periods other than period t . An early example is Hausman and Wise (1979), and a more recent application is Grasdal (2001). Baltagi (2001) and Hsiao (2003) provide further references.

23.6. Transition Data

For concreteness consider panel data on welfare spells. Great interest lies in measuring individual persistence in welfare spells, and determining the extent to which this is due

to true state dependence rather than differences in individual propensities to be on welfare. Since individual propensities may depend in part on unobservables, models with individual-specific effects should be used. For duration data there exists an unusually wide range of modeling approaches, as several types of panel data on transitions are possible. Here we focus on fixed effects models.

Data may be available on whether or not an individual is in a state at several points in time, such as on welfare. Then one can use a binary panel model (see Section 23.4), such as the dynamic fixed effects logit model.

Richer data provide information on the durations of several individual spells. A natural starting point is then a **panel proportional hazards model**

$$\lambda(t_{ij} | \mathbf{x}_{ij}) = \lambda_j(t_{ij}, \gamma_j) \exp(\mathbf{x}'_{ij}\boldsymbol{\beta})\alpha_i, \quad (23.48)$$

where t_{ij} is the completed spell duration for the j th spell of the i th individual and α_i is an individual-specific effect. This is the mixed proportional hazards model, discussed for single-spell data in Chapter 18. The conditions for nonparametric identification of the MPH model with only single-spell data (see Section 18.3) include the assumption that α_i are distributed independently of the regressors. This rules out fixed effects. Once multiple spells become available, however, Honoré (1992) showed that α_i can be a fixed effect if \mathbf{x}_{ij} is constant over j (see Section 19.4.1). For further discussion of the model (23.48), including a dynamic duration model with hazard function for the second spell dependent on the duration of the first spell, see Section 19.4.1.

Chamberlain (1985) presented several approaches for elimination of α_i in various panel duration models. For the MPH model, with baseline hazard $\lambda_j(\cdot)$ the same across spells j , the probability that the second spell is longer than the first spell does not depend on α_i . Conditional ML can be applied to the gamma duration model, since the gamma is an LEF density. For Weibull, gamma and log-normal models the density of t_{i1}/t_{i2} does not depend on α_i .

For more recent references and a detailed discussion, including sensitivity of multiple-spell data to censoring, see Van den Berg (2001).

23.7. Count Data

Hausman et al. (1984) presented estimable fixed effects and random effects models for both panel Poisson and panel negative binomial models. More recent work has emphasized fixed effects in multiplicative effects models, permitting estimation of static and dynamic models under relatively weak distributional assumptions.

23.7.1. Individual-Specific Effects Count Models

We focus on the Poisson model, detailed for cross-section data in Section 20.2, though panel versions of negative binomial are also briefly considered.

The **Poisson individual-specific effects model** specifies that $y_{it} \sim \mathcal{P}[\alpha_i \exp(\mathbf{x}'_{it}\boldsymbol{\beta})]$. Then, assuming conditional independence, the joint density for the i th observation

$\mathbf{y}_i = (y_{i1}, \dots, y_{iT})$ is

$$f(\mathbf{y}_i | \mathbf{X}_i, \alpha_i, \beta) = \prod_{t=1}^T \exp[-\alpha_i \exp(\mathbf{x}'_{it} \beta)] [-\alpha_i \exp(\mathbf{x}'_{it} \beta)]^{y_{it}} / y_{it}! \quad (23.49)$$

A less parametric approach simply models the conditional mean as

$$\begin{aligned} E[y_{it} | \alpha_i, \mathbf{x}_{it}] &= \alpha_i \exp(\mathbf{x}'_{it} \beta) \\ &= \exp(\gamma_i + \mathbf{x}'_{it} \beta). \end{aligned} \quad (23.50)$$

This is both a single-index individual-specific effects model and a multiplicative effects model. Since it is a multiplicative effects model the individual effects α_i can be eliminated by mean differencing or first differencing. Note that the Poisson panel model (23.49) has conditional mean (23.50).

23.7.2. Random Effects Count Models

Assuming gamma-distributed random effects leads to a tractable solution for the marginal density of the random effects model. Assume α_i is $\mathcal{G}[\eta, \eta]$ distributed with mean 1, variance $1/\eta$, and density $g(\alpha_i | \eta) = \eta^\eta \alpha_i^{\eta-1} e^{-\alpha_i \eta} / \Gamma(\eta)$. Then (23.18) for the Poisson model (23.49) becomes

$$\begin{aligned} f(\mathbf{y}_i | \mathbf{X}_i, \beta, \eta) &= \left[\prod_t \frac{\lambda_{it}^{y_{it}}}{y_{it}!} \right] \times \left(\frac{\eta}{\sum_t \lambda_{it} + \eta} \right)^\eta \\ &\quad \times \left(\sum_t \lambda_{it} \right)^{-\sum_t y_{it}} \frac{\Gamma(\sum_t y_{it} + \eta)}{\Gamma(\eta)}, \end{aligned} \quad (23.51)$$

where $\lambda_{it} = \exp(\mathbf{x}'_{it} \beta)$ and derivations are given in Section 23.7.5. The resulting first-order conditions for the Poisson random effects estimator $\hat{\beta}$ can be expressed as

$$\sum_{i=1}^N \sum_{t=1}^T \mathbf{x}_{it} \left(y_{it} - \lambda_{it} \frac{\bar{y}_i + \eta/T}{\bar{\lambda}_i + \eta/T} \right) = \mathbf{0}, \quad (23.52)$$

where $\bar{\lambda}_i = T^{-1} \sum_t \exp(\mathbf{x}'_{it} \beta)$.

The term on the left-hand side of (23.52) has expected value zero if the mean conditional on regressors in all periods $E[y_{it} | \alpha_i, \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}] = \alpha_i \exp(\mathbf{x}'_{it} \beta)$. So despite all the parametric assumptions made, the Poisson random effects estimator is consistent for β under the relatively weak assumption that the conditional mean is that given in (23.50) and that regressors are strongly exogenous. For the density (23.51), $E[y_{it} | \mathbf{x}_i] = \lambda_{it}$ and $V[y_{it} | \mathbf{x}_i] = \lambda_{it} + \lambda_{it}^2 / \delta$, so that overdispersion is of the NB2 form. A sandwich estimate of the variance matrix will permit more flexible models of overdispersion and conditional correlation. The first-order conditions for η (not given) are quite complicated though the information matrix is block diagonal in β and η .

Several alternative estimators are available given random effects. First, the pooled Poisson estimator ignores the random effects and assumes $y_{it} | \mathbf{x}_{it} \sim \mathcal{P}[\exp(\mathbf{x}'_{it} \beta)]$. This

has first-order conditions

$$\sum_{i=1}^N \sum_{t=1}^T \mathbf{x}_{it} (y_{it} - \lambda_{it}) = \mathbf{0}, \quad (23.53)$$

where $\lambda_{it} = \exp(\mathbf{x}'_{it}\beta)$. This estimator is consistent if the conditional mean is (23.50) with $E[\alpha_i | \mathbf{x}_{it}] = 1$. Thus the usual cross-section Poisson MLE is consistent if the true model is one with multiplicative random effects. However, as illustrated in the Section 23.3 example, panel-robust standard errors should be used. Here (23.26) yields

$$\widehat{\text{V}}[\widehat{\beta}_{\text{pool}}] = \left[\sum_{i,t} \widehat{\lambda}_{it} \mathbf{x}_{it} \mathbf{x}'_{it} \right]^{-1} \sum_{i,t,s} \widehat{u}_{it} \widehat{u}_{is} \mathbf{x}_{it} \mathbf{x}'_{it} \left[\sum_{i,t} \widehat{\lambda}_{it} \mathbf{x}_{it} \mathbf{x}'_{it} \right]^{-1}, \quad (23.54)$$

where $\widehat{\lambda}_{it} = \exp(\mathbf{x}'_{it}\widehat{\beta})$, $\widehat{u}_{it} = y_{it} - \widehat{\lambda}_{it}$, $\sum_{i,t}$ denotes $\sum_{i=1}^N \sum_{t=1}^T$ and $\sum_{i,t,s}$ denotes $\sum_{i=1}^N \sum_{t=1}^T \sum_{s=1}^T$. An alternative pooled estimator based on (23.50) is NLS, in which case (23.53) becomes $\sum_i \sum_t \mathbf{x}_{it} \lambda_{it} (y_{it} - \lambda_{it}) = \mathbf{0}$.

Second, more efficient pooled estimation may be possible using the GEE approach of Section 23.2.8, which introduces conditional correlation. The general result (23.30) for $g_{it} = \lambda_{it} = \exp(\mathbf{x}'_{it}\beta)$ becomes

$$\sum_{i=1}^N \mathbf{Z}'_i \Sigma_i^{-1} (\mathbf{y}_i - \boldsymbol{\lambda}_i) = \mathbf{0}, \quad (23.55)$$

where \mathbf{Z}_i is a $T \times K$ matrix with t th row observation $\lambda_{it} \mathbf{x}'_{it}$, and $\boldsymbol{\lambda}_i$ is a $T \times 1$ vector with t th entry λ_{it} . Several different working variance matrices Σ_i for $\text{V}[\mathbf{y}_i | \mathbf{X}_i]$ are possible. The choice $\Sigma_i = \text{Diag}[\lambda_{it}]$ yields the pooled Poisson estimating equations in (23.53). Letting $\Sigma_{i,tt} = \lambda_{it}$ and $\Sigma_{i,ts} = \lambda_{is} = \phi \sqrt{\lambda_{it} \lambda_{is}}$ for $s \neq t$ permits correlation over t that is **equicorrelated** or **exchangeable** since the correlation is a constant ϕ for $s \neq t$.

Third, more efficient pooled estimation may be possible using ML with the negative binomial rather than the Poisson as the starting point. Suppose y_{it} is iid negative binomial with NB2 variance function with parameters $\alpha_i \lambda_{it}$ and ϕ_i (see Section 20.4.1), implying y_{it} has mean $\alpha_i \lambda_{it} / \phi_i$ and variance $(\alpha_i \lambda_{it} / \phi_i) \times (1 + \alpha_i / \phi_i)$. If $(1 + \alpha_i / \phi_i)^{-1}$ is a beta-distributed random variable with parameters (η_1, η_2) , then after some considerable algebra (23.18) reduces to

$$f(\mathbf{y}_i | \mathbf{X}_i, \beta, \eta) = \left(\prod_t \frac{\Gamma(\lambda_{it} + y_{it})!}{\Gamma(\lambda_{it})! \Gamma(y_{it} + 1)!} \right) \times \frac{\Gamma(\eta_1 + \eta_2) \Gamma(\eta_1 + \sum_t \lambda_{it}) \Gamma(\eta_2 + \sum_t y_{it})}{\Gamma(\eta_1) \Gamma(\eta_2) \Gamma(\eta_1 + \eta_2 + \sum_t \lambda_{it} + \sum_t y_{it})}. \quad (23.56)$$

where $\lambda_{it} = \exp(\mathbf{x}'_{it}\beta)$. This is the basis for ML estimation of β , η_1 , and η_2 . This model relies on stronger assumptions than does the Poisson random effects model.

Fourth, analysis need not be restricted to parametric models with closed-form solutions for $f(\mathbf{y}_i | \mathbf{X}_i, \beta, \eta)$. Crépon and Dugeut (1997a) use maximum simulated likelihood methods to estimate hurdle and zero-inflated panel count models with joint normal random effects.

23.7.3. Fixed Effects Count Models

The fixed effects estimator for the Poisson panel model (23.50) can be derived in several ways.

First, the Poisson MLE simultaneously estimates β and $\alpha_1, \dots, \alpha_N$. The log-likelihood based on (23.49) is

$$\begin{aligned}\ln L(\beta, \alpha) &= \ln \left[\prod_i \prod_t \{\exp(-\alpha_i \lambda_{it}) (\alpha_i \lambda_{it})^{y_{it}} / y_{it}!\} \right] \\ &= \sum_i \left[-\alpha_i \sum_t \lambda_{it} + \ln \alpha_i \sum_t y_{it} + \sum_t y_{it} \ln \lambda_{it} - \sum_t \ln y_{it}! \right],\end{aligned}\quad (23.57)$$

where $\lambda_{it} = \exp(\mathbf{x}'_{it} \beta)$. Differentiating with respect to α_i and setting to zero yields $\hat{\alpha}_i = \sum_t y_{it} / \sum_t \lambda_{it}$. Substituting this back into (23.57) yields the **concentrated likelihood function**. Dropping terms not involving β , we get

$$\ln L_{\text{conc}}(\beta) \propto \sum_i \sum_t \left[y_{it} \ln \lambda_{it} - y_{it} \ln \left(\sum_s \lambda_{is} \right) \right]. \quad (23.58)$$

It follows that for the Poisson fixed effects model there is no incidental parameters problem. Consistent estimates of β for fixed T and $N \rightarrow \infty$ can be obtained by maximization of $\ln L_{\text{conc}}(\beta)$ in (23.58). Differentiation of (23.58) with respect to β yields first-order conditions

$$\sum_i \sum_t \left[y_{it} \mathbf{x}_{it} - y_{it} \left[\sum_s \lambda_{is} \mathbf{x}_{is} \right] / \left[\sum_s \lambda_{is} \right] \right] = \mathbf{0},$$

which can be reexpressed as

$$\sum_{i=1}^N \sum_{t=1}^T \mathbf{x}_{it} \left(y_{it} - \frac{\lambda_{it}}{\bar{\lambda}_i} \bar{y}_i \right) = \mathbf{0}, \quad (23.59)$$

where $\lambda_{it} = \exp(\mathbf{x}'_{it} \beta)$ and $\bar{\lambda}_i = T^{-1} \sum_t \exp(\mathbf{x}'_{it} \beta)$; see Blundell, Griffith, and Windmeijer (1995). The Poisson panel model (23.49) and the linear panel model of Section 21.6 are unusual in that simultaneous estimation of β and α provides consistent estimates of β in short panels, so there is no **incidental parameters problem**.

Second, the conditional MLE eliminates the fixed effects by conditioning on a sufficient statistic for α_i . For the Poisson panel model this is $\sum_t y_{it}$. Some algebra given in Section 23.7.5 shows that this leads to a conditional log-likelihood function that is proportional to the concentrated log-likelihood function given in (23.58). It follows that the conditional ML estimator for β in the fixed effects Poisson model solves (23.59). This was the original derivation of the Poisson fixed effects estimator of β by Palmgren (1981) and Hausman et al. (1984).

Third, the mean-differenced transformation (23.14) for the multiplicative effects model (23.50) implies that $E[y_{it} - (\lambda_{it}/\bar{\lambda}_i)\bar{y}_i | \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}] = 0$, and hence

$$E[\mathbf{x}_{it}(y_{it} - (\lambda_{it}/\bar{\lambda}_i)\bar{y}_i)] = \mathbf{0}. \quad (23.60)$$

Using the corresponding sample moment conditions leads to an estimator β that solves (23.59).

The same estimator has been obtained in three different ways. The third derivation makes it clear that the essential assumption for the consistency of the Poisson fixed effects estimator is that regressors are strongly exogenous and (23.50) is correctly specified. Inference should be based on panel-robust standard errors. In particular, if the usual default ML or conditional ML output is used, following the first two derivations, standard errors may be considerably understated owing to failure to control for overdispersion in the count data. The fixed effects estimator leads to some loss of data, as observations i with $\sum_t y_{it} = 0$ do not contribute to the sum in (23.59).

Consistent estimation of β in the presence of fixed effects is also possible for a particular parameterization of the negative binomial model. Hausman et al. (1984) assumed that y_{it} is iid NB1 with parameters $\alpha_i \lambda_{it}$ and ϕ_i , where $\lambda_{it} = \exp(\mathbf{x}'_{it} \beta)$, so y_{it} has mean $\alpha_i \lambda_{it} / \phi_i$ and variance $(\alpha_i \lambda_{it} / \phi_i) \times (1 + \alpha_i / \phi_i)$. The parameters α_i and ϕ_i can only be identified up to the ratio α_i / ϕ_i , and this ratio drops out of the conditional joint density for the i th observation, which after considerable algebra can be shown to be

$$f(y_{i1}, \dots, y_{iT} | \sum_t y_{it}) = \left(\prod_t \frac{\Gamma(\lambda_{it} + y_{it})}{\Gamma(\lambda_{it})\Gamma(y_{it} + 1)} \right) \times \frac{\Gamma(\sum_t \lambda_{it}) \Gamma(\sum_t y_{it} + 1)}{\Gamma(\sum_t \lambda_{it} + \sum_t y_{it})}. \quad (23.61)$$

This distribution for integer λ_{it} is the **negative hypergeometric distribution**. The conditional ML negative binomial fixed effects estimator of β maximizes the log-likelihood function based on (23.61). The Poisson fixed effects model is more commonly used since it is consistent under much weaker distributional assumptions.

23.7.4. Dynamic Count Models

There are several ways to bring dynamics into a count data model. Pure time series models are surveyed in Cameron and Trivedi (1998). For simplicity consider inclusion of one lagged dependent variable. The obvious model is $E[y_t | y_{t-1}, \mathbf{x}_t] = \exp(\gamma y_{t-1} + \mathbf{x}'_t \beta)$, but this can lead to explosive behavior as a result of exponentiation of y_{t-1} . A more stable model may be obtained by instead using $\exp(\gamma \ln y_{t-1} + \mathbf{x}'_t \beta)$, but this then runs into problems when $y_{t-1} = 0$. For this reason an appealing model is the linear feedback model $E[y_t | y_{t-1}, \mathbf{x}_t] = \gamma y_{t-1} + \exp(\mathbf{x}'_t \beta)$. The Poisson integer-valued AR(1) model has this property and in the pure time series case has correlation function $\text{Cor}[y_t, y_{t-k}] = \gamma^k$, similar to that for the AR(1) model (see Al-Osh and Alzaid, 1987).

Thus Blundell, Griffiths, and Windmeijer (1995, 2002) considered the dynamic fixed effects panel data model with

$$E[y_{it} | \alpha_i, y_{i,t-1}, \mathbf{x}_{it}] = \gamma y_{i,t-1} + \alpha_i \exp(\mathbf{x}'_t \beta).$$

Applying the first-difference transformation (23.15) leads to conditional moment restrictions

$$E \left[\frac{\exp(\mathbf{x}'_{i,t-1}\beta)}{\exp(\mathbf{x}'_{it}\beta)} (y_{it} - \gamma y_{i,t-1}) - (y_{i,t-1} - \gamma y_{i,t-2}) | y_{i1}, \dots, y_{i,t-2}, \mathbf{x}_{i1}, \dots, \mathbf{x}_{i,t-1} \right] = 0.$$

These lead to many unconditional moment conditions (see Section 22.5.3 for a similar discussion for the linear model) that can supply the basis for GMM estimation as in Section 23.2.6. Crépon and Dugeut (1997b), Montalvo (1997), and Blundell, Griffith, and Van Reenen (1995, 1999) use similar quasi-differencing methods, with application to the Patents–R&D relationship.

Böckenholt (1999) uses a more parametric model, estimating a Poisson integer-valued AR(1) model with unobserved heterogeneity modeled using a finite mixture distribution (see Section 18.5).

23.7.5. Derivations for Random and Fixed Effects Poisson

First, consider a random effects Poisson model with gamma-distributed random effects. For simplicity suppress the subscript i and let $\lambda_t = \exp(\mathbf{x}'_t\beta)$. The general formula (23.18) for the Poisson model (23.49) and random effects density $g(\alpha|\gamma)$ yields

$$\begin{aligned} f(y_1, \dots, y_T | \mathbf{x}_t) &= \int_0^\infty \left[\prod_t \left(e^{-\alpha \lambda_t} (\alpha \lambda_t)^{y_t} / y_t! \right) \right] g(\alpha|\gamma) d\alpha \\ &= \int_0^\infty \left[\prod_t \lambda_t^{y_t} / y_t! \right] \left(e^{-\alpha \sum_t \lambda_t} \cdot \alpha^{\sum_t y_t} \right) g(\alpha|\gamma) d\alpha \\ &= \left[\prod_t \lambda_t^{y_t} / y_t! \right] \times \int_0^\infty \left(e^{-\alpha \sum_t \lambda_t} \cdot \alpha^{\sum_t y_t} \right) g(\alpha|\gamma) d\alpha. \end{aligned}$$

For $g(\alpha_i|\eta) = \eta^\eta \alpha^{\eta-1} e^{-\alpha\eta} / \Gamma(\eta)$, similar algebra to that in Section 20.4.1 yields the density given in (23.51).

Second, derive the conditional density for the Poisson fixed effects model for observations in all time periods for a given individual, where for simplicity the individual subscript i is dropped. In general the density of y_1, \dots, y_T given $\sum_t y_t$ is

$$\begin{aligned} f(y_1, \dots, y_T | \sum_t y_t) &= f(y_1, \dots, y_T | \sum_t y_t) / f(\sum_t y_t) \\ &= f(y_1, \dots, y_T) / f(\sum_t y_t) \\ &= \frac{\prod_t (\exp(-\mu_t) \mu_t^{y_t} / y_t!)}{\exp(-\sum_t \mu_t) (\sum_t \mu_t)^{\sum_t y_t} / (\sum_t y_t)!} \\ &= \frac{\exp(-\sum_t \mu_t) \prod_t \mu_t^{y_t} / \prod_t y_t!}{\exp(-\sum_t \mu_t) \prod_t (\sum_s \mu_s)^{y_t} / (\sum_t y_t)!} \\ &= \frac{(\sum_t y_t)!}{\prod_t y_t!} \times \prod_t \left(\frac{\mu_t}{\sum_s \mu_s} \right)^{y_t}, \end{aligned}$$

where the second equality arises because knowledge of $\sum_t y_t$ adds nothing given knowledge of y_1, \dots, y_T , the third equality specializes to y_t iid $\mathcal{P}[\mu_t]$ and hence $\sum_t y_t$ is $\mathcal{P}[\sum_t \mu_t]$, and the fourth and fifth equalities simplify. The conditional density is that of the multinomial for $\sum_t y_t$ trials where the t th of T distinct outcomes occurs in any trial with probability $\mu_t / \sum_s \mu_s$. Setting $\mu_{it} = \alpha_i \exp(\mathbf{x}'_{it}\beta)$ and taking logarithms yields conditional likelihood that is proportional to the concentrated log-likelihood given in (23.58).

23.8. Semiparametric Estimation

The **semiparametric** literature for panel data has emphasized models for limited dependent variables since, as for cross-section data, parametric assumptions become much more important when truncation, censoring, or selection are present. Attention focuses on models with fixed effects. We provide a brief summary.

For binary data Manski (1987) extended his maximum score estimator from cross-section models to the panel model with fixed effects given in (23.33) where now the function $F(\cdot)$ is no longer specified. Although this estimator is consistent it converges at rate less than \sqrt{N} and is not asymptotically normal.

For the Tobit model Honoré (1992) extended the censored LAD approach of Powell (1986a) to the panel fixed effects model (23.45) where the distribution of the error term ε_{it} is unspecified. The data are artificially trimmed so that the fixed effect is subsequently eliminated by appropriate differencing. The estimator is \sqrt{N} consistent and asymptotically normal.

For panel data with sample selection Kyriazidou (1997) considered the fixed effects version of the type 2 Tobit model (23.47) where the distribution of the errors ε_{it} and v_{it} is unspecified. She presented a Heckman-type two-step estimator. A smoothed version of the maximum score estimator of Manski (1987) eliminates the fixed effect in the selection equation, although a quite complicated differencing procedure is used in the second stage to eliminate the fixed effect in the outcome equation. This approach can be generalized to other generalized Tobit models. Charlier, Melenberg, and van Soest (2001) provide an application to a panel version of the Roy model or type 5 Tobit model.

Censoring is common in duration models. Section 23.6 focused on panel models with completed spells. When both complete and incomplete spells are observed for an individual, partial likelihood methods are inappropriate, since censoring is not independent given presence of the time-invariant fixed effect. Horowitz and Lee (2004) propose a consistent estimator for the MPH model (23.43) with incomplete spells that does not require specification of the baseline hazard.

23.9. Practical Considerations

As was the case for linear models, if panel data are used then at a minimum inference needs to be based on panel-robust standard errors. These are not provided by a

computer program for cross-section data unless it has an option for clustered standard errors, in which case clustering is specified to be by the individual.

More efficient estimation is available using models that incorporate serial correlation. Econometricians emphasize random effects. Several packages fit models with normally distributed random effects, using Gaussian quadrature to integrate out the effect, as well as the more specialized analytically tractable random effects count data models. Statisticians instead emphasize the GEE approach for GLMs, available in many statistical packages and some econometrics packages.

These preceding methods lead to inconsistent estimation if the random effect is correlated with regressors. Econometricians therefore emphasize the fixed effects approach. Because of the incidental parameters problem, this yields consistent estimates in short panels for only a subset of nonlinear models. Econometrics packages are available for conditional ML estimation of these models, the fixed effects logit and fixed effects count models. If a fixed effects model is infeasible then random effects models richer than the simplest iid random effects model might be used.

Dynamic panel models can also be estimated. These permit distinction between persistence caused by unobserved heterogeneity and persistence caused by true state dependence. Implementation may require writing one's own programs.

23.10. Bibliographic Notes

This chapter provides an overview of a vast and divergent literature and of necessity skips many details. The monographs on panel data by Arellano (2004), Baltagi (2001), Hsiao (2003), and M.-J. Lee (2002) provide considerable treatment of panel models for binary data and censored and selected models. Panel models for counts are presented in Cameron and Trivedi (1998) and M.-J. Lee (2002). Wooldridge (2002) presents panel methods for binary, censored, and count data. The statistical literature for various generalized linear models is summarized in Fahrmeier and Tutz (1994) and Diggle et al. (1994, 2002). Various papers in Mátyás and Sevestre (1995) consider nonlinear panel models. M.-J. Lee (2002) emphasizes GMM estimation. Arellano and Honore (2001) emphasize semiparametric methods for nonlinear panel models. Bayesian estimation with panel data is presented in Koop (2003).

- 23.2** For general discussion of the incidental parameters problem see Lancaster (2002). Key references are Andersen (1970) for conditional ML and Chamberlain (1992) and Wooldridge (1997a) for differencing methods. For random effects models Butler and Moffitt (1982) detail use of Gaussian quadrature to eliminate normally distributed random effects, whereas the statistics literature emphasizes the LEE approach of Liang and Zeger (1986).
- 23.4** For fixed effects logit models key references are Chamberlain (1980) for static models, Chamberlain (1985) for pure time series dynamic models, and Honore and Kyriazidou (2000) for dynamic models with additional regressors. See also Hsiao (1995).
- 23.5** For selection in panel data see the survey by Vella (1998) and the texts by Baltagi (2001) and Wooldridge (2002).
- 23.6** Chamberlain (1985) presents several ways to eliminate fixed effects in various duration models. Van den Berg (2001, section 6) provides a good discussion and many references. **Event history analysis** using multiple-spells data on individuals is more complicated than most panel analysis as the models are intrinsically dynamic.

- 23.7** The classic reference for panel count data models is Hausman et al. (1984). For dynamic models see Blundell et al. (2002).
- 23.8** For a survey of panel semiparametric methods see Arellano and Honore (2001) and also L.-F. Lee (2001).

Exercises

- 23–1** Consider the nonlinear panel data model $y_{it} = \alpha_i + \exp(\mathbf{x}'_{it}\beta) + u_{it}$, where β are parameters to be estimated, α_i , $i = 1, \dots, N$, are individual specific effects, u_{it} are iid $[0, \sigma^2_\varepsilon]$ errors, and the panel is short.
- (a) Suppose that all $\alpha_i = 0$. Can β be consistently estimated? If yes, provide the formula or objective function for a consistent estimator. If no, give a brief explanation of why β cannot be consistently estimated.
 - (b) Suppose the individual-specific effects α_i are random and are iid $[0, \sigma^2_\alpha]$ distributed independently of the regressors. Can β be consistently estimated? If yes, provide the formula or objective function for a consistent estimator. If no, give a brief explanation of why β cannot be consistently estimated.
 - (c) Suppose the individual specific effects α_i are random but are correlated with the regressors. Can β be consistently estimated? If yes, provide the formula or objective function for a consistent estimator. If no, give a brief explanation of why β cannot be consistently estimated.
- 23–2** (Adapted from Chamberlain, 1980) Show that MLE in a binary logit panel model is inconsistent, with plim of 2β in a simple $T = 2$ model.
- 23–3** Use the same model for the Patents–R&D data as in Section 23.3, except vary the dependent variable and model as suggested in the following. In each case estimate random effects models and, if theoretically feasible, a fixed effects model.
- (a) Use a logit model of whether or not the firm has a patent.
 - (b) Use a truncated tobit model of number of log(Patents) with observations of firms with zero patents dropped.
 - (c) Use a Poisson model for number of patents.

PART SIX

Further Topics

In empirical work data frequently present not one but multiple complications that need to be dealt with simultaneously. Examples of such complications include departures from simple random sampling, clustering of observations, measurement errors, and missing data. When they occur, individually or jointly, and in the context of any of the models developed in Parts 4 and 5, identification of parameters of interest will be compromised. Three chapters in Part 6 – Chapters 24, 26, and 27 – analyze the consequences of such complications and then present methods that control for these complications. The methods are illustrated using examples taken from the earlier parts of the book. This feature gives points of connection between Part 6 and the rest of the book.

Chapter 24, which deals with several features of data from complex surveys, notably stratified sampling and clustering, complements various topics covered in Chapters 3, 5, and 16. Chapter 26 which deals with measurement errors in models studied in Chapters 4, 14, and 20. Chapter 27 is a stand-alone chapter on missing data and multiple imputation, but its use of the EM algorithm and Gibbs sampler also gives it points of contact with Chapters 10 and 13, respectively.

Chapter 25 presents treatment evaluation. Treatment is a broad term that refers to the impact of one variable, e.g. schooling, on some outcome variable, e.g. earnings. Treatment variables may be exogenously assigned, or may be endogenously chosen. The topic of treatment evaluation concerns the identifiability of the impact of treatment on outcome, as measured by either the marginal effects or certain functions of the marginal effect. A variety of methods are used including instrumental variables regression and propensity score matching. The problem of treatment evaluation can arise in the context of any model considered in parts 4 and 5. This chapter emphasizes the linear regression model, so may be read early on. However, it does presume familiarity with many other topics covered in the book, including instrumental variables and selection models. For this reason this topic of growing importance is placed in the last part of the book.

Stratified and Clustered Samples

24.1. Introduction

Microeconomics research is usually performed on data collected by survey of a sample of the population of interest. The simplest statistical assumption for survey data is **simple random sampling** (SRS), under which each member of the population has equal probability of being included in the sample. Then it is reasonable to base statistical inference on the assumption that the data (y_i, \mathbf{x}_i) are independent over i and identically distributed. This assumption underlies the small-sample and asymptotic properties of estimators presented in this book, with the notable exception of sample selection models in Chapter 16.

In practice, however, SRS is almost never the right assumption for survey data. Alternative sampling schemes are instead used to reduce survey costs and to increase precision of estimation for subgroups of the population that are of particular interest.

For example, a household survey may first partition the population geographically into subgroups, such as villages or suburbs, with differing sampling rates for different subgroups. Interviews may be conducted on households that are clustered in small geographic areas, such as city blocks. The data (y_i, \mathbf{x}_i) are clearly no longer iid. First, the distribution of (y_i, \mathbf{x}_i) may vary across subgroups, so the identical distribution assumption may be inappropriate. Second, since data may be correlated for households in the same cluster, the assumption that (y_i, \mathbf{x}_i) are independent within the cluster breaks down.

The usual methods employed to obtain the distribution of estimators therefore need to be adapted, and the properties of estimators may depart from results obtained under SRS. This is the subject of this chapter.

The consequences for regression modeling are the following. First, **weighted estimators** that adjust for differences in sampling rates may be necessary if the goal of analysis is prediction of population behavior. Second, such weighting is unnecessary if interest lies in regression of y on \mathbf{x} , provided the conditional model for y given \mathbf{x} is correctly specified and **stratification** is not on the dependent variable. Third, if samples are determined in part by the value of the dependent variable, such as an oversample

of low-income people when income is the dependent variable, weighted estimation is necessary. A range of estimation procedures are possible, with some presented in Chapter 16 in the context of sample selection bias. Fourth, clustering at a minimum leads to standard error estimates that appreciably understate the true standard errors and can even lead to inconsistent parameter estimates unless adjustment is made for clustering using methods similar to those presented in Chapter 21 for panel data analysis.

The most important implication for most microeconomics applications using survey data is the need to control for clustering. Clustering of observations is often found in both cross-section and panel data, as a consequence of (1) sampling design, (2) design of a social experiment, or (3) the nature of the observation method. An example of (1) is a **complex large-scale household survey** in which spatial clusters of households are sampled to reduce the cost of surveys. An example of (2) is a randomized social experiment in which a common treatment is assigned to individuals in a particular location such as an industrial plant or a school. Examples of (3) are regressions with individual cross-section data when regressors also include group averages such as unemployment or tax rates at the state level, use of panel data, and use of siblings data even if there is no clustering of households.

Section 24.2 introduces some of the concepts and terminology of survey sampling. Sections 24.3–24.5 consider the three key features of survey data, respectively, sample weights, stratification, and clustering. Section 24.6 considers hierarchical linear models where both stratification and clustering are present. An application to data is presented in Section 24.7. Complex surveys are considered further in Section 24.8.

24.2. Survey Sampling

Survey sampling has been well researched in the statistics literature, since data collection must be done before any analysis, and surveying can be very expensive. The goal of the survey literature is usually to obtain with minimal cost a sample that can provide unbiased and reasonably precise estimates of population parameters, especially the population mean.

The structure of a multistage survey was described in Section 3.2. The U.S. CPS is a leading example of such a sample design.

24.2.1. Current Population Survey

The CPS is a monthly survey of approximately 56,000 households that is intended to be representative of the civilian noninstitutional population 16 years and older. Households in smaller states are **oversampled** to provide more reliable state-level data. Within states the surveyed households are clustered to reduce interview costs. Specifically, households are interviewed in four consecutive months, rested for eight months, and then interviewed for another four months. Reinterviewing reduces survey costs and the 4–8–4 schedule permits some longitudinal analysis, including one-year differences. There are eight **rotation groups** of similar size, with a new rotation

group being introduced each month. We consider the sampling design for one rotation group.

Specifically, there are 792 strata, where each stratum is a subregion of a state or in some cases an entire state. The 792 strata are split into 2,007 PSUs, where a PSU may be a metropolitan statistical area (MSA), a state–MSA intersection if the MSA covers more than one state, a single county, or two or more contiguous counties, with departures from this scheme when a PSU has low population or large area. On average there are 2.5 PSUs per strata. Of the 792 strata, 432 contain only one PSU, in which case the PSU is called **self-representing** and is always included in the CPS survey. The other 360 strata have more than one PSU, and exactly one PSU is randomly chosen from the strata with probability proportional to the 1990 population.

Within the PSUs there are no intermediate SSUs. The survey directly samples USUs, a geographically compact group of approximately four addresses. The sampling probability increases if there was low probability of drawing the PSU from its strata and usually increases if the PSU is in a small state, to permit oversampling in low-population states. (In this calculation New York and Los Angeles are treated as states.) All households in the USU are surveyed, unless the USU has an unusually high number of households, in which case a subset of households is randomly chosen.

The CPS is designed to be self-weighting by state so that, despite the use of nonrandom sampling, the CPS should provide a representative sample for each state. However, the unweighted sample is not nationally representative because of the oversampling of low population states and because not all PSUs are sampled.

24.2.2. Sampling

Before moving to a more detailed analysis of survey sampling, we provide a brief overview of sampling basics in the absence of complications such as stratification.

Let \mathbf{z} denote a vector of variables, where there is no need to distinguish between dependent and regressor variables. We assume that in the population the variable \mathbf{z} is iid with density $f(\mathbf{z})$. The population is of size N^* and the sample is of size N . The sample is $\{\mathbf{z}_i, i = 1, \dots, N\}$, where i denotes the i th sampling unit. The usual notation in the sampling literature is n for sample size and N for population size. We instead continue to use N for sample size as there is only occasional need to introduce the population size N^* .

Exhaustive Sampling

Under **exhaustive sampling** every element of the population is sampled, so the sample is the population. Such sampling is rare with individual-level data. It does happen in a population census such as the U.S. decennial census. Yet even for the census, subsampling is used for the lengthier questionnaires, researchers may prefer to work with a more manageable census subsample, and in practice the coverage of the census is incomplete. Exhaustive sampling is more common with firm-level data, where, for example, all firms in an industry may be studied.

Exhaustive sampling can lead to debate about whether the usual inferential methods are warranted, as the sample moments then equal the population moments. The usual procedure is to still use the usual inferential methods. This is done by viewing the finite population to be a sample from an infinite **superpopulation**.

For example, suppose interest lies in gender differences in salary at a work site that has a population total of 20 men and 12 women performing similar tasks. Salaries are obtained for all men and women at the work site, so the sample is the population, and mean salary is found to be higher for men than women. It is customary to perform conventional hypothesis tests on the differences in mean salary, rather than to conclude that since the sample mean equals the population mean there is 100% certainty that male salaries are higher. The rationale is that the population at this particular work site is viewed as a sample from a superpopulation of work sites, or from a superpopulation of the particular work site at many points in time.

Exhaustive sampling is expensive and is generally unnecessary for large populations unless the actual population size needs to be determined. Instead, a subset of the population is usually sampled.

Simple Random Sampling

A **simple random sample** is one where observations are drawn from the population at random and with equal probability. Each observation appears in the sample, with probability equal to the sample size divided by the population size, and has the same marginal density $f(\mathbf{z})$. The prefix “simple” is added because more systematic sampling methods still usually have a random element.

Finite-Sample Correction

Most econometric analysis presumes that SRS leads to draws of \mathbf{z} that are independent, so the joint density of the sample under SRS is the product of the individual densities $f(\mathbf{z}_i)$. This is reasonable if the SRS is obtained from an infinite population, as is the case if sampling is viewed to be from a superpopulation, or if it is obtained from a finite population and sampling is with replacement.

In practice for finite populations an SRS is obtained **without replacement**, to ensure that the same observation does not appear in the sample twice. Then observations are no longer independent, even under SRS. To see this, note that under SRS the probability of any particular member of the population appearing in the sample is N/N^* . Given knowledge that this member appears in the sample, however, the probability of any other member appearing in the sample falls to $(N - 1)/(N^* - 1)$. Clearly, the conditional probability differs from the unconditional probability. More formally, one introduces indicator variables for whether each case in the population appears in the sample. These indicator variables are joint multinomial distributed with means π , variances $\pi(1 - \pi)$, and covariances $-\pi(1 - \pi)/(N^* - 1)$, where $\pi = N/N^*$.

The correlation between sample observations is $\rho = -1/(N^* - 1)$, where ρ is called the **intraclass correlation**. Letting z be a scalar, we have that the sample mean $\bar{z} = N^{-1} \sum_i z_i$ has variance $V[\bar{z}] = N^{-2}V[\sum_i z_i]$, which does not simplify to

$N^{-2} \sum_i V[z_i]$ owing to the correlation of the z_i . Some algebra given, for example, in Cochran (1977, pp. 23–24) yields

$$V[\bar{z}] = (1 - f) \frac{S^2}{N},$$

where $f = N/N^*$ is the sampling fraction, and results in this literature are usually simpler to express in terms of $S^2 = (N^* - 1)^{-1} \sum (z_i - \bar{z})^2$ rather than the usual finite population variance $\sigma^2 = N^{*-1} \sum (z_i - \bar{z})^2$.

Thus for sampling without replacement from a finite population, the variance of the sample mean equals the usual S^2/N multiplied by the **finite-sample correction term** $1 - f$. This correction term appears in statistical packages for survey data. Failure to allow for the finite-sample correction term leads to conservative statistical inference as $V[\bar{z}]$ will be overestimated. For regression using data from SRS with replacement, a finite-sample correction is similarly relevant, though the extent and direction of bias in the variance of the OLS estimator now additionally depends on the design matrix.

The finite-sample correction term is usually ignored in microeconomics. This is often reasonable. For example, for household survey data the sample size is small relative to population size so that $f = N/N^* \rightarrow 0$.

24.3. Weighting

Household surveys such as the CPS are usually constructed in a way that leads to different households having different probabilities of inclusion in the sample. Sample weights are assigned to each observation to correct for this.

As explained in the following, provided stratification is exogenous, weights should be used if regression is viewed as a tool to describe population responses but need not be used if the regression model is assumed to be the correct structural model.

24.3.1. Sample Weights

Suppose each household in the population has a probability π_i of appearing in the sample and assume that, unlike SRS, this probability varies across households.

Statistics such as overall sample means that give equal weight to all observations will then tend to give too much weight to households that appear with high probability in the sample. This can be corrected by weighting, using **sample weights** that are inversely proportional to the probability of inclusion in the sample:

$$w_i \propto 1/\pi_i. \tag{24.1}$$

For example, instead of $\bar{z} = N^{-1} \sum_i z_i$ we may use the weighted mean

$$\bar{z}_W = N^{-1} \sum_i w_i z_i / \sum_i w_i.$$

Note that all that matters in (24.1) is proportionality. The weights need not sum to one, provided we divide by the sum of the weights. A common scaling is $\sum_i w_i = N^*$,

in which case a weight of w_i means that the observation represents w_i households in the population. Note that care is needed in using weights. Some references instead define $w_i \propto \pi_i$, and some computer packages compute the weighted mean as $\sum_i(z_i/w_i)/\sum_i(1/w_i)$. It is easy to incorrectly weight by the reciprocal of the sample weights.

For an SRS of size N from a finite population of size N^* , $\pi_i = 1/N^*$, so w_i is constant and $\bar{z}_W = \bar{z}$.

For **simple stratified sampling** with SRS within strata, suppose it is known that a fraction H_s of the population size N^* is in strata s and that N_s observations are from the s th strata. Then $\pi_i = N_s/H_s N^*$. It follows that the sample weights $w_i \propto H_s/N_s$.

For **two-stage sampling without stratification** let π_c be the probability that the c th PSU is selected and π_{jc} be the probability that household j is selected in PSU c . Then the sample weights $w_{jc} \propto 1/(\pi_c N_c \pi_{jc} N)$, where N_c is the number of survey households in the c th PSU and $N = \sum_c N_c$. A two-stage sample is **self-weighting** if the sampling probabilities at each stage are proportional to population numbers, so $\pi_c = N_c^*/N^*$ and $\pi_{jc} = 1/N_c^*$, where N_c^* is the population size for the c th PSU. Then the weights w_{jc} are equal as in SRS, though estimator standard errors may still have to be adjusted for the two-stage sampling as shown in Section 24.8.

For the CPS, which oversamples households in small states, it would appear sufficient to use $w_i \propto H_s/N_s$, where s denotes state. The CPS uses this as a baseweighting, but then adjusts for subsampling within the USU if the USU has too many households. A further complication is that not all PSUs in a strata are surveyed; consequently, the surveyed households in a strata may not be representative of the strata if the sampled PSUs differ considerably from strata norms. This leads to two additional adjustments. First, adjust for unrepresentative racial (black/nonblack) composition at the strata level. Second, adjust weights to ensure that sample estimates for key subgroups (formed by state, race, sex, and age) match independent population data. For details see U.S. Bureau of Census (2002). The CPS sample weights are constructed to permit the CPS to provide nationally representative statistics, controlling for the composition of the CPS differing from that of the U.S. civilian population on the dimensions of state, race, sex, and age.

The actual computation of sample weights for multistage surveys involves estimation procedures that can be quite complicated. The weights can be misestimated; even if they are correctly estimated they may take into account only some of the dimensions of sample nonrepresentativeness.

24.3.2. Weighted Regression

Should one perform weighted regression when sample weights are provided? We consider this issue in detail when the stratification is not on the dependent variable. Stratification on the dependent variable is examined in Section 24.4.

Consider estimation of the linear regression

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + u_i, \quad (24.2)$$

given survey data with sampling weights w_i . Two possible estimators are OLS,

$$\hat{\beta}_{\text{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \quad (24.3)$$

and WLS using the sampling weights,

$$\hat{\beta}_{\text{WLS}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}, \quad (24.4)$$

where $\mathbf{W} = \text{Diag}[w_i]$.

Correctly Specified Conditional Mean

The OLS estimator is appropriate if it is assumed that $E[\mathbf{u}|\mathbf{x}] = \mathbf{0}$ so that the conditional mean is linear in \mathbf{x} ,

$$E[y_i|\mathbf{x}_i] = \mathbf{x}_i'\beta. \quad (24.5)$$

Then OLS is consistent for β . Furthermore, it is second-moment efficient by the Gauss–Markov Theorem if the errors u_i are homoskedastic. The WLS estimator is also consistent for β under these assumptions but will be inefficient if errors are homoskedastic (since the weighting in (24.5) controls for unrepresentativeness of the sample rather than heteroskedasticity).

Incorrectly Specified Conditional Mean

In many applications (24.5) does not hold. Examples include cases with omitted regressors or situations when $E[y|\mathbf{x}]$ is nonlinear in \mathbf{x} or $E[y_i|\mathbf{x}_i] = \mathbf{x}_i'\beta_i$ where some components of β_i are correlated with \mathbf{x}_i . Linear regression can still be interpreted as the best linear prediction of y given \mathbf{x} under squared error loss, though this needs to be adapted to allow for unrepresentative sampling.

In the population, (y_i, \mathbf{x}_i) are iid, and from Section 4.2 we can always write

$$y_i = \mathbf{x}_i'\beta^* + u_i,$$

where $E[u] = 0$ and $\text{Cov}[\mathbf{x}, u] = 0$ and

$$\beta^* = (E[\mathbf{x}\mathbf{x}'])^{-1} E[\mathbf{x}y].$$

Note that it is no longer assumed that $E[\mathbf{u}|\mathbf{x}] = \mathbf{0}$, so it is possible that $E[y|\mathbf{x}] \neq \mathbf{x}'\beta$.

The parameter β^* is called the **census coefficient** by DuMouchel and Duncan (1983). It is the probability limit of the regression coefficient that would be obtained by regression of y on \mathbf{x} using the entire population rather than an unrepresentative sample.

If the conditional mean is nonlinear in \mathbf{x} and the sample is unrepresentative of the population, then the OLS estimator generally does not converge to β^* , since with unrepresentative samples $N^{-1}\mathbf{X}'\mathbf{X}$ does not converge to the population moment $E[\mathbf{x}\mathbf{x}']$ and similarly for $N^{-1}\mathbf{X}'\mathbf{y}$. Intuitively, if the conditional mean is nonlinear in \mathbf{x} then there is no reason to believe that linear regressions using quite different surveys of the same population will yield the same OLS estimates.

However, WLS using sample weights may consistently estimate β^* . Specifically, if the weighting matrix \mathbf{W} is such that

$$\begin{aligned} N^{-1}\mathbf{X}'\mathbf{W}\mathbf{X} &\xrightarrow{P} \mathbf{E}[\mathbf{x}\mathbf{x}'], \\ N^{-1}\mathbf{X}'\mathbf{W}\mathbf{y} &\xrightarrow{P} \mathbf{E}[\mathbf{x}\mathbf{y}], \end{aligned} \quad (24.6)$$

then $\widehat{\beta}_{WLS}$ defined in (24.4) converges to β^* .

Simple Stratified Samples

Much of the analysis of weighted LS estimation is presented for simple stratified sampling with SRS within strata. Then it is clear that (24.6) is satisfied with $w_i \propto H_s/N_s$ if the i th interviewed household is in the s th strata.

This literature also considers the possibility of different regression parameters within strata. It is assumed that $\mathbf{E}[y_i|\mathbf{x}_i] = \mathbf{x}'_i\beta_s$ for household i in strata s . The goal may be to estimate the population-weighted parameter $\beta_W = N^{-1} \sum_s N_s^* \beta_s$. Then in general neither OLS nor WLS converge to β_W , unless the β_s are equal across strata or are iid with constant mean across strata. A notable exception to this result is estimation of the mean of y (regression with $\mathbf{x} = 1$), in which case the weighted average of the strata sample means is unbiased for the population mean. For details see Section 24.4.1 and DuMouchel and Duncan (1983), Deaton (1997), or Ullah and Breunig (1998).

Should One Use Sample Weights?

The preceding analysis can be used to answer the question of whether to use sample weights in estimation, assuming there is **no endogenous stratification**. The discussion considers estimation of (possibly nonlinear) models of $\mathbf{E}[y|\mathbf{x}]$, but it also applies to models of any other feature of the conditional distribution of y given \mathbf{x} such as the median or the density.

If one takes a **structural** or **analytical** approach and assumes that the model of $\mathbf{E}[y|\mathbf{x}]$ is correctly specified, there is no need to use sample weights. Results can be used to analyze effects of changes in \mathbf{x} on $\mathbf{E}[y|\mathbf{x}]$.

If one instead takes a **descriptive** or **data summary** approach then weights should be used. Regression is then interpreted as estimating census coefficients. A major caveat, however, is that in complicated surveys it is not possible to compute weights that so clearly satisfy (24.6) as was the case for stratified sampling with SRS within strata. In practice sampling weights are constructed to match population proportions for some subgroups based on age, sex, and race. There is no guarantee that such weights will satisfy (24.6).

Some data sets, such as relatively small longitudinal surveys of a few thousand households, are developed with a structural modeling approach in mind. Nonetheless, they usually attempt to provide a reasonably representative sample of the population while using clustered sampling to keep down survey costs. Other data sets, such as the CPS, are designed to provide accurate descriptive measures such as national and regional estimates of unemployment rates. Here designers of the survey are taking a

census approach and indeed would prefer a monthly census if it were not so expensive to conduct.

For either sort of data set microeconometricians usually strive to take the **structural modeling approach**. As an example, consider regression of earnings on schooling level and socioeconomic characteristics such as age, sex, and race, but not measures of innate ability.

Most econometricians would only give a descriptive interpretation to the coefficient of schooling in an OLS regression because of the endogeneity of schooling. The interpretation then is that if we hold certain key regressors constant, one more year of schooling is associated with, but does not necessarily cause, a 6% increase, say, in earnings. Here using sample weights in OLS regression is appropriate to permit estimates to be interpreted as measuring associations in the population, rather than merely those in a possibly unrepresentative sample. Even though no causal interpretation is possible, this estimate can be useful as it does measure how income varies across educational groups after controlling for some other key socioeconomic variables. After all, a major goal of statistics is data summary.

A consistent estimate of the schooling coefficient may be obtained using more advanced estimation methods, such as instrumental variables or panel data methods. Then the coefficient can be given a causal interpretation. Weighting by sample weights is no longer necessary, though the usual weighting to improve efficiency if, for example, errors are heteroskedastic, may be appropriate.

Whether a model can be interpreted as correctly specified is a judgement call. If it is correctly specified then sample weighted and unweighted estimates should have the same probability limit, since both are consistent. This suggests testing correct model specification by a Hausman test of the difference between sample-weighted and unweighted regressors, a test proposed by DuMouchel and Duncan (1983) in the case of linear regression.

24.3.3. Prediction

Consider nonlinear regression with correctly specified conditional mean, $g(\mathbf{x}, \beta)$, and no endogeneity. The unweighted NLS estimator consistently estimates β and can be given a causal interpretation. In particular, we can use $\partial g(\mathbf{x}, \hat{\beta}) / \partial \mathbf{x}$ to calculate the causal effect of a one-unit change in \mathbf{x} of the conditional mean.

This predicted effect varies with the evaluation point \mathbf{x} , since $g(\cdot)$ is nonlinear. An estimate of the average response in the population is

$$\widehat{E} \left[\frac{\partial y}{\partial \mathbf{x}} \right] = \sum_{i=1}^N w_i \frac{\partial g(\mathbf{x}_i, \hat{\beta})}{\partial \mathbf{x}_i},$$

where w_i are the sample weights. Similarly, if one instead evaluates the response at the mean of the regressors it may be better to use the weighted sample mean of \mathbf{x} , an estimate of the population mean of \mathbf{x} , rather than the unweighted sample mean of \mathbf{x} .

Even if the parameters are consistently estimated using unweighted estimation, weighting must be used in subsequent impact calculations if one wishes to predict population impacts, rather than sample impacts.

24.4. Endogenous Stratification

Stratification is widely used as it can increase precision of estimation, or equivalently reduce survey costs for a given level of precision. For example, more precise estimation of the mean unemployment rate in low-population states may be obtained by oversampling poor states. For similar reasons minority groups may be oversampled.

One complication, already considered in Section 24.3, is that parameters may vary across strata. For example, the mean unemployment rate may vary across strata. Then a descriptive approach is taken and weighted estimators are used.

Microeconometricians often prefer a structural approach and assume parameters are unchanging across strata. Then from Section 24.3 stratification apparently causes no complications and unweighted regression is used. A major proviso is that problems still arise if stratification is based on the value of the dependent variable. For example, if low-income people are purposely oversampled and income is the dependent variable then the usual regression estimators are inconsistent. Note that there is no problem if stratification is on regressors such as race and this leads indirectly to oversampling of low-income people. Problems only arise if stratification is directly on income.

In this section we define endogenous stratification and analyze the resulting complications. We then present several estimators that are consistent. The simplest is a weighted estimator that can be used if both the sample and population strata probabilities are known. The method is given in Section 24.4.5, which is self-contained.

24.4.1. Stratification Schemes

For general data $\mathbf{z} \in \mathcal{Z}$ the strata are subsets of \mathcal{Z} . Econometric analysis usually partitions the data into dependent variable $\mathbf{y} \in \mathcal{Y}$, where for generality we allow \mathbf{y} to be a vector, and regressor or independent variable $\mathbf{x} \in \mathcal{X}$. The strata \mathcal{C}_s , for $s = 1, \dots, S$, are then defined to be subsets of the sample space $\mathcal{Y} \times \mathcal{X}$. The notation is that used by Imbens and Lancaster (1996), who present some leading examples that are reproduced in Table 24.1.

Sampling within strata is assumed to be random but some strata may be oversampled. From Table 24.1 it is clear that the strata may sum to less than the sample space or more than the sample space. For the fourth and fifth schemes the stratification may be solely on endogenous variables, solely on exogenous variables, or on a mixture of the two.

The econometrics literature has focused on sampling schemes with an endogenous component, since in that case the usual conditional MLE is inconsistent.

Endogenous stratification has already been considered in Chapter 16. As an example, consider **truncated regression**, where we observe y only if $y > 0$, so stratification is purely on y . Then for sampled data the conditional density of y given \mathbf{x} is a zero-truncated density that divides the untruncated density by $\Pr[y > 0 | \mathbf{x}]$ and so

$$f^s(y|\mathbf{x}, \boldsymbol{\theta}) = \frac{f(y|\mathbf{x}, \boldsymbol{\theta})}{1 - F(0|\mathbf{x}, \boldsymbol{\theta})},$$

Table 24.1. Stratification Schemes with Random Sampling within Strata

Stratification Scheme	Definition	Stratum Description
Simple random	$S = 1, \mathcal{C}_1 = \mathcal{Y} \times \mathcal{X}$	One stratum covers entire sample space.
Pure exogenous	$\mathcal{C}_s = \mathcal{Y} \times \mathcal{X}_s$, with $\mathcal{X}_s \subset \mathcal{X}$	Stratify on regressors only, not on dependent variable
Pure endogenous	$\mathcal{C}_s = \mathcal{Y}_s \times \mathcal{X}$, with $\mathcal{Y}_s \subset \mathcal{Y}$	Stratify on dependent variable only, not on regressors
Augmented sample	$S = 2, \mathcal{C}_1 = \mathcal{Y} \times \mathcal{X}$, and $\mathcal{C}_2 \subset \mathcal{Y} \times \mathcal{X}$.	Random sample augmented by extra observations from part of the sample space
Partitioned	$\mathcal{C}_s \subset \mathcal{Y} \times \mathcal{X}, \mathcal{C}_s \cap \mathcal{C}_t = \emptyset$, and $\bigcup_{s=1}^S \mathcal{C}_s = \mathcal{Y} \times \mathcal{X}$.	Sample space split into mutually exclusive strata that fill the entire sample space

where the superscript s is used to distinguish the **sample density** from the **population density** $f(y|\mathbf{x}, \boldsymbol{\theta})$. As discussed in Chapter 16, this sampling scheme tends to drop observations with low realizations of y , given \mathbf{x} . Suppose $E[y|\mathbf{x}] = \beta_1 + \beta_2 x$ and $\beta_2 > 0$. Then for low values of x there will be too many relatively high values of y . The regression will accordingly overpredict $E[y|\mathbf{x}]$ for low values of x , leading to upward bias in the intercept β_1 and downward bias in the slope β_2 .

A second example is **choice-based sampling** for binary or multinomial data where samples are chosen based on the discrete outcome y . For example, if choice is between travel to work by bus or travel by car we may oversample bus riders if relatively few people commute by bus. This example is pursued in the following. It is similar to **case-control studies** in the medical literature where, for example, a complete sample of people who died from a disease ($y = 1$) is contrasted with a similar-sized subsample of the universe of people who did not die of the disease ($y = 0$). The goal is to find whether one or more regressors are able to predict $y = 1$.

A related example is count data on number of visits collected by **on-site sampling** of users, such as sampling at recreational sites or shopping centers or doctor's offices. Then data are truncated, since those with $y = 0$ are not sampled, and additionally high-frequency visitors are oversampled. Shaw (1988) shows that the sampling distribution of the data, $f^s(y|\mathbf{x}, \boldsymbol{\theta})$, is related to the population distribution through the equation

$$f^s(y|\mathbf{x}, \boldsymbol{\theta}) = f(y|\mathbf{x}, \boldsymbol{\theta}) \frac{y}{E[y|\mathbf{x}, \boldsymbol{\theta}]}.$$

In this case the sampling scheme is clearly endogenous though it is not a stratified sampling scheme.

24.4.2. Endogeneity Induced by Stratification

Sampling schemes such as stratification schemes lead to the density in the sample differing from that in the population. If stratification is purely exogenous, then despite this difference the usual MLE is still consistent because the conditional density of y given \mathbf{x} in the sample is the same as that in the population. However, if any aspect of stratification is endogenous, then these conditional densities differ, as illustrated by the preceding examples. We now provide a detailed discussion of this point.

The goal of ML estimation lies in consistently estimating the parameters $\boldsymbol{\theta}$ of $f(y|\mathbf{x}, \boldsymbol{\theta})$. In general the MLE should be based on the likelihood formed from the joint distribution of the data (y, \mathbf{x}) . In practice it is often sufficient to simply form a conditional likelihood from the conditional distribution of y given \mathbf{x} . This simpler approach can lead to consistent estimation under the assumption that \mathbf{x} is **exogenous** with respect to y , in which case the joint density factorizes as

$$g(y, \mathbf{x}|\boldsymbol{\theta}) = f(y|\mathbf{x}, \boldsymbol{\theta}) \times h(\mathbf{x}), \quad (24.7)$$

where the parameters of the density of \mathbf{x} are suppressed as there is no desire to estimate these parameters.

It is always the case that we can write $g(y, \mathbf{x}) = f(y|\mathbf{x}) \times h(\mathbf{x})$. The assumption made in (24.7) is that, upon introduction of parameters, $\boldsymbol{\theta}$ appears in $f(y|\mathbf{x}, \boldsymbol{\theta})$ but does not appear in $h(\mathbf{x})$. In general, rather than (24.7) we may have

$$g(y, \mathbf{x}|\boldsymbol{\theta}) = f(y|\mathbf{x}, \boldsymbol{\theta}) \times h(\mathbf{x}|\boldsymbol{\theta}). \quad (24.8)$$

Then one or more components of \mathbf{x} are **endogenous** with respect to y since there is now feedback – y depends on \mathbf{x} but \mathbf{x} in turn depends on y via the presence of $\boldsymbol{\theta}$ in $h(\mathbf{x}|\boldsymbol{\theta})$. A classic example of this is linear simultaneous equations. In such cases ML estimation should be based on the **joint likelihood**

$$\ln L_{\text{JOINT}}(\boldsymbol{\theta}) = \sum_{i=1}^n \ln f(y_i|\mathbf{x}_i, \boldsymbol{\theta}) + \sum_{i=1}^n \ln h(\mathbf{x}_i|\boldsymbol{\theta}). \quad (24.9)$$

This yields a consistent estimate of $\boldsymbol{\theta}$ if, from Chapter 5,

$$\mathbf{0} = E \left[\frac{\partial \ln g(y, \mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right] = E \left[\frac{\partial \ln f(y|\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right] + E \left[\frac{\partial \ln h(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right]. \quad (24.10)$$

Condition (24.10) is satisfied if the density $g(y, \mathbf{x}|\boldsymbol{\theta})$ is correctly specified and the range of the data does not depend on $\boldsymbol{\theta}$. The conditional MLE instead maximizes the **conditional likelihood**

$$\ln L_{\text{COND}}(\boldsymbol{\theta}) = \sum_i \ln f(y_i|\mathbf{x}_i, \boldsymbol{\theta}).$$

The conditional MLE is consistent if $E[\partial \ln f(y|\mathbf{x}, \boldsymbol{\theta})/\partial \boldsymbol{\theta}] = \mathbf{0}$. This necessary condition is implied by (24.10) if \mathbf{x} is exogenous, since (24.10) simplifies because then $\partial \ln h(\mathbf{x})/\partial \boldsymbol{\theta} = \mathbf{0}$. If instead \mathbf{x} is endogenous this simplification does not occur, as the second term on the right-hand side of (24.10) does not disappear. So the conditional MLE is inconsistent if \mathbf{x} is endogenous.

The problem that arises with stratification and similar sampling schemes is that even if the population joint density satisfies (24.7) and is the same across strata, the sampling schemes can lead to joint density for (y, \mathbf{x}) in the sample that takes the more general form

$$g^s(y, \mathbf{x}|\boldsymbol{\theta}) = f^s(y|\mathbf{x}, \boldsymbol{\theta}) \times h^s(\mathbf{x}|\boldsymbol{\theta}), \quad (24.11)$$

where the superscript s is used to denote dependence on the particular sampling scheme employed. Then the conditional MLE may be inconsistent, even though it would be consistent if the sample was instead an SRS.

Under **pure exogenous sampling** the only difference between sample and population distribution occurs for the marginal distribution of \mathbf{x} . Assuming (24.7) holds in the population, then in the sample

$$g^s(y, \mathbf{x}|\boldsymbol{\theta}) = f(y|\mathbf{x}, \boldsymbol{\theta}) \times h^s(\mathbf{x}).$$

Clearly, the conditional MLE will be consistent as the conditional density is still $f(y|\mathbf{x}, \boldsymbol{\theta})$ and $\boldsymbol{\theta}$ does not appear in $h^s(\mathbf{x})$.

Under **endogenous sampling** the more general result (24.11) holds in the sample even if (24.7) holds in the population. The sample and population conditional distributions of y given \mathbf{x} may differ, with $f^s(y|\mathbf{x}, \boldsymbol{\theta}) \neq f(y|\mathbf{x}, \boldsymbol{\theta})$, and $h^s(\mathbf{x}|\boldsymbol{\theta})$ may possibly depend on $\boldsymbol{\theta}$.

24.4.3. Endogenous Sampling

Under pure endogenous sampling the marginal distribution of y in the sample differs from that in the population. Let $h(y)$ denote the population density of y and $h^s(y)$ denote the sampling density of y . (We are using the convention that g , f , and h denote, respectively, joint, conditional, and marginal distributions. It should be clear to the reader that $h(y)$ differs from $h(\mathbf{x})$.)

The joint distribution of y and \mathbf{x} under pure endogenous sampling is best obtained by first conditioning on \mathbf{x} , rather than y . Then

$$g^s(y, \mathbf{x}) = f(\mathbf{x}|y)h^s(y), \quad (24.12)$$

where simplification has occurred because the conditional distribution of \mathbf{x} given y is unaffected under pure endogenous sampling and so $f^s(\mathbf{x}|y) = f(\mathbf{x}|y)$. We now need to reexpress $f(\mathbf{x}|y)$ in terms of $f(y|\mathbf{x})$. Now

$$\begin{aligned} f(\mathbf{x}|y) &= \frac{g(y, \mathbf{x})}{h(y)} \\ &= \frac{f(y|\mathbf{x})h(\mathbf{x})}{h(y)}. \end{aligned} \quad (24.13)$$

Substituting (24.13) into (24.12) and rearranging yields

$$g^s(y, \mathbf{x}|\boldsymbol{\theta}) = f(y|\mathbf{x}, \boldsymbol{\theta}) \times \frac{h^s(y)}{h(y|\boldsymbol{\theta})} \times h(\mathbf{x}),$$

where

$$\begin{aligned} h(y|\boldsymbol{\theta}) &= \int g(y, \mathbf{x}|\boldsymbol{\theta})d\mathbf{x} \\ &= \int f(y|\mathbf{x}, \boldsymbol{\theta})h(\mathbf{x})d\mathbf{x}. \end{aligned}$$

The conditional MLE using just $f(y|\mathbf{x}, \boldsymbol{\theta})$ will be inconsistent because the term $h(y|\boldsymbol{\theta})$ has been neglected. One instead needs to maximize a joint likelihood that additionally includes $h(y|\boldsymbol{\theta})$.

24.4.4. Endogenously Stratified Samples

We now consider the stratification schemes introduced in Section 24.4.1. The population density is

$$g(y, \mathbf{x}|\boldsymbol{\theta}) = f(y|\mathbf{x}, \boldsymbol{\theta})h(\mathbf{x}).$$

There are S strata where the s th strata is the subset \mathcal{C}_s of $\mathcal{Y} \times \mathcal{X}$.

An important distinction is made between the population probability of an observation being in \mathcal{C}_s and the probability of sampling from \mathcal{C}_s , as the two differ in a stratified sampling scheme. We define

$$\begin{aligned} H_s &= \Pr[\text{Draw an observation from } \mathcal{C}_s], \\ Q_s(\boldsymbol{\theta}) &= \Pr[\text{A randomly drawn observation from the population is in } \mathcal{C}_s]. \end{aligned} \quad (24.14)$$

Here H_s is set by the sample design, whereas

$$Q_s(\boldsymbol{\theta}) = \int_{\mathcal{C}_s} f(y|\mathbf{x}, \boldsymbol{\theta})h(\mathbf{x})dyd\mathbf{x}. \quad (24.15)$$

The strata probabilities may or may not be known. A strata is oversampled if $H_s > Q_s$.

We begin by obtaining the joint density of s , y , and \mathbf{x} , where s is an indicator for the stratum from which the observation was obtained. In the population

$$g(s, y, \mathbf{x}|\boldsymbol{\theta}) = Q_s(\boldsymbol{\theta})g(y, \mathbf{x}|s, \boldsymbol{\theta}).$$

In the sample, the marginal distribution of the strata indicator differs from Q_s , and

$$\begin{aligned} g^s(s, y, \mathbf{x}|\boldsymbol{\theta}) &= H_s g(y, \mathbf{x}|s, \boldsymbol{\theta}) \\ &= H_s \frac{f(y|\mathbf{x}, \boldsymbol{\theta})h(\mathbf{x})}{Q_s(\boldsymbol{\theta})}, \end{aligned}$$

where the second equality holds as $g(y, \mathbf{x}|s)$ equals the density $g(y, \mathbf{x}) = f(y|\mathbf{x})h(\mathbf{x})$ divided by the population probability of being in strata s so that the density integrates over \mathcal{C}_s to one.

It follows that the joint density is

$$g^s(s, y, \mathbf{x}|\boldsymbol{\theta}) = \frac{H_s}{Q_s(\boldsymbol{\theta})} f(y|\mathbf{x}, \boldsymbol{\theta})h(\mathbf{x}), \quad (24.16)$$

where $Q_s(\boldsymbol{\theta})$ is defined in (24.15). The conditional MLE based on the population conditional density $f(y|\mathbf{x}, \boldsymbol{\theta})$ will be inconsistent for $\boldsymbol{\theta}$ since it ignores the term $Q_s(\boldsymbol{\theta})$, which depends on $\boldsymbol{\theta}$.

A variety of consistent estimators have been proposed. Here we consider maximum likelihood estimation, GMM estimation, and a much simpler weighted estimator that can be implemented provided both strata sampling probabilities H_s and population probabilities $Q_s(\theta)$ are known.

Maximum Likelihood Estimation

Performing an ML estimation based on the joint density $g^s(s, y, \mathbf{x}|\theta)$ in (24.16) is complicated because from (24.15) the distribution of $Q_s(\theta)$ depends on $h(\mathbf{x})$. One possible solution is to specify the density $h(\mathbf{x})$. This approach is not taken because econometricians shy away from specifying the distribution of regressors, even if there is a willingness to specify the conditional distribution of the dependent variable.

Instead, a semiparametric approach is taken, with the goal of estimating the parameters of the specified density $f(y|\mathbf{x}, \theta)$, for an unspecified density $h(\mathbf{x})$. For simplicity assume the population strata probabilities H_s are known. Cosslett (1981a) obtained the **MLE with endogenous stratification** by first letting \mathbf{x} be discrete with \mathbf{x}_i occurring with probability w_i , and maximizing the joint likelihood with respect to θ and w_i , $i = 1, \dots, N$. The first-order conditions can be collapsed to yield a concentrated likelihood that involves only $(q + S - 1)$ parameters θ and functions $\lambda_s(\theta)$, $s = 1, \dots, S - 1$. Second, maximizing this concentrated likelihood with respect to θ and λ_s yields the same estimates as maximization with respect to θ and $\lambda_s(\theta)$. Third, since it is valid to treat λ_s as a parameter the same procedure can be used for the case of continuous regressors. A problem of dimension q plus infinite-dimensional unknown density $h(\mathbf{x})$ has been reduced to $q + S - 1$ dimensions.

GMM Estimation

The remarkable results of Cosslett (1981a) are difficult to implement.

Imbens (1992) devised a simpler **GMM estimator with endogenous stratification** that has the same efficiency as Cosslett's MLE. A quite general framework and presentation of this estimator is given by Imbens and Lancaster (1996), for stratified samples obtained by multinomial sampling, standard stratified sampling, or variable probability sampling. The joint density is again $g^s(s, y, \mathbf{x}|\theta)$ in (24.16) and the sample strata probabilities H_s are permitted to be possibly unknown. The GMM analysis is based on $S - 1$ equations for the score of H_s , q equations for θ based on the conditional likelihood function of y given s and \mathbf{x} , $S - 1$ equations for the restrictions on the population strata probabilities $Q_s(\theta)$, and a final restriction that is not necessary if there is a linear restriction on the $Q_s(\theta)$, which happens, for example, if the strata are mutually exclusive and cover the sample space.

24.4.5. Weighted Estimation

Endogenous stratification is easily dealt with when the sample and population strata probabilities, H_s and $Q_s(\theta)$ defined in (24.14), are known, though the estimator is not fully efficient. We begin with ML estimation before considering more general estimators.

Weighted ML Estimation

Manski and Lerman (1977) proposed the **weighted maximum likelihood** (WML) estimator. This maximizes

$$Q_{\text{WML}}(\theta) = \sum_i \frac{Q_i}{H_i} \ln f(y_i | \mathbf{x}_i, \theta), \quad (24.17)$$

where $H_i = H_s$ and $Q_i = Q_s$ if the i th observation is in strata s .

Manski and Lerman (1977) called this estimator the **weighted exogenous sampling estimator** (WESML), since (24.17) multiplies the usual term $\ln f(y_i | \mathbf{x}_i, \theta)$ in the conditional likelihood under exogenous sampling by the weight H_i/Q_i . However, the designation WESML can lead to confusion as the problem here is one of endogeneity – it just turns out that appropriately weighting the usual exogenous estimator leads to consistent estimation.

Along similar lines, the objective function $Q_{\text{WML}}(\theta)$ is not formally a likelihood, since (24.16) does not imply that the sample conditional density of y given \mathbf{x} and s is given by $f^s(y | \mathbf{x}, \theta) = f(y | \mathbf{x}, \theta)^{Q_s/H_s}$. Nonetheless, the WML estimator is consistent. The WML estimator solves the first-order conditions

$$\sum_i \frac{Q_i}{H_i} \frac{\partial \ln f(y_i | \mathbf{x}_i, \theta)}{\partial \theta} = \mathbf{0}. \quad (24.18)$$

This estimator is consistent if the terms in the sum have zero expected value, where expectation is with respect to the sampling density $g^s(s, y, \mathbf{x} | \theta)$ in (24.16). Now

$$\begin{aligned} & E_s \left[\frac{Q_s}{H_s} \frac{\partial \ln f(y | \mathbf{x}, \theta)}{\partial \theta} \right] \\ &= \int \int \frac{Q_s}{H_s} \frac{\partial \ln f(y | \mathbf{x}, \theta)}{\partial \theta} \frac{H_s}{Q_s(\theta)} f(y | \mathbf{x}, \theta) h(\mathbf{x}) dy d\mathbf{x} \\ &= \int \int \frac{\partial \ln f(y | \mathbf{x}, \theta)}{\partial \theta} f(y | \mathbf{x}, \theta) h(\mathbf{x}) dy d\mathbf{x} \\ &= \int E \left[\frac{\partial \ln f(y | \mathbf{x}, \theta)}{\partial \theta} \right] h(\mathbf{x}) d\mathbf{x} \\ &= \mathbf{0}, \end{aligned} \quad (24.19)$$

under the usual regularity condition that in the population the specified density satisfies $E[\partial \ln f(y | \mathbf{x}, \theta) / \partial \theta] = \mathbf{0}$. So the WML estimator is consistent in the presence of endogenous stratification.

The information matrix equality does not hold for objective function $Q_{\text{WML}}(\theta)$ in (24.17), so we need to use the sandwich form $N^{-1} \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1}$ for the asymptotic variance of $\hat{\theta}_{\text{WML}}$, where

$$\mathbf{A}(\theta_0) = \text{plim} \frac{1}{N} \sum_{i=1}^N \frac{Q_i}{H_i} \frac{\partial^2 \ln f(y_i | \mathbf{x}_i, \theta)}{\partial \theta \partial \theta'} \Big|_{\theta_0} \quad (24.20)$$

and

$$\mathbf{B}(\theta_0) = \text{plim} \frac{1}{N} \sum_{i=1}^N \left(\frac{Q_i}{H_i} \right)^2 \frac{\partial \ln f(y_i | \mathbf{x}_i, \theta)}{\partial \theta} \frac{\partial \ln f(y_i | \mathbf{x}_i, \theta)}{\partial \theta'} \Big|_{\theta_0}. \quad (24.21)$$

This estimator is less efficient than the ML estimator of Cosslett or Imbens, but it is relatively straightforward to implement. It does, of course, presume knowledge of the strata probabilities.

Weighted m-Estimation

The weighted ML estimator can be applied to estimators other than conditional ML estimation. For example, Hausman and Wise (1979) consider similar weighted estimation for least-squares regression.

Thus suppose with SRS we would minimize $\sum_i q(y_i | \mathbf{x}_i, \boldsymbol{\theta})$, with first-order conditions $\sum_i \partial q(y_i | \mathbf{x}_i, \boldsymbol{\theta}) / \partial \boldsymbol{\theta} = \mathbf{0}$, and suppose in the population that

$$E[\partial q(y | \mathbf{x}, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}] = \mathbf{0},$$

a necessary condition for consistency. Then if sampling is instead endogenously stratified as in Section 24.2 and the sample and population strata probabilities H_s and Q_s are known, then $\boldsymbol{\theta}$ is consistently estimated by the **weighted m-estimator** $\widehat{\boldsymbol{\theta}}_W$ that minimizes

$$Q_W(\boldsymbol{\theta}) = \sum_i \frac{Q_i}{H_i} q(y_i | \mathbf{x}_i, \boldsymbol{\theta}). \quad (24.22)$$

The proof of consistency follows (24.18) and (24.19) for the WML estimator and the variance matrix is of the form $N^{-1} \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1}$, where \mathbf{A} and \mathbf{B} are given in (24.20) and (24.21) with the sole change being replacement of $\partial \ln f(y_i | \mathbf{x}_i, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ by $\partial q(y_i | \mathbf{x}_i, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}$. Wooldridge (2001) provides a formal proof.

Similarly, for estimation based on the q population moment conditions

$$E[\mathbf{h}(y, \mathbf{x}, \boldsymbol{\theta})] = \mathbf{0},$$

under endogenous stratification, use the **weighted estimating equations estimator** that solves

$$\sum_i \frac{Q_i}{H_i} \mathbf{h}(y_i, \mathbf{x}_i, \boldsymbol{\theta}) = \mathbf{0}.$$

The weighted MLE results apply with $\partial \ln f(y_i | \mathbf{x}_i, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ replaced by $h(y_i | \mathbf{x}_i, \boldsymbol{\theta})$.

Note that the weights Q_i/H_i are the same as those proposed in Section 24.3.2 for estimation of the census parameter under simple exogenous stratified sampling. The motivation, however, is quite different. In the current section it is assumed that conditional moments are correctly specified so that with exogenous stratified sampling it would be consistent and efficient to do unweighted estimation. The weights become necessary if stratification is endogenous.

24.5. Clustering

Sections 24.3 and 24.4 on weighting and stratification covered methods to control for a survey design that leads to a sample distribution that differs from the population distribution. The assumption of independence of sampled observations was maintained.

In fact survey data are usually dependent. This may be due to use of clustered samples to reduce survey costs, such as interviewing several households on the same block. In such cases the data may be correlated within a cluster owing to the presence of a common unobserved cluster-specific term. Such dependence may also arise, however, even with SRS. For example, it may be felt that there is an unobservable effect common to all households in the same state.

There are several different methods for controlling for dependence on unobservables within a cluster. If the within-cluster unobservables are uncorrelated with regressors then only the variances of the regression parameters need to be adjusted. If instead the within-cluster unobservables are correlated with regressors then the regression parameters are inconsistent and suitable alternative estimators are needed. The analysis is further complicated because methods may also vary according to whether there are many small clusters or few large clusters. Additional complex survey complications such as weighting and stratification are deferred to Section 24.6.

The notation and models are presented next, with the key distinction being between random cluster effects and fixed cluster effects, similar to panel data analysis. The various estimators are presented in subsequent sections.

24.5.1. Cluster-Specific Effects Models

Interest lies in estimation of a linear regression model given data $(y_i, \mathbf{x}_i), i = 1, \dots, N$, where i denotes the i th sample observation, such as a household.

The concern is that some aspects of the population regression model vary by cluster c , $c = 1, \dots, C$. Suppose the i th household in the overall sample is the j th household in the c th sampled cluster. A quite general model for clustered data is

$$y_{jc} = \mathbf{x}'_{jc} \boldsymbol{\beta}_c + u_{jc}, \quad j = 1, \dots, N_c, \quad c = 1, \dots, C, \quad (24.23)$$

where $\text{Cov}[u_{jc}, u_{kd}] \neq 0$ though $\text{Cov}[u_{jc}, u_{kd}] = 0$ for $c \neq d$. This model incorporates cluster dependence through both regression parameters that vary across clusters and errors that are correlated within a cluster.

Here we focus on a special case, the **cluster-specific effects model**

$$y_{jc} = \mathbf{x}'_{jc} \boldsymbol{\beta} + \alpha_c + \varepsilon_{jc}. \quad (24.24)$$

Here just the regression intercept α_c varies across clusters, whereas the slope coefficients are assumed to be constant across clusters. In the simplest model ε_{jc} is assumed to be homoskedastic,

$$\varepsilon_{jc} \sim [0, \sigma_\varepsilon^2], \quad (24.25)$$

an assumption that can be relaxed to permit heteroskedasticity and correlation within a cluster. More substantively, different assumptions on α_c lead to two quite different models, which we now present.

Cluster-Specific Random Effects

In the **cluster-specific random effects** (CSRE) **model** the intercepts α_c in (24.24) are purely random with distribution that does not depend on any observables. In the simplest case it is assumed that

$$\alpha_c \sim [0, \sigma_\alpha^2]. \quad (24.26)$$

This model is directly analogous to the random effects model for panel data. The model is just a linear regression of y_{jc} on \mathbf{x}_{jc} , with the complication that the error term $\alpha_c + \varepsilon_{jc}$ is correlated for observations in the same cluster. An OLS estimation is consistent but inefficient. Importantly, the correlation of errors makes it necessary to adjust the usual standard errors of the OLS estimator. A GLS estimation is more efficient.

Given assumptions (24.25) and (24.26) on ε_{jc} and α_c , $V[\alpha_c + \varepsilon_{jc}] = \sigma_\alpha^2 + \sigma_\varepsilon^2$ and $\text{Cov}[\alpha_c + \varepsilon_{jc}, \alpha_c + \varepsilon_{kc}] = \sigma_\alpha^2$, for $k \neq j$. We define the **intraclass correlation coefficient**

$$\rho = \text{Cor}[\alpha_c + \varepsilon_{jc}, \alpha_c + \varepsilon_{kc}] = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2}. \quad (24.27)$$

There is a one-to-one correspondence between $(\sigma_\alpha^2, \sigma_\varepsilon^2)$ and (σ^2, ρ) , where ρ is defined in (24.27) and $\sigma^2 = \sigma_\alpha^2 + \sigma_\varepsilon^2$. The CSRE model is equivalent to a model with constant intraclass correlation coefficient. The model can also be given a Bayesian interpretation, viewing each observation as having its own intercept α_{jc} that is a draw from a univariate distribution and appealing to the **exchangeability criterion** that the subscript in α_{jc} is a purely labeling device and has no substantive consequences. In all cases clustering has the expected effect of inducing positive correlation between error terms within a cluster.

Cluster-Specific Fixed Effects

In the **cluster-specific fixed effects** (CSFE) **model** the intercepts α_c in (24.23) are random unobservables, as for the CSRE model, but may possibly be correlated with the regressors. For identification \mathbf{x}_{jc} no longer includes an intercept term.

This model is directly analogous to the fixed effects model for panel data. The model has conditional mean $E[y_{jc} | \mathbf{x}_{jc}, \alpha_c] = \mathbf{x}'_{jc}\beta + \alpha_c$. The OLS estimator from regression of y_{jc} on \mathbf{x}_{jc} alone is inconsistent for β if the omitted variable α_c is correlated with \mathbf{x}_{jc} . Consistent estimation of β requires consistent estimation of α_c , which is possible if the clusters are large. If clusters are instead small the individual α_c need to be eliminated by a differencing transformation.

Comparison to Panel Data Analysis

The setup and terminology clearly closely parallels that for static panel data analysis presented in Chapters 21 to 23. At the same time there are some departures from panel data analysis.

In the panel case the individual unit of analysis, such as the household, is observed more than once whereas in the cluster case the individual unit of analysis is observed only once. In the panel notation it , the first subscript is the clustering unit if the panel is a short panel, whereas in the clustering notation jc , the second subscript is the clustering unit. In the panel case we focused on balanced panels, but clustered data are usually unbalanced as N_c varies across clusters.

Microeconometrics methods for panel data focus on short panels. This is analogous to having few observations per cluster and many clusters. Then N_c is small and $C \rightarrow \infty$, which we call **small clusters**. In addition, it is not unusual to have **large clusters**, with $N_c \rightarrow \infty$ and C small. For the CSFE model with large clusters there will only be a few parameters α_c to estimate and the incidental parameters problems will not arise.

Unlike as in panel data, the appropriate clustering unit may not always be clear. For example, for the CPS data clustering could be viewed as arising within state, within strata, within PSU, or within USU. This issue is deferred to Section 24.6. The intra-cluster correlation is expected to decrease for clustering at more aggregate levels. If clustering is at the state level then the clusters are large, whereas if clustering is viewed as being at the level of USU then the clusters are small. Moreover, it is possible that a data set does not include necessary clustering information, such as the strata or USU for an observation.

The analogue of dynamic, rather than static, panel data models is a model where y_{jc} depends not only \mathbf{x}_{jc} but also on \mathbf{x}_{kc} , for $k \neq j$. For clustered data it is usually sufficient to specify a **peer-effects model** that more simply includes just the cluster average $\bar{\mathbf{x}}_c$, since the ordering of observations within a cluster usually does not matter.

Overview

The three common estimators for clustering are the OLS, the GLS, and the within estimators presented in Sections 24.5.2–24.5.4. The properties of these estimators, summarized in Table 24.2, vary with the true model. Most importantly, if the true model has cluster-specific fixed effects then OLS and RE estimators are inconsistent, whereas the within estimator yields consistent estimates but only for coefficients of regressors that vary within a cluster. Secondly, even if an estimator is consistent the usual standard errors will often need to be adjusted to control for clustering and possibly heteroskedasticity as detailed in the following.

Table 24.2. Properties of Estimators for Different Clustering Models

Section	Estimator	Cluster Model	Consistent
24.5.2	OLS	Random effects	Yes
		Fixed effects	No
24.5.3	GLS for random effects	Random effects	Yes
		Fixed effects	No
24.5.4	Within for fixed effects	Random effects	Yes
		Fixed effects	Yes

24.5.2. OLS Estimator

We consider the OLS regression

$$y_{jc} = \mathbf{x}'_{jc}\beta + u_{jc}. \quad (24.28)$$

Ordinary LS is inconsistent because of omitted variables bias if the true model is the CSFE model (i.e., $u_{jc} = \alpha_c + \varepsilon_{jc}$) with fixed effect α_c correlated with \mathbf{x}_{jc} . Then the OLS estimator should not be used and instead the CSFE estimators of Section 24.5.4 should be used.

In contrast, OLS is consistent in the CSRE model, where α_c is a random effect uncorrelated with \mathbf{x}_{jc} . More generally, OLS is consistent under richer models for u_{jc} than the CSRE model, provided u_{jc} is uncorrelated with \mathbf{x}_{jc} . We consider the OLS estimator in this case, with focus on obtaining correct standard errors given correlation of the error term u_{jc} within a cluster.

Notation

Stacking observations in (24.28) within a cluster yields

$$\mathbf{y}_c = \mathbf{X}_c\beta + \mathbf{u}_c, \quad (24.29)$$

where \mathbf{y}_c and \mathbf{u}_c are $N_c \times 1$ vectors and \mathbf{X}_c is an $N_c \times K$ matrix. Further stacking over clusters yields

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}, \quad (24.30)$$

where \mathbf{y} and \mathbf{u} are $N \times 1$ vectors and \mathbf{X} is an $N \times K$ matrix, $N = \sum_c N_c$.

The three representations of the CSRE model lead to three equivalent ways of expressing the **OLS estimator** of model (24.28),

$$\begin{aligned} \hat{\beta}_{OLS} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \left(\sum_{c=1}^C \mathbf{X}'_c \mathbf{X}_c \right)^{-1} \sum_{c=1}^C \mathbf{X}'_c \mathbf{y}_c \\ &= \left(\sum_{c=1}^C \sum_{j=1}^{N_c} \mathbf{x}_{jc} \mathbf{x}'_{jc} \right)^{-1} \sum_{c=1}^C \sum_{j=1}^{N_c} \mathbf{x}_{jc} y_{jc}. \end{aligned} \quad (24.31)$$

The second of these representations is especially useful given the assumption of independence of errors across clusters. Then, as before in the panel case, the OLS estimator has limit distribution

$$\sqrt{N}(\hat{\beta}_{OLS} - \beta) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}], \quad (24.32)$$

where

$$\begin{aligned} \mathbf{A} &= \text{plim} N^{-1} \sum_{c=1}^C \mathbf{X}'_c \mathbf{X}_c, \\ \mathbf{B} &= \text{plim} N^{-1} \sum_{c=1}^C \mathbf{X}'_c \mathbf{u}_c \mathbf{u}'_c \mathbf{X}_c, \end{aligned} \quad (24.33)$$

using independence of \mathbf{u}_c over c . Different assumptions on the properties of \mathbf{u}_c lead to different estimates of \mathbf{B} .

OLS Cluster-Robust Standard Errors

If clusters are small then there are many clusters and \mathbf{B} in (24.33) can be consistently estimated by replacing \mathbf{u}_c by $\widehat{\mathbf{u}}_c = \mathbf{y}_c - \mathbf{X}_c\widehat{\beta}$. It follows that $\widehat{\beta}_{OLS}$ is asymptotically normally distributed with **cluster-robust variance matrix**

$$\widehat{V}[\widehat{\beta}_{OLS}] = \left(\sum_{c=1}^C \mathbf{X}'_c \mathbf{X}_c \right)^{-1} \sum_{c=1}^C \mathbf{X}'_c \widehat{\mathbf{u}}_c \widehat{\mathbf{u}}'_c \mathbf{X}_c \left(\sum_{c=1}^C \mathbf{X}'_c \mathbf{X}_c \right)^{-1}. \quad (24.34)$$

This formula places no restriction on heteroskedasticity and correlation within a cluster, as $V[\mathbf{u}_c]$ and hence $V[u_{jc}]$ and $Cov[u_{jc}, u_{kc}]$ are unrestricted. However, it does assume that N_c is small and $C \rightarrow \infty$. Statistical packages often give a degrees-of-freedom correction. Typically one multiplies the estimate in (24.34) by

$$dfc = \frac{N-1}{N-K} \times \frac{C}{C-1},$$

which corrects for both estimation of β and the number of clusters in practice being finite.

To see how (24.34) works, treat the regressors as fixed and note that

$$\begin{aligned} \mathbf{B} &= \lim N^{-1} \sum_{c=1}^C \mathbf{X}'_c E[\mathbf{u}_c \mathbf{u}'_c] \mathbf{X}_c \\ &= \lim N^{-1} \sum_{c=1}^C \sum_{j=1}^{N_c} \sum_{k=1}^{N_c} E[u_{jc} u'_{kc}] \mathbf{x}_{jc} \mathbf{x}'_{kc}. \end{aligned}$$

Then (24.34) is obtained using the estimate

$$\begin{aligned} \widehat{\mathbf{B}} &= N^{-1} \sum_{c=1}^C \mathbf{X}'_c \widehat{\mathbf{u}}_c \widehat{\mathbf{u}}'_c \mathbf{X}_c \\ &= N^{-1} \sum_{c=1}^C \sum_{j=1}^{N_c} \sum_{k=1}^{N_c} \widehat{u}_{jc} \widehat{u}'_{kc} \mathbf{x}_{jc} \mathbf{x}'_{kc}. \end{aligned}$$

For example, consider estimation of $E[y]$ by \bar{y} . This is the regression (24.28) with $\mathbf{x}_{jc} = 1$, $\widehat{\beta}_{OLS} = \bar{y}$, and $\widehat{u}_{jc} = y_{jc} - \bar{y}$. Then (24.34) leads to $\widehat{V}[\bar{y}] = N^{-2} \sum_c (\sum_j (y_{jc} - \bar{y}))^2$, compared to the estimate of $N^{-1} \sum_c \sum_j (y_{jc} - \bar{y})^2$ which additionally assumes independence within clusters.

OLS Standard Errors Assuming the CSRE Model

The cluster-robust estimates (24.34) require many clusters. Alternative estimates that also apply to the case of few clusters can be used if assumptions are made about the variances and covariances of the model error u_{jc} . These alternative estimates also permit analytical results regarding the impact of clustering on estimator variances.

In particular, assume that the CSRE model given by (24.24) to (24.26) is appropriate. Then the error $u_{jc} = \alpha_c + \varepsilon_{jc}$ is independent over c and within a cluster

$$\text{Cov}[u_{jc}, u_{kc}] = \begin{cases} \sigma^2, & j = k, \\ \rho\sigma^2, & j \neq k, \end{cases}$$

where the intraclass correlation coefficient ρ is defined in (24.27). It follows that

$$\Sigma_c = V[\mathbf{u}_c] = \sigma^2[(1 - \rho)\mathbf{I}_c + \rho\mathbf{e}_c\mathbf{e}'_c], \quad (24.35)$$

where \mathbf{I}_c is an $N_c \times N_c$ identity matrix and \mathbf{e}_c is an $N_c \times 1$ vector of ones.

Given Σ_c in (24.35), the general result (24.32) to (24.33) yields

$$V[\hat{\beta}_{OLS}] = \left(\sum_{c=1}^C \mathbf{X}'_c \mathbf{X}_c \right)^{-1} \sum_{c=1}^C \sigma^2 \mathbf{X}'_c [(1 - \rho)\mathbf{I}_c + \rho\mathbf{e}_c\mathbf{e}'_c] \mathbf{X}_c \left(\sum_{c=1}^C \mathbf{X}'_c \mathbf{X}_c \right)^{-1}. \quad (24.36)$$

Provided the intraclass correlation coefficient is constant, this variance matrix estimator is consistent in both the small- and large-cluster cases. Obvious estimators for σ^2 and ρ are

$$\hat{\sigma}^2 = \frac{1}{N - K - 1} \sum_{c=1}^C \sum_{j=1}^{N_c} \hat{u}_{jc}^2$$

and

$$\hat{\rho} = \frac{1}{\sum_c N_c(N_c - 1)} \frac{1}{\hat{\sigma}^2} \sum_{c=1}^C \sum_{j=1}^{N_c} \sum_{k \neq j}^{N_c} \hat{u}_{jc} \hat{u}_{kc}.$$

The estimate of ρ involves many intracluster pairs and a consistent estimate can be obtained using just a subset of these. As written $\sum_c N_c(N_c - 1)$ pairs are used, though in fact each unique within-cluster pair is double counted as both $\hat{u}_{jc}\hat{u}_{kc}$ and $\hat{u}_{kc}\hat{u}_{jc}$ appear in the summations.

If the clusters are large the intracluster correlation can be permitted to vary across clusters. Then (24.35) and (24.36) can be amended to replace σ^2 and ρ by σ_c^2 and ρ_c , respectively. These can be consistently estimated by

$$\hat{\sigma}_c^2 = \frac{1}{N_c - K - 1} \sum_{j=1}^{N_c} \hat{u}_{jc}^2$$

and

$$\hat{\rho}_c = \frac{1}{N_c(N_c - 1)} \frac{1}{\hat{\sigma}_c^2} \sum_{j=1}^{N_c} \sum_{k \neq j}^{N_c} \hat{u}_{jc} \hat{u}_{kc}.$$

Bias of Usual OLS Standard Errors

If data are clustered, then intuitively the usual formula variance estimator for the OLS estimator,

$$V^{\text{Formula}} [\hat{\beta}_{\text{OLS}}] = \sigma^2 \left(\sum_{c=1}^C \mathbf{X}'_c \mathbf{X}_c \right)^{-1},$$

underestimates the true variance matrix of the OLS estimator, assuming positive within-cluster correlation, since each additional observation within a cluster will provide less than one additional piece of independent information. We demonstrate this bias when the error process is that of the CSRE model.

Consider the CSRE model with the same regressors within each cluster, so $\mathbf{x}_{jc} = \mathbf{x}_c$ and $\mathbf{X}_c = \mathbf{e}_c \mathbf{x}'_c$. Then by using $\mathbf{e}'_c \mathbf{e}_c = N_c$, (24.36) becomes

$$V[\hat{\beta}_{\text{OLS}}] = \left(\sum_{c=1}^C N_c \mathbf{x}_c \mathbf{x}'_c \right)^{-1} \sum_{c=1}^C N_c \sigma^2 [1 + \rho(N_c - 1)] \mathbf{x}_c \mathbf{x}'_c \left(\sum_{c=1}^C N_c \mathbf{x}_c \mathbf{x}'_c \right)^{-1},$$

a result presented by Kloek (1981) and Moulton (1986).

Now specialize to **balanced clusters**, and define M to be the average cluster size, so $M = N_c = N/C$ is constant. Then the variance estimate simplifies to

$$V[\hat{\beta}_{\text{OLS}}] = [1 + \rho(M - 1)] \times \sigma^2 \left(M \sum_{c=1}^C \mathbf{x}_c \mathbf{x}'_c \right)^{-1},$$

whereas the formula variance simplifies to $\sigma^2(M \sum_c \mathbf{x}_c \mathbf{x}'_c)^{-1}$. It follows that the true variances are a multiple

$$\tau = [1 + \rho(M - 1)]$$

times the usual OLS variance matrix estimate. Even if ρ is small the correction factor can be quite large. For example, if the average cluster size is $M = 101$ observations, then the usual OLS standard errors should be multiplied by $\sqrt{1 + 100\rho}$. The assumed independence within a cluster will also lead to a biased estimate of σ^2 , but this is of second-order importance. In the balanced-cluster case Kloek shows that $E[\sum_c \sum_j \hat{u}_{cj}^2] = \sigma^2[N - K(1 + \rho(m - 1))]$ so we should normalize by $[N - K(1 + \rho(m - 1))]^{-1}$ rather than $[N - K]^{-1}$.

In practice some regressors may be constant within a cluster and others may vary. Then in the case of regression with intercept and scalar regressor (i.e., $\mathbf{x}'_{jc} \beta = \beta_1 + \beta_2 x_{jc}$) Scott and Holt (1982) show that the usual OLS formula variance for the intercept should be multiplied by $1 + \rho(M - 1)$ as done in the preceding, but for the slope coefficient it should be multiplied by the smaller factor $1 + \hat{\rho}_x \rho(M - 1)$, where $\hat{\rho}_x$ can be viewed as an estimate of the intraclass correlation coefficient of the x_{jc} . In cross-section applications $\hat{\rho}_x$ is relatively small, so the main problem lies with standard errors for cluster-invariant regressors.

Moulton (1986) demonstrated in an application that the bias in standard errors using the incorrect OLS formula variance can be quite appreciable. He estimated a log-wage equation using cross-section CPS data where clustering was on states. For his application $N = 18,946$ and $C = 49$. For his data the estimated intraclass correlation coefficient was $\hat{\rho} = 0.032$, a seemingly small value. However, the clusters are large, and if we ignore the data being unbalanced and as a guide use the preceding formulas with $M = 387$, the average cluster size, then $\hat{\tau} = [1 + \hat{\rho}(M - 1)] = 13.3$. For state-invariant regressors the true OLS standard errors are predicted to be $\sqrt{13.3} = 3.7$ times the usual reported standard errors, a very large bias. (One way to view this is that for OLS estimation of the coefficients of state-invariant regressors, the 18,946 clustered observations have the same precision as $18,946/13.3 = 1,425$ independent observations.) For individual-varying regressors the bias will be much smaller, for example, $[1 + \hat{\rho}_x \hat{\rho}(M - 1)] = 2.23$ if $\hat{\rho}_x = 0.10$. Moulton does not report results for the individual-varying regressors included as regressors. For the state-invariant regressors, variables such as growth rate of employment in the state, the cluster-corrected standard errors for OLS are generally between three and four times the incorrect formula standard errors.

The lesson is that there can be great **downward bias** in the default OLS standard errors for the OLS coefficients of cluster-invariant regressors. For individual-varying regressors there is also bias, but it is much less. Cluster-invariant regressors are often included in applications with clustered data, as it is common to model individual behavior as depending in part on attributes of the cluster. Valid statistical inference requires obtaining standard errors that control for clustering.

24.5.3. Cluster-Specific Random Effects Estimator

If a random effects model is appropriate then the GLS estimator is in general more efficient than the OLS estimator of the previous section. Given independence across clusters the **GLS estimator** of model (24.29) is

$$\widehat{\beta}_{\text{GLS,RE}} = \left(\sum_{c=1}^C \mathbf{X}'_c \Sigma_c^{-1} \mathbf{X}_c \right)^{-1} \sum_{c=1}^C \mathbf{X}'_c \Sigma_c^{-1} \mathbf{y}_c, \quad (24.37)$$

where $\Sigma_c = V[\mathbf{u}_c]$. The feasible GLS estimator replaces Σ_c by a consistent estimate $\widehat{\Sigma}_c$, and assuming correct specification of the model (24.29) and error variance matrix Σ_c , we have

$$V[\widehat{\beta}_{\text{GLS,RE}}] = \left(\sum_{c=1}^C \mathbf{X}'_c \Sigma_c^{-1} \mathbf{X}_c \right)^{-1}.$$

For the CSRE model, Σ_c given in (24.35) can be consistently estimated by $\widehat{\Sigma}_c$, which replaces σ^2 and ρ by the consistent estimates given after (24.36). As in the similar random effects model for panel data, the feasible GLS estimator is asymptotically equivalent to the MLE under the additional assumptions that α_c and ε_{jc} are normally distributed.

An attraction of the CSRE model is that the GLS estimator (24.37) can be simply implemented by OLS estimation of the transformed regression

$$y_{jc} - \theta_c \bar{y}_c = (\mathbf{x}_{jc} - \theta_c \bar{\mathbf{x}}_c)' \boldsymbol{\beta} + (\varepsilon_{jc} - \theta_c \bar{\varepsilon}_c), \quad (24.38)$$

where

$$\theta_c = 1 - \frac{\sqrt{1-\rho}}{\sqrt{1+\rho(N_c-1)}} = 1 - \frac{\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2 + N_c \sigma_\alpha^2}}. \quad (24.39)$$

This result is proven later in this section. To implement it we replace θ_c by consistent estimate $\hat{\theta}_c$. As for the panel data model, it can be shown that usual OLS standard errors from this regression can be used if the errors ε_{jc} in model (24.24) are homoskedastic.

The GLS estimator is at least as efficient as OLS assuming (24.24) to (24.26) hold. In the special case that all regressors are cluster-invariant there is no efficiency gain as GLS then coincides with OLS (Kloek, 1981). More generally, Scott and Holt (1982) give a quite conservative upper bound to the efficiency loss of OLS compared to GLS as

$$\frac{V[\mathbf{c}' \widehat{\boldsymbol{\beta}}_{GLS}]}{V[\mathbf{c}' \widehat{\boldsymbol{\beta}}_{OLS}]} \geq 1 - \left(1 + \frac{4(1-\rho)[1+\rho(N_0-1)]}{N_0^2 \rho^2} \right)^{-1}$$

for arbitrary vector \mathbf{c} and where $N_0 = \max\{N_c\}$ is the sample size of the largest cluster. This bound is increasing in N_0 and ρ , and even for $N_0 = 1,000$ and $\rho = 0.10$, OLS is at most 22% less efficient than GLS.

Given these small efficiency gains to GLS it is more common to focus on OLS estimation with correct standard errors, unless OLS is inconsistent because the CSFE model is appropriate. The main impact of clustering is that OLS is much less efficient compared to the case of no clustering, as is clear from the discussion of calculation of standard errors for the OLS estimator in Section 24.5.2.

If clusters are large, then the CSRE model can be relaxed to permit the error variance and intraclass correlation to vary across clusters. Then in (24.35) for Σ_c we replace σ^2 and ρ by σ_c^2 and ρ_c , respectively, using consistent estimates for σ_c^2 and ρ_c given after (24.36).

If clusters are small then robust standard errors that do not constrain error correlation to be constant within a cluster can be obtained, analogous to (24.34) for OLS. Then

$$\widehat{V}[\widehat{\boldsymbol{\beta}}_{GLS,RE}] = \left[\sum_{c=1}^C \mathbf{X}'_c \widehat{\Sigma}_c^{-1} \mathbf{X}_c \right]^{-1} \sum_{c=1}^C \mathbf{X}'_c \widehat{\Sigma}_c^{-1/2} \widehat{\mathbf{u}}_c \widehat{\mathbf{u}}'_c \widehat{\Sigma}_c^{-1/2} \mathbf{X}_c \left[\sum_{c=1}^C \mathbf{X}'_c \widehat{\Sigma}_c^{-1} \mathbf{X}_c \right]^{-1},$$

where $\widehat{\mathbf{u}}_c = \mathbf{y}_c - \mathbf{X}_c \widehat{\boldsymbol{\beta}}_{GLS,RE}$. This estimate requires N_c small and $C \rightarrow \infty$, and it assumes independence of errors in different clusters.

GLS Implemented as OLS in a Transformed Model

To derive (24.38), note that for Σ_c defined in (24.35)

$$\begin{aligned}\Sigma_c^{-1} &= [\sigma^2[(1-\rho)\mathbf{I}_c + \rho\mathbf{e}_c\mathbf{e}'_c]]^{-1} \\ &= \frac{1}{\sigma^2(1-\rho)}[\mathbf{I}_c - (\rho/\tau_c)\mathbf{e}_c\mathbf{e}'_c]^{-1},\end{aligned}$$

where $\tau_c = 1 + \rho(N_c - 1)$ and hence

$$\Sigma_c^{-1/2} = \frac{1}{\sigma\sqrt{1-\rho}}[\mathbf{I}_c - (\theta_c/N_c)\mathbf{e}_c\mathbf{e}'_c],$$

using the general results that if \mathbf{e} is an $M \times 1$ vector of ones then

$$\begin{aligned}[\mathbf{I} + a\mathbf{e}\mathbf{e}']^{-1} &= \mathbf{I} - [a/(1+aM)]\mathbf{e}\mathbf{e}', \\ [\mathbf{I} + a\mathbf{e}\mathbf{e}']^{1/2} &= \mathbf{I} - M^{-1}(1 - \sqrt{1+aM})M\mathbf{e}\mathbf{e}'.\end{aligned}$$

Now in (24.37) $\mathbf{X}'_c \Sigma_c^{-1} \mathbf{X}_c = (\Sigma_c^{-1/2} \mathbf{X}_c)' \Sigma_c^{-1/2} \mathbf{X}_c$, where

$$\begin{aligned}\Sigma_c^{-1/2} \mathbf{X}_c &= [\mathbf{I}_c - (\theta_c/N_c)\mathbf{e}_c\mathbf{e}'_c]\mathbf{X}_c \\ &= \mathbf{X}_c - \theta_c \mathbf{e}_c \bar{\mathbf{x}}'_c\end{aligned}$$

and where $\bar{\mathbf{x}}_c = N_c^{-1} \sum_j \mathbf{x}_{jc}$ and we ignore the scalar multiple $1/\sigma\sqrt{1-\rho}$ as it will cancel out when we similarly consider $\mathbf{X}'_c \Sigma_c^{-1} \mathbf{y}_c$. The transformed regression model (24.38) follows.

24.5.4. Cluster-Specific Fixed Effects Estimator

The basic idea of the **CSFE model** is straight forward: Let the cluster effect enter the conditional mean function through the intercept term. The model is

$$y_{jc} = \alpha_c + \mathbf{x}'_{jc}\beta + \varepsilon_{jc}, \quad j = 1, \dots, N_c, c = 1, \dots, C, \quad (24.40)$$

where now both β and α_c , $c = 1, \dots, C$, are parameters to be estimated.

In the CSFE model all cluster-invariant regressors must be dropped, as they cannot be separately identified from α_c . For example, if clustering is on the state and a fixed effects model is appropriate then the effect of state-invariant regressors such as state average unemployment cannot be identified. If estimation of the coefficients of state-invariant regressors is desired then OLS or the CSRE estimator need to be used instead. However, one should first use a Hausman test analogous to that presented in Chapter 21 for panel data to confirm the validity of the strong assumption of the CSRE model that α_c is uncorrelated with the regressors.

We consider statistical inference under the assumption

$$\varepsilon_{jc} \sim [0, \sigma_{jc}^2].$$

This permits heteroskedasticity of unknown form but assumes that inclusion of the cluster-specific fixed effect α_c is sufficient to control for any error correlation within

a cluster. This is a departure from panel data analysis where concern about time-series correlation in the errors even after inclusion of individual-specific effects leads to richer models. If desired, however, one can additionally adjust estimator standard errors for correlation within a cluster by methods similar to those in Section 24.5.2.

The main complication in estimation of the CSFE model is that in small clusters there are too many intercepts α_c to estimate.

Cluster Dummy Variables Model

We first consider large clusters, where the number of clusters is small relative to the total sample size. Then the intercepts α_c can be estimated directly by introducing dummy variables for each cluster and estimating by OLS.

Let observation i denote the j th household in the c th cluster. Then (24.40) can be written as the **cluster dummy variables model**

$$y_i = \sum_{c=1}^C \alpha_c d_{ci} + \mathbf{x}'_i \beta + \varepsilon_i, \quad i = 1, \dots, N, \quad (24.41)$$

where the d_{ci} are indicator variables that equal one if the i th observation belongs to cluster c and equal zero otherwise. Thus C cluster indicator variables, such as state dummy variables, are included, and to avoid the dummy variable trap, \mathbf{x} should not contain an intercept term.

An OLS estimation of this model yields consistent estimates of both $\alpha_1, \dots, \alpha_C$ and β , assuming a fixed number of clusters C as $N \rightarrow \infty$. One can use the usual Eicker–White estimate to obtain standard errors that are robust given heteroskedastic errors.

Within-Clusters Estimator

When there are many small clusters we can no longer estimate the model (24.40) by OLS. First, OLS estimation may not be computationally feasible because the number of parameters $(C + K) \rightarrow \infty$ as the number of clusters $C \rightarrow \infty$. Second, and more importantly, because the number of parameters is going to infinity with the sample size, the OLS estimator is inconsistent unless $N_c \rightarrow \infty$.

Interest usually lies in the parameters β in (24.40), with $\alpha_1, \dots, \alpha_C$ viewed as **incidental parameters** or as **nuisance parameters**. Then it is convenient to sweep out the fixed effects by an initial data transformation. Each observation $(y_{jc}, \mathbf{x}_{jc})$ is replaced by deviation from the cluster mean, that is, by $(y_{jc} - \bar{y}_c, \mathbf{x}_{jc} - \bar{\mathbf{x}}_c)$, $i = 1, \dots, N_c$, $c = 1, \dots, C$, where $\bar{y}_c = N_c^{-1} \sum_j y_{jc}$ and $\bar{\mathbf{x}}_c = N_c^{-1} \sum_j \mathbf{x}_{jc}$ are cluster-specific averages. Then the model (24.40) for y_{jc} implies that

$$y_{jc} - \bar{y}_c = (\mathbf{x}_{jc} - \bar{\mathbf{x}}_c)' \beta + \varepsilon_{jc} - \bar{\varepsilon}_c. \quad (24.42)$$

Applying OLS to the transformed regression (24.42) yields a consistent estimate of β . If the CSFE coefficients are also of interest, they can be estimated by $\hat{\alpha}_c = \bar{y}_c - \bar{\mathbf{x}}_c' \beta$, though this estimate is not consistent for small N_c .

Comparison with Chapter 21 shows that this is analogous to the **within estimator** for panel data. As for panel data, the estimate of β from OLS estimation of (24.42) coincides with the estimate of β from OLS estimation of the cluster dummy variables model (24.41).

A **between estimator** can also be proposed analogous to that for linear panel models. In this case \bar{y}_c is regressed on $\bar{\mathbf{x}}_c$, $c = 1, \dots, N_c$. From (24.37), the GLS estimator in the CSRE model involves regression in quasi-differences, where cluster means are multiplied by θ_c (defined in (24.39)) before differencing. The GLS estimator can be shown to be a linear combination of the within and between estimators. It approaches the within estimator for large N_c as then $\theta_c \rightarrow 1$. Note that the within estimator is consistent in the CSRE model.

Caution is necessary in interpreting the standard errors if the regression is applied to the mean-corrected observations. The number of degrees of freedom for this regression is $(N - K - C)$, not $(N - K)$. If software neglects this adjustment then the residual variance from the software should be adjusted by multiplying by the **inflation factor** $(N - K) / (N - K - C)$ and the standard errors should be inflated by the square root of the same.

24.5.5. Diagnostic Tests for Cluster Effects

In linear regression a test of cluster-specific fixed effects under normality of errors is just the standard F -test of linear restrictions hypothesis $H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_C = 0$ in (24.40). This simply involves a comparison of the R^2 statistic for the two regressions with and without the cluster-specific dummy variables.

In the CSRE model a test of cluster effects is a one-sided test of the null hypothesis $H_0 : \sigma_\alpha^2 = 0$ versus $H_1 : \sigma_\alpha^2 > 0$. An equivalent test can also be formulated as a test of $H_0 : \rho = 0$ versus $H_1 : \rho > 0$ using the definition in (24.27). The one-sided LM test statistic of this hypothesis, given by Moulton (1987), is

$$\text{LM} = \frac{\sum_c (N_c \bar{u}_c)^2 - \sum_c \sum_i \hat{u}_{ic}^2}{\hat{\sigma}^2 [2 (\sum_c N_c^2 - N)]^{1/2}}, \quad (24.43)$$

where $\hat{\sigma}^2 = \sum_c \sum_i \hat{u}_{ic}^2 / N$, \hat{u}_{ic} denotes the least-squares residual from the pooled regression of y on \mathbf{x} , and \bar{u}_c is the average residual for cluster c .

24.5.6. Clustering in Nonlinear Models

Nonlinear models with clustered data have not attracted much attention in the econometrics literature. There are numerous published articles in biostatistics, however, with a special focus on binary outcome models (Pendergast et al., 1996). Other models such as the Poisson regression and some models for survival data have also been considered. The hierarchical (multilevel) modeling framework has also been used extensively especially for binary outcome models.

Here we continue to exploit the parallel between clustered and panel data. As in the linear case the data (y_i, \mathbf{x}_i) , $i = 1, \dots, N$, are subscripted as $(y_{jc}, \mathbf{x}_{jc})$, $j = 1, \dots, N_c$,

$c = 1, \dots, C$. We assume independence over c but permit dependence of observations within cluster c .

m-Estimation with Clustering

Consider a nonlinear estimating equations estimator that solves

$$\sum_{c=1}^C \sum_{j=1}^{N_c} \mathbf{h}(y_{jc}, \mathbf{x}_{jc}, \boldsymbol{\theta}) = \mathbf{0}. \quad (24.44)$$

Often these equations are obtained from maximization or minimization of the objective function $\sum_c \sum_j q(y_{jc}, \mathbf{x}_{jc}, \boldsymbol{\theta})$, in which case $\mathbf{h}(y_{jc}, \mathbf{x}_{jc}, \boldsymbol{\theta}) = \partial q(y_{jc}, \mathbf{x}_{jc}, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}$. For example, for quasi-MLE based on the product of marginal densities $\mathbf{h}(y_{jc}, \mathbf{x}_{jc}, \boldsymbol{\theta}) = \partial \ln f(y_{jc} | \mathbf{x}_{jc}, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}$.

We assume that data are clustered, so that $\text{Cov}[\mathbf{h}_{jc}, \mathbf{h}_{kc}] \neq \mathbf{0}$. However, we maintain the assumption that $E[\mathbf{h}(y_{jc}, \mathbf{x}_{jc}, \boldsymbol{\theta})] = \mathbf{0}$, a necessary condition for consistency, which rules out the cluster-specific fixed effects model also presented in the following.

The cluster-robust variance of the OLS estimator (24.34) is easily adapted to the current situation by replacing $\mathbf{x}_{jc}\mathbf{x}'_{jc}$ by $\partial \mathbf{h}_{jc} / \partial \boldsymbol{\theta}'$ and $\mathbf{x}_{jc}\hat{u}_{jc}$ by $\mathbf{h}_{jc}(\hat{\boldsymbol{\theta}})$. Then $\hat{\boldsymbol{\theta}}$ is asymptotically normal with **cluster-robust variance matrix**

$$\hat{\mathbf{V}}[\hat{\boldsymbol{\theta}}] = \left(\sum_{c=1}^C \sum_{j=1}^{N_c} \frac{\partial \mathbf{h}'_{jc}}{\partial \boldsymbol{\theta}} \Big|_{\hat{\boldsymbol{\theta}}} \right)^{-1} \sum_{c=1}^C \sum_{j=1}^{N_c} \sum_{k=1}^{N_c} \mathbf{h}_{jc}(\hat{\boldsymbol{\theta}}) \mathbf{h}_{kc}(\hat{\boldsymbol{\theta}})' \left(\sum_{c=1}^C \sum_{j=1}^{N_c} \frac{\partial \mathbf{h}_{jc}}{\partial \boldsymbol{\theta}'} \Big|_{\hat{\boldsymbol{\theta}}} \right)^{-1}. \quad (24.45)$$

Some computer software provides this as a standard option for many parametric non-linear models.

A leading example is quasi-ML estimation based on the product of marginal densities within a cluster rather than the joint density. Specifically, given dependence over j within cluster c we should maximize the log-likelihood

$$\ln L(\boldsymbol{\theta}) = \sum_{c=1}^C \ln f(y_{1c}, \dots, y_{N_c c}, \mathbf{x}_{1c}, \dots, \mathbf{x}_{N_c c}, \boldsymbol{\theta}).$$

However, the joint density may be difficult to work with or difficult to obtain because for many univariate densities there can be a limited range of multivariate densities. Instead, we may maximize

$$\begin{aligned} Q(\boldsymbol{\theta}) &= \sum_{c=1}^C \ln[f(y_{1c}, \mathbf{x}_{1c}, \boldsymbol{\theta}) \times \dots \times f(y_{N_c}, \mathbf{x}_{N_c}, \boldsymbol{\theta})] \\ &= \sum_{c=1}^C \sum_{j=1}^{N_c} \ln f(y_{jc}, \mathbf{x}_{jc}, \boldsymbol{\theta}), \end{aligned}$$

which is no longer a true likelihood function, unless y_{jc} are independent over j , so the information matrix equality no longer applies. The preceding formulas apply with $\mathbf{h}_{jc}(\boldsymbol{\theta}) = \partial \ln f(y_{jc}, \mathbf{x}_{jc}, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ and $\partial \mathbf{h}_{jc}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}' = \partial^2 \ln f(y_{jc}, \mathbf{x}_{jc}, \boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'$.

This means that within each cluster we do not use the likelihood score for each observation as in the case of independent observations; instead, we replace it by the sum of likelihood scores over all cluster elements.

Nonlinear Cluster-Specific Random Effects

A quite general setup for cluster-specific effects in nonlinear models is to consider the estimator that minimizes or maximizes

$$Q(\beta, \alpha_1, \dots, \alpha_C) = \sum_{c=1}^C \sum_{j=1}^{N_c} q(y_{jc}, \mathbf{x}_{jc}, \beta, \alpha_c), \quad (24.46)$$

where cluster effects enter only via the scalar parameter α_c , $c = 1, \dots, C$.

A simple random effects model assumes that the α_c are iid with parameters δ . Taking expectation with respect to α_c yields the objective function

$$Q(\beta, \delta) = \sum_{c=1}^C \int \sum_{j=1}^{N_c} q(y_{jc}, \mathbf{x}_{jc}, \beta, \alpha_c) f(\alpha_c | \delta) d\alpha_c.$$

Estimation can be complicated, especially if there is no closed-form expression for the integral of the sum.

Often it is easy to obtain the expectation with respect to one observation, $E_{\alpha_c}[q(y_{jc}, \mathbf{x}_{jc}, \beta, \alpha_c)] = q^*(y_{jc}, \mathbf{x}_{jc}, \beta, \delta)$. Then the simpler estimator that ignores clustering and minimizes $Q^*(\beta, \delta) = \sum_c \sum_j q^*(y_{jc}, \mathbf{x}_{jc}, \beta, \delta)$ will be consistent, though the standard errors need to be adjusted for clustering using (24.45).

For example, with count data we can develop a clustered analogue of the panel data Poisson–gamma mixture model. However, the Poisson quasi-MLE that ignores clustering can still be used as it is consistent, though standard errors need to be adjusted for clustering.

Therefore, although random effects versions of nonlinear models can be developed, it is often adequate to estimate parameters by ignoring clustering and then correct the standard errors of estimators for the clustering. There can be little reason for estimation of clustered random effects models, aside from the potential for efficiency gains.

Nonlinear Cluster-Specific Fixed Effects

Nonlinear variants of the cluster-specific fixed effects model again maximize or minimize

$$Q(\beta, \alpha_1, \dots, \alpha_C) = \sum_{c=1}^C \sum_{j=1}^{N_c} q(y_{jc}, \mathbf{x}_{jc}, \beta, \alpha_c),$$

as in (24.34), except now the parameters $\alpha_1, \dots, \alpha_C$ are estimated rather than integrated out.

For large clusters, that is, C small and $N_c \rightarrow \infty$, we simply optimize $Q(\beta, \alpha_1, \dots, \alpha_C)$ with respect to β and $\alpha_1, \dots, \alpha_C$. Assuming that $\alpha_1, \dots, \alpha_C$ completely control for any clustering, inference can be based on standard errors obtained

under the usual iid assumptions. This is the nonlinear analogue of the cluster-specific dummy variable model (24.41).

For small clusters, that is, N_c small and $C \rightarrow \infty$, we have the problem of too many incidental parameters $\alpha_1, \dots, \alpha_C$. Unlike the linear model it is generally not possible to eliminate the parameters $\alpha_1, \dots, \alpha_C$ (Hall and Severini, 1998). However, from Chapter 23 on panel data we see that it is possible in some cases.

For example, the **binary logit model with cluster fixed effects** specifies

$$\Pr[y_{jc} = 1] = \frac{1}{1 + \exp(-\alpha_c - \mathbf{x}'_{jc}\boldsymbol{\beta})}, \quad (24.47)$$

where for identification \mathbf{x}_{jc} cannot include an intercept or cluster-invariant regressors. The fixed effects α_c can be eliminated using the **conditional MLE** that conditions on the sum of responses within a cluster, $\sum_{j=1}^{N_c} y_{jc} = N_c \bar{y}_c$. The joint conditional probability for the c th cluster is

$$\Pr[y_{1c}, \dots, y_{N_c c} | N_c \bar{y}_c] = \frac{\exp\left(\boldsymbol{\beta} \sum_{j=1}^{N_c} \mathbf{x}_{jc} y_{jc}\right)}{\sum_{d \in \tilde{B}_c} \exp\left(\boldsymbol{\beta} \sum_{j=1}^{N_c} \mathbf{x}_{jc} d_{jc}\right)} \times \frac{\Gamma\left[\sum_{j=1}^{N_c} y_{jc} + 1\right] \Gamma\left[N_c - \sum_{j=1}^{N_c} + 1\right]}{\Gamma(N_c + 1)}, \quad (24.48)$$

where $\tilde{B}_c = \{(d_{1c}, \dots, d_{N_c c}) \mid d_{nc} = 0 \text{ or } 1, \text{ and } \sum_j d_{jc} = \sum_j y_{jc}\}$. The conditional likelihood is the product over all clusters of terms such as these, with clusters of size one excluded from the likelihood. The second term on the right-hand side does not depend on the unknown parameters and hence does not affect the maximization of the likelihood, so it can be ignored when considering maximization. The likelihood is awkward to maximize because the set \tilde{B}_c ranges over the many ways of choosing N_c outcomes $y_{jc} = 1$ from $(N_{1c} + N_{0c})$ total outcomes in cluster c . Fortunately, however, a number of popular computer packages provide the **conditional logit** option for estimating this model. The covariance matrix of all unknown parameters is estimated by the inverse of the log-likelihood Hessian.

As another example, consider the **Poisson fixed effects cluster model**, which specifies

$$y_{jc} \sim \mathcal{P}[\mu_{jc} = \alpha_c \exp(\mathbf{x}'_{jc}\boldsymbol{\beta})], \quad c = 1, \dots, C,$$

where $\mathcal{P}[\cdot]$ denotes the Poisson distribution, and \mathbf{x}_{jc} excludes an intercept and any cluster-invariant regressors. This is the usual Poisson model, except that the usual conditional mean $\exp(\mathbf{x}'_{jc}\boldsymbol{\beta})$ is scaled multiplicatively by the cluster-specific fixed effect α_c . For this particular model a variety of approaches, including conditional ML and concentrated ML, lead to elimination of the parameters α_c . Consistent estimates of the parameters $\boldsymbol{\beta}$ can be obtained by solving the estimating equations

$$\sum_{c=1}^C \sum_{j=1}^{N_c} \mathbf{x}_{jc} \left(y_{jc} - \frac{\bar{y}_c}{\lambda_c} \lambda_{jc} \right) = \mathbf{0},$$

where $\lambda_{jc} = \exp(\mathbf{x}'_{jc}\boldsymbol{\beta})$ and $\bar{y}_c = N_c^{-1} \sum_j y_{jc}$ and $\bar{\lambda}_c = N_c^{-1} \sum_j \lambda_{jc}$ are cluster means. For further details see the discussion of this model in the panel data case in Section 23.7.

24.5.7. Further Methods for Clustered Data

The essential feature of clustering is that there is dependence across observations. A related topic is **spatial correlation** (see for example Anselin (2001), Lee (2004)), where the observational unit is a region, such as a state, and observations in regions close to each other are likely to be correlated.

The random effects approach can be generalized to consider slope coefficients as well as the intercept. This is presented in the next section for **hierarchical linear models**. For nonlinear models the issues are similar to those for panel data presented in Chapter 23.

The **bootstrap** can be used to obtain cluster-robust standard errors, in settings where clustering leads to correlation within a cluster but does affect estimator consistency. Intuitively, one should resample with replacement over clusters c , in which case we require small clusters with $C \rightarrow \infty$. At the b th bootstrap replication we draw C clusters with replacement and use all of the households j in these C resampled clusters to estimate the $\hat{\theta}_b$ that solves (24.44). Then one can estimate $V[\hat{\theta}]$ by applying the usual sample variance formula to $\hat{\theta}_1, \dots, \hat{\theta}_B$, where B is the number of bootstrap replication. Note that the resampling is done over clusters rather than households, since it is clusters that are assumed to be iid whereas there is within-cluster dependence.

24.6. Hierarchical Linear Models

Section 24.5 restricted the role of cluster effects in the random effects model to be confined to the regression intercept. A more general random effects model allows clusterwise variation in the slope parameters also. Intercluster variation in a subset of regression parameters could be linked to observable cluster characteristics. Because such models involve several layers of specification, they are called **hierarchical models**.

A standard framework for clustered data in many applied statistics disciplines is that of **hierarchical linear models**, also called **multilevel linear models**, random coefficients models, variance components models, and mixed linear or **mixed effects models**. This class of models brings into the specification additional information. We begin with a presentation of the model for individuals clustered in groups. Then the model is adapted to short panels where repeated measures data are clustered for each individual.

24.6.1. Model Structure

A hierarchical or multilevel model is a model that can be applied to data with a nested structure. Examples are data on individuals within a region, such as a state or country,

or within an organizational unit, such as a school or community, or within a family if siblings data are used. Panel data are also an example, with repeated measures on the same individual interpreted as observations that are nested within an individual.

We begin with a linear model

$$y_{ij} = \mathbf{x}'_{ij}\beta_j + u_{ij}, \quad (24.49)$$

where the innovation is to let the K regression parameters β vary by group (or cluster) j . A concrete example is to consider data on students within schools. Then y_{ij} is an outcome measure such as test score for the i th student in the j th school, and the marginal effect of a change in a regressor such as race of the student varies across schools. Note that the standard hierarchical linear model (HLM) notation, which we use, reverses the subscripts compared to those in Section 24.5 where y_{cj} would be the test score for the j th student in the c th school.

The **two-level hierarchical linear model** specifies the coefficients in the level-one model (24.49) to be determined by a linear function of a random term and level-two variables, here school characteristics. Begin with the scalar parameter β_{kj} , the k th component of the $K \times 1$ vector parameter β_j . Then β_{kj} is modeled as depending on a vector of school characteristics \mathbf{w}_k that take value \mathbf{w}_{kj} for the j th school, with

$$\beta_{kj} = \mathbf{w}'_{kj}\gamma_k + v_{kj}, \quad k = 1, \dots, K, \quad (24.50)$$

where the first component of \mathbf{w}_{kj} is usually a constant. Stacking over all K components of β we have

$$\begin{bmatrix} \beta_{1j} \\ \vdots \\ \beta_{Kj} \end{bmatrix} = \begin{bmatrix} \mathbf{w}'_{1j} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{w}'_{Kj} \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_K \end{bmatrix} + \begin{bmatrix} v_{1j} \\ \vdots \\ v_{Kj} \end{bmatrix}$$

or in obvious matrix notation

$$\beta_j = \mathbf{W}_j\gamma + \mathbf{v}_j. \quad (24.51)$$

The model (24.50) is flexible and nests many models as special cases. These special cases include models with random intercepts and random slopes, but the framework additionally permits regression coefficients to vary with level-two observables \mathbf{w}_j . The range of models is very broad as the following indicates.

The k th level-one coefficient is called a **fixed coefficient** if $\beta_{kj} = \gamma_k$, in which case the coefficient does not vary with level-two regressors or with unobservables. If all level-one coefficients are fixed the model (24.49) reduces to $y_{ij} = \mathbf{x}'_{ij}\gamma + u_{ij}$, in which case estimation by OLS regression is appropriate. Note that the term fixed coefficient has a very different meaning to the term fixed effect used by econometricians in the panel context.

The k th level-one coefficient is said to be a **nonrandomly varying coefficient** if $\beta_{kj} = \mathbf{w}'_{kj}\gamma_k$. Then the coefficient is a linear function of school characteristics. If all level-one coefficients are fixed, except that the intercept is nonrandomly varying, the model (24.49) reduces to $y_{ij} = \mathbf{x}'_{ij}\beta + \mathbf{w}'_{1j}\gamma_1 + u_{ij}$, which is a standard OLS regression of the outcome on individual characteristics and school characteristics.

The k th level-one coefficient is said to be a **randomly varying coefficient** if $\beta_{kj} = \gamma_k + v_{kj}$. Then the coefficient is purely random and does not vary with school characteristics. If all level-one coefficients are randomly varying, so that $\beta_j = \gamma + \mathbf{v}_j$, the model is a **variance components model** or random coefficients model. If all level-one coefficients are fixed, except that the intercept is randomly varying, then the model (24.49) reduces further to $y_{ij} = \mathbf{x}'_{ij}\beta + v_{1j} + u_{ij}$, which is a random intercept model.

In practice some of the level-one coefficients may be both nonrandomly and randomly varying, as in the general case (24.49). If just the level-one intercept follows the general model (24.49) whereas all other level-one coefficients are fixed, the model (24.49) reduces to $y_{ij} = \mathbf{x}'_{ij}\beta + \mathbf{w}'_{1j}\gamma_1 + v_{1j} + u_{ij}$. This is the usual pooled regression model, with error that has two components and is therefore correlated across individuals at the same school.

The HLM framework can be extended to additional levels. For example, individual students (subscript i) may be nested in schools (subscript j), which are nested in a region (subscript k). Then the three-level HLM specifies at the first level the student outcome $y_{ijk} = \mathbf{z}'_{ijk}\pi_{jk} + e_{ijk}$, where the parameters $\pi_{jk} = \mathbf{X}_{jk}\beta_k + \mathbf{u}_{jk}$, and in turn $\beta_k = \mathbf{W}_j\gamma + \mathbf{w}_k$.

The HLM can be reexpressed as a **mixed linear model**, since substituting (24.50) into (24.49) yields

$$y_{ij} = (\mathbf{x}'_{ij}\mathbf{W}_j)\gamma + \mathbf{x}'_{ij}\mathbf{v}_j + u_{ij}. \quad (24.52)$$

The goal is to estimate the regression parameter γ and the variances and covariances of the errors u_{ij} and \mathbf{v}_j . Since the errors are assumed to be independent of regressors pooled OLS estimation of (24.52) yields consistent parameter estimates of γ . The HLM approach uses more efficient estimators that exploit assumptions on the variances and covariances of the errors u_{ij} and \mathbf{v}_j .

In the simplest case v_{kj} are assumed to be iid $\mathcal{N}[0, \sigma^2]$ and \mathbf{v}_j is assumed to be iid $\mathcal{N}[\mathbf{0}, \boldsymbol{\Gamma}]$. Then the model can be represented as

$$\begin{aligned} y_{ij} &\sim \mathcal{N}[\mathbf{x}'_{ij}\beta_j, \sigma^2], \\ \beta_j &\sim \mathcal{N}[\mathbf{W}_j\gamma, \boldsymbol{\Gamma}]. \end{aligned}$$

An early treatment of this was provided in a Bayesian setting by Lindley and Smith (1972), in which γ are called **hyperparameters**, which in more general models can themselves depend in turn on higher level hyper parameters. The parameters γ , σ^2 , and $\boldsymbol{\Gamma}$ can be estimated by maximum likelihood methods or by Bayes methods. Alternatively, ML methods can be used that are essentially the same as those for the mixed linear panel data model presented in Section 21.7. A complete treatment is given in Bryk and Raudenbush (1992, 2002).

24.6.2. HLM for Panel Data

The HLM literature interprets a short panel as repeated measures for an individual. Then the individual becomes level two in the two-level HLM, whereas the individual

was level one in the preceding section. The model (24.28) becomes

$$y_{ti} = \mathbf{x}'_{ti}\beta_i + u_{ti}, \quad (24.53)$$

where, for example, y_{ti} denotes an outcome measure at time t for student i , and the marginal effect of changes in regressors such as specific subjects studied varies across students. The scalar parameter β_{ki} , the k th element of the $K \times 1$ vector parameter β_i , is modeled as depending on a vector of individual characteristics \mathbf{w}_k that takes value \mathbf{w}_{ki} for the i th individual, with

$$\beta_{ki} = \mathbf{w}'_{ki}\gamma_k + v_{ki}, \quad i = 1, \dots, N. \quad (24.54)$$

The individual-specific effects model is the special case that all level-one coefficients are fixed, so $\beta_{ki} = \gamma_k$, except that the intercept term β_{1i} can vary across individuals (the level-two grouping).

The individual-specific *fixed effects* model arises if there is no model for the intercept β_{1i} , but instead β_{1i} is directly estimated. This is an extreme case of a nonrandomly varying coefficient, with $\beta_{1i} = \mathbf{w}'_{1i}\gamma_1$, where \mathbf{w}_{1i} is an $N \times 1$ vector of indicator variables with l th component equal to one if $i = l$ and equal to zero otherwise so that $\beta_{1i} = \gamma_{1i}$. The HLM framework is not designed to accommodate what econometricians call the fixed effects model.

The individual-specific *random effects* model arises if the intercept β_{1i} is a randomly varying coefficient, so that $\beta_{1i} = \gamma_1 + v_{1i}$. Clearly, much more general random effects models can be specified with β_{ki} also depending on regressors \mathbf{w}_{ki} .

As already noted, the HLM is a mixed linear model. For the panel data case the analogue of (24.52) is

$$y_{ti} = (\mathbf{x}'_{ti}\mathbf{W}_i)\gamma + \mathbf{x}'_{ti}\mathbf{v}_j + u_{ti}.$$

The random effects model of Chapter 21 is the specialization to $y_{ti} = \mathbf{x}'_{ti}\gamma + v_j + u_{ti}$.

A standard panel application of the HLM framework is to growth models, where the outcome y_{ti} is individual intelligence or height, which is a function of age, and the marginal effect of age is permitted to vary across individuals. Here the slope coefficient in addition to the intercept is permitted to vary across individuals.

24.7. Clustering Example: Vietnam Health Care Use

In this section we focus on estimation in the presence of clustering, since this is the most common complication of survey data that appears in microeconomics research. The methods in Section 24.5 are implemented.

Both linear and nonlinear regression models are estimated based on individual- and household-level data from the World Bank's Vietnam Living Standards Survey (VLSS) of 1997–1998. The survey collected detailed information on a variety of topics from over 27,700 individuals in approximately 6,000 households distributed over approximately 194 communes. In what follows “commune” is treated as a cluster or a group and it is hypothesized that the observed outcomes are correlated within a commune. Average cluster size in the household sample is about 26, maximum cluster size is 39,

and minimum cluster size is 1. To illustrate linear and nonlinear cluster models three outcomes will be modeled.

First, we consider a (log)linear regression model of total annual household health care expenditure (LNEXP12M), for households with positive expenditure, as a function of the (log of) total household expenditure (HHEXP), controlling for several standard sociodemographic variables, a type of “Engel curve” for health care expenditure. Of interest is the coefficient of total household expenditure, which is an estimate of the household income elasticity of demand for health care.

Second, we use information on individual responses to estimate clustered count models for a type of health care that accounts for a high proportion of aggregate private health care expenditure. In modeling these outcomes we control for recent health status of an individual, household income, health insurance status, and various demographic variables such as age, sex, marital status, and educational attainment of the head of the household. Information about health status was restricted to ILLNESS or INJURY sustained in the survey period, the duration of illness, and number of days of restricted activity. The key coefficients of interest are again the coefficients on the income and insurance status variables.

Table 24.3 provides the definitions and summary statistics for variables used in these examples.

In both cases the key issues are the following: What is the impact of clustering on the estimate of this elasticity? How does the elasticity and its impact vary as different statistical assumptions, models, and estimators are used?

24.7.1. Results and Discussion

Table 24.4 gives the results for the OLS regression, HC t -ratios, fixed effects, and random effects formulations. There is a relatively minor change in standard errors resulting from the use of a heteroskedastic-consistent variance estimator that does not take account of the clusters. However, when the cluster-robust variance estimator (24.34) is used there is a substantial change in the standard errors. The t -ratio for the expenditure elasticity drops from 16.01 to 12.68. All t -ratios become smaller and those for the two variables SEX and HHSIZE fall below 1.96. These results suggest, as expected, that ignoring intracluster correlation causes inflation in the OLS t -ratios.

The F -tests of the null hypothesis that all fixed effects are equal rejects the null. The fixed effects results have essentially the same pattern but note that the t -ratios are even smaller. The point estimate of the income elasticity is now 0.60 compared with 0.67 in the OLS results. However, overall there is no significant shift in the inference about the role of different variables.

A $\chi^2(1)$ score test of the null hypothesis that the random variation in the intercept is zero, based on (24.43), easily rejects the null, indicating that the RE model is an improvement over the restricted regression. However, the estimated RE model also does not result in a significant change in the assessment of the role of different variables. As expected the results presented under the FGLS columns and the RE (GLS) columns are very similar. The minor differences are essentially due to the different values used in the GLS transformation. The FGLS estimates are based on $\hat{\rho} = 0.12$,

Table 24.3. *Vietnam Health Care Use: Data Description*

Household data	Definition	Mean	Standard Deviation
LNEXP12M	Total household health care expenditure for 12 months	6.31	1.59
AGE	Age of head of household	48.01	13.77
SEX	Equals 1 if the head of the household is female, 0 otherwise	0.27	0.44
HHSIZE	Total household size	4.73	1.96
URBAN	Equals 1 if urban household, zero otherwise	0.29	0.45
EDUC	Schooling year of household head	7.09	4.41
HHEXP	Total nominal household expenditure (1998 VN dong)	15273	13020
Individual data			
PHARVIS	Number of direct pharmacy visits	0.51	1.31
LNMEDEXP (> 0)	log (total medical expenditure) for those with positive expenditure (1998 VN dong)	2.14	1.08
AGE	Age in years	29.7	9.67
SEX	Equals 1 if respondent is male	0.51	0.49
MARRIED	Equals 1 for married person	0.40	0.49
EDUC	Completed diploma level	3.38	1.94
ILLNESS	Number of illnesses experienced in past 12 months	0.62	0.90
INJURY	Equals 1 if injured during survey period	0.62	0.90
ILLDAYS	Number of illness days	2.80	5.45
ACTDAYS	Number of days of limited activity	0.06	1.11
INSURANCE	Equals 1 if respondent has health insurance coverage	0.16	0.37
MEDEXP (> 0)	Medical expenditure conditional on positive expenditure	21.04	208
MEDEXP	Medical expenditure (1998 VN dong)	6.13	112.75

an estimate obtained by averaging 100 estimates of ρ obtained using 100 resampled pairs of least-squares residuals.

The absolute differences between FE and RE results are relatively small. Informal comparison does not suggest that the FE and RE formulations yield substantially different results; however, the Hausman test suggests that there is a statistically significant difference between the two sets of estimates.

In summary, these results suggest that it is highly desirable to make some adjustment for intracluster correlation, and how exactly we do so appears to have a relatively small impact on the results.

Next we consider the results for the counted variable, number of pharmacy visits (PHARVIS) by individuals, using the Poisson model. This is an interesting variable because a high proportion of medical expenditure in Vietnam takes the form of self-prescribed medication through the purchase and use of over-the-counter drugs

Table 24.4. Vietnam Health Care Use: FE and RE Models for Positive Expenditure

^a R^2_W , within regression R^2 ; R^2_B , between regression R^2 ; R^2 , overall R^2 .

Table 24.5. Vietnam Health Care Use: Frequencies for Pharmacy Visits

Visits	0	1	2	3	4	5	6	7	8	9	10+
PHARVIS	20639	3827	1716	776	359	174	64	43	16	4	115
PHARVIS (fraction)	.744	.137	.062	.028	.013	.006	.002	.001	.000	.000	.004

purchased directly at pharmacies. This form of health care is assumed to be of lower quality than that obtained under professional supervision. In Vietnam eligible individuals, usually high-income government and private sector employees, are able to purchase health insurance that entitles them to obtain care at government hospitals and to obtain prescribed medications there also. From Table 24.3 observe that 16% of the sampled individuals have such health insurance.

Table 24.5 shows the observed frequency distribution of PHARVIS. About 26% of the individuals have one or more visits in the survey period and around 95% have a total of three or fewer visits.

Table 24.6 presents the results for several variants of the Poisson regression, analogous to those in Table 24.4 for linear regressions. The first column gives the Poisson MLE estimates, and the ordinary unadjusted t -ratios are in the second column. The next column shows robust t -ratios based on heteroskedasticity-consistent variance estimates. These are considerably smaller, in some cases by a factor exceeding 2, than the unadjusted ones. The fourth column gives cluster-adjusted t -ratios based on variances calculated using (24.45). The fact that these are substantially smaller than those in the two preceding columns confirms that there is indeed significant intracluster

Table 24.6. Vietnam Health Care Use: RE and FE Models for Pharmacy Visits

Variables	Poisson		Het Robust	Cluster Robust	Fixed Effects Poisson		Random Effects Poisson	
	Coef.	t	t	t	Coef.	t	Coef.	t
CONS	-1.637	35.78	18.81	12.25	-	-	1.318	19.41
LNHHEXP	.078	5.68	3.08	1.90	-.114	6.01	-.095	4.95
INSURANCE	-.245	9.57	5.68	4.29	-.163	6.17	-.178	6.44
SEX	.084	4.96	2.76	2.73	.098	5.75	.099	5.71
AGE	.024	2.38	1.27	1.06	.003	0.32	.005	0.55
MARRIED	.124	5.92	2.96	2.78	.164	7.59	.158	7.38
ILLDAYS	.042	40.00	14.91	12.91	.046	40.14	.046	40.18
ACTDAYS	.008	1.71	0.43	0.45	.025	4.53	.024	4.35
INJURY	.171	2.30	0.84	0.85	.144	1.80	.143	1.80
ILLNESS	.562	87.15	24.60	21.81	.584	73.45	.585	74.16
EDUC	-.052	11.10	6.47	3.92	-.024	4.18	-.026	4.61
-ln L		25281			22446		23419	
N		27765			27671		27765	

correlation. The average cluster size exceeds 140 observations; hence even a low degree of intracluster correlation is likely to inflate t -ratios substantially and the results confirm that.

We next consider modeling the intracluster correlation using FE and RE models. The FE model is estimated using the conditional MLE. Some clusters that do not have sufficient intracluster variation are dropped. The estimated coefficients lead to dramatically different conclusions from those of the Poisson MLE estimates. First, note that the coefficient of $\ln(\text{HHEXP})$ switches from being significantly positive to being significantly negative. This means that the original regression suggested that a pharmacy visit is a normal good, but the FE estimates suggest that it is an inferior good; that is, individuals avoid this form of self-medication as income rises. This can be rationalized as the fixed effects picking up the influence of omitted variables that are correlated with the observed outcomes. These omitted variables could be the quantity and quality of alternative medical services available to commune residents. These could vary a great deal depending upon the geographical location and economic status of the communes.

The last two columns in Table 24.6 give results based on random effects formulation. Here it is assumed that the intercept in the Poisson distribution varies randomly over clusters, and each cluster “draws” its intercept from a common univariate distribution, specifically a gamma distribution with unit mean. This formulation is attractive because it does not require conditioning. The RE Poisson panel model with gamma-distributed intercept, developed by Hausman et al. (1984), has an analytical likelihood function that can be adapted for clustered data. The estimates obtained for the RE model are qualitatively similar to those from the FE model. However, the estimated coefficient for the key income variable has shifted a long way from that obtained under the simple Poisson assumption.

This example shows that intracluster correlation may have an impact not just on efficiency alone but also on the estimates themselves.

24.8. Complex Surveys

The discussion in preceding sections focused on stratification, weighting, and clustering in isolation. Here we focus on complex surveys that use a stratified multistage cluster sampling design. The intent of such surveys is to present a population summary when population parameters may vary across strata. Then a weighted estimator is used and is viewed as an estimate of the census parameter. The goal is to consistently estimate the variance of the weighted estimator, controlling for clustering that can be more complicated than that in Section 24.5.

24.8.1. Variance Estimation in Complex Surveys

We consider the following setup. The i th observation in the sample is household j in cluster c in strata s . For example, the dependent variable is denoted y_{scj} , though more formally the observation (s, c, j) may be represented as observation (s, c_s, j_c) . The data are $(y_{scj}, \mathbf{x}_{scj}, w_{scj})$, where w_{scj} are sample weights inversely proportional to the

probability of selection of the observation in the sample. The subscripts are ordered in terms of level of disaggregation, a reversal from the notation of Section 24.5.

Two-stage or multistage sampling is used within strata, with households selected as the result of at least two sequential draws. First, a subset of all PSUs within the strata is randomly drawn. Second, a subset of all households in the selected PSUs is drawn, where clustered sampling may be permitted. Further draws within an SSU and so on are also possible.

Variance of a Linear Statistic

The starting point is to consider estimation of the variance of a linear statistic that sums over strata, PSU, and households:

$$\hat{u} = \sum_{s=1}^S \sum_{c=1}^{C_s} \sum_{j=1}^{N_{cs}} u_{scj} = \sum_{s=1}^S \sum_{c=1}^{C_s} u_{sc},$$

where u_{sc} are the totals within a PSU, so

$$u_{sc} = \sum_{j=1}^{N_{cs}} u_{scj}.$$

Examples of u_{scj} such as the weighted mean and weighted regression are given in the following. The variance of u is

$$\text{V}[u] = \sum_{s=1}^S \sum_{c=1}^{C_s} \text{V}[u_{sc}] = \sum_{s=1}^S C_s \sigma_s^2,$$

if we assume that u_{sc} are independent over strata and are iid over PSUs with common variance σ_s^2 . The usual unbiased variance estimate of σ_s^2 can be used, given u_{sc} iid over c , so $\hat{\sigma}_s^2 = (C_s - 1)^{-1} \sum_c (u_{sc} - \bar{u}_s)^2$. It follows that

$$\text{V}[\hat{u}] = \sum_{s=1}^S \frac{C_s}{C_s - 1} \sum_{c=1}^{C_s} (u_{sc} - \bar{u}_s)^2, \quad (24.55)$$

where $\bar{u}_s = C_s^{-1} \sum_c u_{sc}$ is the stratum average of the PSU totals.

This estimator allows for clustering within a PSU, since

$$\begin{aligned} \sum_{c=1}^{C_s} (u_{sc} - \bar{u}_s)^2 &= \sum_{c=1}^{C_s} \left(\sum_{j=1}^{N_{cs}} u_{scj} - \bar{u}_s \right)^2 \\ &= \sum_{c=1}^{C_s} \sum_{j=1}^{N_{cs}} (u_{scj} - \bar{u}_s)^2 + \sum_{c=1}^{C_s} \sum_{j=1}^{N_{cs}} \sum_{k \neq j}^{N_{cs}} (u_{scj} - \bar{u}_s)(u_{sck} - \bar{u}_s). \end{aligned}$$

The first sum is the contribution to the variance under SRS. The second sum will be positive under clustered sampling and leads to a larger variance. No assumption has been made about the nature of the sampling within strata nor about the type of clustering that arises. For example, (24.55) gives correct standard errors even if there is three-stage sampling with further subsampling with SSUs.

The estimator (24.55) does require that at least two PSUs be drawn from each strata. If only one PSU is drawn then one possibility is to collapse the strata that includes the single PSU into another strata that is viewed a priori as being reasonably similar. It is feasible provided $C_s \geq 2$, that is, if there are at least two PSUs per stratum. This will lead to overestimation of $V[u]$ as an upward bias is introduced because of the different means in different strata.¹

In practice PSUs are sampled without replacement so there is some dependence in u_{scj} . Then (24.55) overestimates $V[u]$, similar to the situation in Section 24.2.3. More complicated formulas have been proposed.

Variance of the Weighted Mean

The population mean is estimated by the ratio of the sample-weighted total of y_{scj} , say \hat{y} , to the sum of the sample weights, say \hat{w} . Then

$$\bar{y}_W = \hat{y}/\hat{w} = \sum_{s=1}^S \sum_{c=1}^{C_s} \sum_{j=1}^{N_{cs}} w_{scj} y_{scj} / \left(\sum_{s=1}^S \sum_{c=1}^{C_s} \sum_{j=1}^{N_{cs}} w_{scj} \right).$$

If the sample weights are treated as known, then more simply

$$\bar{y}_W = \sum_{s=1}^S \sum_{c=1}^{C_s} \sum_{j=1}^{N_{cs}} w_{scj}^* y_{scj},$$

where $w_{scj}^* = w_{scj}/\hat{w}$ and $V[\bar{y}_W]$ can be applied using (24.55) with $u_{scj} = w_{scj}^* y_{scj}$.

If the sample weights are treated as unknown then the **delta method** or **linearization method** can be used to obtain $V[\hat{y}/\hat{w}]$ as a function of $V[\hat{y}]$, $V[\hat{w}]$, and $\text{Cov}[\hat{y}, \hat{w}]$. The first two quantities can be estimated using (24.55) with $u_{scj} = w_{scj} y_{scj}$ and $u_{scj} = w_{scj}$. The third quantity can be estimated with $(u_{sc} - \bar{u}_s)^2$ in (24.55) replaced by $(u_{sc} - \bar{u}_s)(v_{sc} - \bar{v}_s)$, where $u_{scj} = w_{scj} y_{scj}$ and $v_{scj} = w_{scj}$. This is an example of a ratio estimator.

For nonlinear statistics such as these ratio estimates, the literature proposes other estimates based on the **jackknife** and **balanced repeated replication**. Because of the nonlinearity the variance estimates are no longer unbiased but can be shown to be consistent if the number of strata $S \rightarrow \infty$ (see Krebski and Rao, 1981). Some results with S fixed and $\sum_{c=1}^{C_s} N_{cs} \rightarrow \infty$ are summarized in Wolter (1985). One can also **bootstrap**, though care is needed. See Rao and Wu (1988) and Shao and Tu (1995).

Variance of Weighted Least-Squares Estimator

From Section 24.3, the weighted regression estimate $\hat{\beta}_W$ of the census regression parameters solve

$$\sum_{s=1}^S \sum_{c=1}^{C_s} \sum_{j=1}^{N_{cs}} w_{scj} \mathbf{x}_{scj} (y_{scj} - \mathbf{x}'_{scj} \hat{\beta}_W) = \mathbf{0}.$$

¹ For the CPS the method here cannot be directly applied as many strata have only one PSU and for other strata only one PSU is collected. Instead, various pseudo-strata are formed and replication methods are used that resample PSUs from the pseudo-strata. See U.S. Census Bureau (2002).

By the usual algebra, we have

$$\widehat{\beta}_W - \beta = \left(\sum_{s=1}^S \sum_{c=1}^{C_s} \sum_{j=1}^{N_{sc}} w_{scj} \mathbf{x}_{scj} \mathbf{x}'_{scj} \right)^{-1} \times \sum_{s=1}^S \sum_{c=1}^{C_s} \sum_{j=1}^{N_{sc}} w_{scj} (y_{scj} - \mathbf{x}'_{scj} \widehat{\beta}_W).$$

This leads to the sandwich form $V[\widehat{\beta}] = \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1}$, where \mathbf{B} is the variance of the second triple sum, which can be estimated using (24.55) with $u_{scj} = w_{scj} \mathbf{x}_{scj} (y_{scj} - \mathbf{x}'_{scj} \widehat{\beta}_W)$.

Variance of Weighted m-Estimator

A quite general framework considers the weighted m-estimator $\widehat{\theta}_W$ that solves

$$\sum_{s=1}^S \sum_{c=1}^{C_s} \sum_{j=1}^{N_{sc}} w_{scj} \mathbf{h}(y_{scj}, \mathbf{x}_{scj}, \widehat{\theta}_W) = \mathbf{0}.$$

Examples include linear regression, $\mathbf{h}_{scj} = \mathbf{x}_{scj} (y_{scj} - \mathbf{x}'_{scj} \beta)$, and quasi-maximum likelihood, $\mathbf{h}_{scj} = \partial \ln f(y_{scj} | \mathbf{x}_{scj}, \theta) / \partial \theta$.

Assuming consistent estimation of θ , which requires that $E[\mathbf{h}(y_{scj}, \mathbf{x}_{scj}, \theta)] = \mathbf{0}$, we can use the usual first-order Taylor series expansion on the estimating equation to get

$$\sqrt{N} (\widehat{\theta}_W - \theta) \xrightarrow{d} \mathcal{N} \left[\mathbf{0}, \mathbf{A}^{-1} \mathbf{B} \mathbf{A}'^{-1} \right],$$

where

$$\mathbf{A} = \text{plim} N^{-1} \sum_{s=1}^S \sum_{c=1}^{C_s} \sum_{j=1}^{N_{sc}} w_{scj} \frac{\partial \mathbf{h}(y_{scj}, \mathbf{x}_{scj}, \theta)}{\partial \theta'}$$

and

$$\mathbf{B} = \text{plim} N^{-1} \sum_{s=1}^S \sum_{c=1}^{C_s} \sum_{j=1}^{N_{sc}} \sum_{k=1}^{N_{sc}} w_{scj} w_{sck} \mathbf{h}(y_{scj}, \mathbf{x}_{scj}, \theta) \frac{\partial \mathbf{h}(y_{sck}, \mathbf{x}_{sck}, \theta)}{\partial \theta'},$$

where the expression for \mathbf{B} assumed independence of \mathbf{h}_{scj} over strata and clusters but permits dependence within a cluster. Estimation of \mathbf{A} is straightforward. For \mathbf{B} use (24.55) with $u_{scj} = w_{scj} \mathbf{h}_{scj}$, so

$$\mathbf{B} = \sum_{s=1}^S \frac{C_s}{C_s - 1} \sum_{c=1}^{C_s} [\bar{z}_{sc} - \bar{z}_s]^2,$$

where $\bar{z}_{sc} = \sum_{j=1}^{N_{sc}} w_{scj} \mathbf{h}(y_{scj}, \mathbf{x}_{scj}, \theta)$ and $\bar{z}_s = C_s^{-1} \sum_{c=1}^{C_s} \bar{z}_s$.

Endogenous Stratification

Sakata (1998) extends these results to endogenous sampling. He takes a census parameter approach and provides asymptotic theory assuming the number of strata $S \rightarrow \infty$. The results are the same as those given in the previous section.

24.9. Practical Considerations

It is most common in microeconomics research to take a structural approach. Unweighted estimators are used, provided there is no endogenous stratification. The main concern is to obtain correct standard errors if clustering is present. If cluster effects are random there is usually little efficiency loss in ignoring clustering in estimation. Some packages may have a cluster robust standard errors option, not to be confused with a heteroskedasticity robust option, which is appropriate if cluster effects are random and there are many clusters. The CSRE and CSFE models can be implemented using OLS, provided in the case of CSFE there are not too many clusters. Alternatively, a panel data module can be used if it supports unbalanced panels. As with panel data most researchers outside econometrics are content to take a random effects approach, but a fixed effects approach may be necessary for consistent estimation.

If a descriptive approach is taken and parameters vary over strata then weighting is necessary. A weighting option within least squares can be used, but it needs to be combined with a cluster-robust standard errors option. Some packages have a survey estimation module that obtains cluster-robust standard errors using the methods of Section 24.6. The package SUDAAN implements many of the methods in this chapter for linear and leading nonlinear regression models.

24.10. Bibliographic Notes

24.2–24.3 The literature on survey sampling is vast. Classic references on sample surveys include Kish (1965) and Cochran (1977, first edition 1953). Skinner (1989) provides a useful overview and Groves (1989) provides a relatively nontechnical treatment that presents the approaches of many of the social sciences to surveying, while raising many useful practical issues. For completeness we have incorporated some of this survey sampling literature, though econometrics studies rarely implement the methods in Section 24.8. There are few econometrics references, with the notable exception of chapters in Pudney (1989) and Deaton (1997) and a book chapter by Ullah and Breunig (1998).

24.4 The main focus of the theoretical econometrics literature has been controlling for endogenous stratification. This literature is challenging and we have merely provided an overview. For detail see Amemiya (1985), who provides many references including Manski and Lerman (1977) for discrete-choice models and Hausman and Wise (1979) for sample selection models. The simple weighted estimator is generally appropriate albeit inefficient. Imbens and Lancaster (1996) present a practical way to implement a fully efficient estimator given specification of the conditional density.

24.5 For microeconomics applications controlling for clustering is of greatest importance. The works by Kloek (1981) and Moulton (1986, 1990) were key in alerting econometricians to this problem. Davis (2002) gives a general treatment of multi-way error component models. Graubard and Korn (1994) provide a useful discussion of linear regression analysis of clustered data. They pay attention to both fixed and random effects models, with emphasis on the assumptions that must be satisfied for the random effects model to be valid. Pendergast et al. (1996) provide an extensive survey of the methods for analyzing clustered binary data. Because the middle term on the right-hand side of (24.34) involves averaging over the number of clusters, the

precision of this estimate depends on the number of clusters. The consequences of using the cluster-robust variance matrix when the number of clusters is small continues to be a topic of research (Donald and Lang, 2001; Angrist and Lavy, 2002). Wooldridge (2003) provides an overview.

- 24.6** Hierarchical linear models have been extensively used in social sciences. Bryk and Raudenbush (2002) provide a comprehensive coverage of binary, ordered, counted, and multinomial outcomes from both likelihood and Bayesian perspectives.
- 24.7** Deaton (1997) examines a number of issues of modeling using data from clustered samples from various Living Standards Surveys conducted in developing economies by the World Bank.
- 24.8** Many standard statistical software packages (e.g., STATA and SUDAAN) accommodate both fixed and random effects formulations of clustering in linear and nonlinear models for cross-section and panel data.

Exercises

- 24–1** (a) Verify the expression for Σ_c given at (24.25).
 (b) Prove the consistency property of the estimators $\hat{\sigma}^2$ and $\hat{\rho}$ in the CSRE model.
 (c) Consider the bias of the standard errors in the balanced cluster CSRE model. Show that in this case $E[\sum_c \sum_j \hat{v}_{cj}^2] = \sigma^2[N - K(1 + \rho(m - 1))]$.
- 24–2** (Adapted from Greenwald, 1983) Consider the linear regression model $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$, where $E[\mathbf{u}] = \mathbf{0}$ and $E[\mathbf{u}\mathbf{u}'] = \sigma^2\Omega^* = \Omega$. By standard results for the OLS estimator $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ (see Section 4.4) we can obtain the correct expression for $V[\hat{\beta}]$ as $\mathbf{V}_2 = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\Omega\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})^{-1}$, whereas $\mathbf{V}_1 = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$ with $\hat{\sigma}^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}}/(N - K)$ is invalid if $\Omega \neq \mathbf{I}$.
- (a) Show that the bias of \mathbf{V}_1 is given by $\mathbf{B} = \mathbf{B}_1 + \mathbf{B}_2$, where $\mathbf{B}_2 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\Omega - \sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ and $\mathbf{B}_1 = (N - K)^{-1} \operatorname{tr}\{\mathbf{B}_2(\mathbf{X}'\mathbf{X})\}(\mathbf{X}'\mathbf{X})^{-1}$. (Greenwald refers to \mathbf{B}_2 as “direct bias.”)
 (b) Evaluate the two terms for the special case of $\mathbf{X}'\mathbf{X} = \mathbf{I}_K$. Show that $\mathbf{B} \rightarrow \mathbf{B}_2$ as $N \rightarrow \infty$.
- 24–3** Consider the OLS cluster-robust variance estimator formula (24.33). Suppose there are two levels of clustering. Specifically, in the context of the empirical example of this chapter, clustering could be at the level of family and commune if multiple members of the family from the same commune are included in the survey. How will the formula be modified if the data have two levels of clustering?
- 24–4** For this exercise use a 50% sample of the VLSMS data. Define $y = 1$ if the subject has at least one pharmacy visit (PHARVIS) and $y = 0$ otherwise. This example presumes access to a program that handles clustering.
- (a) Using the same explanatory variables as those for the Poisson model in Section 24.7, estimate a binary logit model by maximum likelihood, using both the standard estimator and the robust sandwich estimator for the variance.
 (b) Reestimate the specification of part (a) using the cluster-robust standard error option. Explain the differences between the robust standard errors of parts (a) and (b).

24.10. BIBLIOGRAPHIC NOTES

- (c)** Use the “commune” variable as a cluster identifier. Reestimate the logit model using the cluster fixed effects and cluster random effects specification. Compare the estimates and standard errors of the coefficients of LNHHEXP and INSURANCE. Are the conclusions about the significance of these variables affected by clustering in the data?

Treatment Evaluation

25.1. Introduction

The topic of treatment evaluation concerns measuring the impact of interventions on outcomes of interest, with the type of intervention and outcome being defined broadly so as to apply to many different contexts. The treatment evaluation approach and some of its terminology comes from medical sciences where intervention frequently means adopting a treatment regime. Subsequently, one may be interested in measuring the response to the treatment relative to some benchmark, such as no treatment or a different treatment. In economic applications treatment and interventions usually mean the same thing.

Examples of treatments in the economic context are enrollment into a labor training program, being a member of a trade union, receipt of a transfer payment from a social program, changes in regulations for receiving a transfer from a social program, changes in rules and regulations pertaining to financial transactions, changes in economic incentives, and so forth; see Moffitt (1992), Friedlander, Greenberg, and Robbins (1997), and Heckman, Lalonde, and Smith (1999). If the treatment that is applied can vary in intensity or type, we use the term **multiple treatments** when referring to them collectively. Relative to a single type of treatment this does not create complications, but now the choice of a benchmark for comparisons is more flexible.

The term outcome refers to changes in economic status or environment on economic outcomes of individuals. A leading case is one in which the outcome of interest is a continuous variable, say y , whereas the treatment variable is discrete and of on/off variety, say D , where D takes the value 1 if the treatment is applied and is 0 otherwise. An example of an intervention is labor market training, which could affect posttraining wages of the worker. In general, however, either the outcome or treatment can be continuous or discrete or exhibit limited variation. Whereas the details of the analysis will vary, certain key ideas will be relevant in all situations. For simplicity, we will take the case of a continuous outcome and a binary-valued treatment as our leading case. Later we will extend the analysis to other practically relevant situations.

Policy relevance of treatment evaluation is direct because “successful” treatments can be linked to desirable social programs, or improvements in existing programs to attain objectives of social policy. Heckman and Smith (1998) have discussed the relationship between several commonly used measures of treatment impact and traditional cost–benefit analysis.

The standard problem in treatment evaluation involves the inference of a causal connection between the treatment and the outcome. In a canonical single-treatment example we observe (y_i, \mathbf{x}_i, D_i) , $i = 1, \dots, N$, and the impact of a hypothetical change in D on y , holding \mathbf{x} constant, is of interest. Such inference is the main feature of the **potential outcome model**, already introduced in Chapter 2, in which the outcome variable of interest is compared in the treated and nontreated states. However, no individual is simultaneously observed in both states. Hence, the situation is akin to one of missing data, and it can be tackled by methods of causal inference carried out in terms of **counterfactuals**. We ask how the outcome of an average untreated individual would change if such a person were to receive the treatment. That is, a magnitude like $\Delta y / \Delta D$ is of interest. Fundamentally one’s interest lies in the outcomes that result from, or are caused by, such interventions. Here causation is in the sense of *ceteris paribus*, meaning that we hold all other variables constant.

What is the difference between this chapter and earlier ones in which we also considered the identification and estimation of a variety of models? There are many similarities and the differences arise from a shift of emphasis. The main difference stems from the focus on a family of measures of treatment effectiveness. These measures are functions of parameters and data, and they enable comparisons with policy-relevant counterfactuals. An important and interesting result is that not all measures can be constructed, given the data and the estimator. The choice of an estimator and the type of data used in model estimation place restrictions on the counterfactuals that can be identified, and hence on the impact measures that can be consistently estimated.

Another emphasis in the literature on treatment evaluation is on the advantages of identification secured using minimal functional form and exclusion restrictions, (e.g., semiparametric identification). This emphasis is motivated by the desire to produce results that have policy significance but whose validity does not depend on strong assumptions. The feasibility of semiparametric identification is relatively easier to establish for treatment effect estimation in linear models, with continuous support for the dependent variable, than it is in nonlinear models with limited dependent variables.

Section 25.2 discusses identification assumptions. Section 25.3 presents measures of treatment effect that are usually targeted in identification and estimation. Section 25.4 analyzes matching and propensity score estimators. Differences-in-differences estimators of treatment effects that are common in event studies with a quasi-experimental data setup are covered in Section 25.5. Continuing with a quasi-experimental setup, we discuss the regression discontinuity design in Section 25.6, followed by the instrumental variable estimator in Section 25.7. Much of the discussion up to this point is related to linear models. Section 25.8 provides a detailed empirical illustration of the methods developed in the chapter.

25.2. Setup and Assumptions

The methods for estimation of treatment effects rely on assumptions to permit identification of causal effects just as, for example, the linear SEM relies on assumptions to permit causal effects (see Chapter 2). In this section we detail the assumptions that permit use of the key matching and propensity score estimators that are presented later in Section 25.4.

First we consider a framework for estimating causal parameters in treatment evaluation.

25.2.1. Treatment Effects Framework

Let us begin with the setup of randomized treatment assignment in a social experiment as described in Section 3.3. Let there be a target population for the treatment of interest and let N denote the number of randomly selected individuals who are eligible for treatment. Let N_T denote the number of randomly selected individuals who are treated and let $N_C = N - N_T$ denote the number of nontreated individuals who serve as a potential control group.

Random assignment implies that the treatment assignment ignores the possible impact of the treatment on the outcomes. For example, no one is included in the treatment group on the grounds that the benefit of the treatment to that individual would be large, and no one is excluded because the expected benefit is small. Let $(y_i, \mathbf{x}_i, D_i; i = 1, \dots, N)$ be the vector of observations on the scalar-valued **outcome variable** y , a vector of observable variables \mathbf{x} , and a binary indicator of a treatment variable D . For simplicity, we assume that anyone who is assigned treatment gets it, and anyone who is not does not get it. The outcome variable of the treated individual is denoted y_1 and that for the nontreated individual is denoted y_0 . After the experiment is run and data are collected, we would like to obtain a measure of the treatment impact. The most natural way of measuring the effect of the treatment would be to construct a measure that compares the average outcomes of the **treated** and **nontreated** groups.

With one important difference the same data setup could be applied to observational data. The difference is that there is no random assignment mechanism for treatment, perhaps because individuals choose to be treated, or because of some other reason.

It needs to be stated at the outset that most treatment evaluation studies have a partial equilibrium character. Specifically, they assume an absence of general equilibrium effects. By that we mean that the treatment effects are small and do not affect the status of some of the variables that are treated as exogenous. This assumption will not do if one were considering a treatment program that affected an entire sector that was a significant part of the national economy. For example, instituting universal health insurance may have impact on the entire health services sector, which would make it difficult to apply the methods discussed in this chapter.

There are potential pitfalls in constructing estimates of treatment effects. There are also subtle differences of interpretations that arise from variations in the assumptions used to construct such measures. Therefore, we begin by examining these assumptions.

25.2.2. Conditional Independence Assumption

Meaningful comparisons between the outcomes of the two groups require some assumptions. We shall initially list and explain these assumptions and later use them in the discussion of identifiability of certain treatment effects.

An important assumption is the **conditional independence assumption** that states that conditional on \mathbf{x} , the outcomes are independent of treatment, written as

$$y_0, y_1 \perp D | \mathbf{x}. \quad (25.1)$$

Behavioral implication of this assumption is that participation in the treatment program does not depend on outcomes, after controlling for the variation in outcomes induced by differences in \mathbf{x} . Random assignment, properly applied, will validate this assumption. Indeed, under completely random assignment one may even make a stronger assumption

$$y_0, y_1 \perp D, \quad (25.2)$$

because randomization would be over (y, \mathbf{x}) space. The more commonly used assumption (25.1), if valid, can be useful for identification of some impact parameters because it states that once we control for the effects of regressors \mathbf{x} , some of which may be related to D , treatment and outcomes are independent.

The conditional independence assumption is broad and implies the following:

$$\begin{aligned} F(y_j | \mathbf{x}, D = 1) &= F(y_j | \mathbf{x}, D = 0) = F(y_j | \mathbf{x}), j = 0, 1, \\ F(u_j | \mathbf{x}, D = 1) &= F(u_j | \mathbf{x}, D = 0) = F(u_j | \mathbf{x}), j = 0, 1, \end{aligned} \quad (25.3)$$

where u is the regression model error, which means that the participation decision does not affect the **distribution of potential outcomes**.

To see the impact of this assumption let $E[y | \mathbf{x}, D]$ be linear; that is, the outcome-participation equation is

$$y = \mathbf{x}'\beta + \alpha D + u, \quad (25.4)$$

where $E[u | D] = E[y - \mathbf{x}'\beta - \alpha D | D] = 0$. Therefore, D may be treated as an exogenous variable, and there will be no simultaneity bias or selection bias. Under the standard conditions on \mathbf{x} , consistent estimation of regression parameters is possible.

An assumption that is weaker than (25.1) is

$$y_0 \perp D | \mathbf{x}, \quad (25.5)$$

which implies conditional independence of participation and y_0 . This assumption is used in establishing identifiability of a population-average **treatment effect on the treated (ATET)**, as will be seen later.

Assumption (25.5) has other names in the literature. Imbens (2005) refers to it as the **unconfoundedness assumption** and Rubin refers to it as the **ignorability assumption** (Rubin, 1978; Wooldridge, 2001). If valid, the assumption implies that there is no **omitted variable bias** once \mathbf{x} is included in the regression, and hence there will be no confounding. The assumption is tantamount to treatment assignment that ignores outcomes; hence it is appropriate to refer to it as the ignorability assumption.

This assumption is necessary if the treatment variable is to be treated as exogenous, which is essential for simplicity in estimation. If valid, sample selection models or IV methods to handle endogenous treatment variables are not needed, and the methods of Section 25.4 can be applied.

25.2.3. Matching Assumption

A second assumption, referred to as the **overlap** or **matching assumption**, is necessary for identifying some population measures of impact. It states that

$$0 < \Pr[D = 1 | \mathbf{x}] < 1. \quad (25.6)$$

This assumption ensures that for each value of \mathbf{x} there are both treated and nontreated cases. In that sense there is overlap between the treated and untreated subsamples. For each treated individual there is another matched untreated individual with a similar \mathbf{x} . If the assumption were to fail, then we could potentially have individuals with \mathbf{x} vectors who are all treated and those with a different \mathbf{x} who are all untreated. This condition is not required for identifying the treatment parameter for the treated group. For identifying the treatment effect on a randomly selected individual one needs for each participant an analogous nonparticipant. Then the condition $\Pr[D = 1 | \mathbf{x}] < 1$ is sufficient.

25.2.4. Conditional Mean Assumption

A third assumption is the **conditional mean independence assumption**

$$\mathbb{E}[y_0 | D = 1, \mathbf{x}] = \mathbb{E}[y_0 | D = 0, \mathbf{x}] = \mathbb{E}[y_0 | \mathbf{x}], \quad (25.7)$$

which implies that y_0 does not determine participation.

25.2.5. Propensity Scores

When treatment participation is not by random assignment but depends stochastically on a vector of observable variables \mathbf{x} , as in observational data or when the treatment is targeted to some population defined by some observable characteristics (such as age, sex, or socioeconomic status), then the concept of **propensity scores** is useful. This is a conditional probability measure of treatment participation given \mathbf{x} and is denoted $p(\mathbf{x})$, where

$$p(\mathbf{x}) = \Pr[D = 1 | \mathbf{X} = \mathbf{x}]. \quad (25.8)$$

The propensity score measure can be computed given the data (D_i, \mathbf{x}_i) using any of the parametric or semiparametric methods covered in Chapter 14 (e.g., by doing a logit regression).

An assumption that plays an important role in treatment evaluation is the **balancing condition**, which states that

$$D \perp \mathbf{x} | p(\mathbf{x}). \quad (25.9)$$

Table 25.1. Treatment Effects Framework

Symbol	Definition
y_1	Outcome for the treated group
y_0	Outcome for the nontreated group
$p(\mathbf{x})$	Propensity score
N_T	Number of treated cases in the sample

This can be expressed alternatively by saying that for individuals with the same propensity score the assignment to treatment is random and should look identical in terms of their \mathbf{x} vector. The balancing condition is a testable hypothesis.

A useful result about conditional independence given $p(\mathbf{x})$ due to Rosenbaum and Rubin (1983) states that

$$y_0, y_1 \perp D | \mathbf{x} \Rightarrow y_0, y_1 \perp D | p(\mathbf{x}). \quad (25.10)$$

This implies that the conditional independence assumption given \mathbf{x} implies conditional independence given $p(\mathbf{x})$, that is, independence of y_0 , y_1 , and D given $p(\mathbf{x})$.

To obtain this result, note that

$$\begin{aligned} \Pr[D = 1 | y_0, y_1, p(\mathbf{x})] &= E[D | y_0, y_1, p(\mathbf{x})] \\ &= E[E[D | y_0, y_1, p(\mathbf{x}), \mathbf{x}] | y_0, y_1, p(\mathbf{x})] \\ &= E[E[D | y_0, y_1, \mathbf{x}] | y_0, y_1, p(\mathbf{x})] \\ &= E[E[D | \mathbf{x}] | y_0, y_1, p(\mathbf{x})] \\ &= E[p(\mathbf{x}) | y_0, y_1, p(\mathbf{x})] \\ &= p(\mathbf{x}). \end{aligned}$$

Here the second and third lines follow from the law of iterated expectations. The fourth line uses conditional independence. The intuition behind this result is that $p(\mathbf{x})$ is a particular function of \mathbf{x} and, in a sense, contains less information than \mathbf{x} . Hence conditional independence given $p(\mathbf{x})$ is implied for the same given \mathbf{x} . Because by conditioning on \mathbf{x} we get rid of the correlation between \mathbf{x} and D , likewise by conditioning on the propensity score $p(\mathbf{x})$ we also expunge the correlation between \mathbf{x} and D . Thus a regression similar to (25.4) is

$$y = \mathbf{x}'\beta + \alpha p(\mathbf{x}) + u \quad (25.11)$$

$$= \mathbf{x}'\beta + \alpha \hat{p}(\mathbf{x}) + (u + \alpha(p(\mathbf{x}) - \hat{p}(\mathbf{x}))), \quad (25.12)$$

where in the second line the unknown $p(\mathbf{x})$ is replaced by a sample estimate, resulting in the addition of the sampling error to the regression error. The pros and cons of this strategy will be considered later. Table 25.1 summarizes the notation.

25.3. Treatment Effects and Selection Bias

We begin by presenting two-widely used measures of treatment effect – one that averages over all individuals and one that averages over only the treated. We then discuss

in some detail the role of selection into treatment. The methods presented in Sections 25.4–25.6 presume that selection effects directly depend on only measurable observed characteristics of the individual, such as age. If additionally selection effects depend on unobservables then the methods of Chapter 16 must instead be used. The current section includes considerable discussion of selection issues.

25.3.1. Two Key Parameters: ATE and ATET

Define Δ as the difference between the outcome in the treated and untreated states

$$\Delta = y_1 - y_0, \quad (25.13)$$

where we may condition on \mathbf{x} if desired. It is emphasized that Δ is not directly observable because no individual can be observed in both states. Population values of the **average treatment effect** and **average treatment effect on the treated** are defined as

$$\text{ATE} = E[\Delta], \quad (25.14)$$

$$\text{ATET} = E[\Delta | D = 1], \quad (25.15)$$

with sample analogues

$$\widehat{\text{ATE}} = \frac{1}{N} \sum_{i=1}^N [\Delta_i], \quad (25.16)$$

$$\widehat{\text{ATET}} = \frac{1}{N_T} \sum_{i=1}^{N_T} [\Delta_i | D_i = 1], \quad (25.17)$$

where $N_T = \sum_{i=1}^N D_i$. In each of these two cases, computation is straight-forward if Δ_i can be obtained. The procedure is not direct because the formulas have an unobserved component that must be estimated and that step calls for some assumptions.

The ATE measure is relevant when the treatment has universal applicability so that it is reasonable to consider the hypothetical gain from treatment to a randomly selected member of the population. The ATET measure is relevant when we want to consider the average gain from treatment for the treated. See Heckman and Vytlacil (2002).

To understand the treatment evaluation problem consider the average gain from participation given characteristics \mathbf{x} . This is

$$\begin{aligned} \text{ATE} &= E[\Delta | X = \mathbf{x}] \\ &= E[y_1 - y_0 | X = \mathbf{x}] \\ &= E[y_1 | X = \mathbf{x}] - E[y_0 | X = \mathbf{x}] \\ &= E[y_1 | \mathbf{x}, D = 1] - E[y_0 | \mathbf{x}, D = 0], \end{aligned} \quad (25.18)$$

where the last equality uses the conditional independence assumption (25.1).

Given a sample of participants, $E[y_1 | D = 1, \mathbf{x}]$ can be estimated. However, $E[y_0 | \mathbf{x}, D = 0]$ is *not* observable because it is a measure of the average outcomes for the participants had they in fact not participated, and one cannot simultaneously observe the same individuals as both participants and nonparticipants. To make ATE operational we must find an estimator for the second term.

By definition (25.18)

$$\text{ATE} = E[y_1|\mathbf{x}, D = 1] - E[y_0|\mathbf{x}, D = 0] \quad (25.19)$$

$$= \mu_1(\mathbf{x}) - \mu_0(\mathbf{x}) + E[u_1|\mathbf{x}, D = 1] - E[u_0|\mathbf{x}, D = 0]$$

$$= \mu_1(\mathbf{x}) - \mu_0(\mathbf{x}) + E[u_1|\mathbf{x}] - E[u_0|\mathbf{x}]$$

$$= \mu_1(\mathbf{x}) - \mu_0(\mathbf{x}), \quad (25.20)$$

where the first term in the first line on the right-hand side can be estimated using the data from treatment participants, but the second term is not directly observable. The next three lines follow by applying the conditional independence and conditional mean assumption and adopting the specifications $y_1 = \mu_1(\mathbf{x}) + u_1$ for the treated and $y_0 = \mu_0(\mathbf{x}) + u_0$ for the untreated. The second from the last line only requires mean independence rather than full conditional independence.

25.3.2. Sampling and Selection Bias

The crux of the evaluation problem is that $E[y_0|\mathbf{x}, D = 1]$ is unobservable. The solution to the problem depends in part on the type of data available. Social experiments use the eligible participants that are excluded from the treatment group as a proxy for the counterfactual. Observational studies generate a **comparison group** from the same source as the treated group, or from other databases, and essentially end up using some function of $E[y_0|\mathbf{x}, D = 0]$ that can be estimated using data from nonparticipants. The simplicity of the computation when the data come from a well-designed and executed social experiment should be viewed against the background of actual social experiments, which are subject to other problems such as **randomization bias** and **substitution bias** (discussed in Chapter 3).

Suppose that for the treated participants the outcome equation is

$$y_1 = E[y_1|\mathbf{x}] + u_1 \quad (25.21)$$

$$= \mu_1(\mathbf{x}) + u_1 \quad (25.22)$$

and for the nonparticipants the equation is

$$y_0 = E[y_0|\mathbf{x}] + u_0 \quad (25.23)$$

$$= \mu_0(\mathbf{x}) + u_0. \quad (25.24)$$

Note that this specification is of the switching regression type (analogous to the Roy model discussed in Section 16.7) in the sense that the treated and nontreated have different conditional mean functions, $\mu_1(\mathbf{x})$ and $\mu_0(\mathbf{x})$, that are written in a more general notation than necessary for the purely linear model. We assume that $E[u_1|\mathbf{x}] = E[u_0|\mathbf{x}] = 0$, though $E[u_1|\mathbf{x}, D]$ and $E[u_0|\mathbf{x}, D]$ do not necessarily equal zero.

A more common, but restrictive, specification has

$$\mu_1(\mathbf{x}) = \mu_0(\mathbf{x}) + \alpha D, \quad (25.25)$$

in which the treated group has an additional intercept component α , but the slope coefficients of the regressors are unaffected by the treatment.

Table 25.2. Treatment Effects Measures: ATE and ATET

Measure	Treatment Effect	Special Case (25.25)
ATE given \mathbf{x}	$E[\Delta \mathbf{x}] = \mu_1(\mathbf{x}) - \mu_0(\mathbf{x})$	$E[\Delta \mathbf{x}] = \alpha$
ATET with \mathbf{x} and selection effect	$E[\Delta \mathbf{x}, D = 1]$ $= \mu_1(\mathbf{x}) - \mu_0(\mathbf{x})$ $+ E[u_1 - u_0 \mathbf{x}, D = 1]$	$E[\Delta \mathbf{x}, D = 1]$ $= \alpha + E[u_1 - u_0 \mathbf{x}, D = 1]$
Additional benefit to individual with \mathbf{x}	$E[u_1 - u_0 \mathbf{x}, D = 1]$	$E[u_1 - u_0 \mathbf{x}, D = 1]$
Average selection bias	$E[u_0 \mathbf{x}, D = 1]$ $- E[u_0 \mathbf{x}, D = 0]$	$E[u_0 \mathbf{x}, D = 1]$ $- E[u_0 \mathbf{x}, D = 0]$

The observed outcome y is written as

$$y = Dy_1 + (1 - D)y_0. \quad (25.26)$$

Combining these equations we get

$$\begin{aligned} y &= D(\mu_1(\mathbf{x}) + u_1) + (1 - D)(\mu_0(\mathbf{x}) + u_0) \\ &= \mu_0(\mathbf{x}) + D(\mu_1(\mathbf{x}) - \mu_0(\mathbf{x}) + u_1 - u_0) + u_0. \end{aligned} \quad (25.27)$$

Because $D = 1$ or 0 , the second term in the regression “switches” on and off. The second term in (25.27) measures the benefit of participation; the first component $\mu_1(\mathbf{x}) - \mu_0(\mathbf{x})$ measures the average gain to a participant with characteristics \mathbf{x} and the second component $(u_1 - u_0)$ is individual-specific benefit. The second component may be observable by the participant, but not by the investigator.

The expressions for ATE and ATET are given in Table 25.2, for the general case and the specialization (25.25).

Average selection bias is the difference between program participants and nonparticipants in the base state. This effect cannot be attributed to the program. A special case is $E[u_1 - u_0|\mathbf{x}, D = 1] = 0$, which can arise if there are no unobservable components of the benefit or if the best individual estimate of $u_1 - u_0$ is zero.

Selection bias arises when the treatment variable is correlated with the error in the outcome equation. This correlation could be induced by incorrectly omitted observable variables that partly determine D and y . Then the omitted variable component of the regression error will be correlated with D – the case of **selection on observables**. Another source comprises unobserved factors that partly determine both D and y . This is the case of **selection on unobservables**. The conditional independence assumption essentially rules out confounding caused by omitted variables.

25.3.3. Selection on Observables

In observational data the problem of selection on observables is solved using regression and matching methods. Subsequent sections of this chapter present these methods in detail. Before doing so, we note that the two-part model of Section 16.4 is an example, and in this section we discuss a second straightforward method.

The **control function estimator** is motivated by the possibility that a set of observable variables \mathbf{z} that determine D may be correlated with the outcomes. For concreteness let us consider the special case where the outcome equation is

$$y_i = \mathbf{x}'_i \beta + \alpha D_i + u_i \quad (25.28)$$

and the error is such that

$$E[u_i | \mathbf{x}_i, D_i] = E[u_i | \mathbf{x}_i, D_i, \mathbf{z}_i].$$

In the case of **selection on observables** we may have $E[u_i | \mathbf{z}_i] \neq 0$. Let us write

$$E[y_i | \mathbf{x}_i, D_i, \mathbf{z}_i] = \mathbf{x}'_i \beta + \alpha D_i + E[u_i | \mathbf{x}_i, \mathbf{z}_i], \quad (25.29)$$

which motivates the use of a **control function estimator** based on the OLS/GLS estimation of the equation. The essential idea is to introduce into the outcome equation all observable variables that could possibly be correlated with u_i and then estimate the resulting equation by least squares. Specifically,

$$y_i = \mathbf{C}'_i \delta + \alpha D_i + \{u_i - E[u_i | D_i, \mathbf{C}_i]\}, \quad (25.30)$$

where \mathbf{C}_i includes all variables that are included in either \mathbf{x} or \mathbf{z} . The presence of \mathbf{z} in the regression expunges the possible correlation between u and \mathbf{z} . Note that if there is **selection on unobservables**, caused by common unobservable factors that affect both D and u , then we still have a potential identification problem.

This estimator was used by Heckman and Hotz (1989), who also suggested a number of variations on the basic control function estimators.

25.3.4. Selection on Unobservables

We now consider a special linear case in which the treatment participation decision is endogenous. This is an example of a well-known class of models with an “endogenous dummy variable.” The model is empirically very important when working with observational data because in such cases there are several reasons for abandoning the restrictive assumption $y_0, y_1 \perp D | \mathbf{x}$ or $E[u | \mathbf{x}, D] = 0$. The breakdown of the conditional independence assumption implies that the simple least-squares regression cannot identify the ATE, and an alternative identification strategy should be pursued.

The essential elements of the identification strategy we are about to discuss are common to other selection models. The approach involves fairly strong identifying assumptions and is fully parametric. In the special case considered, the specification is analogous to the Roy model. The conditional means in the outcome equations are taken

to be linear. The model is completed by adding a participation (binary) decision equation for D_i . Then

$$\begin{aligned} y_{1i} &= \mathbf{x}'_i \boldsymbol{\beta}_1 + u_{1i}, \\ y_{0i} &= \mathbf{x}'_i \boldsymbol{\beta}_0 + u_{0i}, \\ D_i^* &= \mathbf{z}'_i \boldsymbol{\gamma} + \varepsilon_i, \end{aligned} \quad (25.31)$$

where D_i^* is a latent variable such that

$$D_i = \begin{cases} 1 & \text{iff } D_i^* > 0, \\ 0 & \text{iff } D_i^* \leq 0, \end{cases} \quad (25.32)$$

and it is assumed that $E[u_1 | \mathbf{x}, \mathbf{z}] = E[u_0 | \mathbf{x}, \mathbf{z}] = 0$.

The variables \mathbf{z} may overlap with \mathbf{x} , but it is assumed that at least one component of \mathbf{z} , denoted z_1 , is unique and is a nontrivial determinant of D . That is, there is at least one independent source of variation in D . Hence we may refer to z_1 as an instrumental variable that is correlated with the endogenous variable D , but uncorrelated with the outcomes y_1 and y_0 , except through D .

Next it is assumed that the triple $(u_{1i}, u_{0i}, \varepsilon_i)$ is jointly multivariate normal distributed with zero mean and covariance matrix Σ given by

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{10} & \sigma_{1\varepsilon} \\ \sigma_{10} & \sigma_{00} & \sigma_{0\varepsilon} \\ \sigma_{1\varepsilon} & \sigma_{0\varepsilon} & 1 \end{bmatrix}. \quad (25.33)$$

The nonzero covariance parameters $\sigma_{1\varepsilon}$ and $\sigma_{0\varepsilon}$ reflect the endogeneity of the treatment variable. The covariance parameter σ_{10} reflects the covariance between the outcomes. Because we never observe any individual in both states, this parameter cannot be identified and is usually set to zero. The variance of ε is restricted to 1 for identification.

Given such a fully parametric specification, the model can be estimated by maximum likelihood or by a two-step semiparametric procedure. Most of these issues have been discussed in Chapter 16. Leaving aside the estimation issue, we consider measures of treatment impact.

The benefit of participation, or the ATET, is given by

$$y_{1i} - E[y_{0i} | D_i = 1] = y_{1i} - \mathbf{x}'_i \boldsymbol{\beta}_0 + \sigma_{0\varepsilon} \frac{\phi(\mathbf{z}'_i \boldsymbol{\gamma})}{(1 - \Phi(\mathbf{z}'_i \boldsymbol{\gamma}))}, \quad (25.34)$$

which may also be written as

$$E[y_{1i} | D_i = 1] - E[y_{0i} | D_i = 1] = \mathbf{x}'_i (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + (\sigma_{0\varepsilon} - \sigma_{1\varepsilon}) \frac{\phi(\mathbf{z}'_i \boldsymbol{\gamma})}{\Phi(\mathbf{z}'_i \boldsymbol{\gamma})}, \quad (25.35)$$

where the term $(\sigma_{0\varepsilon} - \sigma_{1\varepsilon}) \phi(\mathbf{z}'_i \boldsymbol{\gamma}) / \Phi(\mathbf{z}'_i \boldsymbol{\gamma})$ denotes the **selection effect**; see Section 16.7.1.

In the special case in which $\mathbf{x}'_i \boldsymbol{\beta}_0 = \mathbf{x}'_i \boldsymbol{\beta}_1$, and the treatment dummy enters the y_1 equation linearly with coefficient α , the mean impact of the program is given by

$$E[y_i | D_i = 1] - E[y_i | D_i = 0] = \alpha + \text{selection term.} \quad (25.36)$$

In some sample situations this identification strategy may be somewhat fragile. For example, the treated and untreated groups may be too different, the multivariate normality assumption may be inappropriate, or the identifying instrumental variable z_1 may be weak or possibly correlated with the error in the outcome equations.

These considerations motivate the use of alternative estimation methods presented in this chapter. These estimators generally presume selection on observables only, though Section 25.7 presents IV methods applicable when selection is additionally on unobservables.

25.4. Matching and Propensity Score Estimators

In observational studies, by definition there are no experimental controls. Therefore, there is no direct counterpart of the ATE calculated as a mean difference between the outcomes of the treated and nontreated groups. In other words, the counterfactual is not identified. As a substitute we may obtain data from a set of potential comparison units that are not necessarily drawn from the same population as the treated units, but for whom the observable characteristics, \mathbf{x} , match those of the treated units up to some selected degree of closeness.

The average outcome for the untreated matched group identifies the mean counterfactual outcome for the treated group in the absence of the treatment. This approach solves the evaluation problem by assuming that selection is unrelated to the untreated outcome, conditional on \mathbf{x} . To make the approach operational it is necessary to define the matching criteria.

25.4.1. Treatment Effect Assumptions

Matching estimators of treatment effects are useful when selection into treatment is on observables only. In addition it is assumed the **overlap (or support) condition** (25.6) applies, which means that for every \mathbf{x} there is a positive probability of nonparticipation. This ensures that we have untreated matches for the treated observations for every \mathbf{x} . Roughly speaking, the control and treated populations have comparable observed characteristics. Generating good matches means ensuring that the support condition does not fail. Further, the key assumption is that unobservable variables play no role in the treatment assignment and outcome determination.

The regression estimator imputes the missing potential outcome using the estimated regression function. If $D_i = 1$, $y_{0,i}$ is imputed using the estimated conditional regression function $\hat{\mu}_0(\mathbf{x}_i)$. Matching estimators impute the missing value using the outcomes of the “nearest neighbors”; the latter are defined by a suitable metric based on some observable characteristics. This is the basis of the analogy between a matching estimator and nonparametric methods based on the number of nearest neighbors, typically just one. The matching estimator typically approximates the difference between the means, and the variance of the estimator is estimated using many of the available results on variance of differences between the means.

Matching is a persuasive and attractive methodology if (1) we can control for a rich set of \mathbf{x} variables, (2) there are many potential controls, and (3) ATET is the parameter of interest. It also requires the “no general equilibrium effects” assumption, or **stable unit treatment value assumption (SUTVA)**, which implies that treatment does not indirectly affect untreated observations. The matching estimator avoids the assumption that the treatment effect enters the conditional mean function linearly. The initial step of establishing the nearest matches for each observation will also clarify whether comparable control observations are available. Unlike the regression approach there is less danger of extrapolation into regions outside the range of the data.

Suppose the treated cases are matched in terms of all observable covariates. In a restricted sense all differences between the treated and untreated groups are controlled. Given the outcomes y_{1i} and y_{0i} , for the treatment and control, respectively, the average treatment effect is

$$\begin{aligned} & E[y_{1i}|D_i = 1] - E[y_{0i}|D_i = 0] \\ &= E[y_{1i} - y_{0i}|D_i = 1] + \{E[y_{0i}|D_i = 1] - E[y_{0i}|D_i = 0]\}. \end{aligned} \quad (25.37)$$

The first term in the second line is the ATET, and the second term in braces is a “bias” term, which will be zero if the assignment to the treatment and control is random. In this case all that is necessary to estimate the ATET is a simple average of the differential due to treatment.

More realistically the data will involve some observed covariates \mathbf{x}_i . It is assumed that the covariates include variables that include the determinants of selection into the treatment group. If treated and nontreated groups are matched on each combination of covariates, then the treatment differential can be easily calculated for each treated case and each \mathbf{x}_i . The average of the differential over all treated individuals and all \mathbf{x}_i measures the average treatment effect. Formally, in this case (see Angrist and Krueger, 2000, p. 1316) the effect of the treatment on the treated is given by

$$\begin{aligned} E[y_{1i} - y_{0i}|D_i = 1] &= E[\{E[y_{1i}|\mathbf{x}_i, D_i = 1] - E[y_{0i}|\mathbf{x}_i, D_i = 0]\}|D_i = 1] \\ &= E[\Delta_{\mathbf{x}}|D_i = 1], \end{aligned} \quad (25.38)$$

where $\Delta_{\mathbf{x}} = E[y_{1i}|\mathbf{x}_i, D_i = 1] - E[y_{0i}|\mathbf{x}_i, D_i = 0]$.

If the \mathbf{x} variables are discrete, then the matching estimator is defined as a weighted sum

$$E[y_{1i} - y_{0i}|D_i = 1] = \sum_{\mathbf{x}} \Delta_{\mathbf{x}} \Pr[\mathbf{x}_i = \mathbf{x}|D_i = 1], \quad (25.39)$$

where $\Pr[\mathbf{x}_i = \mathbf{x}|D_i = 1]$ is the probability mass for \mathbf{x}_i , given $D_i = 1$. Angrist and Krueger (2000) discuss several aspects of this estimator.

25.4.2. Exact Matching

The procedure is to match treated and untreated individuals on their observable characteristics \mathbf{x} .

Exact matching is practicable when the vector of covariates is discrete and the sample contains many observations at each distinct value of \mathbf{x}_i .

If the covariate vector has a high dimension, or if continuous variations among some covariates are present, then exact matching between treated and nontreated groups becomes impractical. This problem motivates **inexact matching** methods. Inexact matching works by mapping \mathbf{x} into a lower dimensional measure, continuous or discrete, usually a scalar $f(\mathbf{x})$ that forms the basis of matching.

25.4.3. Propensity Scores

The method of propensity scores (Rosenbaum and Rubin, 1983) is a popular inexact matching method. Rather than match on the regressors it matches on the propensity score. Even here an exact match is not possible, so the comparison units are those whose propensity scores are sufficiently close to the treated unit.

The **propensity score**, the conditional probability of receiving treatment given \mathbf{x} , denoted $p(\mathbf{x})$, was suggested by Rosenbaum and Rubin (1983) as a matching measure. As noted in Section 25.2.5, if the data justify matching on \mathbf{x} , then matching based on propensity score is also justified.

The propensity score is usually estimated using a parametric model such as a logit or probit but can in principle be estimated using nonparametric methods.

Matching Using Propensity Scores

In the method of propensity scores one controls for the covariates by controlling for a particular function of the covariates, specifically the conditional probability of treatment, $\Pr[D_i = 1|\mathbf{x}_i]$. That is, matching is on the propensity score. This can be easily calculated by (for example) a logit regression. Moreover, one can also control for lagged variables by including them in the vector of covariates. If selection bias is eliminated by controlling for \mathbf{x}_i , it is also eliminated by controlling for the propensity score. Conditioning on the propensity score is often simpler than conditioning on a large dimensional vector \mathbf{x} . Dehejia and Wahba (1998) provide an empirical illustration based on the data previously used by Lalonde (1986).

Implementation Issues

Propensity score methods call for a good model to generate the scores. Our interest is in estimating consistently the participation probability rather than the estimates of parameters in the propensity score function. A better statistical fit for the propensity score is more likely to result from a flexible parametric or nonparametric model.

In implementing matching based on $p(\mathbf{x}_i)$ three issues are relevant: (1) whether to match with or without replacement, (2) the number of units to use in the comparison set, and (3) the choice of the matching method.

Matching without replacement means that any observation in the comparison group is matched to no more than one treated observation, that which is the closest match, whereas with replacement means that there can be multiple matches. If matching without replacement, the smallness of the comparison set would mean that the matches may not be very close in terms of $p(\mathbf{x})$, which will increase the bias of the estimator.

The issue of choosing the number of cases in the comparison set involves trade-off between bias and variance. By using a single closest match to a treated case, one reduces the bias, but by including more matched controls, the variance is reduced whereas bias increases if the additional observations are inferior matches for the treated observations. A partial solution is to use a predefined neighborhood in terms of a radius around the $p(\mathbf{x})$ of the treated observation and to exclude matches that lie outside this neighborhood. In other words, one only uses the better matches. This is called “**caliper matching**.”

Heckman et al. (1997, 1998) study the performance of matching estimators using experimental data from the Job Training Partnership Act (JTPA) combined with samples of comparison groups from three sources. Data quality plays a key role in robust estimation of treatment effects by matching methods. The results are best when the data sources and definitions are comparable for treated and nontreated groups, when the treated and nontreated come from the same labor market, and when the propensity score can be modeled using a rich set of regressors.

The issue of the sensitivity of the results to the chosen method is not amenable to a simple direct answer. The outcome may vary across different samples, depending on the extent of overlap between the treated and untreated observations. If the two groups are similar in the sense that there is a substantial overlap in their propensity scores, and if the comparison group is large, then the matches will be easier to find and matching with replacement will be feasible. If the comparison group is small and disparate from the treated group, then one may run out of satisfactory matches and be unable to use the full treated sample, this being especially likely if matching is without replacement.

The application of Dehejia and Wahba (2002) to the National Supported Work Program data provides an instructive illustration. We examine and illustrate the issues of implementation in Section 25.8 using the Dehejia and Wahba data set.

25.4.4. Measuring Treatment Effects

Denote the comparison group for the treated case i with characteristics \mathbf{x}_i as the set $A_j(\mathbf{x}) = \{j \mid \mathbf{x}_j \in c(\mathbf{x}_i)\}$, where $c(\mathbf{x}_i)$ is the characteristics neighborhood of \mathbf{x}_i . Let N_C denote the number of cases in the comparison group and let $w(i, j)$ denote the weight given to the j th case in making a comparison with the i th treated case, $\sum_j w(i, j) = 1$. Then a **general formula** for the matching ATET estimator is

$$\Delta^M = \frac{1}{N_T} \sum_{i \in \{D = 1\}} [y_{1,i} - \sum_j w(i, j)y_{0,j}], \quad (25.40)$$

where $0 < w(i, j) \leq 1$, and $\{D = 1\}$ is the set of treated individuals, and j is an element of the set of matched comparison units. Different matching estimators are generated by varying the choice of $w(i, j)$.

Matching Methods

Simple matching compares cells with exactly the same discrete \mathbf{x} ,

$$\Delta^M = \sum_k w_k [\bar{y}_{1,k} - \bar{y}_{0,k}], \quad (25.41)$$

where \bar{y}_1 is the mean outcome of the treated and \bar{y}_0 is the mean outcome of the untreated and w_k is the weight of the k th cell (i.e., the fraction of observations in cell k).

A specific example (Dehejia and Wahba, 2002) is

$$\frac{1}{N_T} \sum_i \left(y_i - \frac{1}{N_{C,i}} \sum_{j \in \{D=0\}} y_j \right), \quad (25.42)$$

where N_T is the number in the treated group ($D = 1$) and $N_{C,i}$ is the number in the comparison group corresponding to the i th observation.

The **nearest-neighbor matching** method chooses, for every treated individual i , the set $A_i(\mathbf{x}) = \{j \mid \min_j \|\mathbf{x}_i - \mathbf{x}_j\|\}$, where $\|\cdot\|$ denotes the Euclidean distance between vectors. If $w(i, j) = 1$ in (25.40) when $j \in A_i(\mathbf{x})$, and zero otherwise, then this specification uses only one case to construct the comparison group for the treated cases.

Another estimator is generated by **kernel matching** in which

$$w(i, j) = \frac{K(\mathbf{x}_j - \mathbf{x}_i)}{\sum_{j=1}^{N_{C,i}} K(\mathbf{x}_j - \mathbf{x}_i)},$$

where K is a kernel discussed in Section 9.3.

These methods share the advantage that they avoid functional form assumptions for the outcome equations in estimating ATET and can estimate it at specific values of \mathbf{x} . They have the disadvantage that if \mathbf{x} is high dimensional then the number of matches can become very small. In such cases matching based on a scalar-valued metric has attractions. **Propensity score matching**, discussed previously, is such a method.

Nearest-neighbor and kernel matching can be defined in terms of propensity scores also. For example, for nearest-neighbor matching we can define the matching set as $A_i(p(\mathbf{x})) = \{p_j \mid \min_j \|p_i - p_j\|\}$.

Stratification or interval matching is based on the idea of dividing the range of variation of the propensity score in intervals such that within each interval the treated and control units have, on the average, the same propensity score. One can use the same blocks identified by the algorithm used for computing the propensity scores. Then we compute the difference between the average outcomes of the treated and the control groups. ATET is the weighted average of these differences, with weights being determined by the distribution of the treated units across the blocks. One of the disadvantages of this method is that it discards observations in blocks in which either treated or control units are absent.

Denote by b the blocks defined over intervals of propensity score. Then the treatment effect within the b th block is defined as

$$\text{ATET}_b^S = (N_b^T)^{-1} \sum_{i \in I(b)} Y_{1i} - (N_b^C)^{-1} \sum_{j \in I(b)} Y_{0j},$$

where $I(b)$ is the set of units in block b , N_b^T is the number of treated units in the b th

block, and N_b^C is the number of control units in the b th block. Then the treatment effect based on stratification is defined as

$$\text{ATET}^S = \sum_{b=1}^B \text{ATET}_b^S \times \left[\frac{\sum_{i \in I(b)} D_i}{\sum_{i \in I(b)} D_i} \right], \quad (25.43)$$

where the term in brackets is the weight for each block given by the corresponding fraction of treated units and where B is the total number of blocks.

In **radius matching** the set $A_i(p(\mathbf{x})) = \{p_j \mid \|p_i - p_j\| < r\}$ is based on propensity scores. This means that all control cases with estimated propensity scores falling within radius r are matched to the i th treated case.

We can express ATE and ATET in terms of $p(\mathbf{x})$, assuming the overlap condition $0 < p(\mathbf{x}) < 1$. The two key results are

$$\text{ATE} = E \left[\frac{(D - p(\mathbf{x})) y}{p(\mathbf{x})(1 - p(\mathbf{x}))} \right], \quad (25.44)$$

$$\text{ATET} = E \left[\frac{(D - p(\mathbf{x})) y}{\Pr[D = 1](1 - p(\mathbf{x}))} \right]; \quad (25.45)$$

the last result is due to Dehejia (1997).

The derivations of these results are as follows:

$$\begin{aligned} y &= (1 - D)y_0 + Dy_1 \\ &= y_0 + D(y_1 - y_0), \\ (D - p(\mathbf{x}))y &= (D - p(\mathbf{x}))(y_0 + D(y_1 - y_0)) \\ &= Dy_1 - p(\mathbf{x})y_0 - Dp(\mathbf{x})y_1 + Dp(\mathbf{x})y_0 \\ &= Dy_1 - p(\mathbf{x})(1 - D)y_0 - Dp(\mathbf{x})y_1. \end{aligned} \quad (25.46)$$

Next, taking expectations and noting that $E[D|\mathbf{x}] = p(\mathbf{x})$ we get

$$\begin{aligned} E[(D - p(\mathbf{x}))y|\mathbf{x}] &= p(\mathbf{x})E[y_1] - p(\mathbf{x})(1 - p(\mathbf{x}))E[y_0] - p^2(\mathbf{x})E[y_1] \\ &= p(\mathbf{x})E[y_1 - p(\mathbf{x})y_1] - p(\mathbf{x})(1 - p(\mathbf{x}))E[y_0] \\ &= p(\mathbf{x})(1 - p(\mathbf{x}))E[y_1 - y_0], \end{aligned} \quad (25.47)$$

whence it follows that

$$\text{ATE} = E[y_1 - y_0] = E \left[\frac{(D - p(\mathbf{x})) y}{p(\mathbf{x})(1 - p(\mathbf{x}))} \right].$$

To derive the Dehejia result, we have

$$\begin{aligned} E \left[\frac{(D - p(\mathbf{x})) y}{1 - p(\mathbf{x})} \right] &= E[p(\mathbf{x})E[\mu_1(\mathbf{x}) - \mu_0(\mathbf{x})]] \\ &= E[D(y_1 - y_0)] \\ &= E[D(y_1 - y_0)|D = 1]\Pr[D = 1], \end{aligned} \quad (25.48)$$

where the first line follows from (25.47), the second line is implied by the conditional independence assumption, and the last line expresses joint expectation as a product of

marginal and conditional expectations, which implies

$$\text{ATET} = \frac{\mathbb{E}[D(y_1 - y_0)]}{\Pr[D = 1]}.$$

Using (25.44) and (25.45), consistent estimators, based on a sample of size N , are

$$\widehat{\text{ATE}} = \frac{1}{N} \sum_{i=1}^N \left[\frac{(D_i - \widehat{p}(\mathbf{x}_i)) y_i}{\widehat{p}(\mathbf{x}_i)(1 - \widehat{p}(\mathbf{x}_i))} \right], \quad (25.49)$$

$$\widehat{\text{ATET}} = \left(\frac{1}{N} \sum_{i=1}^N D_i \right)^{-1} \sum_{i=1}^N \left[\frac{1}{N} \frac{(D_i - \widehat{p}(\mathbf{x}_i)) y_i}{(1 - \widehat{p}(\mathbf{x}_i))} \right], \quad (25.50)$$

where $(N^{-1} \sum_{i=1}^N D_i)$ is a consistent estimator of $\Pr[D = 1]$.

25.4.5. Variance of ATET Based on \mathbf{x} and $p(\mathbf{x})$

Under identifiability assumptions given in Section 25.2, $\widehat{\Delta}_{\mathbf{x}}$ and $\widehat{\Delta}_{p(\mathbf{x})}$ are defined as

$$\begin{aligned} \widehat{\Delta}_{\mathbf{x}} &= \frac{1}{N_T} \sum [y_{1i} - \widehat{\mathbb{E}}[y_0 | D = 0, \mathbf{x} = \mathbf{x}_i]] \\ &= \frac{1}{N_T} \sum_{i \in \{D=1\}} [y_{1i} - \sum_{j \in A_i(\mathbf{x})} w_{ij} y_{0,j}] \end{aligned}$$

and

$$\begin{aligned} \widehat{\Delta}_{p(\mathbf{x})} &= \frac{1}{N_T} \sum [y_{1i} - \widehat{\mathbb{E}}[y_0 | D = 0, p(\mathbf{x}) = p(\mathbf{x}_i)]] \\ &= \frac{1}{N_T} \sum_{i \in \{D=1\}} [y_{1i} - \sum_{j \in A_i(p(\mathbf{x}))} w_{ij} y_{0,j}], \end{aligned}$$

where i is the subscript for the treated group, $w_{ij} = 1/N_{C,i}$, and $N_{C,i}$ is the number of cases in the comparison group for the i th treated case. Both are consistent estimators of ATET, $\mathbb{E}[y_1 - y_0 | D = 1, \mathbf{x}]$, the first based on \mathbf{x} , and the second on $p(\mathbf{x})$. A practical issue is whether adjusting for differences by propensity score is better in terms of efficiency than adjusting for differences using \mathbf{x} . Hahn (1998), Heckman et al. (1998), and others have shown that there is no unambiguous ranking of the two estimators in terms of their asymptotic variance, even if we assume that $p(\mathbf{x}_i)$ is known, which in practice will not be the case in observational studies.

Write the asymptotic variances for the two cases as follows:

$$\begin{aligned} V[\widehat{\Delta}_{\mathbf{x}}] &= E[V[y_1 | D = 1, \mathbf{x}] | D = 1] + V[E[y_1 - y_0 | D = 1, \mathbf{x}] | D = 1], \\ V[\widehat{\Delta}_{p(\mathbf{x})}] &= E[V[y_1 | D = 1, p(\mathbf{x})] | D = 1] + V[E[y_1 - y_0 | D = 1, p(\mathbf{x})] | D = 1], \end{aligned}$$

where we use the variance decomposition result given in Section A.8. In general \mathbf{x} is a better predictor than $p(\mathbf{x})$, which implies that

$$E[V[y_1 | D = 1, \mathbf{x}] | D = 1] \leq E[V[y_1 | D = 1, p(\mathbf{x})] | D = 1],$$

$$V[E[y_1 - y_0 | D = 1, \mathbf{x}] | D = 1] \geq V[E[y_1 - y_0 | D = 1, p(\mathbf{x})] | D = 1],$$

because conditioning on \mathbf{x} loses less information than conditioning on $p(\mathbf{x})$, which is a particular function of \mathbf{x} . Thus the second comparison favors the propensity score method whereas the first term comparison favors the use of \mathbf{x} over $p(\mathbf{x})$.

A helpful practical guide and computer programs for implementing the calculations of ATET are provided by Becker and Ichino (2002).

25.5. Differences-in-Differences Estimators

Chapters 2 and 3 discussed the setting of a **natural experiment** or a **quasi-experiment** in which a treatment variable undergoes a change that can be viewed as an exogenous variation in a treatment variable. The treated group can be compared to an untreated comparison group.

In some cases one has data on the treated and the comparison (control) groups both before and after the experiment. Then for the i th treated case the change in the outcome is measured by $[y_{ia} - y_{ib}|D_{ia} = 1]$ and that for the untreated group is measured by $[y_{ia} - y_{ib}|D_{ia} = 0]$. Then the *differences-in-differences measure* $[y_{ia} - y_{ib}|D_{ia} = 1] - [y_{ia} - y_{ib}|D_{ia} = 0]$, where subscripts a and b denote “after” and “before” the experiment occurs, forms the basis of an estimate of the treatment effect. This method has been introduced in Sections 3.4.2 and 22.6.

Consider a model with a fixed effect ϕ_i and a drift term δ_t , where the pre-treatment and post-treatment outcomes are given by, respectively,

$$y_{it,0} = \phi_i + \delta_t + \varepsilon_{it}, \quad (25.51)$$

$$y_{it,1} = y_{it,0} + \alpha, \quad (25.52)$$

so that

$$\begin{aligned} y_{it} &= (1 - D_{it}) y_{it,0} + D_{it} y_{it,1}, \\ &= \phi_i + \delta_t + \alpha D_{it} + \varepsilon_{it}. \end{aligned} \quad (25.53)$$

The preceding equations are for $t = a, b$; (25.51) is for the group that did not get treated and (25.52) is for the group that did get treated. Using the “before” and “after” formulation, we obtain the treatment effect

$$\begin{aligned} \alpha &= E[y_{ia} - y_{ib}|D_{ia} = 1] - E[y_{ia} - y_{ib}|D_{ia} = 0] \\ &= \{E[y_{ia}|D_{ia} = 1] - E[y_{ia}|D_{ia} = 0]\} \\ &\quad - \{E[y_{ib}|D_{ia} = 1] - E[y_{ib}|D_{ia} = 0]\}, \end{aligned} \quad (25.54)$$

where the differencing step eliminates the fixed effect α and the drift δ_t .

There are alternatives to taking differences. One alternative is to control directly for pretreatment outcome difference between treatment and control groups by regression.

For example, replace ϕ_i in (25.51) by $\mathbf{x}'_i \beta + \gamma y_{ib}$ to obtain

$$\begin{aligned} y_{ia,0} &= \mathbf{x}'_i \beta + \gamma y_{ib} + \delta_a + \varepsilon_{ia}, \\ y_{ia,1} &= \mathbf{x}'_i \beta + \gamma y_{ib} + \delta_a + \alpha D_{ia} + \varepsilon_{ia}. \end{aligned} \quad (25.55)$$

Estimates of α are constructed by regressing posttreatment outcomes on a constant, pretreatment outcomes, \mathbf{x}_i , and D_{ia} . The interpretation of α as a causal parameter relies on the assumption that after controlling for \mathbf{x} , and y_b , the treatment effect completely accounts for the posttreatment difference between the treated and control groups. The fixed effect is given a linear functional form, whereas a matching strategy can be based on weaker assumptions.

Our previous results could actually be based on quasi-experimental data. For example, compare people in one state with one law with those in a different state with a different law, and use control functions for the state effects. The new element is the addition of data before the experiment. By the assumption that the two states have the same drift term, we can use the differences-in-differences method to eliminate the state effects for which otherwise we would need control functions.

25.6. Regression Discontinuity Design

Identification of the treatment effect can sometimes be facilitated by either a natural experiment or using data generated in a quasi-experimental setting. Regression-discontinuity (RD) design is an example of a quasi-experimental design in which the probability of receiving a treatment is a discontinuous function of one or more underlying variables. Such a design can arise in circumstances where a treatment is triggered by an administrative or organizational rule. For example, Angrist and Lavy (1999) study the effect of class size on student test scores, taking advantage of the data generated under the operation of “Maimonides Rule,” which stipulates that the class be split when it reaches a specific threshold size. Van der Klaauw (2003) estimates the effect of financial aid offers on the student’s decision to attend a college, exploiting the identifying information provided by a discontinuity in the administrative rule that relates the aid to the student’s SAT score and the grade point average. These econometric applications are predated by Thistlethwaite and Campbell (1960), who analyzed the impact of student scholarships on career aspirations, exploiting the fact that the awards are made only when the student’s test score exceeds a threshold; see also Trochim (1984). The treatment here follows Van der Klaauw (2003).

25.6.1. Discontinuous Treatment Assignment Mechanism

In the case of an RD design, there is additional information about the selection rule: It is known that the treatment assignment mechanism depends (at least in part) on the value of an observed continuous variable relative to a given threshold, or cutoff score, in such a way that the corresponding probability of getting treated (propensity score)

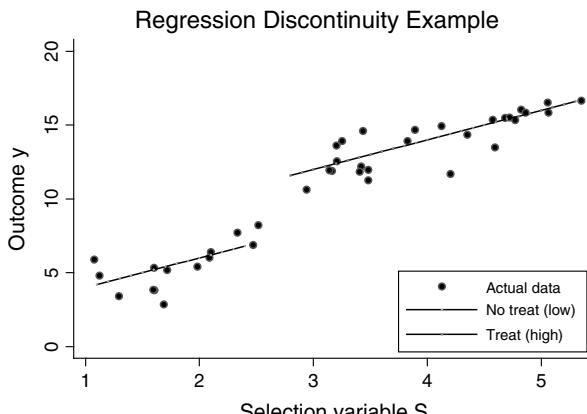


Figure 25.1: Regression-discontinuity design: example.

is a discontinuous function of this variable at the cutoff score. Figure 25.1 illustrates a sample generated by the RD design.

In the simplest RD design, called the **sharp RD design**, individuals are assigned to treatment and control groups solely on the basis of an observed continuous measure S , called the selection or assignment variable. Those falling below the distinct cutoff \bar{S} do not receive treatment and constitute the control group whereas those that are above the cutoff receive treatment ($D = 1$). That is, the treatment assignment occurs through a known and measured deterministic decision rule: $D_i = \mathbf{1}[S_i \geq \bar{S}]$. In Figure 25.2 the sharp RD design is shown as a solid line (see Van der Klaauw, 2003).

In the sharp RD design

$$E[u|D, S] = E[u|S], \quad (25.56)$$

where u denotes the error in the outcome equation. Because S is the only systematic determinant of D , S will capture any correlation between D and u .

With $D_i = D(S_i) = \mathbf{1}[S_i \geq \bar{S}]$, a dependence between D_i and u_i would make OLS an inconsistent estimator of α . As previously mentioned, one approach to estimating the treatment effect in such a case is to specify and to include the conditional mean function $E[u|D, S]$ as a “control function” in the outcome equation. Thus

$$y_i = \beta + \alpha D_i + k(S_i) + \varepsilon_i, \quad (25.57)$$

where $\varepsilon_i = y_i - E[y_i|D_i, S_i]$. If $k(S)$ is correctly specified, the regression will consistently estimate α .

If $k(S)$ is linear then α will be estimated by the distance between the two linear parallel regression lines at the cutoff point, which in this case equals the difference between the two intercepts. It is an unbiased estimate of the common treatment effect if the control function is linear.

In the more general case of varying treatment effects in which the coefficient of D represents $E[\alpha_i|\bar{S}]$, or local LATE discussed in Section 25.7.1, where $k(S)$ is a specification of $E[u|S] + (E[\alpha_i|S] - E[\alpha_i|\bar{S}])\mathbf{1}[S \geq \bar{S}]$, where $\mathbf{1}[S \geq \bar{S}] = 1$ if

the condition in parenthesis is satisfied. Incorrect specification of $k(S)$ leads to inconsistency, and hence a semiparametric specification may be tried, for example, $k(S) = \sum_{j=1}^J \eta_j S^j$, where J may be determined by a suitable method.

The variable S may be related to the outcome y , which would automatically cause (y, D) to be related even when there is no causal link between the two variables. This contrasts with random assignment that avoids such dependence.

Whereas random assignment makes treatment and control groups similar in respects other than the receipt of treatment, the sharp RD design makes them different, at least in terms of their S value. This violates the “**strong ignorability**” assumption of Rosenbaum and Rubin (1983), which also requires the overlap condition, $0 < \Pr[D = 1|S] < 1$, whereas in the sharp RD design model $\Pr[D = 1|S] \in [0, 1]$.

25.6.2. Identification and Estimation under RD Design

The main intuition is that the sample of individuals in the small neighborhood of the cutoff will be similar to a randomized experiment at the cutoff point because they have essentially the same S value. Those just below the cutoff are expected to be very similar to those just above it. A comparison of the average y value of those just above and those just below the cutoff will produce an estimate of the average treatment effect.

Increasing the interval around the cutoff will bias the estimate of the treatment effect, especially if the assignment variable was itself related to the outcome variable conditional on treatment status. If an assumption about the functional form of this relationship can be made then one can “use more observations and extrapolate from above and below the cutoff point to what a tie-breaking randomized experiment would have shown. This double extrapolation, combined with exploitation of the ‘randomized experiment’ around the cutoff point, has been the main idea behind regression-discontinuity analysis” (Van der Klaauw, 2003, p. 1258).

Observe that in this RD design,

$$\lim_{S \downarrow \bar{S}} E[y|S] - \lim_{S \uparrow \bar{S}} E[y|S] = \alpha + \lim_{S \downarrow \bar{S}} E[u|S] - \lim_{S \uparrow \bar{S}} E[u|S]. \quad (25.58)$$

A more formal way of assuming that, in the absence of treatment, individuals in a small interval around \bar{S} would have similar average outcomes is to specify the following:

Assumption A1. The conditional mean function $E[u|S]$ is continuous at \bar{S} .

Assumption A2. The mean treatment effect function $E[\alpha_i|S]$ is right continuous at \bar{S} :

$$y_i = \beta + \alpha D_i + k(S_i) + \varepsilon_i, \quad (25.59)$$

where $\varepsilon_i = y_i - E[y_i|D_i, S_i]$.

Then the result in (25.58) follows.

25.6.3. Fuzzy RD Design

Here the treatment assignment depends on the selection variable in a stochastic manner. The relation between the propensity score $\Pr[D = 1|S]$ is known to have a discontinuity at \bar{S} . A possible consequence of misassignment relative to the cutoff value is a fuzzy design, with values of S near the cutoff point appearing both in the treatment and control groups. Alternatively, the assignment may be based on additional variables observed by the treatment administrator but unobserved by the program evaluator. So relative to the sharp RD design, the **fuzzy RD design** selection depends on both observables and nonobservables. In Figure 25.2 the fuzzy RD design is shown as a dashed line.

We can still exploit the discontinuity in the selection rule to identify the treatment effect under assumption A1. If $E[u|S]$ is continuous at \bar{S} , then $\lim_{S \downarrow \bar{S}} E[y|S] - \lim_{S \uparrow \bar{S}} E[y|S] = \alpha[\lim_{S \downarrow \bar{S}} E[D|S] - \lim_{S \uparrow \bar{S}} E[D|S]]$. Therefore, the treatment effect α is identified by

$$\frac{\lim_{S \downarrow \bar{S}} E[y|S] - \lim_{S \uparrow \bar{S}} E[y|S]}{\lim_{S \downarrow \bar{S}} E[D|S] - \lim_{S \uparrow \bar{S}} E[D|S]}, \quad (25.60)$$

where the denominator $\lim_{S \downarrow \bar{S}} E[D|S] - \lim_{S \uparrow \bar{S}} E[D|S] \neq 0$ because of the known discontinuity of $E[D|S]$ at \bar{S} .

In the case of **heterogeneous treatment responses** we need additional assumptions.

Assumption A2*. The average treatment effect function $E[\alpha_i|S]$ is continuous at \bar{S} .

Assumption A3. D_i is independent of α_i conditional on S near \bar{S} :

$$y_i = \beta + \alpha E[D_i|S_i] + k(S_i) + \varepsilon_i, \quad (25.61)$$

where $\varepsilon_i = y_i - E[y_i|D_i, S_i]$ and $k(S_i)$ is a specification of $E[u_i|S_i]$.

25.6.4. A Two-Stage Estimator

If $\text{Cov}[D, u] \neq 0$, OLS regression will produce a biased estimate of α . However, the following can lead to a consistent estimator. Consider

$$y_i = \beta + \alpha E[D_i|S_i] + k(S_i) + \varepsilon_i, \quad (25.62)$$

where $\varepsilon_i = y_i - E[y_i|S_i]$ and $k(S_i)$ is a specification of $E[u_i|S_i]$.

Stage 1: Specify propensity score function for a fuzzy RD design as

$$E[D_i|S_i] = f(S_i) + \gamma 1[S_i \geq \bar{S}], \quad (25.63)$$

where $f(S_i)$ is some continuous function of S that is continuous at \bar{S} . By specifying the functional form of f (or by estimating f semi- or nonparametrically) we can estimate γ , the discontinuity in the propensity score function at \bar{S} .

Stage 2: The control function-augmented outcome equation is then estimated with D_i replaced by the first-stage estimate of $E[D_i|S_i] = \Pr[D_i = 1|S_i]$; this estimate will be discontinuous in S whereas the included control function for $k(S)$ would be

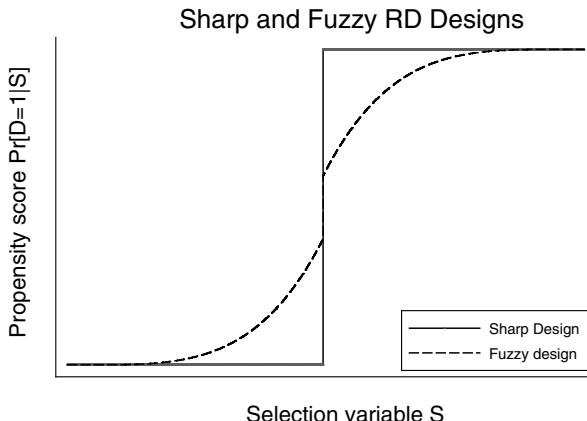


Figure 25.2: Regression Discontinuity Design; treatment assignment in sharp (solid) and fuzzy (dashed) designs.

continuous in S at \bar{S} . Under correct specification of $f(S_i)$ and $k(S_i)$ the two-stage procedure is consistent.

25.7. Instrumental Variable Methods

In recent years instrumental variable methods have been strongly advocated as an alternative to MLE and other strongly parametric methods (Angrist, Imbens, and Rubin, 1996). Such an identification strategy is attractive in models with **selection on unobservables** (see Section 25.3.4). In many applications such a model consists of a linear equation for a continuous outcome variable whose conditional mean and variance structure is specified, without any additional distributional assumptions. A leading case has a continuous outcome dependent upon a vector of regressors \mathbf{x} and a single endogenous treatment (dummy) variable (D) that represents the decision to participate in the treatment. This equation is called the participation or selection equation. In a more general setting, one may have a limited dependent or discrete outcome and there may be multiple treatment variables.

The discussion that follows overlaps with the coverage of IV estimation in several places in this book and with that of selection models. The IV approach allows us to develop another “local” variant of the ATE parameter.

25.7.1. Local ATE (LATE)

We reconsider the simple linear formulation. The outcome equation is a linear function of observable variables \mathbf{x} and a participation indicator D :

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \alpha D_i + u_i, \quad (25.64)$$

and the participation decision depends on a single variable z , referred to as an instrument,

$$D_i^* = \gamma_0 + \gamma_1 z_i + v_i, \quad (25.65)$$

where D_i^* is a latent variable with its observable counterpart D_i generated by

$$D_i = \begin{cases} 0 & \text{if } D_i^* \leq 0, \\ 1 & \text{if } D_i^* > 0. \end{cases} \quad (25.66)$$

There are two assumptions:

1. There is a variable z that appears in the equation for D that does not appear in the equation for y . It may be continuous or discrete, and in a special case it is binary. The exclusion of regressors \mathbf{x} from the participation equation is a simplification. The simultaneous presence of z in the participation equation and its exclusion from the outcome equation is referred to as the **exclusion restriction**. This model structure is familiar from Chapter 16 on selection models.
2. $\text{Cov}[z, v] = \text{Cov}[u, z] = \text{Cov}[\mathbf{x}, u] = 0$, and

$$\text{Cov}[D, z] \neq 0.$$

Together with the first assumption, this assumption implies, as previously emphasized, that y depends on z only through D , and D depends on z in a nontrivial fashion. Hence we use the notation $D(z)$ to emphasize the dependence of D on z .

Under these assumptions IV estimation of (25.64) yields consistent estimates of (β, α) . Let $z' = z + \delta$, $\delta \neq 0$. Then noting that $E[D|\mathbf{x}, D(z)] = \Pr[D(z) = 1]$ and taking expectations we obtain

$$\begin{aligned} E[y|\mathbf{x}, D(z)] &= \mathbf{x}'\beta + \alpha \Pr[D(z) = 1], \\ E[y|\mathbf{x}, D(z')] &= \mathbf{x}'\beta + \alpha \Pr[D(z') = 1], \end{aligned}$$

where, after subtraction, we have

$$E[y|\mathbf{x}, z'] - E[y|\mathbf{x}, z] = \alpha [\Pr[D(z') = 1] - \Pr[D(z) = 1]].$$

Solving the equation for α yields the expression for the **local average treatment effect** (LATE), analyzed by Imbens and Angrist (1994):

$$\begin{aligned} \alpha_{\text{LATE}} &= \frac{E[y|\mathbf{x}, z'] - E[y|\mathbf{x}, z]}{\Pr[D(z') = 1] - \Pr[D(z) = 1]}, \\ &= \frac{\int_{R(\mathbf{x})} [E[y|\mathbf{x}, z'] - E[y|\mathbf{x}, z]] dF(\mathbf{x}|\mathbf{x} \in R(\mathbf{x}))}{\int_{R(\mathbf{x})} [\Pr[D(z') = 1] - \Pr[D(z) = 1]] dF(\mathbf{x}|\mathbf{x} \in R(\mathbf{x}))}, \\ &= \frac{E[y|z'] - E[y|z]}{\Pr[D(z') = 1] - \Pr[D(z) = 1]}, \end{aligned} \quad (25.67)$$

where the second line involves averaging over \mathbf{x} , whose support is denoted by $R(\mathbf{x})$. This expression is well defined if $\Pr[D(z') = 1] - \Pr[D(z) = 1] \neq 0$. The sample analogue of this expression is the ratio of the mean difference between the treated and the nontreated divided by the change in the proportion treated owing to the change in z .

This estimator is an IV estimator. Using the results on the asymptotic normality of the IV estimator, we can obtain confidence intervals for the LATE parameter.

The qualifier “local” in LATE is justified because it measures the treatment effect on the “compliers” that are induced to participate in the treatment as a result of the change in z . LATE depends on the particular values of z used to evaluate the treatment and on the particular instrument chosen. The group of “movers” may not be representative of the whole treated population, let alone the whole population. Consequently, the LATE parameter may not be informative about the consequences of large policy changes brought about by changes in instruments different from those historically observed.

For binary instrument the LATE and the IV estimates are equivalent, as shown in Angrist et al. (1996, p. 447). If more than one instrument appears in the participation equation, as when there exist overidentifying restrictions, the LATE parameter estimated for each instrument will in general differ. However, a weighted average may be constructed.

The foregoing analysis applies when the treatment effect does not vary with individuals. If, however, the treatment effect is **heterogeneous**, then there is a potential for confounding the variation induced by z : Is the observed variation due to z -differences or α -differences? Under heterogeneity the idiosyncratic component of the treatment effect,

$$u_{i,1} = u_{i,0} + D_i(\alpha_i(\mathbf{x}_i) - \alpha(\mathbf{x}_i)),$$

is a function of $\alpha_i(\mathbf{x}_i) - \alpha(\mathbf{x}_i)$, see (25.27). Then the previous assumptions are not enough to determine ATE or ATET. A solution to this difficulty is the addition of the **monotonicity assumption** as an additional identifying condition. Essentially this says that the instrument affects participation in a monotonic fashion, so that if on average participation is more likely given $Z = w$ than given $Z = z$, then anyone who would participate given $Z = z$ must also participate given $Z = w$.

25.7.2. Relation to Other Measures

The IV estimator of α is the same as what we would estimate by using a two-stage least-squares procedure in which we first estimate the probability of receiving treatment, $E[D = 1|\mathbf{x}, \mathbf{z}]$, and then run a regression of the outcome y on \mathbf{x} and the fitted probability, assuming of course that the treatment effect is additive. Consider a special case of the IV estimator in which \mathbf{x} is a scalar and equals one, and \mathbf{z} is a scalar dummy variable that denotes eligibility to participate in the treatment; $z = 1$ implies eligibility and $z = 0$ implies noneligibility.

We can partition the population into four categories: **compliers** (C), **always-takers** (A), **never-takers** (N), and **defiers** (D). Compliers are induced to receive treatment by being eligible, always-takers will receive treatment whether or not they are eligible, never-takers refuse treatment regardless of eligibility, and defiers are contrarians who refuse treatment if eligible and take treatment if not. Assume that there are no defiers, so there are just three categories.

The **Wald estimator** of the treatment effect is defined by

$$TE_{WALD} = \frac{E[y_i|z_i = 1] - E[y_i|z_i = 0]}{E[D_i|z_i = 1] - E[D_i|z_i = 0]}, \quad (25.68)$$

whose numerator, expressed as a weighted average of treatment effects on the three categories, with weights equal to the probability of being in each category, is

$$\begin{aligned} & \Pr[C]\{E[y_i|z_i = 1, C] - E[y_i|z_i = 0, C]\} \\ & + \Pr[A]\{E[y_i|z_i = 1, A] - E[y_i|z_i = 0, A]\} \\ & + \Pr[N]\{E[y_i|z_i = 1, N] - E[y_i|z_i = 0, N]\} \\ & = \Pr[C]\{E[y_i|z_i = 1, C] - E[y_i|z_i = 0, C]\}. \end{aligned}$$

The result in the final line follows because the terms corresponding to always-takers and never-takers are identically zero. The denominator in (25.68) is the probability of compliance, $\Pr[C]$. Therefore,

$$TE_{WALD} = E[y_{1,i}|z_i = 1, C] - E[y_{0,i}|z_i = 0, C]. \quad (25.69)$$

If we compare TE_{WALD} with the LATE measure, we find that LATE is a measure of the effect of treatment on the subgroup of those at the margin of participating, denoted as compliers.

In empirical economic applications the concept of a marginal impact caused by variation in a continuous variable, measured by a partial derivative, is well entrenched and is replaced by a discrete analogue when the variation in the causal variables is discrete. Thus a **marginal treatment effect (MTE)** measure conditional on \mathbf{x} is defined as

$$MTE = \left. \frac{\partial E[y|\mathbf{x}, z]}{\partial \Pr[D = 1|\mathbf{x}, Z]} \right|_{Z=z}. \quad (25.70)$$

Heckman and Vytlacil (2002) show that ATE, ATET, and LATE are all averages of MTE taken over different subsets of the Z support, or subpopulations. ATE is the expected value of MTE over the full support of z , including where participation rate is zero or one. ATET excludes the support of z where participation does not occur. LATE is the average of MTE over an interval of z where participation rates differ.

25.7.3. IV Estimation in a Model with Heterogeneous Treatment Effect

We now consider a model that allows for selection on unobservables and heterogeneous treatment effect. The context is of a linear model with an endogenous treatment variable whose coefficient is random, see Bjorklund and Moffitt (1987). Such a model, which can be motivated by the consideration that the treatment effect is not constant across the treated, has been considered by Wooldridge (1997) and Heckman and Vytlacil (1998).

We write the model as a simultaneous equations model with the outcome variable y_1 that depends upon treatment variable y_2 . For simplicity the treatment variable y_2 is

taken to be continuous. Given instrument z and exogenous variable \mathbf{x}_i , the model is as follows:

$$y_{1,i} = (\alpha + v_i)y_{2i} + \mathbf{x}'_i \boldsymbol{\beta}_1 + \varepsilon_i \quad (25.71)$$

$$\begin{aligned} &= \alpha y_{2i} + \mathbf{x}'_i \boldsymbol{\beta}_1 + \varepsilon_i + v_i y_{2i} \\ &= v_i \bar{y}_2 + \alpha y_{2i} + \mathbf{x}'_i \boldsymbol{\beta}_1 + w_i, \end{aligned}$$

$$y_{2i} = \gamma z_i + \mathbf{x}'_i \boldsymbol{\beta}_2 + \eta_i, \quad (25.72)$$

where $w_i = \varepsilon_i + v_i(y_{2i} - \bar{y}_2)$. The marginal response of y_1 with respect to a change in y_2 is $(\alpha + v_i)$, which varies across individuals, thus permitting a **heterogeneous treatment effect**.

Suppose $E[\varepsilon_i | \mathbf{x}_i, y_{2i}] = E[v_i | \mathbf{x}_i, y_{2i}] = 0$. Then $E[\varepsilon_i + v_i y_{2i} | \mathbf{x}_i, y_{2i}] = 0$, and $V[\varepsilon_i + v_i y_{2i} | \mathbf{x}_i, y_{2i}]$ depends on \mathbf{x}_i and hence is heteroskedastic. Then the least-squares estimator of $(\alpha, \boldsymbol{\beta}_1)$ is consistent but not efficient. This follows from the assumed exogeneity of y_2 .

We next consider the case where the treatment variable is endogenous. The following assumptions are made:

$$E[\varepsilon_i | \mathbf{x}_i, z_i] = E[\eta_i | \mathbf{x}_i, z_i] = E[v_i | \mathbf{x}_i, z_i] = 0, \quad (25.73)$$

$$E[\varepsilon_i^2 | \mathbf{x}_i, z_i] = \sigma_\varepsilon^2; \quad E[v_i^2 | \mathbf{x}_i, z_i] = \sigma_v^2; \quad E[\eta_i^2 | \mathbf{x}_i, z_i] = \sigma_\eta^2. \quad (25.74)$$

Endogeneity is introduced by permitting correlation between v and η . Specifically, assume that $E[v_i | \eta_i] = \rho \eta_i$, which would hold if (v, η) were bivariate normal distributed. Under these assumptions, z is a valid instrument, and \mathbf{x} is exogenous. The exclusion of z from the y_1 equation is an identifying restriction. Therefore instrumental variable estimation of (25.71) with instruments (z, \mathbf{x}) is a natural estimator. Note, however, that the condition for consistent estimation is $E[w_i | \mathbf{x}_i, z_i] = 0$. The first component of w_i , ε_i , is uncorrelated with z_i by assumption; the second component of w_i is $v_i(y_{2i} - \bar{y}_2)$, which may at first sight seem to be correlated with z_i on which y_{2i} depends. If so, the IV estimator would be inconsistent. However, it can be shown that the IV estimator is consistent under the preceding assumptions. The key step in the argument involves showing that $E[v_i y_{2i} | z_i] = E[v_i y_{2i}]$, a result established in Wooldridge (1997) by applying the law of iterated expectations; thus,

$$\begin{aligned} E[v y_2 | z] &= E[E[v y_2 | z, \eta] | z] \quad (25.75) \\ &= E[y_2 E[v | z, \eta] | z] = E[\rho \eta y_2 | z] \\ &= \rho E[\eta^2 | z] = \rho \sigma_\eta^2 = E[v y_2]. \end{aligned}$$

Although the IV estimator is consistent under the assumptions given here, it is not efficient because of the heteroskedastic error. Hence heteroskedastic-consistent standard errors should be used. Finally, we have not tackled the issue of sensitivity of estimated treatment effects to the choice of instruments when the response to treatment is heterogeneous.

25.7.4. Endogenous Treatment in Nonlinear Models

Consider how the analyses of Sections 25.3 and 25.7 change if the outcome of a job training program were employment rather than earnings, or was duration to job placement. Alternatively, suppose that posttraining a significant proportion remains unemployed and has zero earnings, so that the sample is a mixture of those with zero and positive earnings and hence will be nonnormal. How should one extend the previous methods to handle the complications of nonlinearity and nonnormality?

The specification and estimation of nonlinear, nonnormal models of treatment and outcome with selection is an issue that occurs frequently in microeconomics. As in linear models, a major focus in such models is on the effect of an endogenous treatment variable on an economic outcome. The model specification comprises an outcome equation with a structural-causal interpretation and other equations that model the generating process of treatment variables. There are two broad approaches to this problem, a parametric one that relies on likelihood-based (including Bayesian) methods and a semiparametric one that relies on GMM or linearized IV methods.

The typical setup is illustrated by the following selected examples. In labor economics, Bingley and Walker (2001) examine the effect of duration of husbands' unemployment on wives' discrete labor supply choices. Here the treatment variable is nonnegative and possibly censored or truncated. Pitt and Rosenzweig (1990) study the effect of endogenous health status of infant children on their mothers' main daily activity; here the treatment variable is discrete and the outcome is continuous. Carrasco (2001) examines the effect of childbirth on labor force participation of women. In treatment-outcome models related to fertility, Jensen (1999) examines the effect of contraceptive use, a discrete variable, on duration between births, a limited dependent variable. Olsen and Farkas (1989) examine the effect of childbirth on the hazard of dropping out of school. In health economics, Kenkel and Terza (2001) examine the effect of physician advice (discrete) on the consumption of alcohol (continuous and nonnegative). Gowrisankaran and Town (1999) study the effect of hospital choice on the hazard of death in a hospital. In health economics the impact of health insurance choice on health care utilization, sometimes measured as an expenditure variable and sometimes as a count of number of units of some specific type of service such as doctor visits or hospital admissions, is frequently studied using the framework of a two-part model (Deb and Trivedi 1997). Terza (1998) and van Ophem (2000) model the effect of household vehicle ownership on counts of trips. Many other examples can be cited.

These models share many statistical features. First, both treatment and outcome processes are nonnormal and nonlinear: multinomial, count, discrete, or censored. Second, in each model the treatment is endogenous. Finally, investigators often have good a priori reasons for choosing particular parametric marginal models for both treatments and outcomes. However, the transition from given marginal distributions to a joint model for treatment and outcome is an essential step that is potentially problematic when nonnormal multivariate distributions are involved. Often the marginal models have no (or very restrictive) tractable multivariate counterparts (e.g., in models of counts and durations). In others, treatment and outcome are from different statistical families (e.g., treatment being a multinomial and the outcome being a hazard rate) and so analytically

tractable multivariate distributions often do not exist. Because of the specialized nature of applications in this area, this topic is not pursued any further here.

25.8. Example: The Effect of Training on Earnings

The National Supported Work (NSW) demonstration project, conducted in the 1970s, measured the impact of training on earnings by a randomized experiment that assigned some individuals to receive training (a treatment group) and others to receive no training (a control group). The effect of training could then be measured by direct comparison of sample means of posttreatment earnings for the treatment and control groups.

As was discussed in Chapter 3, randomized experiments are relatively rare in the social sciences. More often an observational sample is used with some individuals observed to receive a treatment while others do not. Comparison of the treated with the nontreated must then control for differences in observed characteristics, and possibly in unobserved characteristics.

To determine the adequacy of standard microeconometric methods for observational data, Lalonde (1986) contrasted outcomes for the NSW treated group with those for control groups drawn from two national surveys. He obtained results that differed substantially from the experimental results that contrasted the NSW treated and control groups, and he concluded that the observational methods were unreliable.

Dehejia and Wahba (1999, 2002) reanalyzed a subset of the Lalonde data using alternative matching methods, which they argued led to conclusions from observational data that were considerably closer to those from experimental data. In this section we use their data from Dehejia and Wahba (1999) to illustrate the application of methods introduced in Sections 25.2 to 25.5 that control only for selection on observables.

25.8.1. Dehejia and Wahba Data

The treated sample is one of 185 males who received training during 1976–1977. The control group consists of 2,490 male household heads under the age of 55 who are not retired, drawn from the PSID. Dehejia and Wahba (1999) call these two samples the RE74 subsample (of the NSW treated) and the PSID-1 sample (of nontreated). The treatment indicator variable D is defined as $D = 1$ if training is received (so the observation is in the treated sample) and $D = 0$ if no training was received (and the observation is in the control sample).

Summary statistics for key variables are given in Table 25.3. The treated group differs considerably from the control group, being disproportionately black (84%) with less than a high school degree (71%) and unemployed in the pre-treatment year 1975 (71%). Estimates of the effect of training should control for these differences.

25.8.2. Control Function Approach

Various estimates of the effect of training on earnings are given in Table 25.4.

The outcome of interest is posttreatment earnings, RE78. One possible measure of the effect of training is the mean difference in RE78 between NSW treated and PSID

Table 25.3. Training Impact: Sample Means in Treated and Control Samples^a

Variable	Definition	Treated	Control
AGE	Age in years	25.82	34.85
EDUC	Education in years	10.35	12.12
NODEGREE	1 if EDUC < 12	0.71	0.31
BLACK	1 if race is black	0.84	0.25
HISP	1 if Hispanic	0.06	0.03
MARR	1 if married	0.19	0.87
U74	1 if unemployed in 1974	0.60	0.10
U75	1 if unemployed in 1975	0.71	0.09
RE74	Real earnings in 1974 (in 1982 \$)	2,096	19,429
RE75	Real earnings in 1975 (in 1982 \$)	1,532	19,063
RE78	Real earnings in 1978 (in 1982 \$)	6,349	21,554
D	1 if received training (treatment)	1.00	0.00
Sample size		185	2,490

^a Data are the same as in table 1 of Dehejia and Wahba (1999). The treated group is the RE74 subsample of the NSW. The control group is the PSID-1 sample of male household heads under 55 years and not yet retired. Treatment occurred in 1976–1977.

control individuals, leading to the estimate $\$6,349 - \$21,554 = -\$15,205$. This is called a **treatment-control comparison** estimator as it mimics the analysis in an experimental setting. It can equivalently be computed as the coefficient of the treatment indicator D in OLS regression of RE78 on an intercept and D , using a combined treatment-control sample.

The large treatment estimate is misleading as it mostly reflects the difference in the types of individuals in the two samples – the control sample individuals are not good controls. This difference can be controlled for by including pretreatment characteristics as regressors, and estimating by OLS

$$\text{RE78}_i = \mathbf{x}'_i \boldsymbol{\beta} + \alpha D_i + u_i, \quad i = 1, \dots, 2675. \quad (25.76)$$

This leads to a much smaller estimated treatment effect $\hat{\alpha} = \$218$ when, following Dehejia and Wahba, the regressors \mathbf{x} are specified to be an intercept, AGE, AGESQ, EDUC, NODEGREE, BLACK, HISP, RE74, and RE75. This approach is called the **control function estimator** in Section 25.3.3.

25.8.3. Differences in Differences

A second approach is a **before-after comparison**, which looks at the difference between posttreatment earnings RE78 and pretreatment earnings RE75. Using mean earnings for the treated group leads to the difference estimate $\$6,349 - \$1,532 = \$4,817$.

This estimate may be misleading as it reflects all changes over this time period, such as an improved economy, and not just training. The **difference-in-differences estimator**, considered in Section 25.5, additionally calculates a similar quantity for the control group, $\$21,554 - \$19,063 = \$2,491$, and uses this as a measure of

Table 25.4. Training Impact: Various Estimates of Treatment Effect

Method	Definition	Estimate	St. Error ^a
Treatment-control comparison	$\overline{RE78}_{D=1} - \overline{RE78}_{D=0}$	-15,205	656
Control function estimator	$\hat{\alpha}$ from OLS regression (25.76)	218	768
Before-after comparison	$\overline{RE78}_{D=1} - \overline{RE75}_{D=1}$	4,817	625
Differences-in-differences	$\hat{\alpha}$ from OLS regression (25.77)	2,326	749
Propensity score	See Section 25.8.4	995	-

^a Standard errors for the first four estimates are computed using heteroskedastic-consistent standard errors from the appropriate OLS regression.

nontreatment related changes over time in earnings, so that the change over time solely due to treatment is \$4,817 – \$2,491 = \$2,326.

The DID estimator can be shown to be equivalent to the estimate of α in the OLS regression

$$RE_{it} = \phi + \delta D78_{it} + \gamma \alpha D_i + \alpha D78_{it} \times D_i + u_i, \quad i = 1, \dots, 2675, t = 75, 78. \quad (25.77)$$

Here $RE_{i,75}$ denotes earnings in the pretreatment period and $RE_{i,78}$ denotes earnings in the posttreatment period, so the regression is one with 5,350 earnings observations. The indicator variable $D78_{it}$ equals one in the posttreatment period, the indicator variable D_i equals one if the individual is in the treated sample, and the interaction term $D78_{it} \times D_i$ equals one for treated individuals in the posttreatment period.

More generally, the intercept ϕ in (25.77) can be replaced by $\mathbf{x}'_i \boldsymbol{\beta}$. This makes no difference in this example where regressors are time-invariant so that $\mathbf{x}_{it} = \mathbf{x}_i$. The method can be applied to repeated cross-section data (see Section 22.6.2) as it does not require that individuals in the treated and control groups be observed in both 1975 and 1978.

25.8.4. Simple Propensity Score Estimate

A third approach compares the outcome $RE78$ for a treated individual with a counterfactual prediction of $RE78$ if the same treated individual had not in fact received the treatment. The initial treatment-control estimate of \$15,205 is an oversimplified example that uses as counterfactual the average of $RE78$ in the control group (\$21,554). Better counterfactuals can be generated by specifying a regression model. For example, the regression (25.76) specifies $E[RE78|\mathbf{x}]$ to equal $\mathbf{x}'\boldsymbol{\beta} + \alpha$, if treated, with counterfactual $\mathbf{x}'\boldsymbol{\beta}$, if not treated. This places restrictions on both the effect of regressors \mathbf{x} and on the effect of treatment, which, conditional on \mathbf{x} , is assumed to be constant across individuals.

The treatment effects literature emphasizes counterfactuals that do not rely on such strong assumptions. An obvious approach is to compare treated and untreated individuals with the same value of \mathbf{x} , but in practice such **matching on regressors** is not possible if several regressors are felt to be relevant and these regressors take a number of different values.

Post-treatment Earnings against Propensity Score

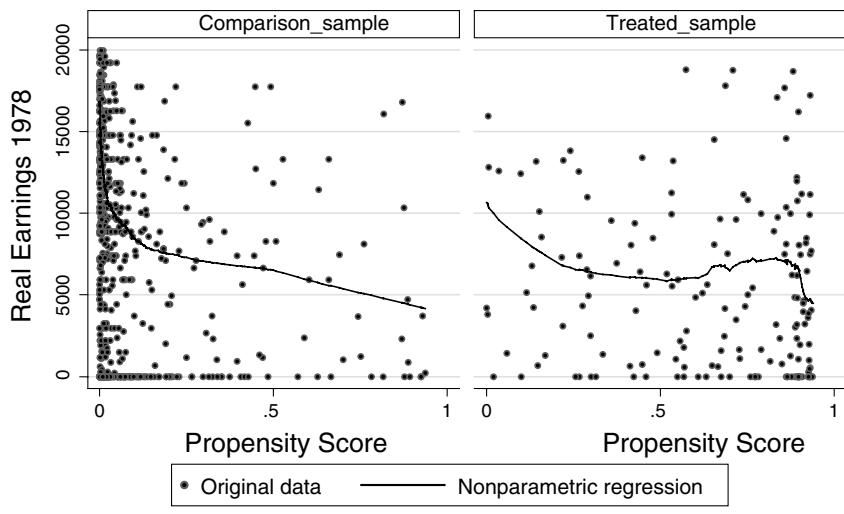


Figure 25.3: Training impact: post-treatment earnings plotted against propensity score by treatment status. Only observations with common support for the propensity score are included. Observations with earnings over \$20,000 are excluded from the scatter plot, for readability, though they are included in the nonparametric regression.

Instead, it can be sufficient, given assumptions detailed in Sections 25.3 and 25.4, to **match on the propensity score**, defined as the conditional probability of treatment $\Pr[D = 1 | \mathbf{x}]$. For this example we estimate using only data for the initial year 1975 the logit model

$$\Pr[D_i = 1 | \mathbf{x}_i] = \Lambda(\mathbf{x}'_i \boldsymbol{\beta}), \quad i = 1, \dots, 2675, \quad (25.78)$$

where, from Section 14.2, $\Lambda(z) = e^z / (1 + e^z)$, and following Dehejia and Wahba (1999) the regressors chosen are AGE, AGESQ, EDUC, EDUCSQ, NODEGREE, BLACK, HISP, MARR, RE74, RE75, RE74SQ, RE75SQ, and U74*BLACK.

Figure 25.3 plots posttreatment earnings RE78 against the propensity score, separately for the treated and control samples. Considering just the propensity score (x axis) it is clear that most observations in the control sample have very low propensity score, an expected result given the Table 25.3 data that treated individuals were disproportionately black, unemployed, low-education individuals.

Turning to the posttreatment outcome RE78 (y axis), we see that the treatment effect is estimated as the difference between a given treated individual ($D = 1$) and a control sample individual ($D = 0$) with the same (predicted) propensity score. Each panel in Figure 25.3 includes a fitted nonparametric regression of RE78 on the propensity score. The treatment effect is less than one thousand dollars over much of the range of propensity score, though it is considerably larger and positive for propensity score around 0.80.

There are many ways to implement this approach of comparing individuals with similar propensity score and then averaging over all treated individuals. One strategy

is to match a treated individual with the control-sample individual who has the closest propensity score. This approach was labeled as the nearest-neighbor matching in Section 25.4.4. A simpler strategy is to stratify data by propensity score, denoted $p(\mathbf{x})$, and let the counterfactual be the within-strata average of RE78 for the control group. For example, if a treated observation has propensity score $p(\mathbf{x}) = 0.35$ then the counterfactual may be the average of $p(\mathbf{x})$ for control group observations with $0.30 < p(\mathbf{x}) \leq 0.40$. The total effect is then $\sum_s w_s (\overline{\text{RE78}}_{s,D=1} - \overline{\text{RE78}}_{s,D=0})$, where $\overline{\text{RE78}}_{s,D=1}$ and $\overline{\text{RE78}}_{s,D=0}$ denote the strata s averages of RE78 for, respectively, the treated and control observations, and the weights w_s equal the fraction of treated observations in each stratum. A simple stratification scheme uses, say, 10 equally spaced strata with $0.0 < p(\mathbf{x}) \leq 0.1$, $0.1 < p(\mathbf{x}) \leq 0.2$, and so on. This was referred to as stratification matching in Section 25.4.4. This procedure should be restricted to cases where the propensity scores for the treated and control samples overlap, see Section 25.4.3. Here the propensity score ranges from 0.0005 to 0.9420 for the treated sample and from 0.0000 to 0.9371 for the control sample, leading to dropping of 1,423 control group individuals and 8 treated individuals. The resulting estimated total effect is \$995 given in Table 25.4.

25.8.5. Matching Using Propensity Scores

As mentioned in Section 25.4, other matching strategies include radius and kernel matching, which are also relatively easy to implement. The remainder of this chapter details these and other approaches, with emphasis on propensity score methods.

Fitted Propensity Score

The fitted propensity score is obtained using two different logit specifications, from Dehejia and Wahba (1999) and Dehejia and Wahba (2002), respectively. The specifications for propensity scores are detailed at the bottom of Table 25.6. In the only departure from Dehejia and Wahba (1999, 2002), a constant term is included in our logit models. The estimated coefficients, not presented to save space, show an expected sign pattern.

Matching Algorithms and Balancing

An important practical issue is the choice of an appropriate matching algorithm based on propensity scores that ensures that balancing condition (25.9) is met. Dehejia and Wahba (2002, p. 161) provide an algorithm that starts with a parsimonious logit model to estimate $p(\mathbf{x})$. The algorithm works as follows. The data are sorted according to $\hat{p}(\mathbf{x})$. The sample observations are stratified such that within a stratum the $\hat{p}(\mathbf{x})$ for treated and control units are close. For example, initially a rough grid with equal ranges may be used. Within each stratum the equality of means between treated and control units should be tested for each covariate. If there is no statistically significant difference, then the regressors are balanced between the treated and control groups and one can stop. If, for some stratum, there is no balance, then for the **unbalanced stratum** a finer grid is used to achieve balance. If there are many unbalanced strata, then the original logit model is reestimated with an improved specification that includes interaction and higher order terms among the regressors.

Table 25.5. *Training Impact: Distribution of Propensity Score^a for Treated and Control Units Using Dehejia and Wahba's (1999) Specification*

Minimum $\hat{p}(x)$	Treated	Untreated	Total
0.000364	9	960	969
0.10	10	56	66
0.20	14	33	47
0.40	24	22	46
0.60	33	7	40
0.80	95	8	103
Total	185	1086	1271

^a From the second row, for example, the propensity score lies between 0.10 and 0.20 for 10 treated and 56 untreated individuals.

Using the software of Becker and Ichino (2002), Dehejia and Wahba's (2002) algorithm is used to compute the propensity scores. In all of the cases noted, the propensity score computation has been restricted to the common support region by testing the **balancing property** using those observations whose propensity scores lie in the intersection of the supports of the propensity score of the treated and the control units. This restriction reduces the original sample significantly. The size of the control group drops from 2,490 units to 1,086 for the Dehejia and Wahba (2002) specification.

Table 25.5 displays the number of treated and control units in different blocks after the balancing is carried out by the procedure just outlined. The reported results differ from those of Dehejia and Wahba (2002) because the latter exclude control units from NSW-PSID composite samples not on the basis of common support region but on the basis of whether the estimated propensity score of a sample unit is less than the minimum of the estimated propensity score for the treated units. The tables show that the proportion of treated units to control units is very low for the first blocks, compared with the remaining blocks.

A similar exercise for the Dehejia and Wahba (1999) specification, not tabulated for brevity, leads to similar results. The control group has 1,146 observations. The boundary values for blocking $\hat{p}(x)$ are then 0.0006526, 0.05, 0.10, 0.20, 0.40, 0.60, and 0.80.

ATET Estimates by Matching Methods

A selection of results for various matching methods are summarized in Table 25.6. The nearest neighbor estimate of ATET for the Dehejia and Wahba (2002) specification is \$2,385, and for the Dehejia and Wahba (1999) specification, it is approximately \$560. The performance of stratification and kernel matching is also mixed, the estimates of ATET ranging from \$1,452 to \$2,156.

For comparison, Dehejia and Wahba's (2002) ATET estimates are reproduced in Table 25.7. We also note that the benchmark estimate of the treatment effect is \$1,794. It is obtained by regressing RE78 on D for the Dehejia and Wahba's (2002) version of

Table 25.6. Training Impact: Estimates of ATET

Matching Procedure	Number Treated	Number in Control	ATET	Standard Error	% of \$1794
Dehejia and Wahba (2002) specification ^a					
Nearest neighbor	185	53	2385	1209 ^c	133
Radius, $r = 0.001$	54	517	-7815	1118 ^d	-436
Radius, $r = 0.0001$	24	92	-9333	2282 ^d	-520
Radius, $r = 0.00001$	15	19	-2200	2986 ^d	-123
Stratification	185	1086	1452	1041 ^c	81
Kernel	185	1058	1309	975 ^c	73
Dehejia and Wahba (1999) specification ^b					
Nearest neighbor	185	57	560	1098 ^c	31
Radius, $r = 0.001$	57	583	-9358	997 ^d	-522
Radius, $r = 0.0001$	27	76	-7847	2066 ^d	-437
Radius, $r = 0.00001$	16	13	223	4551 ^d	12
Stratification	185	1146	2156	814 ^c	120
Kernel	185	1146	1518	890 ^c	85

^a Logit Model: $\text{Pr}[\text{treat} = 1] = h(\text{CONSTANT}, \text{AGE}, \text{AGE}^2, \text{EDU}, \text{EDU}^2, \text{MARRIED}, \text{NODEGREE}, \text{BLACK}, \text{HISPANIC}, \text{RE74}, \text{RE74}^2, \text{RE75}, \text{U74}, \text{U75}, \text{U74}^*\text{HISPANIC})$.

^b Logit Model: $\text{Pr}[\text{treat} = 1] = h(\text{CONSTANT}, \text{AGE}, \text{AGE}^2, \text{EDU}, \text{EDU}^2, \text{MARRIED}, \text{NODEGREE}, \text{BLACK}, \text{HISPANIC}, \text{RE74}, \text{RE74}^2, \text{RE75}, \text{RE75}^2, \text{RE74}^*\text{RE75}, \text{U74}^*\text{BLACK})$.

^c Bootstrapped standard errors with 200 replications.

^d Analytical standard errors.

^e ATET/1794 × 100.

the NSW sample of both participants and nonparticipants. It is clear that the reported ATET estimates in this table differ significantly from those of Dehejia and Wahba (2002), as well as from the benchmark actual experimental estimate. For the Dehejia and Wahba (2002) specification, the nearest-neighbor estimator is very close to the benchmark estimate and is even better than the results of Dehejia and Wahba (2002) in terms of reduced bias.

For stratification and kernel estimators, the bias is larger. For the radius matching estimator, this bias is worse, and gives negative estimates of the treatment effect as opposed to the positive estimates that Dehejia and Wahba (2002) found using caliper matching. The difference between our radius matching and the caliper matching of Dehejia and Wahba (2002) is that in the latter scheme, when a given treated unit does not have a match within the given caliper, matching is then done with the nearest comparison unit outside of the given caliper. In our case, if such a situation arises, we ignore treated units that have no match in the prespecified radius. This illustrates the sensitivity of the matching estimators to assumptions.

The robustness of ATET estimates across specifications can be evaluated in terms of the ratio of ATET and the benchmark estimate, given in the last column of Table 25.6. With the exception of the stratification matching estimator, the ratio varies widely over the two specifications. For example, the nearest-neighbor estimator is 133% of the benchmark estimator in the Dehejia and Wahba (2002) specification, but only 31% in

Table 25.7. Training Evaluation: Dehejia and Wahba's (2002) Estimates of ATET

Matching Procedure	ATET	Standard Error
Nearest neighbor	1890	1202
Radius, $r = 0.001$	1824	1187
Radius, $r = 0.0001$	1973	1191
Radius, $r = 0.00005$	1928	1196
Radius, $r = 0.00001$	1893	1198

the Dehejia and Wahba (1999) specification. Similarly, except for the kernel estimator, the ATET estimates are sensitive to the propensity score used.

Whether matching methods work well depends on the suitability of the propensity score model for the treatment and control groups (Dehejia and Wahba, 2002). However, there is clearly an interaction between the methods and the propensity score model.

25.9. Bibliographic Notes

Early economic applications of matching and differences-in-differences methods to program evaluation include Ashenfelter (1978) and Ashenfelter and Card (1985). Treatment evaluation is currently a very active and fast-moving area of econometrics research.

- 25.2 Angrist et al. (1996) make useful connections between the concepts and terminology in the medical and the econometrics literature.
- 25.3 Heckman and Robb (1985) consider the estimation of program impacts in a variety of data settings, in the presence of selection. See also Björklund and Moffitt (1987). Heckman and Hotz (1989) also argue strongly that one needs to subject the results to several specification tests to assess their robustness and to evaluate the impact of selection bias. For example, they suggest the use of multiple comparison groups to evaluate the sensitivity of the results based on a single control group. Most of this earlier work is parametric in approach. More recently nonparametric methods have been used also.
- 25.4 Heckman, Ichimura, and Todd (1997) and Heckman et al. (1998) study and apply matching estimators. The important result concerning conditioning on the propensity score is given in Rosenbaum and Rubin's (1983, theorem 2). Efficient estimation of ATE using estimated propensity scores is analyzed in Hirano, Imbens, and Ridder (2003). Dehejia and Wahba (2002) apply propensity score matching methods to a variant of the Lalonde (1986) data set. The experimental data are matched with observations from the CPS and the PSID. Smith and Todd (2004) reanalyze the data used by Dehejia and Wahba using a number of different variants of propensity score estimators. They highlight the biases associated with alternative propensity score estimators and emphasize the importance of high-quality data in bias minimization. Becker and Ichino (2002) provide an overview of some propensity score matching estimators. They also provide a set of STATA programs, with illustration, that can be used for estimating ATET. The February 2004 issue of the *Quarterly Journal of Economics* includes a symposium on the econometrics of matching.
- 25.6 Hahn, Todd, and Van der Klaauw (2001) analyze identification of treatment effects in the RD model under weak assumptions.

25.7 Imbens and Angrist (1994) analyze the properties of the LATE estimator. Angrist et al. (1996) discuss the use of IV methods and make a connection with the LATE measure of treatment impact. The article is followed by a lively discussion that gives a spectrum of views on the IV estimator as well as literature connections, see also Heckman (1997). Angrist (2001) discusses some simple strategies for dealing with endogenous dummies in nonlinear outcome models with nonnormal outcomes. The paper is followed by discussion and comments that analyze the pros and cons of the linearized IV approach. There is lack of consensus on the most promising among the competing approaches. Heckman, Tobias, and Vytlacil (2003) develop estimators for treatment effects within a latent variable framework. Vella and Verbeek (1999) compare the IV approach with a control function approach that includes a selection bias correction term.

Exercises

- 25–1** (Adapted from Heckman, 1996) Consider the treatment–outcome model $y = \mathbf{x}'\beta + \alpha d + \varepsilon$, where d is a binary indicator variable taking the value 1 if treatment is assigned randomly and 0 if treatment is not assigned (also randomly).
- Is randomized treatment a sufficient condition for identification of α ?
 - Is randomized treatment a sufficient condition for identification of α and β ?
- 25–2** In the previous problem randomization refers to treatment. Here we consider randomized eligibility for receiving the treatment. Now $e = 1$ means that an individual is randomly made eligible and $e = 0$ means randomly made ineligible. Show that in this case, given $\Pr[d = 1|\mathbf{x}] \neq 0$, the treatment effect is given by $E[y|e = 1, \mathbf{x}] - E[y|e = 0, \mathbf{x}] / \Pr[d = 1|\mathbf{x}]$.
- 25–3** Consider the nonlinear treatment outcome model $E[y|\mathbf{x}, d] = \exp(\mathbf{x}'\beta + \alpha d)$, where d is a binary treatment indicator. Suppose that we have available consistent estimates of (β, α) and an estimated covariance matrix $\hat{V}[\hat{\beta}, \hat{\alpha}]$. Assume that the estimator is asymptotically normal. Outline a bootstrap or a Monte Carlo algorithm for estimating the ATE parameter and its asymptotic variance given (\mathbf{x}_i, d_i) , $i = 1, \dots, N$.
- 25–4** Consider the nonlinear treatment outcome model $E[\ln y|\mathbf{x}, d] = \mathbf{x}'\beta + \alpha d$, where d is a binary treatment indicator. Suppose that we have available consistent estimates of (β, α) and an estimated covariance matrix $\hat{V}[\hat{\beta}, \hat{\alpha}]$. Suppose we are interested in estimating the ATE in terms of y rather than $\ln y$. Suggest an estimation method and discuss its consistency property.
- 25–5** In this chapter the empirical illustration used the PSID control group and the NSW treatment group. Dehejia and Wahba (2002) used two control groups. There is another control group available based on the CPS. In this exercise you will be asked to replicate some of the calculations reported here using the CPS control group in place of the PSID sample.
- Generate a table similar to Table 25.3. Compare the NSW group with the CPS controls in terms of age, ethnic composition, educational attainment, and pretreatment earnings.
 - The differences between the treatment and control groups can be viewed using the estimated propensity score, as was done in Section 25.8. Using the approach of Section 25.8.4 estimate the propensity score for the

NSW-CPS composite sample, incorporating the covariates linearly and with higher order terms, as in Dehejia and Wahba (2002). Ignoring those comparison units whose propensity scores are less than the minimum of the treated units, compare the two sets of propensity scores using a histogram. Comment on the goodness of match with comparison units in different propensity score intervals (“bins”).

- (c) Using the matching methods described and implemented in Sections 25.8.4 and 25.8.5 (especially nearest-neighbor, stratification, or interval matching, kernel matching, and radius matching), construct a table similar to Table 25.6. Comment on the estimates of ATET and compare them with those based on the PSID comparison group.

Measurement Error Models

26.1. Introduction

Problems of measurement error pervade all econometrics. In microeconomics, a common source of the measurement error problem comes from incorrect response to a survey question, incorrect coding of a correct response, and the use of a correctly measured variable as a proxy for another theoretically valid but unobserved variable (e.g., using observed income as a proxy for “normal income”). Questions that seek sensitive information may elicit partial or incorrect responses. That is, a measurement error is triggered by unobservables (or latent variables) when such variables are replaced by proxy variables.

Here are some examples. Consider the problem of testing for the presence of gender bias in a study of earnings. The obvious approach is to regress a measure of earnings on a categorical gender variable while controlling for qualifications, age, experience, and so forth. However, the most relevant variable may be an individual’s on-the-job productivity, which may not be directly observed and a proxy may be used instead. Therefore, the impact of measurement error on inferences about the gender discrimination is an important issue. Studies of individual demand for goods and services feature concepts such as “economic cost” or “full price of a service.” However, these are rarely directly measured in published data and must be constructed by the econometrician prior to model estimation. Inevitably their measurement is subject to error.

There are virtually no models discussed in this book that are protected from the problem of measurement errors. Binary outcome endogenous or exogenous variables are potentially subject to classification errors; transition and count data collected from retrospective surveys are affected by recall errors; data on relatively unambiguous variables such as hourly earnings and household expenditure are distorted by deliberate exaggerations and/or reporting errors. Unlike aggregate data where aggregation may result in some cancellation of measurement errors, for individual-level data measurement errors persist.

In the first part of this chapter we study the consequences of measurement errors and estimation strategies for remedying the consequences. Both linear and nonlinear

models are considered. Although it is more realistic to acknowledge that the problem usually occurs in combination with others, it is convenient for exposition to suppose that the only problem confronting the econometrician is measurement error.

Broadly speaking the consequence of errors of measurement is a failure to identify the parameter of interest. The issue of fixing the problem is complex. One may consider simply omitting the relevant variable in the model or substituting a proxy for the true measure. There are at least two important reasons for not doing so except in extreme cases. First, if the variable is of central interest, then omission lends to serious omitted variable bias, so one is substituting one type of problem for another, and identification is still not possible. Second, in a linear regression, using a proxy for the latent variable will have smaller asymptotic bias than simply omitting the variable from the model, provided the measurement errors are random and independent of the true regressor (McCallum, 1972). Ignoring the variable provides inferior estimates. However, using the proxy still gives inconsistent estimates even though the biases are smaller.

The essential insight underlying the solution of the measurement error problem is that to recover the parameter of the latent variable and to identify the model, it is necessary to have extraneous information in the form of additional assumptions about the measurement error or obtain additional data and to use this information after invoking plausible assumptions. This is a popular approach. However, when additional data are unavailable, an econometric model makes a good alternative.

Measurement errors have potentially very serious consequences since in many cases they lead to regression parameters becoming unidentified. For example, Card (2001) reviews empirical evidence on the coefficient of schooling on earnings and finds that the typical downward bias is of the order of 25–35%. The precise consequences of measurement errors may depend on the functional form of the model, how the errors enter the model (e.g., additively or multiplicatively), and the data structure under consideration. The solution of the problem resulting from measurement errors typically requires introduction of additional information into the model, either in the form of additional data or additional assumptions.

It is convenient to organize the discussion of measurement error models into separate sections on linear and nonlinear models, and then to consider special cases. Sections 26.2 and 26.3 are devoted to linear regression. Section 26.4 covers nonlinear regression. Section 26.5 discusses some Monte Carlo examples. Essential insights provided by linear models provide a useful basis for understanding the results for nonlinear models. In all cases clearer results are usually available for specific models.

26.2. Measurement Error in Linear Regression

Measurement error in the regressors, also called **error-in-variables**, is an important topic as it leads to inconsistency of the OLS estimator even if the measurement error has zero mean. Measurement error in the regressors is often said to lead to bias, but we use the stronger term inconsistency as the bias does not disappear as the sample size goes to infinity.

Measurement error models have a broad scope and they cover situations in which the measurement error affects the right-hand-side variables (“regressors”), or the left-hand-side variable (“outcome”), or both. Hausman (2001) refers to them as “problems from the right” and “problems from the left.” In the latter case, usually referred to as the classic errors-in-variable model, the relationship of interest is between the outcome y and covariates $(\mathbf{W}, \mathbf{X}^*)$, where \mathbf{W} is measured without error and \mathbf{X}^* is not observed but a proxy for it, denoted \mathbf{X} , is available. The question of interest is whether an estimated relation between y and (\mathbf{W}, \mathbf{X}) provides a satisfactory basis for inference regarding \mathbf{X}^* .

In the statistical literature it is conventional to distinguish between the **functional** and **structural approaches** to measurement error models. If \mathbf{X}^* denotes the true unobserved covariates, then the functional approach regards these as unknown fixed constants (parameters). In the structural approach they are treated as random variables. Carroll, Ruppert, and Stefanski (1995) further distinguish between *functional modeling* in which only minimal assumptions are made about the \mathbf{X} s, regardless of whether they are fixed or random, and *structural modeling* in which parametric assumptions are made regarding the distribution of the \mathbf{X} s. Functional measurement error models are examples of models with infinitely many nuisance parameters for which the maximum likelihood method has well-known deficiencies (discussed in the panel data chapters). This distinction is less common in the econometrics literature.

The magnitude of the inconsistency can be substantial in applications. There is a particularly extensive discussion of measurement error, and ways to control for it, in econometric studies of the determinants of individual earnings.

26.2.1. Classical Measurement Error Model

The standard measurement error model has a continuous dependent variable y that is a linear function of K true regressors \mathbf{x}^* . An additive measurement error in y may cause no problems if it is uncorrelated with the regressors because it can be absorbed into the error on the equation. If \mathbf{x}^* were observed then parameters could be consistently estimated by OLS regression of y on \mathbf{x}^* ,

$$y_i = \mathbf{x}_i^{*\top} \boldsymbol{\beta} + u_i,$$

where u_i are iid $[0, \sigma^2]$. Instead, the observed data are $\mathbf{x} \neq \mathbf{x}^*$, and y is regressed on \mathbf{x} rather than on \mathbf{x}^* . The relationship between the true and observed regressors is postulated to be

$$\mathbf{x}_i = \mathbf{x}_i^* + \mathbf{v}_i, \quad i = 1, \dots, N, \tag{26.1}$$

where the additive measurement errors are assumed to be distributed as

$$\mathbf{v}_i \sim [\mathbf{0}, \Sigma_{\mathbf{vv}}]. \tag{26.2}$$

The unobserved true regressors are assumed to have mean zero, so variables are measured as deviations from mean and to have variance matrix

$$\text{V}[\mathbf{x}_i^*] = \Sigma_{\mathbf{x}^*\mathbf{x}^*}. \tag{26.3}$$

Note that \mathbf{x} is an unbiased estimate of \mathbf{x}^* , since the measurement error \mathbf{v} is assumed to have mean zero. The measurement error is assumed to be independent of both \mathbf{x}^* and the regression error u ,

$$E[\mathbf{v}_i | \mathbf{x}_i^*] = E[\mathbf{v}_i | u_i] = 0. \quad (26.4)$$

26.2.2. Inconsistency of OLS

To consider the consequences of measurement error it is helpful to write the assumed dgp for the classical measurement error model in matrix notation as

$$\begin{aligned} \mathbf{y} &= \mathbf{X}^* \boldsymbol{\beta} + \mathbf{u}, \\ \mathbf{X} &= \mathbf{X}^* + \mathbf{V}, \end{aligned} \quad (26.5)$$

where u , the equation error, obeys the conditions $E[\mathbf{u} | \mathbf{X}^*] = \mathbf{0}$ and $E[\mathbf{u}\mathbf{u}' | \mathbf{X}^*] = \sigma^2 \mathbf{I}_N$. Substituting the second equation into the first yields

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + (\mathbf{u} - \mathbf{V}\boldsymbol{\beta}). \quad (26.6)$$

An OLS regression of \mathbf{y} on \mathbf{X} will lead to an inconsistent estimate of $\boldsymbol{\beta}$, since the error term $(\mathbf{u} - \mathbf{V}\boldsymbol{\beta})$ is correlated with the regressor \mathbf{X} through the measurement error \mathbf{V} .

Formally, we have

$$\begin{aligned} \text{plim } N^{-1} \mathbf{X}'(\mathbf{u} - \mathbf{V}\boldsymbol{\beta}) &= \text{plim } N^{-1} (\mathbf{X}^* + \mathbf{V})'(\mathbf{u} - \mathbf{V}\boldsymbol{\beta}) \\ &= -\Sigma_{vv}\boldsymbol{\beta} \\ &\neq \mathbf{0}, \end{aligned}$$

using $N^{-1} \mathbf{V}' \mathbf{V} = N^{-1} \sum_i \mathbf{v}_i \mathbf{v}_i'$ and \mathbf{v}_i iid $[\mathbf{0}, \Sigma_{vv}]$. This is the essential source of inconsistency. Now

$$\begin{aligned} \text{plim } N^{-1} \mathbf{X}' \mathbf{X} &= \text{plim } N^{-1} (\mathbf{X}^* + \mathbf{V})' (\mathbf{X}^* + \mathbf{V}) \\ &= \Sigma_{x^*x^*} + \Sigma_{vv}, \end{aligned}$$

where we have used the iid property of \mathbf{x}_i^* with mean zero and $V[\mathbf{x}_i^*] = \Sigma_{x^*x^*}$. Also,

$$\begin{aligned} \text{plim } N^{-1} \mathbf{X}' \mathbf{y} &= \text{plim } N^{-1} (\mathbf{X}^* + \mathbf{V})' (\mathbf{X}^* \boldsymbol{\beta} + \mathbf{u}) \\ &= \Sigma_{x^*x^*} \boldsymbol{\beta} \\ &\neq \mathbf{0}, \end{aligned}$$

so that, applying Slutsky's theorem (Appendix A, Theorem A.3), we get

$$\begin{aligned} \text{plim } \widehat{\boldsymbol{\beta}} &= (\text{plim } N^{-1} \mathbf{X}' \mathbf{X})^{-1} \text{plim } N^{-1} \mathbf{X}' \mathbf{y} \\ &= (\Sigma_{xx})^{-1} (\Sigma_{xx} - \Sigma_{vv}) \boldsymbol{\beta} \\ &= \boldsymbol{\beta} - (\Sigma_{x^*x^*} + \Sigma_{vv})^{-1} \Sigma_{vv} \boldsymbol{\beta}. \end{aligned} \quad (26.7)$$

Clearly, OLS is inconsistent as long as there are measurement errors and $\Sigma_{vv} \neq \mathbf{0}$.

For later reference note that if we have available a consistent estimate of Σ_{vv} , denoted \mathbf{S}_{vv} , and if $(\mathbf{X}' \mathbf{X} - \mathbf{S}_{vv})$ is positive definite, then the adjusted least-squares

estimator $\widehat{\beta}_a = (\mathbf{X}'\mathbf{X} - \mathbf{S}_{vv})^{-1}\mathbf{X}'\mathbf{y}$ can be computed. This formula can also be used to study the impact of hypothetical values of measurement error variances on the least-squares estimator.

26.2.3. Measurement Error with a Scalar Regressor

A special case of this model that routinely features in textbooks involves the case of a single true or unobserved regressor x^* with variance $\sigma_{x^*}^2$, observed value x , zero-mean measurement error v , and associated variance σ_v^2 . That is, the regression is $y = \beta x^* + u$, where $E[u|x^*] = 0$, $V[u|x^*] = \sigma_u^2$, and $Cov[v, u] = 0$, but in estimating the regression x^* is replaced by the observed variable x .

In this case, (26.7) specializes to

$$\begin{aligned}\text{plim } \widehat{\beta} &= \frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_v^2} \beta \\ &= \frac{1}{1 + \sigma_v^2/\sigma_{x^*}^2} \beta \\ &= \beta [1 - s/(1 + s)],\end{aligned}\tag{26.8}$$

where $s = \sigma_v^2/\sigma_{x^*}^2$ is often referred to as the the **noise-to-signal ratio** and the entire term $(1 + s)^{-1}$ is referred to as the **reliability ratio**. Asymptotically $\widehat{\beta}$ is downward biased toward zero to an extent that depends directly on the noise-to-signal ratio. This bias is also called **attenuation bias**. The terminology is intuitive since it suggests that a researcher's estimate of the marginal impact of change in x^* on y is attenuated by the presence of measurement error in x^* .

Note also that

$$V[y|x] = \sigma_u^2 + \frac{\beta^2 \sigma_v^2 \sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_v^2} \geq \sigma_u^2.$$

This implies that measurement errors not only cause attenuation bias but they also inflate the equation error variance. Unambiguously, a reduction in the variance of the measurement error will reduce the residual variance of the equation.

Had an intercept term been included in the bivariate regression just presented, this would bias upward the least-squares estimator of the intercept, $\bar{y} - \widehat{\beta}\bar{x}$, where (\bar{y}, \bar{x}) are sample averages that are still consistent estimates of the respective population means. Cragg (1994) suggests the term "**contamination bias**" for this effect of measurement error on another regression parameter in the equation.

As an example, consider regression of log hourly wage on years of schooling. Suppose years of schooling x^* are measured with error, and assume that the standard deviation of true years of schooling is 2 and the standard deviation of the measurement error is 1, so that $\sigma_{x^*}^2 = 4$, $\sigma_v^2 = 1$, and $\sigma_x^2 = 5$. Then $\text{plim } \widehat{\beta} = 0.8 \times \beta$. For example, an OLS estimated slope coefficient of 0.04 means that one more year of school is actually associated with a 5% higher wage rather than a 4% higher wage.

26.2.4. Extensions

In extensions and generalizations of this simple but elegant result, researchers often ask if attenuation bias is a general feature of measurement error models, and what if anything is attenuated. Although the result does not necessarily carry over to more general models, it does provide a useful benchmark. Hausman (2001) has called the attenuation bias caused by measurement error the “Iron Law of Econometrics.”

If the measurement error is assumed to be uncorrelated with the true unobserved value, the measurement error is said to be “classical.” Although convenient, this assumption may not hold. Indeed in some cases it cannot hold. For example, if x is a binary 0/1 variable, the measurement error will be a classification error. If, owing to misclassification, a 0 is measured as a 1, and vice versa, then the measurement error must be correlated with the true value.

When there is more than one regressor, let $\mathbf{X}^* = [\mathbf{x}^* \ \mathbf{Z}]$, where as in the preceding case we assume that only one regressor is observed with measurement error, that is, $x = x^* + v$. Then the expression for the least-squares estimator of the coefficient of x becomes

$$\text{plim } \hat{\beta}_{x|\mathbf{Z}} = \beta \left[1 - \frac{\sigma_v^2}{\sigma_{x^*}^2 (1 - R_{x^*, \mathbf{Z}}^2) + \sigma_v^2} \right], \quad (26.9)$$

where $R_{x^*, \mathbf{Z}}^2$ denotes the R^2 in the auxiliary regression of \mathbf{x}^* on \mathbf{Z} . The formula (26.9) is essentially the same as (26.9), provided we reinterpret the variance of x^* to mean the variance after controlling for or removing the linear influence of \mathbf{Z} on \mathbf{x}^* . Once again the inconsistency of the least-squares estimator is toward zero, though by a smaller multiple of β than in the single regressor case. The coefficients of the regressors measured without error are also inconsistent, in a direction that depends on $\Sigma_{\mathbf{x}^*\mathbf{x}^*}$ (Levi, 1973). This effect can once again be thought of as contamination bias. The attenuation bias that is demonstrated in these special cases depends critically on the assumption of additive measurement errors.

When more than one regressor is measured with error general results on the direction of the inconsistency are no longer available, though in any given problem they can be determined given knowledge of $\Sigma_{\mathbf{x}^*\mathbf{x}^*}$ and Σ_{vv} . Most studies consider measurement error in only one regressor, in which case the inconsistency is toward zero. The intuition from the foregoing examples is that if the measurement errors on different regressors are independent, then each source will contribute to the attenuation bias of its “own” coefficient, and all will contribute to the inflation bias of the conditional variance. Cragg (1994) analyzes a multiple regression model with measurement errors and shows the interactions among biases from different sources.

26.2.5. Measurement Error in Linear Panel Models

The effects of measurement error in regressors are compounded when panel data are used.

26.3. IDENTIFICATION STRATEGIES

Assume a pooled panel model $y_{it} = \beta x_{it}^* + u_{it}$, where we observe $x_{it} = x_{it}^* + v_{it}$, and a scalar regressor is assumed for simplicity. The preceding results still hold if we estimate a single cross section. However, if we estimate using more than one year of data for each individual we need to adapt the previous results, since the regressor x_{it}^* will most likely be positively correlated, rather than independent over t for given i . For example, if we do the first-differences regression

$$\begin{aligned}\Delta y_{it} &= \beta \Delta x_{it}^* + \Delta u_{it} \\ &= \beta \Delta x_{it} + \Delta u_{it} - \beta \Delta v_{it}\end{aligned}$$

(see Section 21.6) and define $\rho = \text{Cor}[x_{it}^*, x_{i,t-1}^*]$, then

$$\begin{aligned}\text{plim } \widehat{\beta} &= \beta + \left(\text{plim} \frac{1}{N} \sum_{i=1}^N (\Delta x_{it})^2 \right)^{-1} \left(\text{plim} \frac{1}{N} \sum_{i=1}^N (\Delta x_{it} \Delta u_{it} - \beta \Delta x_{it} \Delta v_{it}) \right) \\ &= \beta - \frac{2\beta\sigma_v^2}{2(1-\rho)\sigma_{x^*}^2 + 2\sigma_v^2} \\ &= \beta - \frac{\beta\sigma_v^2}{(1-\rho)\sigma_{x^*}^2 + \sigma_v^2},\end{aligned}$$

using $V[\Delta v_{it}] = 2V[v_{it}]$ and $V[\Delta x_{it}^*] = 2(1-\rho)V[x_{it}^*]$.

The inconsistency is larger than in the cross-section case if $\rho > 0$. Moreover, as $\rho \rightarrow 1$, as can happen with panel data, the inconsistency becomes very large. This inconsistency can be decreased by using differences that are $m > 1$ lags apart because $\text{Cor}[x_{it}^*, x_{i,t-m}^*]$ will be decreasing in m .

26.3. Identification Strategies

It is conventional to say that without additional assumptions the errors-in-variables model is not identified. This statement can be interpreted as follows in the context of the special case of the bivariate model. An estimated value of $\widehat{\beta}$, or more precisely its probability limit, is consistent with infinitely many different combinations of β and s , the noise-to-signal ratio. If, however, additional assumptions or information can be brought to bear on the problem, it may be possible to rule out some combinations of the underlying parameters that are consistent with the observed data distribution. If the additional restrictions are just sufficient to obtain a unique solution, the model is said to be exactly identified. If the additional restrictions are more than sufficient to uniquely identify the model parameters, the model is said to be overidentified.

A general identification strategy for the measurement error model is to obtain bounds rather than point estimates of the parameters of interest if there is no further a priori information or data. If additional data and/or information about measurement error are available then additional identification strategies, such as instrumental variables

estimation or identification through moment restrictions, become feasible. Additional information about the measurement error is a broad concept that includes one of the oldest identification strategies, one using instrumental variables that link the true unobserved variables to their observable counterparts. For example, additional information may yield a consistent estimator for the attenuation factor, $\sigma_{x^*}^2 / (\sigma_{x^*}^2 + \sigma_v^2)$, making it possible to adjust the inconsistent estimate for the bias. Finally, replicated data or validation data may be available, and these can yield useful information about the moments of measurement error. These possibilities are analyzed in the following.

26.3.1. Setting Bounds on Regression Parameters

Reconsider the multiple regression problem of Section 26.2. The model given there is subject to the requirement that the variances $\Sigma_{x^*x^*}$, Σ_{vv} , and σ^2 must be positive semidefinite. This together with the orthogonality conditions of estimation can be used to place some bounds on the region in which the coefficients must lie. Klepper and Leamer (1984) and Wansbeek and Meijer (2000) consider the problem in some generality. A more accessible special case of the bounds approach is the reverse regression approach considered next.

Reverse Regression

In a simple bivariate regression model with variables (y, x) , **direct regression** refers to the regression of y on x , whereas **reverse regression** refers to the regression of x on y . In the general multivariate regression case with K covariates, there is only one direct regression but there are K reverse regressions. Each reverse regression has a mismeasured exogenous variable on the left-hand side and the remaining exogenous variables and y on the right-hand side. In the bivariate regression case with measurement errors, it is easy to show that the estimated slope coefficients from the direct and reverse regressions place lower and upper bounds on the value of the true slope coefficient. This is a potentially useful result in analyzing the effects of measurement errors. Leamer (1978) provides an excellent discussion of the logic of reverse regression.

First, we consider the logic of reverse regression by reference to a simple bivariate regression model with measurement errors:

$$\begin{aligned} y &= \beta x^* + u, \\ x &= x^* + v, \end{aligned} \tag{26.10}$$

where u is the regression error and v is the measurement error that accounts for the difference in the observed variable x and the error-free measure x^* that enters the regression. We will assume that $u \sim \mathcal{N}[0, \sigma_u^2]$ and $v \sim \mathcal{N}[0, \sigma_v^2]$.

26.3. IDENTIFICATION STRATEGIES

Following the structural approach of Solari (1969) (and Leamer, 1978), treat x^* as unknown parameters in the likelihood function. The joint likelihood given data (\mathbf{y}, \mathbf{x}) is

$$L(\mathbf{x}^*, \beta, \sigma_u^2, \sigma_v^2) \propto (\sigma_u^2)^{-N/2} \exp \left[-\frac{1}{2\sigma_u^2} (\mathbf{y} - \beta\mathbf{x})' (\mathbf{y} - \beta\mathbf{x}) \right] \\ \times (\sigma_v^2)^{-N/2} \exp \left[-\frac{1}{2\sigma_v^2} (\mathbf{x}^* - \mathbf{x})' (\mathbf{x}^* - \mathbf{x}) \right]. \quad (26.11)$$

This function is not defined at points that satisfy the conditions $\sigma_u^2 = 0$ and $\mathbf{x}^* = \mathbf{x}$, or the conditions $\sigma_v^2 = 0$ and $\mathbf{y} = \beta\mathbf{x}^*$. If we simply minimize the well-defined parts of this likelihood subject to the constraints we get two scalar regression parameters, $\hat{\beta}_D = \mathbf{y}'\mathbf{x}/\mathbf{x}'\mathbf{x}$ for the direct regression and $\hat{\beta}_R = \mathbf{y}'\mathbf{x}/\mathbf{y}'\mathbf{y}$ for the reverse regression. To aid intuition, notice that if \mathbf{x} is measured without error then \mathbf{y} is stochastic and \mathbf{x} is not, so direct regression has a meaningful conditional expectation interpretation, and if only \mathbf{x} is stochastic (measured with error), then the conditional expectation $E[\mathbf{x}|\mathbf{y}]$ is meaningful, because the two-equation system then reduces to $x = (1/\beta)y - u/\beta + v$. That is, the reverse regression produces the least-squares estimate $\widehat{(1/\beta)}$. It is straightforward to verify that

$$r_{xy}^2 \hat{\beta}_R = \hat{\beta}_D, \quad (26.12) \\ \hat{\beta}_D < \beta < \hat{\beta}_R,$$

where r_{xy}^2 is the simple squared correlation between x and y ; the bounds indicate that $\hat{\beta}_D$ is a downward biased estimate of β and $\hat{\beta}_R$ is an upward biased estimate. Note that these bounds can be very broad in microeconomic data where $r_{xy}^2 < 0.5$ is almost always the case and even $r_{xy}^2 < 0.1$ is quite common.

Leamer (1978) considers the model in which (y, x^*) has a bivariate normal distribution with mean $(\beta\bar{x}^*, \bar{x}^*)$ and covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_u^2 + \beta^2\sigma_{x^*}^2 & \beta\sigma_{x^*}^2 \\ \beta\sigma_{x^*}^2 & \sigma_{x^*}^2 + \sigma_v^2 \end{bmatrix}. \quad (26.13)$$

He shows (Leamer, 1978, pp. 239–240) that the likelihood function for this model attains its maximum at any value of β between the direct regression estimator $\hat{\beta}_D$ and the reverse regression estimator $\hat{\beta}_R$.

The foregoing analysis suggests that even though β is not identified, consistent bounds can be placed on its value. This is a potentially useful application of **bounds identification**. The result can be extended in a straightforward manner to the case of multiple regression in which only one regressor is measured with error (Bollinger, 2003). Klepper and Leamer (1984) consider an extension to the multiple regression case of K regressors, all of which are measured with error. There is one direct regression and K reverse regressions. After estimation each reverse fitted regression is renormalized with a unit coefficient for y on the left-hand side. Then $\hat{\beta}_D$ is the estimated vector from the direct regression, and $\hat{\beta}_{R,j}$ ($j = 1, \dots, K$) is the vector from the j th reverse regression. By the results of Klepper and Leamer (1984), if the direct

and reverse regression coefficient vectors are all in the same orthant then the set of feasible values of β is the convex hull of the direct and reverse regressions; that is, $\beta \in \{\widehat{\beta} | \widehat{\beta} = \lambda_D \widehat{\beta}_D + \lambda_1 \widehat{\beta}_{R,1} + \dots + \lambda_k \widehat{\beta}_{R,k}\}$, where the λ -weights are nonnegative and sum to one. The smallest coefficient in the direct and reverse regression vectors is the lower bound, and the largest coefficient is the upper bound. These bounds do not exist if the coefficient changes its sign.

In addition to the work of Klepper and Leamer (1984), there are several studies that use these ideas in applied contexts. Greene (1983) and Goldberger (1984) apply reverse regression to measurement of salary discrimination. Bollinger (2003) analyzes measurement of the black–white wage gap in a model of wages and human capital. Bollinger (1996) applies the bounds approach to the case of regression on a categorical dummy variable in which observation categories are misclassified.

26.3.2. Identification Using Instrumental Variables

One solution to the identification problem is to introduce one or more moment restrictions that constitute further identifying information. A moment restriction typically states that there is available an instrumental variable that is correlated with, or causally related to, the variable that is measured with error. Moreover, this variable is uncorrelated with, or causally unconnected with, the outcome that is being modeled. Adding this restriction to the original model helps in principle to solve the identification problem.

Historically, the IV estimator was suggested as a potential solution for the measurement error problem in linear models (Reiersøl, 1941; Durbin, 1954). The IV approach is similarly motivated when one or more variables on the right-hand side are endogenous and hence correlated with the regression error. The linear simultaneous equation model and the linear measurement error model are isomorphic and hence the use of IV-type estimators in the context of measurement errors is natural.

Reconsidering the linear IV model of Sections 4.8 and 6.4, where $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$ and $E[\mathbf{u}|\mathbf{X}] \neq \mathbf{0}$, we can use the 2SLS estimator if a valid set of instruments \mathbf{Z} , $\dim[\mathbf{Z}] \geq \dim[\mathbf{X}]$ is available.

One can test for the presence of measurement error using a Hausman test of endogeneity of regressors, see Section 8.3. Several variants of the test are available, and one variant was given in Section 8.4.

A major problem in implementing the IV estimator lies in the practical difficulty of finding valid instruments. Good instruments have two properties: zero correlation with equation errors (for consistency) and high correlation with variables being instrumented (for efficiency). Such instruments are not typically easy to find. Although ideally one should explicitly derive valid instruments from detailed specification of relationships between regressors and covariates, in practice ad hoc methods are common. Unlike the full system specification approach, the ad hoc method is simpler and less demanding. Notice that the conditions for the validity of instruments do not create an automatic procedure for selecting one. These technical conditions could be satisfied by a variable that is causally unconnected with the phenomenon under study. One has to think of a variable that correlates strongly with the regressor(s) and is uncorrelated

with the equation error. A number of interesting applications of this idea are available in the literature; see, for example, Angrist (1990). If selected, the use of such an instrumental variable may be controversial and puzzling.

We consider several possible instruments for the cross-section regression of earnings on schooling example. First, if data are available on siblings then the schooling level of a sibling may be used as an instrument, since the education levels of siblings are likely to be correlated. Consistency of the IV estimate then requires no correlation between the measurement error v and any measurement error in schooling of the sibling. Second, more generally other variables related to schooling such as parents' educational level or income may be used. Casting a broader net, however, runs the risk of leading to instruments that are only weakly correlated with x , leading to imprecision and possible poor finite-sample properties of the IV estimator. Third, more than one question on schooling level may have been asked in the survey, or schooling level may be available from surveys in other years if data are from a panel study. Such instruments are likely to be highly correlated with x , but the assumption of no correlation between measurement errors in x and z may be more difficult to believe in this example.

Lagged variables are frequently used as instruments, but these too will have measurement errors, so the approach is minimally satisfactory only if serial correlation in measurement error is not a problem.

The effect of measurement error can be large in the panel context. Since panel data provide measures of x_{it}^* in multiple periods, instrumental variables estimation can be used to provide consistent parameter estimates assuming uncorrelated measurement errors across the time periods. See Hsiao (1986, pp. 63–65).

26.3.3. Identification via Additional Moment Restrictions

Distributional assumptions about the equation and measurement errors (u , v) can secure identification. There is one important case in which the identification is aided instead by information or assumption about the distribution of the unobserved true value of the mismeasured variable. The assumption of joint multivariate normality of (y, x, x^*) , together with the assumption that the measurement error v and equation error u are, respectively, iid $\mathcal{N}[0, \sigma_v^2]$ and iid $\mathcal{N}[0, \sigma_u^2]$, are not sufficient to identify the measurement error model. However, the assumption that the first four moments of (x^*, u, v) exist and that the third moments of each and the third cross-moments are not zero, indicating a departure from normality, is sufficient to secure identification, as we now demonstrate.

Let us reconsider the model (26.10)

$$\begin{aligned} y &= \beta x^* + u, \\ x &= x^* + v, \end{aligned}$$

whose reduced form $y = \beta x + \varepsilon$, where $\varepsilon = u - \beta v$, is to be estimated by an instrumental variables procedure. However, we now add a new piece of information: that the distribution of x^* is nonnormal in the sense that it is both skewed and has nonnormal (excess) kurtosis Cragg (1997) Dagenais and Dagenais, 1997; Wansbeek and Meijer,

2000). These assumptions imply the following six conditions:

$$\begin{aligned} E[(xy)x] &= \beta E[x^{*3}], & E[(xy)u] &= 0, \\ E[(x^2)x] &= E[x^{*3}] + E[v^3], & E[(x^2)u] &= -\beta E[v^3], \\ E[(y^2)x] &= \beta^2 E[x^{*3}], & E[(y^2)u] &= -\beta E[\varepsilon^3]. \end{aligned}$$

The first row implies that the product variable $x_i y_i$ is a valid instrument if $E[x_i^{*3}] \neq 0$. The second row implies that x_i^2 is a valid instrument if $E[x_i^{*3}] \neq 0$, but $E[v_i^3] = 0$; that is, if x^* is nonnormal but v has a symmetric distribution. Indeed, the greater the skewness the better is the instrument. However, because x^* is unobservable, any inferences about it will need to be based on x itself. The last row implies that y_i^2 is a valid instrument if the third moment of x^* is nonzero but the third moment of ε is zero.

Given these moment conditions, the IV approach can be applied to consistently estimate the model parameters. This example illustrates how additional moment assumptions can help generate useful instruments even when no data other than (y_i, x_i) are available.

26.3.4. Replicated Data

An alternative solution is possible if the measurement error variances can be estimated. The basic idea here is that we can adjust the sample second-moment matrix $\mathbf{X}'\mathbf{X}$ of the regressors by an amount that depends on the variance and covariances of measurement errors. Notice that we do not try to adjust the observations themselves. Instead, the sample moments are adjusted because the estimator is a function of those sample moments. This key idea generalizes to more complex models also.

When the measurement error variance Σ_{vv} is known, a consistent estimate of β can be obtained using

$$\tilde{\beta} = (\mathbf{X}'\mathbf{X} - N\Sigma_{vv})^{-1}\mathbf{X}'\mathbf{y}, \quad (26.14)$$

where N is the sample size. This is consistent since

$$\begin{aligned} \tilde{\beta} &= \text{plim}(N^{-1}\mathbf{X}'\mathbf{X} - \Sigma_{vv})^{-1} \text{plim} N^{-1}\mathbf{X}'\mathbf{y} \\ &= (\Sigma_{x^*x^*} + \Sigma_{vv} - \Sigma_{vv})^{-1} \Sigma_{x^*x^*} \beta \\ &= \beta, \end{aligned}$$

where $\text{plim } N^{-1}\mathbf{X}'\mathbf{y} = \Sigma_{x^*x^*} \beta$ is obtained using $\mathbf{X} = \mathbf{X}^* + \mathbf{V}$ and $\mathbf{y} = \mathbf{X}\beta + (\mathbf{u} - \mathbf{V}\beta)$. For a detailed account of ways to estimate Σ_{vv} in a substantive application, see Krashinsky (2004).

Data replication is a situation in which an unbiased estimate of the unobserved \mathbf{X}^* is available. Suppose that the measurement error is additive and we have an observable \mathbf{X} :

$$\mathbf{X} = \mathbf{X}^* + \mathbf{V}.$$

If \mathbf{X} is an unbiased estimate of \mathbf{X}^* , then $E[\mathbf{V}|\mathbf{X}] = \mathbf{0}$. If data are replicated, this simply means that we have at least two measurements available on \mathbf{X} . It also means that

with multiple measurements we can obtain estimates of the moments of \mathbf{V} , assuming the measurement errors for multiple measures are uncorrelated.

Suppose there are two scalar measurements (replicates) $X_{(1)}$ and $X_{(2)}$, such that $X_{(j)} = X^* + V_{(j)}$, $j = 1, 2$. Then $V[V_{(j)}] = E[X_{(j)}^2] - E[X_{(1)}X_{(2)}]$, which can be estimated by the sample average $N^{-1} \sum_i [X_{(j),i}^2 - X_{(1),i}X_{(2),i}]$. Then the regression parameters can be estimated using Equation (26.14).

For example, suppose we wish to predict grade point average (GPA) in the first year of college using performance on the SAT exam taken in high school. It is known that observed SAT scores for a given person vary across different takes of the exam. Let x^* denote the true SAT score, and let x_1 and x_2 denote the observed SAT score on two separate SAT exams. Then $x_1 = x^* + v_1$, $x_2 = x^* + v_2$, and it is assumed that v_1 and v_2 are independent with equal variance σ_v^2 . It follows that $\text{Cov}[x_1, x_2] = \sigma_{x^*}^2$, $V[x_1] = V[x_2] = \sigma_{x^*}^2 + \sigma_v^2$, and $\text{Cor}^2[x_1, x_2] = \sigma_{x^*}^2 / (\sigma_{x^*}^2 + \sigma_v^2)$. Studies find the tests to have a reliability of 0.9, which means that the correlation from one test to the next is 0.9 and the squared correlation is 0.81. Thus $\sigma_{x^*}^2 / (\sigma_{x^*}^2 + \sigma_v^2) = 0.81$. It follows from (26.8) that $\text{plim } \hat{\beta} = 0.81 \times \beta$, so that because of measurement error SAT scores are as stronger a predictor of first-year college GPA than OLS regression suggests.

26.3.5. Validation Data

Sometimes a validation sample is also collected as an additional check on the original responses. Although the **validation sample** pertains to the population of interest, it may come from a different independent source. For example, patients may respond to a questionnaire about medical services received, and providers of services may respond to a validation survey. Another example is that of employees who may provide some information about an event, and the information may be validated by the same information obtained from the employers. A leading example in economics is the PSID validation study of Bound et al. (1994).

Let \mathbf{X} be an $N \times K$ matrix of observations on regressors measured with error, and let \mathbf{X}_v be an $M \times K$ matrix of validation data. We can use validation data by regressing the columns of \mathbf{X}_v on \mathbf{X} , and generating “predicted” values $\mathbf{X}[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{X}_v$ that replace the error-contaminated matrix \mathbf{X} . For nonlinear models more complex procedures are used, see Lee and Sepanski (1995).

The use of generated regressors that are substituted into the regression of interest can be a practical useful strategy if the predictions come from a well-fitting regression. Generated regressors are estimates of the true values and hence subject to estimation uncertainty. As such this uncertainty should be taken into account in estimating the sampling variance of the regression coefficients. The relevant theory was covered in Section 6.8.

26.4. Measurement Errors in Nonlinear Models

Nonlinear models, as should by now be abundantly clear, comprise a bewildering array of models. Obtaining general results, such as attenuation bias, that apply to a broad

class of models poses a major challenge. Not unusually, general results are obtained under simplifying assumptions, whereas more specific results can pay more attention to complexity and specificity of particular data situations. Therefore, it is not surprising that the development of this topic in the literature has produced many procedures and approaches that are specific to particular models. For example, in dealing with binary outcome models with left-hand-side measurement error it is natural to focus on the problem of misclassification; in dealing with count models also with left-hand-side measurement error it is equally natural to focus on the issues of under- and over-reporting. Motivated by this difficulty, Hsiao (1992) recommends shifting attention from providing solutions for a general model to a specific type of question. In covering model-specific results, there is a danger of being compendious and of losing sight of general results. We therefore begin with some selected general results.

26.4.1. Identification through Instrumental Variables

A general technique in the linear errors-in-variables model is the instrumental variables method. For the nonlinear (in regressors) regression model, Y. Amemiya (1985) showed that the IV estimator is generally inconsistent, being consistent only under the assumption of a shrinking error variance–covariance matrix.

A simple exposition of the aforementioned point is based on the regression equation

$$y = \beta_0 + \beta_1 f(x^*) + \varepsilon, \quad (26.15)$$

where $f(x^*)$ is a smooth, differentiable, and bounded function of an error-free scalar regressor x^* . The observed variable $x = x^* + v$, where v is a measurement error. Substituting for x^* and taking a Taylor expansion of $f(x - v)$ around x yields

$$y = \beta_0 + \beta_1 f(x) + \varepsilon - \beta_1 f^{(1)}(x)v + \beta_1 \sum_{j=2}^{\infty} f^{(j)}(x)(-v)^j/j!, \quad (26.16)$$

where $f^{(j)}(\cdot)$ denotes the j th derivative of $f(\cdot)$. Consider the quadratic case $f(x) = x^2 + \gamma x$, so $f^{(1)}(x) = 2x + \gamma$, $f^{(2)}(x) = 2$, and $f^{(j)}(x) = 0$, $j > 2$. Then

$$\begin{aligned} y &= \beta_0 + \beta_1 (x^2 + \gamma x) + \varepsilon - \beta_1 (2x + \gamma) v + \beta_1 2v^2/2 \\ &= \beta_0 + \beta_1 x^2 + \beta_1 \gamma x + (\varepsilon - \beta_1 xv - \beta_1 \gamma v + \beta_1 v^2), \end{aligned} \quad (26.17)$$

so valid instrumental variables should be correlated with x^2 and x , but uncorrelated with $u = (\varepsilon - \beta_1 xv + \beta_1 \gamma v + \beta_1 v^2)$. Clearly it is not enough that v and ε are individually uncorrelated with the instruments. This means that the instrumental variable for $f(x)$ has to satisfy more stringent properties than in the linear case.

More generally, Y. Amemiya has shown, using Taylor approximation, that the instrumental variable does not yield consistent estimates for nonlinear errors-in-variables models because the residual term involves both measurement error and an observed error-contaminated variable. Therefore it is not possible to find an instrumental variable that is highly correlated with the observed variable but uncorrelated with residual term. Furthermore, from a practical viewpoint, it is not easy to verify the validity of

an instrumental variable in estimation because of limited information about the latent variable (x^*) and measurement error.

26.4.2. Identification Using Replicated Data

Faced with the difficulty of implementing an IV-type estimation method, there are two alternatives.

The first is to make very strong distributional assumptions about the conditional distribution of the unobserved x^* given the observed x . Such assumptions, augmented by other technical conditions, make it possible to identify the parameters of the model. This approach has been followed by Y. Amemiya (1985) and Hsiao (1989), among others.

A second approach is to consider the possibility of having a large number of measurements of each unobserved x^* , denoted $x^{(j)}$. Then the average of the replicated measures for each x^* is substituted for the unobserved regressor. Consistent estimation of the nonlinear regression then follows because the covariance matrix of measurement errors shrinks to zero as the number of replicates grows; see Y. Amemiya (1985). Unfortunately, such a scenario is rarely encountered in econometrics.

Since there does not exist common structural information in nonlinear measurement error models that can be used to identify and estimate regression models, we consider some specific nonlinear regression models.

Hausman, Newey, and Powell (1995) analyze polynomial Engel curves using Consumer Expenditure Survey data. Their polynomial function is linear in parameters. They prove that, under regularity conditions, both an instrumental variable and an additional measurement can be used to obtain consistent and asymptotically normally distributed estimates. In this application, an adjacent quarter is treated as a replication and an instrumental variable. They further propose that a general nonlinear function can be approximated by a polynomial function. However, they admit that the IV method cannot be implemented in this case and an additional measure of true regressors is needed.

Li (2002) proposes a general two-stage approach to the nonlinear errors-in-variables problem, which relies on repeated measurements. In the first stage, based on empirical characteristic functions and the inverse Fourier transform, a nonparametric estimator is obtained for the conditional density of the latent variables. With this estimator available, a semiparametric nonlinear least-squares estimator is constructed using a minimum distance criterion. He establishes the estimator's consistency. This estimator is also robust in the sense that it does not require any knowledge of the functional form of the latent variables. Li's approach can be applied to any nonlinear errors-in-variables situation if replicated measurements are available. However, the asymptotic distribution of the estimator has not been established.

26.4.3. Measurement Errors in Dependent Variables

In a linear regression model the measurement errors in the dependent variable inflate the standard errors of regression parameters but do not lead to inconsistency of the estimator. In a nonlinear model there are additional consequences.

One class of applications has considered misclassification of responses in qualitative choice models. This has generated a literature on reporting errors.

Discrete Choice Models

Poterba and Summers (1995), in a study of the effects of unemployment insurance on the duration of unemployment using the CPS data, generalize a probabilistic model to allow for misclassification in labor market status transition. Specifically, they focus on potential classification errors in three classes: employed, unemployed, and not in the labor force. They develop a multinomial logit model with a special feature of the data set: that all of the individuals are assumed to correctly report as unemployed in the first survey month. Their results show that unemployment insurance increases unemployment spells and that correction for labor market status misclassification strengthens the apparent effect of unemployment insurance on spell durations. However, their model is based on an assumption that the probability of reporting errors is fixed and uncorrelated with individual characteristics, which, as the authors agree, is “unlikely to hold in practice.” Although the authors claim that the parameter estimates are consistent, Hausman, Abrevaya, and Scott-Morton (1998) argue that the standard errors are inconsistently estimated because of ignorance of sampling variability of the estimated error probability and a non-block-diagonal form of information matrix.

Hausman et al. (1998) propose a parametric method for estimating a binary choice model with misclassification. However, their parametric method requires knowledge of the error distribution. They emphasize that parameter estimates may be inconsistent if the distribution does not have the assumed parametric distribution. They further introduce a two-stage semiparametric method. The key condition in the model for identification is that the expected value of the observed dependent variable is an increasing function of the underlying index, which they show is weaker than the condition for identification of a parametric model. Compared to the approach of Poterba and Summers (1995), theirs is robust in the sense that the misclassification probability is a function of individual characteristics. Using the CPS and PSID, they show that serious misclassification exists in a job-change variable.

Klein and Sherman (1997) develop an “**Orbit model**” (with features of ordered choice model and Tobit model) for the estimation of projected demand for a potential new video product. They find evidence that potential consumers exaggerate demand. The Orbit model is a two-stage procedure with the first stage estimating the parameters of a standard Tobit model for actual future demand and the second stage estimating the mapping function between current projected demand and actual future demand. They further establish consistency and asymptotic normality of Orbit estimators. However, the identification of the model requires the assumption that the projected zero demand will be exact zero demand in future as well. This may be a strong assumption.

Hsiao and Sun (1999) use market survey data on the demand for an advanced electronic device. They argue that respondents may report biased demands. They propose a randomized response model and a one-sided response bias model for overreporting, in which different parametric probabilities are assigned to the truth and alternative

choices (including the truth) with logit or probit density function for the truly revealed preference. They find that “there is a substantial response bias in the data and the revised market take rates and price elasticities appear more reasonable than the estimates obtained based on the assumption that the respondents truly indicate their preference.”

Count Regression

In the nonlinear count regression context, Cameron and Trivedi (1998) suggest an approach for modeling count data subject to probabilistic **underrecording**. The approach generates compound Poisson and negative binomial count models by allowing for a binary recording outcome. Specifically, for each single occurrence of an event, a Bernoulli trial is used to determine whether the event is recorded. Given a positive probability that an event may not be recorded, the distribution of the recorded events has a smaller mean and variance than the distribution of the actual events. They further investigate estimation of the models by ML, quasi-generalized pseudo maximum likelihood, and moment-based methods. Based on a Monte Carlo study, they find that the performance of the ML estimator is good for samples of size 500 or more.

Jordan et al. (1997) give an application of the errors-in-variables method in the Poisson regression model. In a study of death from stomach cancer in five Japanese counties, they notice that a covariate (e.g., plasma lycopene level) is unknown and is estimated from a randomly chosen collective and, therefore, is subject to sampling error. With the assumption that the measurement error is distributed normally, they implement a Bayesian technique by obtaining the posterior distributions of the parameters using Gibbs sampling. The results indicate that the corrected model gives more accurate estimates of the parameters even when the original sample is small.

26.4.4. Poisson Regression with Measurement Errors in Covariates

We now consider in greater detail one specific example of a nonlinear regression model with additive measurement errors in covariates. This example illustrates both the consequences of such measurement errors and also feasible estimation strategies.

Guo and Li (2002) have shown that measurement errors in covariates in general lead to the overdispersion in the observed data. They also show using Monte Carlo simulations that biases will occur if the overdispersion caused by measurement errors is incorrectly modeled as arising from unobserved heterogeneity. Therefore, one should not conclude from the presence of overdispersion that a model with unobserved heterogeneity is warranted.

Stefanski (1989) and Nakamura (1990) propose a **corrected score estimator** that is consistent if measurement errors are present. In particular, Nakamura (1990) gives a closed form of corrected score function when the measurement errors are normally distributed and replicated data are also available. By contrast, Guo and Li (2002) have generalized Nakamura (1990).

Measurement Errors and Overdispersion

In this section, we consider the Poisson regression model in which the discrete random variable y follows the Poisson distribution with parameter $\mu = \exp(\mathbf{x}^* \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a $K \times 1$ parameter. As is well known, the Poisson regression model has an equidispersion property that

$$E[y|\mathbf{x}^*] = V[y|\mathbf{x}^*]. \quad (26.18)$$

If the measurement errors are additive, then

$$\mathbf{x} = \mathbf{x}^* + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ are assumed to be independent of unobserved latent variable \mathbf{x}^* , with mean zero and variance–covariance matrix Σ_{ε} . This notation covers the case where all or some of the explanatory variables are measured with errors.

Measurement errors increase dispersion (see Chesher, 1991). This applies to the Poisson regression, in the sense that although (26.18) holds for the conditional mean and variance of y given \mathbf{x}^* , conditioning on \mathbf{x} changes the result. Instead, we get $E[y|\mathbf{x}] < V[y|\mathbf{x}]$, in part because $E[y|\mathbf{x}^*] \neq E[y|\mathbf{x}]$, and $V[y|\mathbf{x}^*] \neq V[y|\mathbf{x}]$.

If $g(\mathbf{x}^*|\mathbf{x})$ denotes the conditional density of \mathbf{x}^* given \mathbf{x} , then Guo and Li show that

$$\begin{aligned} E[y|\mathbf{x}] &= \int E[y|\mathbf{x}^*]g(\mathbf{x}^*|\mathbf{x})d\mathbf{x}^* \\ &= \int E[y^2|\mathbf{x}^*]g(\mathbf{x}^*|\mathbf{x})d\mathbf{x}^* - \int (E[y|\mathbf{x}^*])^2g(\mathbf{x}^*|\mathbf{x})d\mathbf{x}^*, \end{aligned} \quad (26.19)$$

and using (26.18) the conditional variance of y given \mathbf{x} is given by

$$V[y|\mathbf{x}] = \int E[y^2|\mathbf{x}^*]g(\mathbf{x}^*|\mathbf{x})d\mathbf{x}^* - \left[\int E[y|\mathbf{x}^*]g(\mathbf{x}^*|\mathbf{x})d\mathbf{x}^* \right]^2. \quad (26.20)$$

A comparison of (26.19) and (26.20) shows that the first term inside the brackets of (26.19) is the same as the first term in (26.20). Using this Guo and Li show that

$$\left[\int E[y|\mathbf{x}^*]g(\mathbf{x}^*|\mathbf{x})d\mathbf{x}^* \right]^2 \leq \int (E[y|\mathbf{x}^*])^2g(\mathbf{x}^*|\mathbf{x})d\mathbf{x}^*, \quad (26.21)$$

which is interpreted to mean that measurement errors lead to overdispersion.

Estimation of Errors-in-Variables Model

When \mathbf{x} are contaminated by measurement errors ML estimation or NLS based on the observables (y, \mathbf{x}) does not provide consistent estimates. Replacement of covariate \mathbf{x}^* by \mathbf{x} in estimation is referred to as a “naive” model.

There are two issues to consider. First, why does ML give inconsistent estimates when measurement errors are present? Second, is consistent estimation possible? The answer to the second question is “yes” if we adopt, following Stefanski (1989) and Nakamura (1990), the method of **corrected score estimation** for the generalized linear models.

The idea underlying the corrected score estimator is that the conditional distribution of the corrected estimate with respect to \mathbf{x} , given the true independent variables \mathbf{x}^* and the dependent variables y , is centered around the ML estimate, which provides a consistent estimate of the true value of the parameter of interest.

Inconsistent and Consistent Estimators

Suppose that N observations (y_i, \mathbf{x}_i^*) , $i = 1, \dots, N$, are generated from a Poisson distribution with probability mass function

$$\Pr[Y_i = y_i | \mathbf{x}_i^*] = \frac{e^{-\mu_i(\beta_0)} \mu_i(\beta_0)^{y_i}}{y_i!},$$

where $\mu_i(\beta_0) = \exp(\mathbf{x}_i^{*\prime} \beta_0)$. Given observations (y_i, \mathbf{x}_i^*) , $i = 1, \dots, N$, the MLE $\widehat{\beta}$ is consistent since the probability limit of the average log-likelihood function

$$\begin{aligned} \text{plim } N^{-1} \ln L(\beta) &= N^{-1} \sum_i \{-e^{\mathbf{x}_i^{*\prime} \beta} + y_i \mathbf{x}_i^{*\prime} \beta - \ln y_i!\} \\ &= E_{y, \mathbf{x}^*}[-e^{\mathbf{x}^{*\prime} \beta} + y \mathbf{x}^{*\prime} \beta - \ln y!] \end{aligned} \quad (26.22)$$

is maximized at $\beta = \beta_0$.

Suppose we observe \mathbf{x}_i rather than \mathbf{x}_i^* , where $\mathbf{x}_i = \mathbf{x}_i^* + \varepsilon_i$ and $\varepsilon_i \sim \mathcal{N}[\mathbf{0}, \Sigma_\varepsilon]$ independent of \mathbf{x}_i^* . Then $y_i | \mathbf{x}_i$ is not Poisson distributed. If one nevertheless uses the **“naive” Poisson model**, the resulting estimator $\widetilde{\beta}$ maximizes

$$Q(\beta) = N^{-1} \sum_i \{-e^{\mathbf{x}_i^{*\prime} \beta} + y_i \mathbf{x}_i^{*\prime} \beta - \ln y_i!\}. \quad (26.23)$$

This misspecified log-likelihood function converges to

$$\text{plim } Q(\beta) = E_{y, \mathbf{x}^*}[-e^{\mathbf{x}^{*\prime} \beta} + y \mathbf{x}^{*\prime} \beta - \ln y!] + E_{\mathbf{x}^*}[-e^{\mathbf{x}^{*\prime} \beta}] (E_\varepsilon[e^{\varepsilon' \beta}] - 1), \quad (26.24)$$

which in general is not maximized at $\beta = \beta_0$. So $\widetilde{\beta}$ is inconsistent for β_0 .

A suitably modified objective function yields consistent estimates. Equations (26.22) and (26.24) imply that

$$\{\text{plim } Q(\beta) - E_{\mathbf{x}^*}[-e^{\mathbf{x}^{*\prime} \beta}] (E_\varepsilon[e^{\varepsilon' \beta}] - 1)\} = \text{plim } N^{-1} \ln L(\beta).$$

This suggests maximizing the objective function

$$Q^+(\beta) = N^{-1} \sum_i \{-e^{\mathbf{x}_i^{*\prime} \beta} + y_i \mathbf{x}_i^{*\prime} \beta - \ln y_i!\} - E_{\mathbf{x}^*}[-e^{\mathbf{x}^{*\prime} \beta}] (E_\varepsilon[e^{\varepsilon' \beta}] - 1),$$

since $Q^+(\beta)$ converges to $\text{plim } N^{-1} \ln L(\beta)$. Now, given independence of \mathbf{x}^* and ε ,

$$E_{\mathbf{x}^*}[-e^{\mathbf{x}^{*\prime} \beta}] E_\varepsilon[e^{\varepsilon' \beta}] = E_{\mathbf{x}^*, \varepsilon}[-e^{(\mathbf{x}^* + \varepsilon)' \beta}] = -E_\mathbf{x}[e^{\mathbf{x}' \beta}],$$

which is consistently estimated by $-N^{-1} \sum_i e^{\mathbf{x}_i^{*\prime} \beta}$. It follows after some cancellation that maximizing $Q^+(\beta)$ is equivalent to maximizing

$$Q^{++}(\beta) = N^{-1} \sum_i \{y_i \mathbf{x}_i^{*\prime} \beta - \ln y_i!\} - E_{\mathbf{x}^*}[e^{\mathbf{x}^{*\prime} \beta}]. \quad (26.25)$$

This yields a consistent estimate of β_0 . Implementation requires a suitable estimate of $E_{\mathbf{x}^*}[e^{\mathbf{x}'\beta}]$, which is possible if replicated data are available. If the distribution of explanatory variables is specified up to unknown parameters, then these unknown parameters can be estimated by the replicated measurements. Therefore, $E_{\mathbf{x}^*}[e^{\mathbf{x}'\beta}]$ can be estimated.

The estimator $\hat{\beta}_C$ that maximizes (26.25) is termed the **corrected score estimator** by Guo and Li (2002) because it is the root of the corrected score function $\sum_i(y_i \mathbf{x}_i - E_{\mathbf{x}^*}[\mathbf{x}^* e^{\mathbf{x}'\beta}]) = \mathbf{0}$. Guo and Li also establish the asymptotic normality of this estimator. The estimated asymptotic covariance matrix $\hat{V}[\hat{\beta}_C] = N^{-1}\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}$, where

$$\begin{aligned}\hat{\mathbf{A}} &= E_{\mathbf{x}^*}[e^{\mathbf{x}'\hat{\beta}_C} \mathbf{x}^* \mathbf{x}^{*\prime}], \\ \hat{\mathbf{B}} &= N^{-1} \sum_i (y_i \mathbf{x}_i - E_{\mathbf{x}^*}[e^{\mathbf{x}'\hat{\beta}_C} \mathbf{x}^*])(y_i \mathbf{x}_i - E_{\mathbf{x}^*}[e^{\mathbf{x}'\hat{\beta}_C} \mathbf{x}^*])'.\end{aligned}$$

Nakamura (1990) made the stronger assumption that the measurement errors ε are normally distributed as $\mathcal{N}[\mathbf{0}, \Omega]$. Then

$$\exp(\mathbf{x}'\beta) = E_{\mathbf{x}|\mathbf{x}^*} [\exp(\mathbf{x}'\beta - (\beta'\Omega\beta/2))].$$

By the law of iterated expectations

$$E_{\mathbf{x}^*}[\exp(\mathbf{x}'\beta)] = E_{\mathbf{x}} [\exp(\mathbf{x}'\beta - (\beta'\Omega\beta/2))],$$

which can be consistently estimated by $N^{-1} \sum_i [\exp(\mathbf{x}'_i\beta - (\beta'\Omega\beta/2))]$. Consequently, for $Q(\beta)$ in (26.23) the probability limit given in (26.24) reduces to

$$\text{plim } Q(\beta) = N^{-1} \sum_i [y_i \mathbf{x}'_i \beta - \ln y_i! - \exp(\mathbf{x}'_i \beta - (\beta'\Omega\beta/2))].$$

This is the **corrected log-likelihood** function given in Nakamura (1990). Maximization with respect to β yields a consistent estimate of β_0 .

Nakamura's approach reminds one of the estimation of the linear regression with measurement errors (see (26.14)) given an estimate of the covariance matrix of measurement errors. As in that case, to maximize Nakamura's corrected log-likelihood function one requires knowledge of Ω , the covariance matrix of measurement errors. This can come from replicated data. However, if the covariates are predominantly discrete, then the normality of measurement error is not a sensible assumption. In such cases the estimator of Guo and Li is more attractive.

For the case of multivariate \mathbf{x}^* , the computation of $E[\exp(\mathbf{x}'\beta)]$ is not straightforward, even if the distribution of \mathbf{x}^* is known, because multiple integrals are involved. Simulation-based methods (Li, 2002) provide one possible approach to this problem.

Implementation of several other nonlinear errors in variable models also require replicated observations; for example, see Hsiao (1992) and Hausman, Newey, and Powell (1995). Panel data could provide replicated observations at the level of an individual. For example, consider the case of a scalar regressor x^* for which two replications of x are available, because $x_{ij} = x_i + \varepsilon_{ij}$ for $i = 1, \dots, N$ and $j = 1, 2$. Then a moment-based consistent estimator of σ_ε^2 is $\hat{\sigma}_\varepsilon^2 = \sum_i(x_{i1}^2 + x_{i2}^2 - 2x_{i1}x_{i2})/2N$. Thus both the mean and variance of \mathbf{x}^* can be estimated.

26.5. Attenuation Bias Simulation Examples

Analytical results for the linear model are given in Section 26.2, but results are much more difficult to obtain in nonlinear models. Here we present two simulation examples, one for the logit model and one for a linear-in-logs model, that illustrate attenuation bias in nonlinear regression with measurement error in the regressor.

In the first example, the dgp is the logit model with

$$\begin{aligned} y^* &= \alpha^* + \beta^* x^* + \varepsilon, \\ x^* &\sim \mathcal{U}[0, 1], \quad \varepsilon \sim \text{logistic}, \\ y &= \begin{cases} 0 & \text{if } y^* \leq 0, \\ 1 & \text{if } y^* > 0. \end{cases} \end{aligned}$$

The complication is that x^* is measured with error, so that

$$\begin{aligned} x &= x^* + v, \\ v &\sim \mathcal{N}[0, \sigma_v^2]. \end{aligned}$$

Since $x^* \sim \mathcal{U}[0, 1]$ it has variance $\sigma_{x^*}^2 = 1/12$, and the noise-to-signal ratio is $s = 12\sigma_v^2$. A logit regression of y on x rather than of y on x^* is estimated.

To conduct a simulation exercise we carry out a logit regression of y on x , for six different values of the noise-to-signal ratio including the value of zero, which benchmarks the model. The sample size is fixed at 1,000, and 100 simulation replications are used.

Table 26.1 shows the average values of $(\hat{\alpha}, \hat{\beta})$ in 100 replications, where $\hat{\alpha}$ and $\hat{\beta}$ are the estimated intercept and slope from logit regression of y on x , rather than the correct logit regression of y on x^* , for sample size $N = 1,000$ and for six different values of σ_v^2 leading to six different noise-to-signal ratios s . The first column with $s = 0$ benchmarks the model. Recall that for OLS linear regression in the same setup the multiplicative bias in the slope coefficient is $1/(1 + s)$, or 0.96, 0.8, 0.5, 0.2, and 0.1, respectively. Here the biases have a similar direction, except for logit regression they are considerably larger.

The second example is a bivariate linear-in-logs multiplicative model with $\alpha = 4$, $\beta = 0.4$, and additive measurement errors in both variables. In this case the setup is

Table 26.1. Attenuation Bias in a Logit Regression with Measurement Error

Noise/Signal	0	0.04	0.25	1	4	9
Average $\hat{\alpha}$	0.785	1.062	1.406	1.548	1.570	1.596
Average $\hat{\beta}$	1.799	1.224	0.446	0.125	0.037	0.012

Table 26.2. Attenuation Bias in a Nonlinear Regression with Additive Measurement Error

$\sigma_x^2/\sigma_{x^*}^2$	0.00025	0.0025	0.025	0.25	2.5	25
Average $\hat{\beta}$	0.393	0.383	0.341	0.217	0.063	0.020

as follows:

$$\begin{aligned} y^* &= 4x^{*0.4} u, \quad u \sim \mathcal{N}[10, 0.0001], \\ x^* &= 100 + \mathcal{U}[0, 1], \\ y &= y^* + \varepsilon_y, \quad \varepsilon_y \sim \mathcal{N}[0, \sigma_y^2], \\ x &= x^* + \varepsilon_x, \quad \varepsilon_x \sim \mathcal{N}[0, \sigma_x^2]. \end{aligned}$$

In the simulation the sample size is 1,000, and the number of replications is 100. We vary the value of the variance of x^* from experiment, to experiment, resulting in the following values of $\sigma_x^2/\sigma_{x^*}^2$: 0.001, 0.01, 0.1, 1, 5, 10, 50, 100, 1,000, and 5,000.

The upper row of Table 26.2 gives the average values of slope coefficients across different experiments in which the noise-to-signal ratio varies. Once again the attenuation bias is obvious.

Both examples produce results that are consistent with the hypothesis underlying the “Iron Law of Econometrics.”

26.6. Bibliographic Notes

Wansbeek and Meijer (2000) is the most up to date and comprehensive work on measurement errors written from an econometric perspective. It covers in depth most of the topics in this chapter, with emphasis on linear models. The authors also include several chapters linking measurement error models with factor models, latent variable models, and structural equation models. In discussing results the authors eschew the phrase “it can be shown” in favor of deriving them in detail. Again from the econometric perspective Hausman (2000) provides a survey of the recent results obtained in his and his collaborator’s research. Bound, Brown, and Mathiowetz (2001) for a survey of measurement error issues in labor markets.

The topic of measurement errors is well established in the statistics literature. Fuller (1987) is a useful reference; see, in particular, his treatment of the **orthogonal regression** approach that is applicable when the noise-to-signal ratio is known. Although our analysis of the linear model is very standard in the econometrics literature, the reader should also be aware of the alternative **Berkson error model**, in which the unobserved true variable is assumed constant but the imperfectly measured variable is subject to error, and the **nonclassical measurement error** model discussed in Angrist and Krueger (1999). Madansky (1959) provides a survey of numerous early results and approaches. See also Stefanski (2000).

26.2 Panel data models with measurement errors are analyzed in Bjorn (1992).

26.3 The intriguing topic of reverse regression is analyzed by Goldberger (1984) and Greene (1983) in their commentary on Conway and Roberts (1983). Leamer (1978) provides an insightful discussion of reverse regression from a Bayesian perspective. Hahn and Hausman (2002) use the reverse regression idea to construct a specification test for the validity of the IV approach to the measurement error problem. The concern is that the

available instruments may be weak, leading to poor estimates. The Hahn–Hausman idea is to carry out IV estimation of the direct regression in which the mismeasured variable appears on the right-hand side of the equation. The reverse regression has the same mismeasured variable on the left-hand side. This regression is estimated also by instrumental variables using the same instrumental variables as the direct regression.

- 26.4** The literature on measurement errors in nonlinear models is more diffuse. Y. Amemiya (1985) is especially useful to econometricians. From a statistical viewpoint, Carroll et al. (1995) consider nonlinear models, especially in the generalized linear class, with additive measurement errors in regressors, using a variety of methods, including a number that can be used if replicated data are available. Li, Trivedi, and Guo (2003) develop and apply a measurement error variable model in which the counted response variable has measurement error.

Exercises

- 26–1** Consider the attenuation bias result for the slope parameter of the bivariate errors-in-variables model (Equation (26.9) in Section 26.2.3). Extend the model to include an intercept term.
- (a) Derive a parallel result for the measurement error bias of the intercept term.
 - (b) Derive a parallel identification-by-bounds result for the least-squares intercept estimate, similar to Equation (26.12) in Section 26.3.1.
- 26–2** (Adapted from Bollinger, 2003) Consider a linear multiple regression model with scalar regressor x that is measured with error and a vector of other regressors \mathbf{z} that are free of measurement error.
- (a) Maintaining the assumptions regarding measurement errors in the bivariate errors-in-variables model, extend the attenuation bias result and the identification-by-bounds result to this case.
 - (b) Check that the new results specialize to those for the bivariate case.
- 26–3** (Adapted from Wansbeek and Meijer, 2000) Consider the quadratic regression model $y = \alpha + \beta x^* + \gamma x^{*2} + \varepsilon$, where the regressor $x^* = x + v$, with x observed and v a measurement error. Assume that (x^*, ε, v) are mutually uncorrelated and normally distributed and that all variables have zero mean.
- (a) Compare the bias of the least-squares estimator of β and γ .
 - (b) Is the model identified? Compare the latter result with that from the bivariate linear errors-in-variable model.
- 26–4** The literature on intergenerational mobility uses the following model (Solon, 1992; Zimmerman, 1992):

$$Y_i^{\text{son}} = \alpha + \beta Y_i^{\text{father}} + \varepsilon_i^{\text{son}}, \quad (26.26)$$

with $\varepsilon_i \sim \text{iid } \mathcal{N}[0, \sigma^2]$. Here Y is a measure of permanent status (such as permanent income) and β measures the degree of regression toward the mean in economic status. Suppose that permanent status is not observed. Instead, current status Y_{it} is observed with $Y_{it} = Y_i + \gamma X_{it} + w_{it}$, so that Y_{it} is composed of an individual fixed effect Y_i , referred to as the permanent status, a systematic factors X_{it} , and a transitory error component w_{it} . Let $\hat{\gamma}$ denote the fitted

least-squares coefficient, and let

$$Y_{it} - \hat{\gamma} X_{it} = Y_i + (\gamma - \hat{\gamma}) X_{it} + w_{it} = Y_i + v_{it}.$$

- (a) Let $\bar{Y}_i^{\text{father}} = T^{-1} \sum_{t=1}^T Y_{it}^{\text{father}}$ denote an average of father's status used as the independent variable, a proxy, for the unobserved permanent status in (26.26). Let $\hat{\beta}_{\text{avg}}$ denote the corresponding regression coefficient. Show that $\text{plim } \hat{\beta}_{\text{avg}} = \beta P_Y$, where $P_Y = \sigma_Y^2 / (\sigma_Y^2 + T^{-1} \sigma_\varepsilon^2)$.
- (b) Assume that the transitory component of father's earnings follows an autoregressive scheme, $v_{it}^{\text{father}} = \rho v_{it}^{\text{father}} + \xi_{it}$, where $\xi_i \sim \mathcal{N}[0, \sigma_\xi^2]$, $i = 1, \dots, T$. Show that now $\text{plim } \hat{\beta}_{\text{avg}} = \beta P_Y^*$, where $P_Y^* = \sigma_Y^2 / (\sigma_Y^2 + T^{-1} V)$ and $V = \sigma_\xi^2 [T(1 - \rho^2)]^{-1} [(1 + 2\rho\{T - (1 - \rho^T)/(1 - \rho)\}) / T(1 - \rho)]$.

Missing Data and Imputation

27.1. Introduction

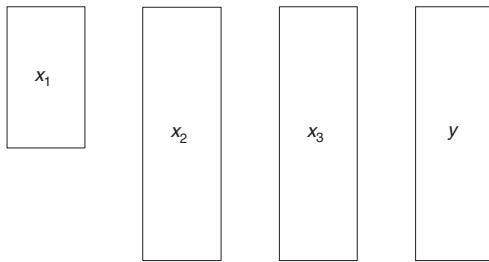
The problem of **missing data** in survey data is one of long standing, arising from nonresponse or partial response to survey questions. Reasons for nonresponse include unwillingness to provide the information asked for, difficulty of recall of events that occurred in the past, and not knowing the correct response. **Imputation** is the process of estimating or predicting the missing observations.

In this chapter we deal with the regression setup with data vector (y_i, \mathbf{x}_i) , $i = 1, \dots, N$. For some of the observations some elements of \mathbf{x}_i or of both (y_i, \mathbf{x}_i) are missing. A number of questions are considered. When can we proceed with an analysis of only the complete observations, and when should we attempt to fill the gaps left by the missing observations? What methods of imputation are available? When imputed values for missing observations are obtained, how should estimation and inference then proceed?

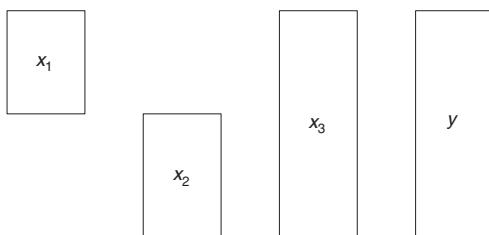
If a data set has missing observations, and if these gaps can be filled by a statistically sound procedure, then benefit comes from a larger and possibly more representative sample and, under ideal circumstances, more precise inference. The cost of estimating missing data comes from having to make (possibly wrong) assumptions to support a procedure for generating proxies for the missing observations, and from the approximation error inherent in any such procedure. Further, statistical inference that follows data augmentation after imputed values replace missing data is more complicated because such inference must take into account the approximation errors introduced by imputation.

Gaps in data as the result of survey nonresponse and attrition from panels occur frequently. Imputation of missing values may be done by agencies for creating and maintaining the public-use survey databases or by those who use the data for modeling. In the former case the agency may have more extensive information, including confidential information, that can be harnessed in the imputation process. In the latter case the modeler may have a specific modeling framework that can be exploited in the imputation process. In both cases model-based imputation procedures are feasible.

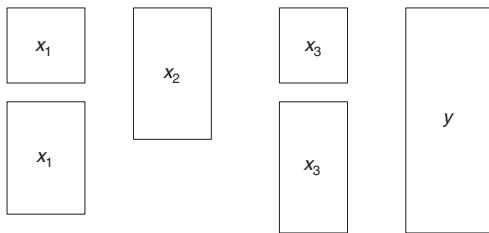
MISSING DATA AND IMPUTATION



A: Univariate missing data pattern



B: Special Pattern of missing data on x_1 and x_2



C: General pattern of missing data

Figure 27.1: Missing data: examples of missing regressors.

An interesting example of missing data arises in the context of the Survey of Consumer Finances (Kennickell, 1998). Because of the sensitivity of the issue of consumer finances the survey exhibits many gaps in information on income and wealth. Analysts at the U.S. Federal Reserve have developed and implemented complex imputation algorithms for continuous and discrete variables using both publicly available survey information on income and wealth as well as confidential information from census data.

Figure 27.1 shows some potential patterns of missing data on the regressors. The data set has a scalar dependent variable y and three regressors: x_1 , x_2 , and x_3 for each observation, then stacked as (y, x_1, x_2, x_3) . In panel A, there are complete data on (y, x_1, x_2, x_3) but a block of observations on x_1 are missing. In panel B there are complete data on (y, x_3) but there are missing blocks of data on (x_1, x_2) such that x_1 and x_2 are never simultaneously observed. In panel C there is a general pattern of missing

observations with missing observations on all three regressors, but there is no particular pattern of missingness.

The simplest way of handling missing data is to delete them and analyze only the reduced sample of “complete” observations. For example, in the case of panel A, the complete sample would be the subset of $(\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ formed by all available data on \mathbf{x}_1 and the corresponding observations on $(\mathbf{y}, \mathbf{x}_2, \mathbf{x}_3)$. In the case of panel B, however, following this approach one would leave no usable observations, unless one excluded $(\mathbf{x}_1, \mathbf{x}_2)$ from the analysis. In panel C the complete data set is formed after deleting any observation that contains a missing data point on any of the three regressors.

The procedure just described is called **listwise deletion**. It is widely followed and is often a default option in statistical software. It is not necessarily innocuous; the consequences depend on the missing data mechanism, and the conclusions drawn from such studies might be seriously flawed. Of course, in general throwing away data means throwing away information, and that reduces efficiency in estimation. Hence, provided the gaps attributed to missing data can be filled without creating distortion, listwise deletion seems worth trying. This chapter will study alternative approaches and their limitations.

Broadly, there are two approaches to imputation, one that is **model-based** and one that is not. The modern approach prefers model-based approaches. These use a model to impute the missing observations and then use the subsequent full data set to obtain better estimates of the model parameters. The process is iterative. Single and multiple imputation are feasible. A key feature of the modern approach is to regard missing data as random variables and then to replace them with multiple draws from the assumed underlying distribution; the process is called **multiple imputation**. Simulation methods may be used to approximate such a distribution.

This topic warrants a separate short introductory chapter as imputation is an important aspect of microeconometric work. Survey data inevitably include missing data, and the common practice of listwise deletion is an imputation method. Better imputation methods are available. An important caveat, however, is that all imputation methods are based on assumptions that in some applications may be too strong.

Most of the chapter deals with model-based approaches. Section 27.2 provides an introduction to the terminology and assumptions that are firmly entrenched in the imputation literature. Section 27.3 gives a brief treatment of missing data methods that do not use models. Section 27.4 begins with the first of the model-based methods, maximum likelihood. Section 27.5 considers the regression framework and EM-type methods of imputation. Sections 27.6 and 27.7 present approaches to imputation using the Bayesian concepts of data augmentation and MCMC. Section 27.8 provides an illustrative example. Sections 27.6–27.8 provide a nice application of the Bayesian methods of Chapter 13.

27.2. Missing Data Assumptions

Some of the basic terminology and formal definitions widely used in the imputation literature are due to Rubin (1976), who introduced two key missing data

mechanisms, missing at random and missing completely at random, that serve as useful benchmarks.

Rubin's setup involves \mathbf{Y} , an $N \times p$ matrix consisting of a complete data set, which may not be fully observed. Denote by \mathbf{Y}_{obs} the observed part and by \mathbf{Y}_{mis} the nonobserved (missing) part. In the context of a regression model \mathbf{Y} refers to both the regressors and the response (dependent) variables. Therefore, the analysis covers the general case of missing data. Let \mathbf{R} denote an $N \times p$ matrix of indicator variables whose elements are zero or one depending on whether corresponding values in the \mathbf{Y} matrix are missing or observed.

For regression with single dependent variable, \mathbf{Y} contains data on the response variable \mathbf{y} and the $(p - 1)$ regressors \mathbf{X} . The probability that x_{ki} , the i th observation on variable x_k , is missing may be (i) independent of its realized value, (ii) dependent on its realized value, (iii) dependent on x_{kj} , $j \neq i$, or (iv) dependent on x_{lj} , $j \neq i$, $l \neq k$.

Assumptions about the structure of missingness follow.

27.2.1. Missing at Random

Suppose x_i ($i = 1, \dots, N$) is an observation on a variable in the data set under study. The **missing at random (MAR) assumption** is that the “missingness” in x_i does not depend on its value but may depend on the values of x_j ($j \neq i$). Formally,

$$\begin{aligned} x_i \text{ is MAR} &\Rightarrow \Pr[x_i \text{ is missing} | x_i, x_j \forall j \neq i] \\ &= \Pr[x_i \text{ is missing} | x_j \forall j \neq i]. \end{aligned} \tag{27.1}$$

After controlling for other observations on x , the probability of missingness of x_i is unrelated to the value of x_i .

Rubin's (1976) even more formal definition states the following: The MAR assumption implies that the probability model for the indicator variable \mathbf{R} does not depend on \mathbf{Y}_{mis} , that is,

$$\Pr[\mathbf{R} | \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}, \psi] = \Pr[\mathbf{R} | \mathbf{Y}_{\text{obs}}, \psi],$$

where ψ is the underlying (vector) parameter of the missingness mechanism.

Under MAR no nonresponse bias is induced in a likelihood-based inference that ignores the missing data mechanism, although the resulting estimates may be inefficient. If the MAR assumption fails, however, the probability of missingness depends on the unobserved missing values. The MAR restriction is not testable because the values of the missing data are unknown. Because MAR is a strong assumption, sensitivity analyses based on different assumptions about missingness are desirable.

A separate issue is whether the pattern of missing data is purely random. In practice, we might expect that observations missing inside clusters of data, in the sense of Chapter 24, may be correlated. However, this issue is not related to that of nonresponse bias resulting from the missingness being connected to data values.

27.2.2. Missing Completely at Random

Missing completely at random (MCAR) is a special case of MAR. It means that \mathbf{Y}_{obs} is a simple random sample of all potentially observable data values (Schafer, 1997).

Again suppose x_i is an observation on a variable in the data set under study. Then the data on x_i is said to be MCAR if the probability of missing data on x_i depends neither on its own values nor on the values of other variables in the data set. Formally,

$$\begin{aligned} x_i \text{ is MCAR} &\Rightarrow \Pr[x_i \text{ is missing} | x_i, x_j \forall j \neq i] \\ &= \Pr[x_i \text{ is missing}]. \end{aligned} \quad (27.2)$$

For example, MCAR is violated if (a) those who do not report income are younger, on average, than those who do or if (b) typically small (large) values are missing.

For cases (i)–(iv) mentioned at the outset in this section, case (i) satisfies both MCAR and MAR, cases (iii) and (iv) satisfy MAR, and (ii) does not satisfy MAR.

MCAR implies that the observed data are a random subsample of the potential full sample. If the assumptions were valid no biases would result from ignoring incomplete observations, that is, observations with missing values.

The corollary is that the failure of MCAR implies a sample selection type of bias. MAR is a weaker assumption that still aids imputation as it assumes that the missing data mechanism depends only on observed quantities.

27.2.3. Ignorable and Nonignorable Missingness

A missing data mechanism is said to be **ignorable** if (a) the data set is MAR and (b) the parameters for the missing data-generating process, ψ , are unrelated to the parameters θ that we want to estimate.

This condition, which is similar to that of **weak exogeneity** discussed in Chapter 2, implies that the parameters θ of the model are distinct from parameters ψ of the missingness mechanism. Thus, if the missing data are ignorable, then there is no need to model the dgp for missing data as an essential part of the modeling exercise. MAR and “ignorability” are often treated as equivalent under the assumption that condition (b) for ignorability is almost always satisfied (Allison, 2002).

A **nonignorable** missing data mechanism arises if the MAR assumption is violated for (y, x) , but it would not be violated if MAR is violated only for x . In that case the dgp for missing data must be modeled along with the overall model to obtain consistent estimates of the parameters θ . To avoid the possibility of selection bias, estimators such as Heckman’s two-stage procedure (see Chapter 16) must be used.

The imputation literature focuses on ignorable missingness. If additionally the data set is MCAR then missing data cause no problem, aside from efficiency loss that might be reduced by imputation. If instead the data set is only MAR then imputation methods may be needed to ensure consistency, as well as to increase efficiency.

27.3. Handling Missing Data without Models

If no models are to be used, then one can simply analyze the available data or one can analyze data after non-model-based imputation.

27.3.1. Using Available Data Only

Listwise deletion or complete case analysis means the deletion of the observations (cases) that have missing values on one or more of the variables in the data set. Under the MCAR assumption, the remaining sample after listwise deletion remains a random sample from the original population; therefore the estimates based on it are consistent. However, the standard errors will be inflated because less information is used. If the number of regressors is large, then the total effect of listwise deletion can lead to very substantial reduction in the total number of observations. This might encourage one to leave out of the analysis variables with a high proportion of missing observations, but the results generated by such practice are potentially misleading.

If MCAR is not satisfied and the missing data are only MAR, then the estimates will be biased. Thus listwise deletion is not robust to the violations of MCAR. However, listwise deletion is robust to the violations of MAR among the independent variables (regressors) in regression analysis, that is, when the probability of missing data on any regressor does not depend on the values of the dependent variable. Briefly, listwise deletion is acceptable if incomplete cases attributable to missing data comprise a small percentage, say 5% or less, of the number of total cases (Schafer, 1996). It is important that the sample after listwise deletion is representative of the population under study.

Pairwise deletion or available-case analysis is often considered a better method than listwise deletion. The idea here is to use all possible pairs of observations (x_{1i}, x_{2i}) in estimating joint sample moments of (x_1, x_2) and to use all observations on an individual variable in estimating marginal moments. Thus, in a linear regression, under pairwise deletion we would estimate $(\mathbf{X}'\mathbf{X})$ and $(\mathbf{X}'\mathbf{y})$ using all possible pairs of regressors, whereas under listwise deletion we would estimate the same after deleting all cases with *any* missing observations. It is clear that we lose less information under pairwise deletion. The proposal here is to use maximum information to estimate individual summary statistics such as means and covariances and then to use these summary statistics to compute the regression estimates.

There are two important limitations of pairwise deletion: (1) Conventionally estimated standard errors and test statistics are biased and (2) the resulting regressor covariance matrix $(\mathbf{X}'\mathbf{X})$ may not be positive definite.

27.3.2. Imputation without Models

There are a number of ad hoc or weakly justified procedures often implemented in statistical software.

Mean imputation or **mean substitution** involves replacing missing observations by the average of the available values. It is mean-preserving but will have impact on the marginal distribution of the data. It is obvious that the probability mass in the center

of the marginal distribution will increase. It will also affect the covariances and correlations with other variables.

Simple hot deck imputation involves replacement of the missing value by a randomly drawn value from the available observed values of that variable, somewhat like a bootstrap procedure. It preserves the marginal distribution of the variable, but it distorts the covariances and correlations between variables.

In a regression setting neither of these two well-known approaches are attractive despite their simplicity.

27.4. Observed-Data Likelihood

The modern approach to missing data is to impute values for missing observations by making single or multiple draws from the estimated distribution based on the postulated observed data model and the model for the missing data mechanism. The Bayesian variants of this procedure make the draws from the posterior distribution, which uses both the likelihood and the prior distribution of the parameters.

The first important issue involves the role played by the missing data mechanism in the imputation procedure and especially whether the missing data mechanism is ignorable.

Let θ denote the parameters of the dgp for $\mathbf{Y} = (\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$ and let ψ denote the parameters of the missing data mechanism. For convenience of notation it is assumed that $(\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$ are continuous variables. Then the joint distribution of $(\mathbf{R}, \mathbf{Y}_{\text{obs}})$ is given by

$$\begin{aligned}\Pr [\mathbf{R}, \mathbf{Y}_{\text{obs}} | \theta, \psi] &= \int \Pr [\mathbf{R}, \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}} | \theta, \psi] d\mathbf{Y}_{\text{mis}} \\ &= \int \Pr [\mathbf{R} | \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}, \psi] \Pr [\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}} | \theta] d\mathbf{Y}_{\text{mis}} \\ &= \Pr [\mathbf{R} | \mathbf{Y}_{\text{obs}}, \psi] \int \Pr [\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}} | \theta] d\mathbf{Y}_{\text{mis}} \\ &= \Pr [\mathbf{R} | \mathbf{Y}_{\text{obs}}, \psi] \Pr [\mathbf{Y}_{\text{obs}} | \theta].\end{aligned}\tag{27.3}$$

The first equality derives the joint probability of $(\mathbf{R}, \mathbf{Y}_{\text{obs}})$ by integrating out (or averaging over) \mathbf{Y}_{mis} from the joint probability of all data and \mathbf{R} . The second line factors the joint probability into conditional and marginal components, the conditioning being with respect to \mathbf{Y}_{obs} and \mathbf{Y}_{mis} . The third line separates the missing data mechanism from the observed data mechanism; this step is justified by the MAR assumption. The last line means that θ and ψ are distinct parameters and hence inference about θ can ignore the missing data mechanism and depends on \mathbf{Y}_{obs} alone.

The **observed-data likelihood** is proportional to the last factor in the fourth line:

$$L[\theta | \mathbf{Y}_{\text{obs}}] \propto \Pr [\mathbf{Y}_{\text{obs}} | \theta].\tag{27.4}$$

It involves only the observed data \mathbf{Y}_{obs} even though the parameters θ appear in the dgp for all observations (observed and missing). As in Chapter 13, the constant of proportionality does not appear in (27.4).

Under the MAR assumption the **joint posterior probability** of (θ, ψ) is written as the product of $\Pr[\mathbf{R}, \mathbf{Y}_{\text{obs}} | \theta, \psi]$ and the joint prior distribution $\pi(\theta, \psi)$ as follows:

$$\begin{aligned}\Pr[\theta, \psi | \mathbf{Y}_{\text{obs}}, \mathbf{R}] &= k \Pr[\mathbf{R}, \mathbf{Y}_{\text{obs}} | \theta, \psi] \pi(\theta, \psi) \\ &\propto \Pr[\mathbf{R} | \mathbf{Y}_{\text{obs}}, \psi] \Pr[\mathbf{Y}_{\text{obs}} | \theta] \pi(\theta, \psi) \\ &\propto \Pr[\mathbf{R} | \mathbf{Y}_{\text{obs}}, \psi] \Pr[\mathbf{Y}_{\text{obs}} | \theta] \pi_\theta(\theta) \pi_\psi(\psi),\end{aligned}\tag{27.5}$$

where k in the first line is a constant of proportionality free of (θ, ψ) . The second line uses the factorization given in (27.3), and the third line uses the assumption of independent priors for θ and ψ .

As our main interest is in θ , we derive the marginal posterior for θ by integrating out ψ from the joint posterior. This yields the **observed-data posterior**

$$\begin{aligned}\Pr[\theta | \mathbf{Y}_{\text{obs}}, \mathbf{R}] &= \int \Pr[\theta, \psi | \mathbf{Y}_{\text{obs}}, \mathbf{R}] d\psi \\ &\propto \Pr[\mathbf{Y}_{\text{obs}} | \theta] \pi_\theta(\theta) \int \Pr[\mathbf{R} | \mathbf{Y}_{\text{obs}}, \psi] \pi_\psi(\psi) d\psi \\ &\propto L[\theta | \mathbf{Y}_{\text{obs}}] \pi_\theta(\theta),\end{aligned}\tag{27.6}$$

where the second line separates θ and ψ , and the last line absorbs the integral expression into the constant of proportionality. Therefore, the last line does not involve ψ and is independent of the missing data mechanism \mathbf{R} .

27.5. Regression-Based Imputation

In this section we consider a least-squares based imputation. The key component is use of the EM algorithm, previously introduced and discussed in Section 10.3.7.

The EM algorithm consists of the expectation step and the maximization step. The structure of the EM algorithm is closely related to Bayesian MCMC and data augmentation methods. Therefore, rather than providing a fully operational method for handling missing data, we will introduce an example that brings out the motivation behind modern multiple imputation techniques and suggests the major features of such an approach.

27.5.1. Linear Regression Example with Missing Data on a Dependent Variable

In practice one can have missing observations on dependent (endogenous) variables and/or explanatory variables. We consider a regression example that has missing data on the dependent variable, with

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_{\text{mis}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix},\tag{27.7}$$

where $E[\mathbf{u}|\mathbf{X}] = \mathbf{0}$ and $E[\mathbf{u}\mathbf{u}'|\mathbf{X}] = \sigma^2\mathbf{I}_N$. The complication is that a block of observations on the dependent variable \mathbf{y} , denoted \mathbf{y}_{mis} , is missing. We assume that the available complete observations are a random sample from the population, so that the missing data are assumed to be MAR though not MCAR.

Given the MAR assumption and $N_1 > K$, the first block of N_1 observations can be used to consistently estimate the K -dimensional parameter β and σ^2 . The maximum likelihood estimates of (β, σ^2) under Gaussian errors are $\hat{\beta} = [\mathbf{X}'_1 \mathbf{X}_1]^{-1} \mathbf{X}'_1 \mathbf{y}_1$ and $s^2 = (\mathbf{y}_1 - \mathbf{X}_1 \hat{\beta})'(\mathbf{y}_1 - \mathbf{X}_1 \hat{\beta})/N_1$. By standard theory, and under the normality assumption, $\hat{\beta}|\text{data} \sim \mathcal{N}[\beta, \sigma^2 [\mathbf{X}'_1 \mathbf{X}_1]^{-1}]$ and $s^2/\sigma^2|\hat{\beta} \sim (N_1 - K)\chi^2_{N_1 - K}$.

First, consider a naive single-imputation procedure for generating the missing observations. Conditional on \mathbf{X}_2 , the predicted values of \mathbf{y}_{mis} , denoted $\hat{\mathbf{y}}_{\text{mis}}$, are given by $\mathbf{X}_2 \hat{\beta}$, where $\hat{\beta}$ is the preceding estimate obtained using only the first N_1 observations. Then

$$\begin{aligned}\hat{E}[\mathbf{y}_{\text{mis}}|\mathbf{X}_2] &= \hat{\mathbf{y}}_{\text{mis}} = \mathbf{X}_2 \hat{\beta}, \\ \hat{V}[\hat{\mathbf{y}}_{\text{mis}}] &\equiv \hat{V}[\hat{\mathbf{y}}|\mathbf{X}_2] = s^2(\mathbf{I}_{N_2} + \mathbf{X}_2 [\mathbf{X}'_1 \mathbf{X}_1]^{-1} \mathbf{X}'_2),\end{aligned}\tag{27.8}$$

where $s^2 \mathbf{I}_{N_2}$ is an estimate of $\mathbf{V}_{[\mathbf{u}_2]}$.

In the naive method one would generate the N_2 predicted values of \mathbf{y}_{mis} , and then apply standard regression methods to the full sample of $N = N_1 + N_2$ observations.

The two steps in the naive method correspond to the two steps of the **EM algorithm**. The prediction step is the **E-step**, and the second-step application of least squares to the augmented sample is the **M-step**.

However, this solution has flaws. First, consider the data augmentation step. Because the generated values $\hat{\mathbf{y}}_{\text{mis}}$ lie *exactly* on the least-squares fitted plane, the addition of $(\hat{\mathbf{y}}_{\text{mis}}, \mathbf{X}_2)$ to the sample to produce a new estimate, $\hat{\beta}_A$, does not change the previous estimate $\hat{\beta}$:

$$\begin{aligned}\hat{\beta}_A &= [\mathbf{X}'_1 \mathbf{X}_1 + \mathbf{X}'_2 \mathbf{X}_2]^{-1} [\mathbf{X}'_1 \mathbf{y}_1 + \mathbf{X}'_2 \hat{\mathbf{y}}_{\text{mis}}] \\ &= [\mathbf{X}'_1 \mathbf{X}_1 + \mathbf{X}'_2 \mathbf{X}_2]^{-1} [\mathbf{X}'_1 \mathbf{X}_1 \hat{\beta} + \mathbf{X}'_2 \mathbf{X}_2 \hat{\beta}] \\ &= \hat{\beta}.\end{aligned}$$

Second, the estimate of σ^2 obtained by the standard formula to the residuals from the augmented sample yields an estimate that is too small because the additional N_2 residuals are zero by construction,

$$\begin{aligned}s_A^2 &= (\mathbf{y} - \mathbf{X} \hat{\beta}_A)'(\mathbf{y} - \mathbf{X} \hat{\beta}_A)/N \\ &= (\mathbf{y}_1 - \mathbf{X}_1 \hat{\beta})'(\mathbf{y}_1 - \mathbf{X}_1 \hat{\beta})/N < s^2\end{aligned}\tag{27.9}$$

where s^2 correctly divides by N_1 rather than N .

Finally, as can be seen from the expression for the sampling variance of $\hat{\mathbf{y}}_{\text{mis}}$, the generated predictions are heteroskedastic, unlike the \mathbf{y}_1 , and hence the variance of $\hat{\beta}_A$ cannot be estimated using the least-squares formula in the usual way. The observations $\hat{\mathbf{y}}_{\text{mis}}$ are draws from a distribution with a different variance. The naive method does not make allowance for the uncertainty attached to the estimates of $\hat{\mathbf{y}}_{\text{mis}}$.

To fix these problems modifications are needed. First, the estimation of $\widehat{\mathbf{y}}_{\text{mis}}$ should take account of uncertainty regarding $\widehat{\beta}$. This may be done by adjusting $\widehat{\mathbf{y}}_{\text{mis}}$ and adding some “noise” to the generated predictions such that the estimates of missing data more closely mimic a draw from the (estimated or conditional) distribution of \mathbf{y}_1 . A standardization step can use the fact that an estimate of $V[\widehat{\mathbf{y}}_{\text{mis}}]$, $\widehat{\mathbf{V}}$, is available from (27.8). Hence the components of the transformed variable $\widehat{\mathbf{V}}^{-1/2}\widehat{\mathbf{y}}_{\text{mis}}$ have unit variance. To mimic the distribution of \mathbf{y}_1 , we can make a Monte Carlo draw from $\mathcal{N}[0, s^2]$ distribution and multiply it by $\widehat{\mathbf{V}}^{-1/2}\widehat{\mathbf{y}}_{\text{mis}}$.

The revised algorithm is as follows.

1. Estimate $\widehat{\beta}$ using the N_1 complete observations as before.
2. Generate $\widehat{\mathbf{y}}_{\text{mis}} = \mathbf{X}_2\widehat{\beta}_2$.
3. Generate adjusted values of $\widehat{\mathbf{y}}_{\text{mis}}^a = (\widehat{\mathbf{V}}^{-1/2}\widehat{\mathbf{y}}_{\text{mis}}) \odot \mathbf{u}_m$ of $\widehat{\mathbf{y}}_{\text{mis}}$, where \mathbf{u}_m is a Monte Carlo draw from the $\mathcal{N}[0, s^2]$ distribution and \odot denotes element-by-element multiplication.
4. Using the augmented sample obtain a revised estimate of $\widehat{\beta}$.
5. Repeat steps 1–4 where in step 1 the revised estimate of $\widehat{\beta}$ is used.

The revised algorithm, also an EM-type algorithm, continues until it converges in the sense that the changes in the coefficients or the changes in regression residual sum of squares become arbitrarily small.

To make connection with later discussion we give the algorithm a different interpretation. Step 3 is a draw from the conditional distribution of \mathbf{y} given β , and step 4 is a draw from the conditional distribution of β given s^2 , \mathbf{X} . The approach may be refined further by adding a step that involves a draw from the distribution of s^2 . We do not go through all the steps of this approach because they will become clearer in our later discussion of imputation.

Alternative models for missing data on the dependent variable were presented in Chapter 16. These relaxed the MAR assumption and specified nonignorable missingness. Then the preceding EM algorithm leads to inconsistent estimation of β . The censored Tobit model specifies that data are missing for observations with $\mathbf{x}'\beta + u \leq 0$ and a consistent estimator is the Tobit MLE (see Section 16.3). Amemiya (1985, pp. 376–378) details the EM algorithm for the Tobit model.

27.6. Data Augmentation and MCMC

The general structure of the Bayesian approach to missing data is to use the following type of iterative algorithm that uses imputation and prediction steps.

The **imputation step (I-step)** makes a draw from the conditional predictive distribution of \mathbf{Y}_{mis} . Given an r th round estimate,

$$\mathbf{Y}_{\text{mis}}^{(r+1)} \sim \Pr[\mathbf{Y}_{\text{mis}} | \mathbf{Y}_{\text{obs}}, \theta^{(r)}]. \quad (27.10)$$

This expression denotes a random draw of $\mathbf{Y}_{\text{mis}}^{(r+1)}$ from the predictive conditional distribution of \mathbf{Y}_{mis} given the current estimate $\theta^{(r)}$ and the observed data \mathbf{Y}_{obs} . Notice that \mathbf{Y}_{mis} is in general a matrix so that this notation refers to (in principle) a series of draws.

The **prediction step (P-step)** is executed by making a draw from the **complete data posterior**

$$\boldsymbol{\theta}^{(r+1)} \sim \Pr[\boldsymbol{\theta} | \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}^{(r+1)}]. \quad (27.11)$$

That is, \mathbf{Y}_{obs} is augmented by an imputed value $\mathbf{Y}_{\text{mis}}^{(r+1)}$ drawn from the predictive distribution of \mathbf{Y}_{mis} , and a draw is made from the posterior distribution of $\boldsymbol{\theta}$. The steps (27.10) and (27.11) can then be repeated.

Sequential sampling from the two distributions generates a Markov chain. This process, which strongly resembles the EM algorithm, is essentially the Gibbs sampler of Section 13.5.2, but in the missing data literature it is referred to as **data augmentation**. Under appropriate conditions, and by a theorem cited in Section 13.5.1, the sequential draws will converge to a stationary distribution for a sufficiently large value of r , which is the length of the chain. When the chain is terminated we have one imputation of \mathbf{Y}_{mis} . Then we can regard $\boldsymbol{\theta}^{(r)}$ as an approximate draw from $\Pr[\boldsymbol{\theta} | \mathbf{Y}_{\text{obs}}]$ and $\mathbf{Y}_{\text{mis}}^{(r+1)}$ as an approximate draw from $\Pr[\mathbf{Y}_{\text{mis}} | \mathbf{Y}_{\text{obs}}]$. As with any MCMC application the chain has to run sufficiently long to ensure that successive imputations are free of statistical dependence. These issues have been discussed in Chapter 13.

After convergence we would have accomplished the joint objectives of imputing the missing values based on the model specified for the data and estimating the model using both observed and imputed values. Postconvergence we would have the data necessary to compute the posterior moments of $\boldsymbol{\theta}$ and any interesting functions of $\boldsymbol{\theta}$ and \mathbf{Y} using the ideas discussed in Chapter 13.

As a specific illustration of this procedure we reconsider the missing data regression example of the previous section. The steps in the MCMC algorithm are as follows:

1. Using observed data calculate $\widehat{\boldsymbol{\beta}} = [\mathbf{X}'_1 \mathbf{X}_1]^{-1} \mathbf{X}'_1 \mathbf{y}_1$, and $\widehat{\mathbf{u}} = (\mathbf{y}_1 - \mathbf{X}_1 \widehat{\boldsymbol{\beta}})$.
2. Generate σ^2 as $\widehat{\mathbf{u}}'\widehat{\mathbf{u}}/\sigma^2$ divided by a draw from $\chi^2_{N_1-K}$ distribution.
3. Draw $\boldsymbol{\beta}|\sigma^2 \sim \mathcal{N}[\widehat{\boldsymbol{\beta}}, \sigma^2 [\mathbf{X}'_1 \mathbf{X}_1]^{-1}]$.
4. Draw $\mathbf{y}_{\text{mis}} \sim \mathcal{N}[\mathbf{X}_2 \widehat{\boldsymbol{\beta}}, \sigma^2]$.
5. Using \mathbf{y} instead of \mathbf{y}_1 , and \mathbf{X} instead of \mathbf{X}_1 , repeat steps 1–4 after appropriate adjustments.

The justification for step 2 is that, under an uninformative prior for $(\boldsymbol{\beta}, \sigma^2)$, the conditional posterior distribution of $\widehat{\mathbf{u}}'\widehat{\mathbf{u}}/\sigma^2$ is $\chi^2_{N_1-K}$ if only the observed data are used. After data augmentation this changes to χ^2_{N-K} . The justification for step 3 is that, under an uninformative prior, the conditional posterior distribution is $\mathcal{N}[\widehat{\boldsymbol{\beta}}, \sigma^2 [\mathbf{X}'_1 \mathbf{X}_1]^{-1}]$. After data augmentation this changes to $\mathcal{N}[\widehat{\boldsymbol{\beta}}, \sigma^2 [\mathbf{X}' \mathbf{X}]^{-1}]$. Step 4 is the imputation step using the conditional predictive density $\mathcal{N}[\mathbf{X}_2 \widehat{\boldsymbol{\beta}}, \sigma^2]$. These steps can be appropriately modified if we use, for example, an informative normal–gamma prior for $(\boldsymbol{\beta}, \sigma^2)$. The conditional posterior distributions for this case are given in Section 13.3.

27.7. Multiple Imputation

The analysis of the preceding section explains how a full MCMC run will generate a single imputation. However, a single imputation does not adequately handle the missing-data uncertainty. This is the essential rationale for using a multiple imputation procedure. The conditional predictive distribution of $\mathbf{Y}_{\text{mis}} | \mathbf{Y}_{\text{obs}}, \boldsymbol{\theta}$ is obtained by averaging over the observed-data posterior of $\boldsymbol{\theta}$:

$$\Pr[\mathbf{Y}_{\text{mis}} | \mathbf{Y}_{\text{obs}}] = \int \Pr[\mathbf{Y}_{\text{mis}} | \mathbf{Y}_{\text{obs}}, \boldsymbol{\theta}] \Pr[\boldsymbol{\theta} | \mathbf{Y}_{\text{obs}}] d\boldsymbol{\theta}.$$

Proper multiple imputations from a Bayesian viewpoint reflect uncertainty about \mathbf{Y}_{mis} , given the uncertainty about parameters of the model.

After **multiple imputation** the missing data \mathbf{Y}_{mis} are replaced by simulated/imputed values $\mathbf{Y}_{\text{mis}}^{(1)}, \mathbf{Y}_{\text{mis}}^{(2)}, \mathbf{Y}_{\text{mis}}^{(3)}, \dots, \mathbf{Y}_{\text{mis}}^{(m)}$. Each of the complete data sets is then analyzed as if it were complete. The results from the m analyses will show variation that reflects the uncertainty resulting from the missing data. With m different data sets questions arise about how one should determine an appropriate value for m and how the m sets of parameter estimates and covariance matrices should be combined. We address both of these questions using results from the literature but without providing detailed justification.

In considering how to combine the results based on multiply imputed data the key result, stated for an arbitrary statistic Q , is

$$\Pr[Q | \mathbf{Y}_{\text{obs}}] = \int \Pr[Q | \mathbf{Y}_{\text{mis}}, \mathbf{Y}_{\text{obs}}] \Pr[\mathbf{Y}_{\text{mis}} | \mathbf{Y}_{\text{obs}}] d\mathbf{Y}_{\text{mis}}, \quad (27.12)$$

which states that the actual posterior distribution of Q , is obtained by averaging over the complete-data posterior distribution of Q . This means averaging over the results of multiple imputations of missing observations (Rubin, 1996).

Equation (27.12) implies that the final estimate of Q is given by the law of iterated expectations,

$$E[Q | \mathbf{Y}_{\text{obs}}] = E[E[Q | \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}] | \mathbf{Y}_{\text{obs}}]. \quad (27.13)$$

The posterior mean of Q is the average of Q_r using complete data after repeated imputation of missing data.

The final variance of Q is given by the formula

$$V[Q | \mathbf{Y}_{\text{obs}}] = E[V[Q | \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}] | \mathbf{Y}_{\text{obs}}] + V[E[Q | \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}] | \mathbf{Y}_{\text{obs}}], \quad (27.14)$$

using the variance decomposition formula given in Section A.8.

Rubin (1996) also gives the following rules for combining moment information, stated in terms of a scalar parameter. For an arbitrary scalar parameter, suppose \widehat{Q}_r is a point estimate at the r th imputation and \widehat{U}_r is a variance estimate. Then define the

Table 27.1. *Relative Efficiency of Multiple Imputation*

Number of Imputations (m)	Observations Missing (λ)		
	10%	30%	50%
3	0.967	0.909	0.857
10	0.990	0.970	0.952
20	0.995	0.985	0.975

averages of the point and variance estimate, respectively, as

$$\bar{Q} = m^{-1} \sum_{r=1}^m \hat{Q}_r, \quad (27.15)$$

$$\bar{U} = m^{-1} \sum_{r=1}^m \hat{U}_r \quad (27.16)$$

and the between-imputation variance as

$$B = (m - 1)^{-1} \sum_{r=1}^m (\hat{Q}_r - \bar{Q})^2 \quad (27.17)$$

and the total variance as

$$T = \bar{U} + (1 + m^{-1}) B. \quad (27.18)$$

The results (27.15, 27.16) follow from (27.13); Equation (27.18) follows from (27.14). Schafer (1997) gives results for combining p -values and likelihood ratio statistics and provides additional references.

Postimputation inference regarding individual coefficients or subsets of coefficients can be carried out using the final estimates, since the standard central limit theorem and the associated large-sample results can be extended to cover this case.

The following is a measure of the relative efficiency of m multiple imputations:

$$reff = (1 + (\lambda/m))^{-1}, \quad (27.19)$$

where λ is the fraction of missing observations. Efficiency is measured relative to no missing data. The arithmetical calculations in Table 27.1 show that with as few as three imputations the efficiency can be as high as 97% with 10% missing data, and 86% with 50% missing data. With 10 or more imputations the relative efficiency exceeds 95% with 50% missing data. Thus, as emphasized by Schafer (1997), the number of imputations need not be very high.

27.8. Missing Data MCMC Imputation Example

This section gives two illustrative applications of missing data imputation: the model-free methods of listwise deletion and mean imputation (see Section 27.2), and the

model-based method of data augmentation using the MCMC algorithm (see Section 27.6). Only data on regressors are missing and the missing mechanism is MAR.

The first application involves simple multiple regression, and the second involves a logit regression. For clarity and simplicity we use artificially generated data with a known dgp.

27.8.1. Linear Regression with Missing Data on Regressors

For the linear regression example the dgp is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i, \quad i = 1, 2, \dots, N, \quad (27.20)$$

with $u_i | x_{1i}, x_{2i} \sim \mathcal{N}[0, \sigma^2]$ and (x_{1i}, x_{2i}) bivariate normally distributed with

$$\begin{bmatrix} x_{1i} \\ x_{2i} \end{bmatrix} \sim \mathcal{N}\left[\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right], \quad (27.21)$$

so that $x_{2i} | x_{1i} \sim \mathcal{N}[\rho x_{1i}, 1 - \rho^2]$. Also, we set $\beta' = [1 \ 1 \ 1]$, $N = 1,000$, and the proportion of randomly missing data on x_1 and x_2 to either 10% or 25%. For any i , either x_1 or x_2 , or both, may be missing. We also use two different values of ρ , 0.36 and 0.64.

For the Markov chain we use 500 iterations for the burn-in phase. The Markov chain calculations are implemented using the SAS MI Proc algorithm, which uses an uninformative prior. For demonstration purposes only, the number of imputations is fixed at 10 but the length of the chain after the burn-in phase varies from 10 to 10,000. Proc MI combines the results from multiple imputations using Equations (27.15)–(27.18).

Tables 27.2 and 27.3 present results for high ρ and low and high rates of missing data. There are no dramatic differences among methods. Because the MAR assumption applies, point estimates from listwise deletion and the full sample remain close, but as expected the standard errors are larger under listwise deletion. Under mean imputation the point estimate of β_2 diverges relatively more, but the observed variation is well within the bounds of sampling error. It appears that in both cases the Markov chain attains stationarity rather rapidly, there being very little difference between the

Table 27.2. Missing Data Imputation: Linear Regression Estimates with 10% Missing Data and High Correlation Using MCMC Algorithm

No Data Missing	Listwise Deletion	Mean Impute	Length of the Markov Chain			
			10	1,000	5,000	10,000
$\hat{\beta}_0$	0.919 (0.104)	0.913 (0.113)	0.899 (0.105)	0.910 (0.102)	0.911 (0.101)	0.909 (0.103)
$\hat{\beta}_1$	1.097 (0.138)	1.067 (0.151)	1.053 (0.141)	1.196 (0.148)	1.205 (0.155)	1.199 (0.144)
$\hat{\beta}_2$	1.000 (0.132)	1.072 (0.145)	1.112 (0.135)	1.042 (0.140)	1.051 (0.146)	1.041 (0.143)
R^2	0.240	0.254	0.226			

Table 27.3. Missing Data Imputation: Linear Regression Estimates with 25% Missing Data and High Correlation Using MCMC Algorithm

No Data Missing	Listwise Deletion	Mean Impute	Length of the Markov Chain			
			10	1,000	5,000	10,000
$\hat{\beta}_0$	0.919 (0.104)	0.863 (0.167)	0.984 (0.108)	0.899 (0.108)	0.898 (0.105)	0.925 (0.111)
$\hat{\beta}_1$	1.097 (0.138)	1.048 (0.167)	1.062 (0.150)	1.028 (0.152)	1.047 (0.166)	1.082 (0.161)
$\hat{\beta}_2$	1.000 (0.132)	1.129 (0.161)	1.156 (0.148)	1.071 (0.152)	1.085 (0.144)	1.024 (0.172)
R^2	0.240	0.268	0.203			

results with 10 and 10,000 iterations. This is probably due to having set the number of burn-in iterations at 500, which may be higher than needed for this relatively simple case.

In Table 27.4 the simulation exercise is repeated for the “worst-case” scenario of low ρ and 25% missing data. The divergence between the point estimates from the full sample and those from listwise deletion and mean imputation cases is overall relatively greater than that for the MCMC cases. However, even in this case there are no really dramatic differences between estimates from the full sample. Once again we see that the benefit of running a long Markov chain are not apparent in this example.

27.8.2. Logit Regression with Missing Data on Regressors

We next consider an example of a nonlinear model with missing data on regressors using simulated data. In this simulation example we retain the dgp given before but change the dependent variable into a discrete dichotomous variable. First, reinterpret

Table 27.4. Missing Data Imputation: Linear Regression Estimates with 10% Missing Data and Low Correlation Using MCMC Algorithm

No Data Missing	Listwise Deletion	Mean Impute	Length of the Markov Chain			
			10	1,000	5,000	10,000
$\hat{\beta}_0$	1.121 (0.099)	1.162 (0.130)	1.142 (0.103)	1.149 (0.104)	1.155 (0.103)	1.154 (0.104)
$\hat{\beta}_1$	1.099 (0.107)	0.930 (0.134)	1.052 (0.121)	1.026 (0.127)	1.020 (0.128)	1.004 (0.124)
$\hat{\beta}_2$	1.102 (0.107)	1.122 (0.134)	1.215 (0.124)	1.130 (0.128)	1.157 (0.129)	1.137 (0.129)
R^2	0.243	0.235	0.186			

Table 27.5. Missing Data Imputation: Logistic Regression Estimates with 10% Missing Data and High Correlation Using MCMC Algorithm

No Data Missing	Listwise Deletion	Mean Impute	Length of the Markov Chain			
			10	1,000	5,000	10,000
$\hat{\beta}_0$	-0.447 (0.070)	-0.498 (0.078)	-0.439 (0.070)	-0.527 (0.073)	-0.534 (0.073)	-0.531 (0.072)
$\hat{\beta}_1$	-0.597 (0.096)	-0.658 (0.108)	-0.602 (0.098)	-0.620 (0.106)	-0.673 (0.102)	-0.681 (0.101)
$\hat{\beta}_2$	-0.444 (0.092)	-0.474 (0.103)	-0.523 (0.094)	-0.597 (0.107)	-0.540 (0.103)	-0.536 (0.099)

the simulation design given for the linear regression example, so $y = y^*$, a latent variable. Let the dgp be

$$y_i^* = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i, \quad i = 1, 2, \dots, N. \quad (27.22)$$

Then a dichotomous y_i is generated according to the following rule:

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0, \\ 0 & \text{if } y_i^* \leq 0. \end{cases} \quad (27.23)$$

We will model the probability that $y_i = 0$ using the logit model, even though the dgp is that for the probit model. As discussed in Section 14.4.1, the logit model identifies the parameter vector β/σ , where the variance $\sigma^2 = \pi^2/3$. With all elements of β set equal to one, the logit model will provide estimates of the true parameter value of approximately -0.551. The MCMC estimation is set up as before with a noninformative prior.

Tables 27.5 covers the favorable case with 10% missing data and high correlation between x_1 and x_2 , and Table 27.6 covers the less favorable case with 25% missing data and low correlation between x_1 and x_2 .

In the first case, even with no missing data the estimate $\hat{\beta}_2$ is substantially off its expected value. The MCMC point estimates change somewhat when the length of the Markov chain is increased from 10 to 1,000. However, more when simulations are implemented, there is only slight change in point estimates, a result that we can interpret as an indication of convergence of the chain to its stationary distribution.

For the second example involving a less favorable simulation design, the results are as shown in Table 27.6. The main difference is that the divergence between the expected point estimates and the estimated values is somewhat larger for the previous case. However, broadly speaking the performance of the multiple imputation method in the logistic regression is similar to that in the linear regression.

Table 27.6. Missing Data Imputation: Logistic Regression Estimates with 25% Missing Data and Low Correlation Using MCMC Algorithm

No Data Missing	Listwise Deletion	Mean Impute	Length of the Markov Chain			
			10	1,000	5,000	10,000
$\hat{\beta}_0$	-0.447 (0.070)	-0.658 (0.097)	-0.582 (0.070)	-0.605 (0.074)	-0.609 (0.074)	-0.609 (0.073)
$\hat{\beta}_1$	-0.597 (0.096)	-0.434 (0.100)	-0.470 (0.085)	-0.447 (0.090)	-0.470 (0.094)	-0.471 (0.094)
$\hat{\beta}_2$	-0.444 (0.092)	-0.593 (0.108)	-0.648 (0.089)	-0.634 (0.084)	-0.615 (0.086)	-0.576 (0.086)

27.9. Practical Considerations

A major implication of the analysis of this chapter for practice is that analysis of multiply, rather than singly, imputed data has theoretical advantages. Moreover, model-based approaches are less ad hoc than mechanical approaches such as mean imputation or hot deck. In many realistic applications devising an MCMC-type imputation procedure may pose a significant challenge, however, compared to the relative simplicity of the examples given in the last section.

A distinction may be drawn between multiple imputations where the end product is the data and one in which the end product consists of estimated coefficients for inference. Although both procedures may be model based the second may involve more complex econometric models. Examples are provided by Brownstone and Valetta (1996), Stinebrinkner (1999), Kennickell (1998), and Davey, Shanahan, and Schafer (2001).

Even when the primary object is imputation, without extensive modeling the problem may be far from simple. For example, in his study of the 1995 Survey of Consumer Finances, Kennickell (1998, p. 5) remarks:

[When] the survey contains a very large number of variables, there is substantial missing or partially missing (range) information, the patterns of missing information are highly heterogeneous, the distributions of many of the variables are highly skewed, and the data have a complex structure, [then], analysis of the survey in the absence of imputation would be a formidable task. Moreover, anyone using the public version of the data set would lack key frame data that turn out to be important for understanding the distributions of the missing data. Thus, even on pure efficiency grounds, there is a good case for imputing the missing data.

Despite the complexity of the problem Kennickell was able to use imputation procedures similar to those discussed in this chapter.

Stinebrinkner (1999), also facing a missing data situation in which listwise deletion “would leave the econometrician with too little data to estimate the model of interest,” develops a two-stage simulated likelihood-based procedure for estimating the joint

distribution of the missing data and estimating duration model for the first teaching spell.

For relatively simple cases software such as the SAS package Proc MI may be used. S-Plus and SOLAS also provide software support. A helpful guide and survey of computer software packages is given in Horton and Lipsitz (2001). For additional information see the relevant Web sites.

Most of the analysis of the chapter is based on assuming an ignorable missing data mechanism. From an econometric viewpoint this might be a major simplification. For example, see Lillard, Smith, and Welch (1986), who critique the Census hot deck method for imputing missing wages. How should one proceed if the mechanism is nonignorable? In the notation of Section 27.4, a nonignorable missing data mechanism would imply that parameters θ and ψ are not distinct. Then one must specify the missing data mechanism explicitly, as in the case of **selection models** and models of **attrition bias** (see Chapter 16 and Section 23.5.2). Schafer (1997, p. 28) provides some relevant references to the literature.

27.10. Bibliographic Notes

Important early references include Little and Rubin (1987) and Rubin (1987). Allison (2002) provides a relatively nontechnical but lucid introduction to the missing data problem and literature. Rubin (1996) provides a survey with historical perspective. Schafer (1997) provides a more complete analysis that covers categorical data, mixed discrete–continuous data, and data from complex surveys.

- 27.2** Meng (2000) provides a historical perspective on the missing data mechanism.
27.5 Little (1988, 1992) provides a good review of the literature on linear regression with missing regressors, covering both non-model-based and model-based approaches.

Exercises

- 27–1** Consider any regression model, linear or nonlinear, with dependent variable y and exogenous variables \mathbf{x} , and iid errors ε . Show that if the probability of missing data on \mathbf{x} is independent of y , then the regression based on listwise deletion will provide a consistent estimate of the conditional mean function. [Hint: Show that the conditional distribution of y given \mathbf{x} is not affected by missing observations.]
- 27–2** (Adapted from Gouriéroux and Monfort, 1981). Consider the regression model $\mathbf{y} = \beta_1 \mathbf{x} + \mathbf{Z}\beta_2 + \mathbf{u}$, where y is an $N \times 1$ vector, \mathbf{Z} is an $N \times K$ matrix, and \mathbf{x} is an $N \times 1$ vector of a scalar regressor, some of whose elements are missing. Assume that observations are missing at random and $E[\mathbf{u}|\mathbf{x}, \mathbf{Z}] = \mathbf{0}$ and $E[\mathbf{u}\mathbf{u}'|\mathbf{x}, \mathbf{Z}] = \sigma^2 \mathbf{I}_N$. Both \mathbf{y} and \mathbf{Z} are fully observed. The following approach is proposed to deal with the missing data. Assume a linear regression model relating \mathbf{x} to \mathbf{Z} , $\mathbf{x} = \mathbf{Z}\gamma + \varepsilon$, where $E[\varepsilon|\mathbf{Z}] = \mathbf{0}$ and $E[\varepsilon\varepsilon'|\mathbf{Z}] = \sigma_\varepsilon^2 \mathbf{I}_N$. Then let $\hat{\gamma} = [\mathbf{Z}'\mathbf{Z}_c]^{-1}\mathbf{Z}'\mathbf{x}_c$, where the subscript c refers to “complete data.” Impute values of $\hat{\mathbf{x}}_m = \mathbf{Z}_m[\mathbf{Z}'_c\mathbf{Z}_c]^{-1}\mathbf{Z}'_c\mathbf{x}_c$, where \mathbf{x}_m refers to the missing observations and \mathbf{Z}_m to the corresponding values of \mathbf{Z} . The original regression is then reestimated

using the full set of N observations after replacing the missing values of \mathbf{x} by imputed values.

- (a) Explain why the OLS regression estimator based on complete and imputed observations might be biased in finite samples.
 - (b) What additional assumptions are required to prove that the OLS estimator based on complete plus imputed values is consistent?
 - (c) Is the OLS estimator efficient?
- 27–3** Consider the point that when estimation of a model is undertaken after data imputation the precision of the estimates is likely to be overstated if no adjustment is made for the imputation step. In other words, imputed data may be regarded as generated variables and hence subject to the problem of the sequential two-step estimator discussed in Section 6.6. Explain whether an adjustment related to imputation of missing data is necessary asymptotically.

APPENDIX A

Asymptotic Theory

A.1. Introduction

In this appendix we consider the behavior of a **sequence of random variables** b_N as $N \rightarrow \infty$.

In applications the index N is the sample size and the sequence b_N is an estimator, such as $\hat{\beta}$ or $\hat{\theta}$, or a component of an estimator, such as $N^{-1} \sum_i x_i^2$ or $N^{-1} \sum_i x_i u_i$ in the case of OLS with one regressor and no intercept, or a test statistic.

For estimation theory it is sufficient to focus on two aspects of the behavior of the **sequence** b_N as $N \rightarrow \infty$. First, we consider **convergence in probability** of b_N to a limit b , a constant or random variable that is very close to b_N in a probabilistic sense defined in the following. Second, if the limit b is a random variable, which may require a rescaling of the original sequence, we consider the **limit distribution**.

Estimators are usually functions of **averages** or **sums**. Then it is easiest to derive limiting results by invoking results on the behavior of averages, notably **laws of large numbers** and **central limit theorems**. The notation used is to consider an average $\bar{X}_N = N^{-1} \sum_i X_i$, where X_i here is generic notation for a random variable being averaged and should not be confused with the use of \mathbf{x}_i to denote the regressor vector. For example, for OLS with one regressor and no intercept we will apply a law of large numbers to the average of $X_i = x_i^2$ and a central limit theorem to the average of $X_i = x_i u_i$.

Table A.1 summarizes the definitions and theorems presented in the remainder of this appendix. These are stated without proof but with some discussion. The focus is on results used to obtain asymptotically normal estimators, the usual case when cross-section data are used. Additional results are needed for application to nonparametric estimation, to parametric estimation when the support of the data depends on parameters, and to time series estimation when data have unit roots.

The first key concept, convergence in probability, is presented in Section A.2. This is established using laws of large numbers given in Section A.3. The other key concept, convergence in distribution, is presented in Section A.4. Convergence to the normal distribution is established using central limit theorems given in Section A.5. Further results and common terminology for limit multivariate normal distributions are given

Table A.1. Asymptotic Theory: Definitions and Theorems

Definition	Theorem	Name	Equation
A.1		Convergence in Probability	(A.1)
A.2		Consistency	(A.2)
	A.3	Slutsky	(A.3)
A.4		Mean-Square Convergence	(A.4)
	A.5	Chebychev's Inequality	(A.5)
A.6		Almost Sure Convergence	(A.6)
A.7		Law of Large Numbers	(A.7)
	A.8	Kolmogorov LLN	
	A.9	Markov LLN	
A.10		Convergence in Distribution	(A.9)
	A.11	Continuous Mapping	(A.10)
	A.12	Transformation	(A.11)
A.13		Central Limit Theorem	(A.13)
	A.14	Lindeberg–Levy CLT	
	A.15	Liapounov CLT	
	A.16	Cramer–Wold Device	
	A.17	Limit Normal Product Rule	(A.15)
A.18		Asymptotic Distribution	(A.17)
A.19		Asymptotic Variance	(A.18)
A.20		Estimated Asymptotic Variance	(A.19)
A.21		Asymptotic Efficiency	
A.22		Stochastic Order of Magnitude	

in Section A.6. Stochastic order of magnitude, a convenient notation commonly used in asymptotic analysis, is presented in Section A.7. Section A.8 presents some useful properties of expectations.

A.2. Convergence in Probability

Because of the intrinsic randomness of a sample we can never be certain that a sequence b_N , such as an estimator $\hat{\theta}$ (often denoted $\hat{\theta}_N$ to make clear that it is a sequence), will be within a given small distance of its limit, even if the sample is infinitely large. However, we can be almost certain. Different ways of expressing this near certainty correspond to different types of convergence of a sequence of random variables to a limit. The one most used in econometrics is convergence in probability.

A.2.1. Convergence in Probability

Recall that a sequence of nonstochastic real numbers $\{a_N\}$ converges to a if, for any $\varepsilon > 0$, there exists $N^* = N^*(\varepsilon)$ such that, for all $N > N^*$,

$$|a_N - a| < \varepsilon.$$

For example, if $a_N = 2 + 3/N$, then the limit is $a = 2$ since $|a_N - a| = |2 + 3/N - 2| = |3/N| < \varepsilon$ for all $N > N^* = 3/\varepsilon$.

When more generally we have a sequence of random variables we cannot be certain of being within ε of the limit, even for large N , because of intrinsic randomness. Instead, we require that the probability of being within ε is arbitrarily close to one. Thus we require

$$\lim_{N \rightarrow \infty} \Pr[|b_N - b| < \varepsilon] = 1,$$

for any $\varepsilon > 0$. A formal definition is the following:

Definition A.1 (Convergence in Probability): A sequence of random variables $\{b_N\}$ **converges in probability** to b if, for any $\varepsilon > 0$ and $\delta > 0$, there exists $N^* = N^*(\varepsilon, \delta)$ such that, for all $N > N^*$,

$$\Pr[|b_N - b| < \varepsilon] > 1 - \delta. \quad (\text{A.1})$$

We write $\text{plim } b_N = b$, where plim is shorthand for **probability limit**, or $b_N \xrightarrow{P} b$.

Note that b may be a constant or a random variable. Convergence in probability includes as a special case the usual definition of convergence for a sequence of real variables.

Definition A.1 is for a sequence of *scalar* random variables. The extension to **vector random variables**, such as a parameter vector estimator, is straightforward. We can either apply the theory for each element of \mathbf{b}_N , or replace $|b_N - b|$ by the scalar $(\mathbf{b}_N - \mathbf{b})'(\mathbf{b}_N - \mathbf{b}) = (b_{1N} - b_1)^2 + \dots + (b_{KN} - b_K)^2$ or its square root $\|\mathbf{b}_N - \mathbf{b}\|$.

When the sequence $\{\mathbf{b}_N\}$ is a sequence of parameter estimates $\hat{\boldsymbol{\theta}}$, we have the following large sample analogue of unbiasedness.

Definition A.2 (Consistency): An estimator $\hat{\boldsymbol{\theta}}$ is **consistent** for $\boldsymbol{\theta}_0$ if

$$\text{plim } \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0. \quad (\text{A.2})$$

The subscript 0 on $\boldsymbol{\theta}$ is explained in Section 5.2.3. Note that unbiasedness need not imply consistency. Unbiasedness states only that the expected value of $\hat{\boldsymbol{\theta}}$ is $\boldsymbol{\theta}_0$, and it permits variability around $\boldsymbol{\theta}_0$ that need not disappear as the sample size goes to infinity. Also, a consistent estimator need not be unbiased. For example, adding $1/N$ to an unbiased and consistent estimator produces a new estimator that is biased but still consistent.

Although the sequence of vector random variables $\{\mathbf{b}_N\}$ may converge to a random variable \mathbf{b} , in many econometric applications $\{\mathbf{b}_N\}$ converges to a constant. For example, we hope that an estimator of a parameter will converge in probability to the parameter itself. One should be aware that some of the results that follow apply only if the limit value \mathbf{b} is a constant.

Theorem A.3 (Slutsky's Theorem): Let \mathbf{b}_N be a finite-dimensional vector of random variables, and $g(\cdot)$ be a real-valued function continuous at a constant

vector point \mathbf{b} . Then

$$\mathbf{b}_N \xrightarrow{p} \mathbf{b} \Rightarrow g(\mathbf{b}_N) \xrightarrow{p} g(\mathbf{b}). \quad (\text{A.3})$$

Proof is given in Amemiya (1985, p. 79). Ruud (2000) presents a related result (see also Rao, 1973, p. 124) that lets the limit \mathbf{b} be a random variable, at the expense of restricting $g(\cdot)$ to be continuous everywhere. Note that some authors instead refer to Theorem A.12 below as Slutsky's Theorem.

Theorem A.3 is one of the major reasons for the prevalence of asymptotic results versus finite-sample results in econometrics. It states a very convenient property that does not hold for expectations. For example, $\text{plim}(b_{1N}, b_{2N}) = (b_1, b_2)$ implies $\text{plim}(b_{1N}b_{2N}) = b_1b_2$, whereas $E[b_{1N}b_{2N}]$ generally differs from $E[b_1]E[b_2]$.

A.2.2. Alternative Modes of Convergence

It is often easier to establish alternative modes of convergence, which in turn imply convergence in probability.

These alternative modes are given for completeness. Laws of large numbers, given in the next section, are used much more often.

Definition A.4 (Mean-Square Convergence): A sequence of random variables $\{b_N\}$ is said to **converge in mean square** to a random variable b if

$$\lim_{N \rightarrow \infty} E[(b_N - b)^2] = 0. \quad (\text{A.4})$$

We write $b_N \xrightarrow{m} b$. Convergence in mean square is useful because $b_N \xrightarrow{m} b$ implies $b_N \xrightarrow{p} b$ (see Rao, 1973, p. 110) and is often easy to prove. This does require existence of the variance of b_N , however. If $E[b_N] = b$, then we need to show that the variance of b_N goes to zero as $N \rightarrow \infty$. If b_N is instead biased for b then we require that the sum of the variance and bias squared goes to zero.

Another result that can be used to show convergence in probability is Chebychev's inequality.

Theorem A.5 (Chebyshev's Inequality): For any random variable Z with mean μ and variance σ^2 ,

$$\Pr[(Z - \mu)^2 > k] \leq \sigma^2/k, \quad \text{for any } k > 0. \quad (\text{A.5})$$

For a proof see Rao (1973, p. 95). The generalized Chebychev's inequality replaces $(Z - \mu)^2$ in Theorem A.5 by any nonnegative function $g(Z)$ and shows that $\Pr[g(Z) > k] \leq E[g(Z)]/k$, for any $k > 0$. See Amemiya (1985, p. 87).

Theorem A.5 can be used to verify convergence in probability by replacing Z with b_N . The theorem requires the mean and variance of b_N , which are easily obtained for estimators that involve an average of independent random variables. However, in such cases we can often take an even easier route and directly apply a law of large numbers to the average to obtain the probability limit.

A conceptually more difficult type of convergence is almost sure convergence.

Definition A.6 (Almost Sure Convergence): A sequence of random variables $\{b_N\}$ is said to **converge almost surely** to b if

$$\Pr[\lim_{N \rightarrow \infty} b_N = b]. \quad (\text{A.6})$$

This is denoted $b_N \xrightarrow{\text{as}} b$. Almost sure convergence implies convergence in probability (see Rao, 1973, p. 111). Convergence in probability allows more erratic behavior in b_N than does almost sure convergence.

Almost sure convergence is also called **strong consistency** for b , to distinguish it from convergence in probability, which is called **weak consistency** for b . Convergence in probability is easier to understand and is sufficient for most econometric applications.

A.3. Laws of Large Numbers

Laws of large numbers are theorems for *convergence in probability* (or almost surely) in the special case where the sequence $\{b_N\}$ is a *sample average*, that is, $b_N = \bar{X}_N$, where

$$\bar{X}_N = \frac{1}{N} \sum_{i=1}^N X_i. \quad (\text{A.7})$$

Note that X_i here is general notation for a random variable, and in the regression context it does not necessarily denote the regressor variables.

A law of large numbers provides a much easier way to establish the probability limit of a sequence $\{b_N\}$ than the alternatives of brute-force use of the (δ, ε) definition given in (A.1) or use of alternative modes of convergence that imply convergence in probability.

Definition A.7 (Law of Large Numbers): A **weak law of large numbers** (LLN) specifies conditions on the individual terms X_i in \bar{X}_N under which

$$(\bar{X}_N - \mathbb{E}[\bar{X}_N]) \xrightarrow{P} 0. \quad (\text{A.8})$$

For a **strong law of large numbers** the convergence is instead almost surely.

It can be helpful to think of a LLN as establishing that \bar{X}_N goes to its expected value, even though strictly speaking it implies the weaker condition that \bar{X}_N goes to the *limit of its expected value*, since (A.8) implies that

$$\text{plim } \bar{X}_N = \lim E[\bar{X}_N].$$

If the X_i have common mean μ , then this simplifies to $\text{plim } \bar{X}_N = \mu$.

Two leading examples of laws of large numbers are the following:

Theorem A.8 (Kolmogorov LLN): Let $\{X_i\}$ be iid (independent and identically distributed). If and only if $\mathbb{E}[X_i] = \mu$ exists and $\mathbb{E}[|X_i|] < \infty$, then $(\bar{X}_N - \mathbb{E}[\bar{X}_N]) \xrightarrow{\text{as}} 0$.

Theorem A.9 (Markov LLN): Let $\{X_i\}$ be iid (independent but not identically distributed) with $E[X_i] = \mu_i$ and $V[X_i] = \sigma_i^2$. If $\sum_{i=1}^{\infty} (E[|X_i - \mu_i|^{1+\delta}] / i^{1+\delta}) < \infty$, for some $\delta > 0$, then $(\bar{X}_N - E[\bar{X}_N]) \xrightarrow{as} 0$.

See White (2001a, p. 32 and p. 35) for statements of these theorems and Rao (1973, pp. 114–116) for proofs. Both laws give the stronger result of almost sure convergence, which implies the desired convergence in probability. Rao (1973) calls Theorem A.8 Kolmogorov LLN2 and presents Theorem A.9 for the special case $\delta = 1$, which he calls Kolmogorov LLN1.

The Kolmogorov LLN allows the variance of X_i to not even exist, at the expense of requiring an identical distribution. It simplifies to $\bar{X}_N \xrightarrow{as} \mu$, where $\mu = E[X]$. A weak version of this law, sufficient for most econometrics applications, is *Khinchine's Theorem*, which states that for $\{X_i\}$ iid the existence of $E[X]$ implies convergence in probability.

The Markov LLN no longer requires an identical distribution, but it does require existence of an absolute moment beyond the first. An obvious choice of δ is $\delta = 1$. Then the variance is needed and the side condition is that $\sum_{i=1}^{\infty} (\sigma_i^2 / i^2) < \infty$. The variance can vary and even grow with i , provided it does not grow so fast that (σ_i^2 / i^2) has infinite sum. The side condition is satisfied if $\sigma_i^2 = \sigma^2$, since $\sum_{i=1}^{\infty} 1/i^2$ converges, but is not satisfied if $\sigma_i^2 = i\sigma^2$, since $\sum_{i=1}^{\infty} 1/i$ diverges.

In most microeconomics applications, including regression with *stratified sampling* or with *fixed regressors*, the more complicated Markov LLN is needed.

Laws of large numbers are appealing because they require assumptions on the individual components X_i , rather than the sequence of averages \bar{X}_N . They are the most common way econometricians prove convergence in probability, since most estimators and test statistics are functions of averages of the data and unobserved random variables.

A.4. Convergence in Distribution

Given consistency, the estimator $\hat{\theta}$ has a *degenerate distribution* that collapses on θ_0 as $N \rightarrow \infty$. We need to magnify or *rescale* $\hat{\theta}$ to obtain a random variable that has *nondegenerate distribution* as $N \rightarrow \infty$. Often the appropriate scale factor is \sqrt{N} , in which case we consider the behavior of the sequence of random variables $b_N = \sqrt{N}(\hat{\theta} - \theta_0)$.

In general, the N th random variable in the sequence b_N has an extremely complicated cumulative distribution function (cdf) F_N . Like any other function F_N , this may have a limit function where convergence is in the usual mathematical sense.

Definition A.10 (Convergence in Distribution): A sequence of random variables $\{b_N\}$ is said to **converge in distribution** to a random variable b if

$$\lim_{N \rightarrow \infty} F_N = F, \quad (\text{A.9})$$

at every continuity point of F , where F_N is the distribution of b_N , F is the distribution of b , and convergence is in the usual mathematical sense.

We write $b_N \xrightarrow{d} b$, and we call F the **limit distribution** of $\{b_N\}$.

Convergence in probability implies convergence in distribution; that is, $b_N \xrightarrow{p} b$ implies $b_N \xrightarrow{d} b$ (see Rao, 1973, p. 122).

In general, the converse is not true. For example, let $b_N = X_N$, the N th realization of $X \sim \mathcal{N}[\mu, \sigma^2]$. Then $b_N \xrightarrow{d} b \sim \mathcal{N}[\mu, \sigma^2]$, but clearly $(b_N - b)$ has variance that does not disappear as $N \rightarrow \infty$, so b_N does not converge in probability to b .

In the special case where b is a constant, however, $b_N \xrightarrow{d} b$ implies $b_N \xrightarrow{p} b$ (see Rao, 1973, p. 120). In this case the limit distribution is degenerate, with all its mass at b .

To extend limit distribution to **vector random variables** simply define F_N and F to be the respective cdfs of vectors \mathbf{b}_N and \mathbf{b} .

Theorem A.11 (Continuous Mapping Theorem): *Let \mathbf{b}_N be a finite-dimensional vector of random variables, and let $g(\cdot)$ be a continuous real-valued function. Then*

$$\mathbf{b}_N \xrightarrow{d} \mathbf{b} \Rightarrow g(\mathbf{b}_N) \xrightarrow{d} g(\mathbf{b}). \quad (\text{A.10})$$

For proof see Rao (1973, p. 124). Theorem A.11 is the convergence in distribution analogue of Theorem A.3 for convergence in probability.

The following theorem considers the effect of transforming a sequence with limit distribution by addition of, or multiplication by, or division by a sequence that converges in probability to a constant.

Theorem A.12 (Transformation Theorem): *If $a_N \xrightarrow{d} a$ and $b_N \xrightarrow{p} b$, where a is a random variable and b is a constant, then*

- (i) $a_N + b_N \xrightarrow{d} a + b$,
 - (ii) $a_N b_N \xrightarrow{d} ab$, and
 - (iii) $a_N/b_N \xrightarrow{d} a/b$, provided $\Pr[b = 0] = 0$.
- (A.11)

For proof see Rao (1973, p. 122). Theorem A.12 is also referred to as *Cramer's Theorem*. It is also called Slutsky's Theorem, the name we have applied to Theorem A.3.

Theorem A.12 is exceptionally useful because it permits one to separately find the limit distribution of a_N and the probability limit of b_N , rather than having to consider the joint behavior of a_N and b_N . Result (ii) is especially useful and is sometimes called the *Product Rule*.

A.5. Central Limit Theorems

Central limit theorems are theorems on *convergence in distribution* when the sequence $\{b_N\}$ is a *sample average*. A central limit theorem provides a simpler way to obtain the limit distribution of a sequence $\{b_N\}$ than the alternatives such as brute-force use of (A.9).

From a law of large numbers, the sample average has a degenerate distribution as it converges to a constant, $\lim E[\bar{X}_N]$. So we scale $(\bar{X}_N - E[\bar{X}_N])$ by its standard deviation

to construct a random variable with unit variance that may converge to a nondegenerate distribution.

Definition A.13 (Central Limit Theorem): Let

$$Z_N = \frac{\bar{X}_N - E[\bar{X}_N]}{\sqrt{V[\bar{X}_N]}}, \quad (\text{A.12})$$

where \bar{X}_N is a sample average. A **central limit theorem** (CLT) specifies the conditions on the individual terms X_i in \bar{X}_N under which

$$Z_N \xrightarrow{d} \mathcal{N}[0, 1], \quad (\text{A.13})$$

that is, under which Z_N converges in distribution to a standard normal random variable.

By construction Z_N has mean 0 and variance 1, so what needs to be proved is the normality. Formal proofs of a CLT do this by obtaining the characteristic function, a generalization of the moment-generating function, of Z_N and showing that it converges as $N \rightarrow \infty$ to the characteristic function of the standard normal distribution.

Note that if \bar{X}_N satisfies a central limit theorem, then so too does $h(N)\bar{X}_N$ for functions $h(\cdot)$ such as $h(N) = \sqrt{N}$, since

$$Z_N = \frac{h(N)\bar{X}_N - E[h(N)\bar{X}_N]}{\sqrt{V[h(N)\bar{X}_N]}}.$$

In many applications it is convenient to apply the central limit theorem to the normalization $\sqrt{N}\bar{X}_N = N^{-1/2} \sum_{i=1}^N X_i$, since $V[\sqrt{N}\bar{X}_N]$ is finite.

Examples of central limit theorems include the following:

Theorem A.14 (Lindeberg–Levy CLT): Let $\{X_i\}$ be iid with $E[X_i] = \mu$ and $V[X_i] = \sigma^2$. Then $Z_N \xrightarrow{d} \mathcal{N}[0, 1]$.

For a proof, see Rao (1973, p. 127).

This is the CLT that usually appears in introductory statistics texts and is useful in the iid case. Since X_i is iid $[0, \sigma^2]$, Z_N simplifies to the more familiar

$$Z_N = \frac{\bar{X}_N - \mu}{\sigma/\sqrt{N}}.$$

Note that in the iid case only the existence of μ is required to ensure that $\bar{X}_N \xrightarrow{p} \mu$, whereas to obtain a limiting normal distribution requires the additional assumption that σ^2 exists.

In applications such as OLS with fixed regressors the iid assumption is inappropriate. One can apply a CLT for $\{X_i\}$ inid, though additional assumptions need to be made.

Theorem A.15 (Liapounov CLT): Let $\{X_i\}$ be independent with $E[X_i] = \mu_i$ and $V[X_i] = \sigma_i^2$. If $\lim(\sum_{i=1}^N E[|X_i - \mu_i|^{2+\delta}])/(\sum_{i=1}^N \sigma_i^2)^{(2+\delta)/2} = 0$, for some choice of $\delta > 0$, then $Z_N \xrightarrow{d} \mathcal{N}[0, 1]$.

This variant of the Liapounov CLT is proved in White (2001a, p. 119). Rao (1973, p. 128) presents the special case $\delta = 1$.

The main additional assumption in the Liapounov CLT is the existence of an absolute moment of order higher than two. Note also the additional assumptions compared to the corresponding LLN for iid data. For X_i iid

$$Z_N = \frac{\sum_{i=1}^N X_i - \sum_{i=1}^N \mu_i}{\sqrt{\sum_{i=1}^N \sigma_i^2}}.$$

Theorems A.14 and A.15 are special cases of the more general Lindeberg–Feller CLT (see Rao, 1973, p. 128). The Lindeberg–Feller CLT has a side condition that can be difficult to verify.

In most microeconomics applications, including regression with *stratified sampling* or with *fixed regressors* the more complicated Liapounov CLT is used.

A.6. Multivariate Normal Limit Distributions

In this section we focus on the typical microeconomics case of estimators with multivariate normal limit distributions.

A.6.1. Multivariate Normal Limit Distributions

The central limit theorems presented were for sequences of scalar random variables. They can be extended to sequences of vector random variables using the following result.

Theorem A.16 (Cramer–Wold Device): *Let $\{\mathbf{b}_N\}$ be a sequence of random $k \times 1$ vectors. If $\lambda' \mathbf{b}_N$ converges to a normal random variable for every $k \times 1$ constant nonzero vector λ , then \mathbf{b}_N converges to a multivariate normal random variable.*

Rao (1973, p. 128) gives a more general result that is not restricted to normal distributions.

The advantage of this result is that, if \mathbf{b}_N is a vector of averages, then $\lambda' \mathbf{b}_N = \lambda_1 b_{1N} + \dots + \lambda_k b_{kN}$ will be a scalar average and we can apply a scalar central limit theorem given in the previous section. This will yield

$$\frac{\lambda' \mathbf{b}_N - \lambda' \mu_N}{\sqrt{\lambda' \mathbf{V}_N \lambda}} \xrightarrow{d} \mathcal{N}[0, 1],$$

where $\mu_N = \mathbb{E}[\mathbf{b}_N]$ and $\mathbf{V}_N = \mathbb{V}[\mathbf{b}_N]$, in which case we conclude that

$$\mathbf{V}_N^{-1/2}(\mathbf{b}_N - \mu_N) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{I}]. \quad (\text{A.14})$$

This result is explained further in Subsection A.6.3.

A.6.2. Linear Transformation

Microeconometric estimators can often be expressed as $\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \mathbf{H}_N \mathbf{a}_N$, where $\text{plim } \mathbf{H}_N$ exists and \mathbf{a}_N has a limit normal distribution. The distribution of this product, or linear transformation of a_N , can be obtained directly from part (ii) of Theorem A.12 (Transformation Theorem). We restate it in a form that arises for many estimators.

Theorem A.17 (Product Limit Normal Rule): If a vector $\mathbf{a}_N \xrightarrow{d} \mathcal{N}[\boldsymbol{\mu}, \mathbf{A}]$ and a matrix $\mathbf{H}_N \xrightarrow{P} \mathbf{H}$, where \mathbf{H} is positive definite, then

$$\mathbf{H}_N \mathbf{a}_N \xrightarrow{d} \mathcal{N}[\mathbf{H}\boldsymbol{\mu}, \mathbf{H}\mathbf{A}\mathbf{H}']. \quad (\text{A.15})$$

Theorem A.17 can be directly applied to an estimator. For example, the OLS estimator

$$\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \left(\frac{1}{N} \mathbf{X}' \mathbf{X} \right)^{-1} \frac{1}{\sqrt{N}} \mathbf{X}' \mathbf{u}$$

is treated as the product of $\mathbf{H}_N = (N^{-1} \mathbf{X}' \mathbf{X})^{-1}$ and $\mathbf{a}_N = N^{-1/2} \mathbf{X}' \mathbf{u}$ and we find the plim of \mathbf{H}_N and the limit distribution of \mathbf{a}_N .

Theorem A.17 can also be used to justify *replacement* of a limit distribution variance matrix by a consistent estimate without changing the limit distribution. If we have shown that

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{B}],$$

then it follows by Theorem A.17 that

$$\mathbf{B}_N^{-1/2} \times \sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{I}]$$

for any \mathbf{B}_N that is a consistent estimate for \mathbf{B} and is positive definite.

A.6.3. Limit Variance Matrix

A formal multivariate CLT yields a notationally cumbersome result such as (A.14). Premultiplying by $\mathbf{V}_N^{1/2}$ and applying Theorem A.17, we can reexpress this in the simpler form

$$\mathbf{b}_N - \boldsymbol{\mu}_N \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{V}],$$

where $\mathbf{V} = \text{plim } \mathbf{V}_N$ and we assume \mathbf{b}_N and \mathbf{V}_N are appropriately scaled so that \mathbf{V} exists and is positive definite.

Different authors express the **limit variance matrix \mathbf{V}** in different ways.

A general definition is simply

$$\mathbf{V} = \text{plim } \mathbf{V}_N.$$

This is the most common way that results are presented and is the form used in this text. In the fixed regressors case it simplifies to $\mathbf{V} = \lim \mathbf{V}_N$.

In microeconomics estimation examples the matrix \mathbf{V}_N is often a matrix average, say

$$\mathbf{V}_N = \frac{1}{N} \sum_{i=1}^N \mathbf{S}_i,$$

where \mathbf{S}_i is a square matrix that is a function of parameters and data for the i th observation. Given independence over i a law of large numbers can usually be applied so that $\mathbf{V}_N - E[\mathbf{V}_N] \xrightarrow{P} \mathbf{0}$. Then

$$\mathbf{V} = \lim E[\mathbf{V}_N] = \lim \frac{1}{N} \sum_{i=1}^N E[\mathbf{S}_i].$$

This is the type of expression used by Amemiya (1985).

If the \mathbf{S}_i are iid then $E[\mathbf{S}_i] = E[\mathbf{S}]$ is the same for all observations. So simple random sampling leads to the simpler expression

$$\mathbf{V} = E[\mathbf{S}],$$

a form used for example by Newey and McFadden (1994) and Wooldridge (2002).

As an example, consider the OLS estimator with homoskedastic error, so that $\sqrt{N}(\hat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \sigma^2 \mathbf{M}_{\mathbf{x}\mathbf{x}}^{-1}]$. Then $\mathbf{M}_{\mathbf{x}\mathbf{x}} = \text{plim } N^{-1} \sum_i \mathbf{x}_i \mathbf{x}_i'$ can be re-expressed as $\mathbf{M}_{\mathbf{x}\mathbf{x}} = \lim N^{-1} \sum_i E[\mathbf{x}_i \mathbf{x}_i']$ if a law of large numbers applies, and as $\mathbf{M}_{\mathbf{x}\mathbf{x}} = E[\mathbf{x}\mathbf{x}']$ under simple random sampling.

More complicated forms of \mathbf{V} arise, such as the sandwich form \mathbf{ABA}' . The preceding discussion is then applied to each component. For example, $\mathbf{B} = \text{plim } \mathbf{B}_N$ may be expressed as $\mathbf{B} = \lim E[\mathbf{B}_N]$ or as $\mathbf{B} = E[\mathbf{S}]$ under random sampling if $\mathbf{B} = N^{-1} \sum_i \mathbf{S}_i$.

A.6.4. Asymptotic Distribution and Variance

To obtain the limit distribution of an estimator we work with the sequence $b_N = \sqrt{N}(\hat{\theta} - \theta_0)$ for theoretical reasons to ensure a nonzero variance of b_N as $N \rightarrow \infty$. Then the limit distribution of b_N is a normal distribution, and many authors say that b_N is **asymptotically normal** and call the limit variance matrix the **asymptotic variance** of b_N .

It can be convenient to reexpress results in terms of the distribution and variance matrix of $\hat{\theta}$ itself.

Definition A.18 (Asymptotic Distribution of $\hat{\theta}$): If

$$\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{B}], \quad (\text{A.16})$$

then we say that *in large samples* $\hat{\theta}$ is **asymptotically normally distributed** with

$$\hat{\theta} \sim \mathcal{N}[\theta_0, N^{-1} \mathbf{B}], \quad (\text{A.17})$$

where the term “in large samples” means that N is large enough for (A.16) to be a good approximation but not so large that the variance in (A.17) goes to zero.

The result (A.17) follows from (A.16) since dividing a random variable by \sqrt{N} leads to division of its variance by N .

A shorthand notation is to implicitly presume asymptotic normality and use the following terminology.

Definition A.19 (Asymptotic Variance of $\hat{\theta}$): If (A.16) holds then we say that the **asymptotic variance matrix** of $\hat{\theta}$ is

$$V[\hat{\theta}] = N^{-1} \mathbf{B}. \quad (\text{A.18})$$

Definition A.20 (Estimated Asymptotic Variance of $\hat{\theta}$): If (A.16) holds then we say that the **estimated asymptotic variance matrix** of $\hat{\theta}$ is

$$\widehat{V}[\hat{\theta}] = N^{-1} \widehat{\mathbf{B}}, \quad (\text{A.19})$$

where $\widehat{\mathbf{B}}$ is a consistent estimate of \mathbf{B} .

Some authors use $\text{Avar}[\hat{\theta}]$ and $\widehat{\text{Avar}}[\hat{\theta}]$ in Definitions A.19 and A.20 to avoid potential confusion with the variance operator $V[\cdot]$. It should be clear that here $V[\hat{\theta}]$ means asymptotic variance of an estimator since few estimators in this book have closed-form expressions for the finite-sample variance.

As an example of Definitions A.18–A.20, if $\{X_i\}$ are iid $[\mu, \sigma^2]$ then the Lindeberg–Levy central limit theorem leads to $\sqrt{N}(\bar{X}_N - \mu)/\sigma \xrightarrow{d} \mathcal{N}[0, 1]$, or equivalently that $\sqrt{N}\bar{X}_N \xrightarrow{d} \mathcal{N}[\mu, \sigma^2]$. We say that asymptotically $\bar{X}_N \sim \mathcal{N}[\mu, \sigma^2/N]$; the asymptotic variance of \bar{X}_N is σ^2/N ; and the estimated asymptotic variance of \bar{X}_N is s^2/N , where s^2 is a consistent estimator of σ^2 such as $s^2 = \sum_i (X_i - \bar{X}_N)^2/(N - 1)$.

A.6.5. Asymptotic Efficiency

In finite samples the Cramer–Rao lower bound for the variance–covariance matrix of unbiased estimators is $-(E[\partial^2 \ln L_N / \partial \theta \partial \theta' |_{\theta_0}])^{-1}$. This result extends to consistent estimators that are asymptotically normal.

Definition A.21 (Asymptotic Efficiency): A consistent asymptotically normal estimator $\hat{\theta}$ of θ is said to be **asymptotically efficient** if it has an asymptotic variance–covariance matrix equal to the Cramer–Rao lower bound.

A.7. Stochastic Order of Magnitude

A useful notation for rates of convergence of sequences of variables is the order of magnitude of a sequence using *(O, o) notation*, or big-O, little-o notation.

A sequence of nonstochastic real numbers a_N is $O(g(N))$, if $\lim(a_N/g(N))$ is finite nonzero, and is $o(g(N))$, if $\lim(a_N/g(N))$ is zero. Thus a_N is $O(g(N))$ if it is of the same order of magnitude as the function $g(N)$ and is $o(g(N))$ if it is of smaller order

of magnitude than $g(N)$. For example, $(3/N) + (5/N)^2$ is $O(1/N)$ or $O(N^{-1})$, as it behaves for large N like a constant times N^{-1} and is $o(N^{-1/2})$ but larger than $o(N^{-1})$.

This notation has been extended to **stochastic orders of magnitude** of sequences of random variables. The notation becomes (O_p, o_p) notation.

Definition A.22 (Stochastic Order of Magnitude): A sequence of random variables b_N is $O_p(g(N))$ if

$$0 < \text{plim} \frac{b_N}{g(N)} < \infty$$

and is $o_p(g(N))$ if

$$\text{plim} \frac{b_N}{g(N)} = 0.$$

Most often $g(N) = N^{-c}$ for some constant $c \geq 0$. An estimator $\widehat{\theta}$ *consistent* for θ_0 can be written as $\widehat{\theta} = \theta_0 + o_p(1)$, since it equals θ_0 plus a term that goes to zero in probability. An estimator $\widehat{\theta}$ that is additionally *root-N consistent* for θ_0 can be written as $\widehat{\theta} = \theta_0 + O_p(N^{-1/2})$, since then $N^{1/2}(\widehat{\theta} - \theta_0) = O_p(1)$.

A.8. Other Results

This section contains some key finite sample results on conditional expectation and on the interchange of expectations and transformation.

Theorem (Law of Iterated Expectations): *For random variables Y and X*

$$E[Y] = E_X[E_{Y|X}[Y|X]],$$

where $E[\cdot]$ denotes the unconditional or marginal mean of Y , $E_X[\cdot]$ denotes unconditional expectation with respect to the marginal cdf of X , and $E_{Y|X}[\cdot|X]$ denotes conditional expectation with respect to the conditional distribution of Y given X .

This result means that if we first obtain the conditional mean of Y given X , and then take the expected value over X , we will obtain the unconditional mean of Y . See Rao (1973, p. 97) for a proof. For example, if $E[u|\mathbf{x}] = 0$ then $E[u] = E_{\mathbf{x}}[E[u|\mathbf{x}]] = E_{\mathbf{x}}[0] = 0$.

Theorem (Decomposition of Variance): *For random variables Y and X*

$$V[Y] = E_X[V_{Y|X}[Y|X]] + V_X[E_{Y|X}[Y|X]],$$

where $V[Y]$ denotes the unconditional variance of Y , $E_X[\cdot]$ denotes unconditional expectation with respect to the marginal cdf of X , $V_{Y|X}[Y|X]$ denotes the conditional variance of Y given X , $V_X[\cdot]$ denotes variance with respect to the unconditional distribution of X , and $E_{Y|X}[\cdot|X]$ denotes conditional expectation with respect to the conditional distribution of Y given X .

In words, the unconditional variance of Y equals the sum of (1) the expected value (over X) of the conditional variance and (2) the variance (over X) of the conditional mean. A simple way to remember this is to recognize that the unconditional variance equals EV plus VE. See Rao (1973, p. 97) for a proof.

Theorem (Jensen's Inequality): *If Z is a random variable such that $E[Z]$ exists, and $g(\cdot)$ is a convex function, then*

$$g(E[Z]) \leq E[g(Z)].$$

If instead $g(\cdot)$ is a concave function then

$$g(E[Z]) \geq E[g(Z)].$$

This result, proved in Rao (1973, p. 58), is very important for nonlinear models. It emphasizes the difference between behavior of the average individual and average behavior. For example, suppose an exponential model is appropriate, with $E[y|\mathbf{x}] = \exp(\mathbf{x}'\boldsymbol{\beta})$. Then since the exponential function is concave, Jensen's Inequality implies that $\exp(E[\mathbf{x}'\boldsymbol{\beta}]) \geq E[\exp(\mathbf{x}'\boldsymbol{\beta})]$. The conditional mean evaluated at the individual with average characteristics $\mathbf{x} = E[\mathbf{x}]$ exceeds the unconditional mean $E[y] = E[E[y|\mathbf{x}]] = E[\exp(\mathbf{x}'\boldsymbol{\beta})]$.

A.9. Bibliographic Notes

A classic source with proofs is Rao (1973, pp. 108–130), who we cite wherever possible. The results summarized also draw heavily on the books by Amemiya (1985, Chapter 3) and White (2001a).

Graduate-level textbooks such as Greene (2003) provide summaries of key results. More advanced texts by Davidson and MacKinnon (1993), Hendry (1995), Ruud (2000), and Wooldridge (2002) provide treatments at least as detailed as that here. Davidson (1994) provides a book-length treatment of stochastic theory for the econometrician. As already noted terminology can differ across references, especially in the use of Slutsky's Theorem and Cramer's Theorem.

APPENDIX B

Making Pseudo-Random Draws

In this appendix we state the density or probability mass functions and first two moments of leading univariate distributions and present methods to generate random draws from these distributions.

Table B.1. *Continuous Random Variable Densities and Moments^a*

Random Variable	pdf $f(x)$	Mean; Variance
Uniform $\mathcal{U}[a, b]$	$1/(b - a),$	$\frac{(a+b)}{2}, \frac{(a-b)^2}{12}$
Normal $\mathcal{N}[\mu, \sigma^2]$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$	$\mu; \sigma^2$
Exponential $\mathcal{E}[\lambda]$	$\lambda e^{-\lambda x}, \quad \lambda > 0$	$1/\lambda; 1/\lambda^2$
Gamma $\mathcal{G}[a,b]$	$\frac{1}{\Gamma(a)b^a} x^{a-1} e^{-\frac{x}{b}}$	$ab; ab^2$
Beta $\mathcal{B}[a,b]$	$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1},$	$\frac{a}{a+b}; \frac{ab}{(a+b)^2(a+b+1)}$
Logistic $\mathcal{L}[a,b]$	$e^{-\frac{x-a}{b}} / [b(1 + e^{-\frac{x-a}{b}})^2], \quad -\infty < a < \infty$	$a; (b\pi)^2/3$
Chi-Square $\chi^2(n)$	$\frac{x^{n/2-1} e^{-x/2}}{\Gamma(n/2)2^{n/2}}$	$n; 2n$
t $t(v)$	$f(x) = \frac{\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})\sqrt{v\pi}} (1 + \frac{x^2}{v})^{-\frac{v+1}{2}}$	$0; \frac{v}{v-2}, \text{ for } v > 2$
F $F(w, v)$	$f(x) = \frac{\Gamma(\frac{w+v}{2})(v/w)^{v/2}}{\Gamma(\frac{w}{2})\Gamma(\frac{v}{2})} x^{w/2-1} \\ \times (x + \frac{v}{w})^{-\frac{w+v}{2}}$	$\frac{v}{v-2}, \text{ for } v > 2;$ $\frac{2v^2(v+w-2)}{w(v-4)(v-2)^2}, \text{ for } v > 4$

^a All parameters are restricted as follows: $b > a$ for the Uniform; μ unrestricted, $\sigma^2 > 0$ for the Normal; $\lambda > 0$ for the Exponential; $a, b > 0$ for the gamma; $a, b > 0$ for the Beta; a unrestricted and $b > 0$ for the Logistic; v is an integer for the t-distribution; for the F-distribution v and w must be integers.

Table B.2. Continuous Random Variable Generators

Random Variable	Range of Values	Random Variable Generator
Uniform $\mathcal{U}[a, b]$	$a \leq x \leq b$	$x = a + (b - a)r, r \sim \mathcal{U}[0, 1]$
Normal $\mathcal{N}[\mu, \sigma^2]$	$-\infty < x_1, x_2 < \infty$	$\begin{cases} x_1 = \mu + \sigma \sqrt{-2 \ln(r_1)} \cos(2\pi r_2) \\ x_2 = \mu + \sigma \sqrt{-2 \ln(r_1)} \sin(2\pi r_2) \end{cases}$ <p>[$r_1, r_2 \sim \mathcal{U}[0, 1]$; the resulting pair x_1 and x_2 are independent random variables.]</p>
Exponential $\mathcal{E}[\lambda]$	$0 \leq x < \infty$	$x = -\frac{1}{\lambda} \ln(r)$
Gamma $\mathcal{G}[a,b]$	$0 \leq x < \infty$	$\begin{cases} (i) x = -\frac{1}{\lambda} \ln(\prod_{i=1}^a r_i) \text{ or} \\ x = \sum_{i=1}^a E_i \\ (ii) x = -\frac{1}{\lambda} [\ln(\prod_{i=1}^m r_i) - y_1 y_2] \end{cases}$ <p>$\left\{ \begin{array}{l} (i) r_i \sim \mathcal{U}[0, 1]; a \text{ is integer. } E_i \text{ s are iid exponential random variates.} \\ \text{When } a = 1, \text{ we have an exponential random variable} \\ (ii) a \text{ is non-integer. } a = m + q, 0 < q < 1, m = \text{integer,} \\ y_1, y_2 \text{ are independent } \mathcal{B}(q, 1 - q) \text{ and } \mathcal{E}(1). \end{array} \right.$</p>
Beta $\mathcal{B}[a,b]$	$0 \leq x \leq 1$	$\begin{cases} (i) x = y_1/(y_1 + y_2) \\ (ii) x = r_1^{\frac{1}{a}} / (r_1^{\frac{1}{a}} + r_2^{\frac{1}{b}}), (r_1^{\frac{1}{a}} + r_2^{\frac{1}{b}}) \leq 1 \end{cases}$ <p>$\left\{ \begin{array}{l} (i) a, b \text{ are integers. } y_1 \text{ is } \mathcal{G}(k, a), y_2 \text{ is } \mathcal{G}(k, b). \\ k \text{ can be chosen arbitrarily.} \\ (ii) a, b \text{ are non-integer } r_i \sim \mathcal{U}[0, 1]; \text{ successive pairs of } r_1 \text{ and } r_2 \text{ are} \\ \text{generated until } (r_1^{\frac{1}{a}} + r_2^{\frac{1}{b}}) \leq 1. \end{array} \right.$</p>
Logistic $\mathcal{L}[a,b]$	$-\infty < x < \infty$	$x = a + b \ln(\frac{r}{1-r})$ $[r \sim \mathcal{U}[0, 1]]$
Chi-Square $\chi^2(n)$	$0 \leq x$	$\sum_{i=1}^n y_i^2$ $[n \text{ is an integer; } y_i \text{ s are independent } \mathcal{N}(0, 1).]$
t t(v)	$-\infty < x < \infty$	$x = y_1 / \sqrt{y_2/v}$ $[y_1 \text{ is } \mathcal{N}(0, 1); y_2, \text{ independent of } y_1, \text{ is } \chi^2(v).]$
F F(w, v)	$0 \leq x$	$x = (y_1/w)/(y_2/v)$ $[y_2, \text{ and } y_1, \text{ are independent } \chi^2(v), \chi^2(w) \text{ respectively.}]$

Table B.3. Discrete Random Variable Probability Mass Functions and Moments

Random Variable ^a	pmf $f(x)$	Mean; Variance
Binomial $Bi[n, p]$	$\binom{n}{x} p^x (1-p)^{n-x}$	$np; np(1-p)$
Poisson $\mathcal{P}[\lambda]$	$e^{-\lambda} \lambda^x / x!$	$\lambda; \lambda$
Negative binomial $NB[n, p]$	$\binom{n+x-1}{x} p^n (1-p)^x$	$\frac{n(1-p)}{p}; \frac{n(1-p)}{p^2}$

^a For the binomial $0 \leq p \leq 1$ and n is a positive integer; for the Poisson $\lambda > 0$; and for the negative binomial $0 < p < 1, n > 1$.

Table B.4. Discrete Random Variable Generators

Random Variable	Range of Values	Random Variable Generator
Binomial $Bi(n, p)$	$x = 0, 1, \dots, n$	set $x = 0$; do the loop n times generate r uniform on $[0,1]$ if $r \leq p$, then $x = x + 1$ output x
Poisson $\mathcal{P}(\lambda)$	$x = 0, 1, \dots$	set $x = 0; t = 0$ do the loop until $t < \lambda$ generate exponential random variable y set $t = t + y$ $x = x + 1$ output x
Negative binomial $NB(n, p)$	$x = 0, 1, \dots$	generate λ from $\mathcal{G}(n, \frac{1-p}{p})$ generate x from $\mathcal{P}(\lambda)$ output x

References

- Abe, M. (1999), “A Generalized Additive Model for Discrete-Choice Data,” *Journal of Business and Economics Statistics*, 17, 271–284.
- Abowd, J. M., and D. Card (1989), “On the Covariance Structure of Earnings and Hours Changes,” *Econometrica*, 57, 411–445.
- Abramowitz, M., and I. A. Stegun (1971), *Handbook of Mathematical Functions*, New York, Dover Press.
- Ahn, S. C., and P. Schmidt (1995), “Efficient Estimation of Models for Dynamic Panel Data,” *Journal of Econometrics*, 68, 5–27.
- Ahn, S. C., and P. Schmidt (1999), “Estimation of Linear Panel Data Models Using GMM,” in *Generalized Method of Moments Estimation*, L. Mátyás (Ed.), 211–247 Cambridge, UK, Cambridge University Press.
- Aitchison, J., and S. D. Silvey (1958), “Maximum-Likelihood Estimation of Parameters Subject to Restraints,” *Annals of Mathematical Statistics*, 29, 813–828.
- Aitken, M., and D. B. Rubin (1985), “Estimation and Hypothesis Testing in Finite Mixture Models,” *Journal of the Royal Statistical Society, B*, 47, 67–75.
- Akaike, H. (1973), “Information Theory and an Extension of the Maximum Likelihood Principle,” in *Second International Symposium on Information Theory*, B. N. Petrov and F. Csaki (Eds.), 267–281, Budapest, Akadémiai Kaidó.
- Albert, J. H. (1988), “Computational Methods for Using a Bayesian Hierarchical Generalized Linear Model,” *Journal of the American Statistical Association*, 83, 1037–1045.
- Albert, J. H., and S. Chib (1993), “Bayesian Analysis of Binary and Polychotomous Response Data,” *Journal of the American Statistical Association*, 88, 669–679.
- Allen, R. G. D., and A. L. Bowley (1935), *Family Expenditure*, London, P. S. King and Son.
- Allenby, G. M., and P. E. Rossi (1991), “There Is No Aggregate Bias: Why Macro Logit Models Work,” *Journal of Business and Economic Statistics*, 9, 1–14.
- Allison, P. D. (1984), *Event History Analysis: Regression for Longitudinal Event Data*, Beverly Hills, CA, Sage Publications.
- Allison, P. D. (1995), *Survival Analysis Using the SAS System: A Practical Guide*, Cary, NC, SAS Institute Inc.
- Allison, P. D. (2002), *Missing Data*, Beverly Hills, CA, Sage Publications.
- Al-Osh, M. A., and A. A. Alzaid (1987), “First Order Integer Valued Autoregressive (INAR(1)) Process,” *Journal of Time Series Analysis*, 8, 261–275.

REFERENCES

- Altman, N. S. (1992), "An Introduction to Kernel and Nearest-Neighbour Nonparametric Regression," *The American Statistician*, 46, 175–185.
- Altonji, J. G., and L. M. Segal (1996), "Small Sample Bias in GMM Estimation of Covariance Structures," *Journal of Business and Economic Statistics*, 14, 353–366.
- Amemiya, T. (1973), "Regression Analysis When the Dependent Variable is Truncated Normal," *Econometrica*, 41, 997–1016.
- Amemiya, T. (1974), "The Nonlinear Two-Stage Least Squares Estimator," *Journal of Econometrics*, 2, 105–110.
- Amemiya, T. (1979), "The Estimation of a Simultaneous Equation Tobit Model," *International Economic Review*, 20, 169–181.
- Amemiya, T. (1980), "Selection of Regressors," *International Economics Review*, 21, 331–345.
- Amemiya, T. (1981), "Qualitative Response Models: A Survey," *Journal of Economic Literature*, 1483–1536.
- Amemiya, T. (1983), "Nonlinear Regression Models," in *Handbook of Econometrics*, Z. Griliches and M. D. Intriligator (Eds.), Volume 1, 333–389, Amsterdam, North-Holland.
- Amemiya, T. (1984), "Tobit Models: A Survey," *Journal of Econometrics*, 24, 3–61.
- Amemiya, T. (1985), *Advanced Econometrics*, Cambridge, MA, Harvard University Press.
- Amemiya, T., and T. E. MacCurdy (1986), "Instrumental Variable Estimation of An Error Component Model," *Econometrica*, 54, 869–880.
- Amemiya, T., and Q. H. Vuong (1987), "A Comparison of Two Consistent Estimators in the Choice-Based Sampling Qualitative Response Model," *Econometrica*, 55, 699–702.
- Amemiya, Y. (1985), "Instrumental Variable Estimator for the Nonlinear Error in Variables Model," *Journal of Econometrics*, 28, 273–289.
- Andersen, E. B. (1970), "Asymptotic Properties of Conditional Maximum Likelihood Estimators," *Journal of the Royal Statistical Society, B*, 32, 283–301.
- Andersen, P. K., O. Borgan, R. D. Gill, and N. Keiding (1993), *Statistical Models Based on Counting Processes*, New York, Springer-Verlag.
- Anderson, T. W. (1971), *The Statistical Analysis of Time Series*, New York, John 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 H. Rubin (1949), "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *Annals of Mathematical Statistics*, 20, 46–63.
- Andrews, D. W. K. (1988a), "Chi-Square Diagnostic Tests for Econometric Models: Theory," *Econometrica*, 56, 1419–1453.
- Andrews, D. W. K. (1988b), "Chi-Square Diagnostic Tests for Econometric Models: Introduction and Applications," *Journal of Econometrics*, 37, 135–156.
- Andrews, D. W. K. (1989), "Power in Econometric Applications," *Econometrica*, 57, 1059–1090.
- Andrews, D. W. K. (1991), "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models," *Econometrica*, 59, 307–345.
- Andrews, D. W. K. (1997), "A Conditional Kolmogorov Test," *Econometrica*, 65, 1097–1128.
- Andrews, D. W. K. (2001), "Testing When a Parameter Is on the Boundary of the Maintained Hypothesis," *Econometrica*, 69, 683–734.
- Andrews, D. W. K., and M. Buchinsky (2000), "A Three-Step Method for Choosing the Number of Bootstrap Replications," *Econometrica*, 68, 23–51.
- Andrews, D. W. K., and M. M. A. Schafgans (1998), "Semiparametric Estimation of the Intercept of a Sample Selection Model," *Review of Economic Studies*, 65, 497–517.
- Angrist, J. D. (1990), "Lifetime Earnings and Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, 80, 313–335.

REFERENCES

- Angrist, J. D. (2001), "Estimation of Limited Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice," *Journal of Business and Economic Statistics*, 19, 2–28.
- Angrist, J. D., G. W. Imbens, and A. B. Krueger (1999), "Jackknife Instrumental Variables Estimation," *Journal of Applied Econometrics*, 14, 57–67.
- Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455.
- Angrist, J. D., and A. B. Krueger (1991), "Does Compulsory School Attendance Affect Schooling and Earnings," *Quarterly Journal of Economics*, 106, 979–1014.
- Angrist, J. D., and A. B. Krueger (1995), "Split Sample Instrumental Variables Estimates of the Return to Schooling," *Journal of Business and Economic Statistics*, 13, 225–235.
- Angrist, J. D., and A. B. Krueger (1999), "Empirical Strategies in Labor Economics," in *Handbook of Labor Economics*, O. C. Ashenfelter and D. E. Card (Eds.), Volume 3A, 1277–1397, Amsterdam, North-Holland.
- Angrist, J. D., and V. Lavy (1999), "Using Maimonides Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics*, 114, 533–575.
- Angrist, J. D., and V. Lavy (2002), "The Effect of High School Matriculation Awards: Evidence from Randomized Trials," MIT Discussion Paper.
- Anselin, L. (2001), "Spatial Econometrics," in *A Companion to Theoretical Econometrics*, B. H. Baltagi (Ed.), 310–330, Oxford, Blackwell.
- Arellano, M. (1987), "Computing Robust Standard Errors for Within-Group Estimators," *Oxford Bulletin of Economics and Statistics*, 49, 431–434.
- 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, 277–298.
- Arellano, M., and O. Bover (1995), "Another Look at Instrumental Variables Estimation of Error Components Models," *Journal of Econometrics*, 68, 29–51.
- Arellano, M., and B. Honoré (2001), "Panel Data Models: Some Recent Developments," in *Handbook of Econometrics*, J. J. Heckman and E. E. Leamer (Eds.), Volume 5, 3229–3296, Amsterdam, North-Holland.
- Ashenfelter, O. (1978), "Estimating the Effect of Training Programs on Earnings," *Review of Economics and Statistics*, 60, 47–57.
- Ashenfelter, O., and D. Card (1985), "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *Review of Economics and Statistics*, 68, 648–660.
- Ashenfelter, O., and A. B. Krueger (1994), "Estimates of the Economic Returns to Schooling from a New Sample of Twins," *American Economic Review*, 84(5) 1157–1173.
- Athey, S., and G. Imbens (2002), "Identification and Inference in Nonlinear Difference-In-Differences Models," working paper, Department of Economics, University of California, Berkeley.
- Avery, R. B. (1977), "Error Components and Seemingly Unrelated Regressions," *Econometrica*, 45, 199–209.
- Avery, R. B., L. P. Hansen, and V. J. Hotz (1983), "Multiperiod Probit Models and Orthogonality Condition Estimation," *International Economic Review*, 24, 21–35.
- Baker, M., and A. Melino (2000), "Duration Dependence and Nonparametric Heterogeneity: A Monte Carlo Study," *Journal of Econometrics*, 96, 357–393.
- Baker, M., and G. Solon (2003), "Earnings Dynamics and Inequality among Canadian Men 1976–1992: Evidence from Longitudinal Income Tax Records," *Journal of Labor Economics*, 21, 289–321.

REFERENCES

- Balestra, P., and M. Nerlove (1966), "Pooling Cross Section and Time Series Data in the Estimation of a Dynamic Model: The Demand for Natural Gas," *Econometrica*, 34, 585–612.
- Baltagi, B. H. (1980), "On Seemingly Unrelated Regressions with Error Components," *Econometrica*, 45, 1547–1552.
- Baltagi, B. H. (1981), "Simultaneous Equations with Error Components," *Journal of Econometrics*, 17, 189–200.
- Baltagi, B. H. (1995, 2001), *Econometric Analysis of Panel Data*, 1st and 2nd editions, New York, John Wiley.
- Baltagi, B. H. (1999), "The Relative Efficiency of the Between Estimator with Respect to the Within Estimator," Problem, *Econometric Theory*, 15, 630–631. Solution by S. Gurmu (2000), *Econometric Theory*, 16, 454–456.
- Baltagi, B. H., and Q. Li (1991), "A Transformation That Will Circumvent the Problem of Autocorrelation in an Error-Component Model," *Journal of Econometrics*, 48, 385–393.
- Baltagi, B. H., and D. Li (1999), "The Overdispersion Test in Count Data as a Gauss–Newton Regression," Problem, *Econometric Theory*, 15, 428. Solution by M. D. Berg (2000), *Econometric Theory*, 16, 297–299.
- Basman, R. L. (1957), "A Generalized Classical Method of Linear Estimation of Coefficients in a Structural Equation," *Econometrica*, 25, 77–83.
- Bayes, T. (1764), "An Essay towards Solving a Problem in the Doctrine of Chances," *Philosophical Transactions of the Royal Society of London*, 53, 370–418. Reprinted in "Studies in the History of Probability and Statistics: IX. Thomas Bayes's Essay towards Solving a Problem in the Doctrine of Chances," *Biometrika*, 1958, 45, 293–315.
- Becker, S. O., and A. Ichino (2002), "Estimation of Average Treatment Effects Based on Propensity Scores," *The Stata Journal*, 2, 358–377.
- Becketti, S., Gould, W., Lillard, L., and Welch, F. (1988), "The Panel Study of Income Dynamics after Fourteen Years: An Evaluation," *Journal of Labor Economics*, 6, 472–492.
- Beggs, S., S. Cardell, and J. Hausman (1981), "Assessing the Potential Demand for Electric Cars," *Journal of Econometrics*, 16, 1–19.
- Begun, J. M., W. J. Hall, W. Huang, and J. A. Wellner (1983), "Information and Asymptotic Efficiency in Parametric–Nonparametric Models," *Annals of Statistics*, 11, 432–452.
- Bekker, P. A. (1994), "Alternative Approximations to the Distributions of Instrumental Variables Estimation," *Econometrica*, 92, 657–681.
- Bellew, C., B. Melenberg, and A. van Soest (2002), "Semiparametric Models for Satisfaction with Income," Discussion Paper No. 2002-87, CentER, Tilburg University.
- Ben-Akiva, M., and S. R. Lerman (1985), *Discrete Choice Analysis: Theory and Application to Travel Demand*, Cambridge, MA, MIT Press.
- Bera, A. K., and Y. Bilias (2002), "The MM, ME, ML, EL, EF and GMM Approaches to Estimation: A Synthesis," *Journal of Econometrics*, 107, 51–86.
- Bera, A. K., and A. Ghosh (2002), "Neyman's Smooth Test and Its Applications in Econometrics," in *Handbook of Applied Econometrics and Statistical Inference*, A. Ullah, A. T. K. Wan, and A. Chaturvedi (Eds.), New York, Marcel Dekker.
- Bera, A. K., and M.-J. Yoon (1993), "Specification Testing with Locally Misspecified Alternatives," *Econometric Theory*, 9, 649–658.
- Beran, R. (1987), "Prepivoting to Reduce Level Error of Confidence Sets," *Biometrika*, 74, 457–468.
- Berkson, J. (1951), "Why I Prefer Logits to Probits," *Biometrics*, 7, 327–339.
- Berndt, E., B. Hall, R. Hall, and J. Hausman (1974), "Estimation and Inference in Nonlinear Structural Models," *Annals of Economic and Social Measurement*, 3/4, 653–665.
- Berndt, E., and N. E. Savin (1975), "Estimation and Hypothesis Testing in Singular Equation Systems with Autoregressive Disturbances," *Econometrica*, 43, 937–957.

REFERENCES

- Berndt, E., and N. E. Savin (1977), "Conflict among Criteria for Testing Hypotheses in the Multivariate Linear Regression Model," *Econometrica*, 45, 1263–1277.
- Berry, S. T. (1994), "Estimating Discrete-Choice Models of Product Differentiation," *Rand Journal of Economics*, 25, 242–262.
- Bertrand, M., E. Duflo, and S. Mullainathan (2004), "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics*, 119, 249–275.
- Bhargava, A., and J. D. Sargan (1983), "Estimating Dynamic Random Effects Models from Panel Data Covering Short Periods," *Econometrica*, 51, 1635–1659.
- Bhat, C. R. (2001), "Quasi-Random Maximum Simulated Likelihood Estimation of the Mixed Multinomial Logit Model," *Transportation Research: Part B: Methodological*, 35, 677–693.
- Bickel, P. J., and D. A. Freedman (1981), "Some Asymptotic Theory for the Bootstrap," *Annals of Statistics*, 9, 1196–1217.
- Bickel, P. J., F. Gotze, and W. R. van Zwet (1997), "Resampling Fewer Than N Observations: Gains, Losses, and Remedies for Losses," *Statistica Sinica*, 7, 1–32.
- Bierens, H. J. (1987), "Kernel Estimators of Regression Functions," in *Advances in Economic Theory: Fifth World Congress*, T. F. Bewley (Ed.), Volume 1, 99–144, Cambridge, UK, Cambridge University Press.
- Bierens, H. J. (1990), "A Consistent Conditional Moment Test of Functional Form," *Econometrica*, 58, 1443–1458.
- Bierens, H. J. (1993), *Topics in Advanced Econometrics: Estimation, Testing, and Specification of Cross-Section and Time-Series Models*, Cambridge, UK, Cambridge University Press.
- Binder, M., C. Hsiao, and M. H. Pesaran (2003), "Estimation and Inference in Short Panel Vector Autoregressions with Unit Roots and Cointegration," revision of University of Cambridge DAE Working Paper No. 0003.
- Bingley, P., and I. Walker (2001), "Household Unemployment and the Labor Supply of Married Women," *Economica*, 68, 157–85.
- Björklund, A., and R. Moffitt (1987), "The Estimation of Wage Gains and Welfare Gains in Self-Selection," *Review of Economics and Statistics*, 69, 42–49.
- Björn, E. (1992), "Panel Data with Measurement Errors," in *The Econometrics of Panel Data*, L. Mátyás and P. Sevestre (Eds.), 152–195, Dordrecht, Kluwer.
- Blake, D., A. Lunde, and A. Timmermann (1999), "The Hazards of Mutual Fund Underperformance: A Cox Regression Analysis," *Journal of Empirical Finance*, 6, 121–152.
- Bliss, C. I. (1934), "The Method of Probits," *Science*, New Series, 79(2037), 38–39.
- Blomquist, S., and M. Dahlberg (1999), "Small Sample Properties of LIML and Jackknife IV Estimators: Experiments with Weak Instruments," *Journal of Applied Econometrics*, 14, 69–88.
- Blundell, R., and S. Bond (1998), "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models," *Journal of Econometrics*, 87, 115–143.
- Blundell, R., R. Griffith, and J. Van Reenen (1995), "Dynamic Count Data Models of Technological Innovation," *Economic Journal*, 105, 333–344.
- Blundell, R., R. Griffith, and J. Van Reenen (1999), "Market Share, Market Value and Innovation in a Panel of British Manufacturing Firms," *Review of Economic Studies*, 66, 529–554.
- Blundell, R., R. Griffith, and F. Windmeijer (1995), "Individual Effects and Dynamics in Count Data," Discussion Paper 95-03, Department of Economics, University College London.
- Blundell, R., R. Griffith, and F. Windmeijer (2002), "Individual Effects and Dynamics in Count Data Models," *Journal of Econometrics*, 102, 113–131.
- Blundell, R. W., and T. E. MaCurdy (2000), "Labor Supply: A Review of Alternative Approaches," in *Handbook of Labor Economics*, O. C. Ashenfelter and D. E. Card (Eds.), Volume 3A, 1559–1695, Amsterdam, North-Holland.

REFERENCES

- Blundell, R. W., and J. L. Powell (2004), "Endogeneity in Semiparametric Binary Response Models," *Review of Economic Studies*, 71, 655–679.
- Blundell, R. W., and R. J. Smith (1989), "Estimation in a Class of Limited Dependent Variable Models," *Review of Economic Studies*, 56, 37–58.
- Böckenholt, U. (1999), "Mixed INAR(1) Poisson Regression Models: Analyzing Heterogeneity and Serial Dependencies in Longitudinal Count Data," *Journal of Econometrics*, 89, 317–338.
- Bolduc, D. (1999), "A Practical Technique to Estimate Multinomial Probit Models in Transportation," *Transportation Research B*, 33, 63–79.
- Bollinger, C. R. (1996), "Bounding Mean Regressions When a Binary Regressor Is Mismeasured," *Journal of Econometrics*, 73, 387–399.
- Bollinger, C. R. (2003), "Measurement Error in Human Capital and the Black–White Wage Gap," *Review of Economics and Statistics*, 85, 578–585.
- Bond, S., and F. Windmeijer (2002), "Finite Sample Inference for GMM Estimators in Linear Panel Data Models," cemmap Working Paper CWP04/02, University College London.
- Borsch-Supan, A. (1987), *Econometric Analysis of Discrete Choice: With Applications on Demand for Housing in the U. S. and West-Germany*, Berlin, Springer-Verlag.
- Bound, J., C. Brown, G. J. Duncan, and W. L. Rogers (1994), "Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data," *Journal of Labor Economics*, 12, 345–368.
- Bound, J., C. Brown, and N. Mathiowetz (2001), "Measurement Error in Survey Data," in *Handbook of Econometrics*, J. J. Heckman and E. E. Leamer (Eds.), Volume 5, 3705–3843, Amsterdam, North-Holland.
- 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.
- Bradley, P., B. L. Fox, and L. E. Schrage (1983), *A Guide to Simulation*, New York, Springer-Verlag.
- Brännäs, K., and P. Johansson (1996), "Panel Data Regression for Counts," *Statistical Papers*, 37, 191–213.
- Breitung, J., and M. Lechner (1999), "Alternative GMM Methods in Nonlinear Panel Data Models," in *Generalized Method of Moments Estimation*, L. Mátyás (Ed.), 248–274, Cambridge, UK, Cambridge University Press.
- Breslow, N. E. (1996), "Statistics in Epidemiology: The Case-Control Study," *Journal of the American Statistical Association*, 91, 14–28.
- Breusch, T. S., G. E. Mizon, and P. Schmidt (1989), "Efficient Estimation Using Panel Data," *Econometrica*, 57, 695–700.
- Breusch, T. S., and A. R. Pagan (1979), "A Simple Test for Heteroscedasticity and Random Coefficient Variation," *Econometrica*, 47, 1287–1294.
- Breusch, T. S., and A. R. Pagan (1980), "The Lagrange Multiplier Test and Its Applications to Model Specification in Econometrics," *Review of Economic Studies*, 47, 239–254.
- Bronars, S. G., and J. Grogger (1994), "The Economic Consequences of Unwed Motherhood: Using Twin Births as a Natural Experiment," *American Economic Review*, 84, 1141–1156.
- Brooks, J. C., A. C. Cameron, and C. A. Carter (1998), "Political Action Committee Contributions and U.S. Congressional Voting on Sugar Legislation," *American Journal of Agricultural Economics*, 80, 441–454.
- Brown, C., G. J. Duncan, and F. P. Stafford (1996), "The Panel Study of Income Dynamics," *Journal of Economic Perspectives*, 10, 155–168.
- Browning, M., A. Deaton, and M. Irish (1985), "A Profitable Approach to Labor Supply and Commodity Demands over the Life-Cycle," *Econometrica*, 53, 503–543.

REFERENCES

- Brownstone, D., and C. Kazimi (1998), "Applying the Bootstrap," manuscript, Department of Economics, University of California, Irvine.
- Brownstone, D., and R. Valletta (1996), "Modeling Earnings Measurement Error: A Multiple Imputation Approach," *Review of Economics and Statistics*, 78, 705–717.
- Brundy, J. M., and D. W. Jorgenson (1971), "Efficient Estimation of Simultaneous Equations by Instrumental Variables," *Review of Economics and Statistics*, 53, 207–205.
- Bryk, A. S., and S. W. Raudenbush (1992, 2002), *Hierarchical Linear Models*, Newberry Park, Sage Publications.
- Buchinsky, M. (1994), "Changes in the U.S. Wage Structure 1963–1987: Application of Quantile Regression," *Econometrica*, 62, 405–458.
- Buchinsky, M. (1998), "Recent Advances in Quantile Regression Models: A Practical Guideline for Empirical Research," *Journal of Human Resources*, 33, 88–126.
- Bunch, D. (1991), "Estimability in the Multinomial Probit Model," *Transportation Research, Part B, Methodological*, 25B, 1–12.
- Burtless, G. (1995), "The Case for Randomized Field Trials in Economic and Policy Research," *Journal of Economic Perspectives*, 9, 63–84.
- Buse, A. (1982), "The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note," *The American Statistician*, 36, 153–157.
- Butler, J. S., K. H. Anderson, and R. V. Burkhauser (1989), "Work and Health after Retirement: A Competing Risks Model with Semiparametric Unobserved Heterogeneity," *Review of Economics and Statistics*, 71, 46–53.
- Butler, J. S., and R. Moffitt (1982), "A Computationally Efficient Quadrature Procedure for the One-Factor Multinomial Probit Model," *Econometrica*, 50, 761–764.
- Cameron, A. C. (1990), "Aggregation in Discrete Choice Models: An Illustration of Nonlinear Aggregation," in *Disaggregation in Economic Modelling*, T. S. Barker and H. M. Pesaran (Eds.), 206–234, London, Routledge.
- Cameron, A. C., and P. Johansson (1997), "Count Data Regressions Using Series Expansions with Applications," *Journal of Applied Econometrics*, 12, 203–223.
- Cameron, A. C., and P. Johansson (1998), "Bivariate Count Data Regression Using Series Expansions: With Applications," manuscript, Department of Economics, University of California, Davis.
- Cameron, A. C., T. Li, P. K. Trivedi, and D. M. Zimmer (2004), "Modelling the Differences in Counted Outcomes Using Bivariate Copula Models With Application to Mismeasured Counts," *Econometrics Journal*, 7, 566–584.
- Cameron, A. C., and P. K. Trivedi (1986), "Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators," *Journal of Applied Econometrics*, 1, 29–53.
- Cameron, A. C., and P. K. Trivedi (1990), "Regression Based Tests for Overdispersion in the Poisson Model," *Journal of Econometrics*, 46, 347–364.
- Cameron, A. C., and P. K. Trivedi (1998), *Regression Analysis for Count Data*, Econometric Society Monograph No. 30, Cambridge, UK, Cambridge University Press.
- Cameron, A. C., P. K. Trivedi, F. Milne, and J. Piggott (1988), "A Microeconometric Model of the Demand for Health Care and Health Insurance in Australia," *Review of Economic Studies*, 55, 85–106.
- Cameron, A. C., and F. A. G. Windmeijer (1996), "R-Squared Measures for Count Data Regression Models with Applications to Health Care Utilization," *Journal of Business and Economic Statistics*, 14, 209–220.
- Cameron, A. C., and F. A. G. Windmeijer (1997), "An R-squared Measure of Goodness of Fit for Some Common Nonlinear Regression Models," *Journal of Econometrics*, 77, 329–342.
- Campbell, D. (1969), "Reforms as Experiments," *American Psychologist*, 24, 409–429.

REFERENCES

- Cappè, O., and C. P. Robert (2000), "Markov Chain Monte Carlo: 10 Years and Still Running," *Journal of the American Statistical Association*, 95, 1282–1286.
- Card, D. E. (1995), "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in *Aspects of Labor Market Behavior: Essays in Honor of John Vanderkamp*, L. N. Christofides, E. K. Grant and R. Swidinsky (Eds.), 201–222, Toronto, University of Toronto Press.
- Card, D. E. (1999), "The Causal Effect of Education on Earnings," in *Handbook of Labor Economics*, O. C. Ashenfelter and D. E. Card (Eds.), Volume 3A, 1801–1863, Amsterdam, North-Holland.
- Card, D. E. (2001), "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica*, 69, 1127–1160.
- Card, D. E., and A. B. Krueger (1994), "Minimum Wages and Employment," *American Economic Review*, 84, 772–793.
- Carrasco, R. (2001), "Binary Choice with Binary Endogenous Regressors in Panel Data: Estimating the Effect of Fertility on Female Labor Participation," *Journal of Business and Economic Statistics*, 19, 385–394.
- Carroll, R. J. (1982), "Adapting for Heteroskedasticity in Linear Models," *Annals of Statistics*, 10, 1224–1233.
- Carroll, R. J., D. Ruppert, and L. A. Stefanski (1995), *Measurement Error in Nonlinear Models*, London, Chapman and Hall.
- Casella, G., and E. George (1992), "Explaining the Gibbs Sampler," *The American Statistician*, 46, 167–174.
- Chamberlain, G. (1980), "Analysis of Covariance with Qualitative Data," *Review of Economic Studies*, 47, 225–238.
- Chamberlain, G. (1982), "Multivariate Regression Models for Panel Data," *Journal of Econometrics*, 18, 5–46.
- Chamberlain, G. (1984), "Panel Data," in *Handbook of Econometrics*, Z. Griliches and M. Intriligator (Eds.), Volume 2, 1247–1318, Amsterdam, North-Holland.
- Chamberlain, G. (1985), "Heterogeneity, Omitted Variable Bias and Duration Dependence," in *Longitudinal Analysis of Labor Market Data*, J. J. Heckman and B. Singer (Eds.), 3–38, Cambridge, UK, Cambridge University Press.
- Chamberlain, G. (1987), "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics*, 34, 305–334.
- Chamberlain, G. (1992), "Comment: Sequential Moment Restrictions in Panel Data," *Journal of Business and Economic Statistics*, 10, 20–26.
- Charlier, E., B. Melenberg and A. van Soest (2001), "An Analysis of Housing Expenditure Using Semiparametric Models and Panel Data," *Journal of Econometrics*, 101, 71–107.
- Chay, K. Y., and B. E. Honoré (1998), "Estimation of Semiparametric Censored Regression Models: An Application to Changes in Black-White Earnings Inequality during the 1960s," *Journal of Human Resources*, 33, 4–38.
- Chay, K. Y., H. Hoynes, and D. Hyslop (2001), "A Nonexperimental Analysis of True State Dependence in Monthly Welfare Participation Sequences," manuscript, Department of Economics, University of California, Berkeley.
- Chay, K. Y., and D. Hyslop (2000), "Identification and Estimation of Dynamic Binary Response Models: Empirical Evidence Using Alternative Approaches," manuscript, Department of Economics, University of California, Berkeley.
- Chay, K. Y., and J. L. Powell (2001), "Semiparametric Censored Regression Models," *Journal of Economic Perspectives*, 15, 29–42.
- Chernozhukov, V., and C. Hansen (2005), "An IV Model of Quantile Treatment Effects," *Econometrica*, forthcoming.

REFERENCES

- Chesher, A. (1984), "Testing for Neglected Heterogeneity," *Econometrica*, 52, 865–872.
- Chesher, A. (1991), "The Effect of Measurement Error," *Biometrika*, 78, 451–462.
- Chesher, A., and M. Irish (1987), "Residual Analysis in the Grouped and Censored Normal Linear Model," *Journal of Econometrics*, 34, 33–62.
- Chib, S. (1992), "Bayes Regression for the Tobit Censored Regression Model," *Journal of Econometrics*, 58, 79–99.
- Chib, S. (1995), "Marginal Likelihood From the Gibbs Output," *Journal of the American Statistical Association*, 90, 1313–1321.
- Chib, S. (2001), "Markov Chain Monte Carlo Methods: Computation and Inference," in *Handbook of Econometrics*, J. J. Heckman and E. E. Leamer (Eds.), Volume 5, 3570–3649, Amsterdam, North-Holland.
- Chib, S., and E. Greenberg (1995), "Understanding the Metropolis-Hastings Algorithm," *The American Statistician*, 49, 327–335.
- Chib, S., and E. Greenberg (1996), "Markov Chain Monte Carlo Simulation Method in Econometrics," *Econometric Theory*, 12, 409–431.
- Chib, S., and I. Jeliazkov (2001), "Marginal Likelihood from the Metropolis-Hastings Output," *Journal of the American Statistical Association*, 96, 270–281.
- Chib, S., and R. Winkelmann (2001), "Markov Chain Monte Carlo Analysis of Correlated Count Data," *Journal of Business and Economic Statistics*, 19, 428–435.
- Cincera, M. (1997), "Patents, R&D, and Technological Spillovers at the Firm Level: Some Evidence from Econometric Count Models for Panel Data," *Journal of Applied Econometrics*, 12, 265–280.
- Clayton, D. G., and J. Cuzick (1985), "Multivariate Generalization of the Proportional Hazards Model (with Discussion)," *Journal of the Royal Statistical Society, A*, 148, 82–117.
- Cleveland, W. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 829–36.
- Cleves, M. R., W. W. Gould, and R. G. Guitirrez (2002), *An Introduction to Survival Analysis Using Stata*, College Station, TX, STATA Press.
- Cochran, W. G. (1977), *Sampling Techniques*, New York, John Wiley.
- Collado, M. D. (1997), "Estimating Dynamic Models from Time Series of Independent Cross-Sections," *Journal of Econometrics*, 82, 37–62.
- Conway, D., and H. V. Roberts (1983), "Reverse Regression, Fairness, and Employment Discrimination," *Journal of Business and Economic Statistics*, 1, 75–85.
- 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.
- Cosslett, S. R. (1981a), "Maximum Likelihood Estimator for Choice-Based Samples," *Econometrica*, 49, 1289–1316.
- Cosslett, S. R. (1981b), "Efficient Estimation of Discrete Choice Models," in *Structural Analysis of Discrete Data with Econometric Applications*, C. F. Manski and D. McFadden (Eds.), 2–50, Cambridge, MA, MIT Press.
- Cox, D. R. (1961), "Tests of Separate Families of Hypotheses," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 105–123, Berkeley, University of California Press.
- Cox, D. R. (1962a), *Renewal Theory*, London, Methuen.
- Cox, D. R. (1962b), "Further Results on Tests of Separate Families of Hypotheses," *Journal of the Royal Statistical Society, B*, 24, 406–424.
- Cox, D. R. (1972), "Regression Models and Life Tables (with Discussion)," *Journal of the Royal Statistical Society, B*, 34, 187–220.
- Cox, D. R. (1975), "Partial Likelihood," *Biometrika*, 62, 269–276.
- Cox, D. R., and D. Oakes (1984), *Analysis of Survival Data*, London, Methuen.

REFERENCES

- Cox, D. R., and E. J. Snell (1968), "A General Definition of Residuals (with Discussion)," *Journal of the Royal Statistical Society, B*, 30, 248–275.
- Cragg, J. G. (1971), "Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods," *Econometrica*, 39, 829–844.
- Cragg, J. G. (1994), "Making Good Inferences from Bad Data," *Canadian Journal of Economics*, 27, 776–800.
- Cragg, J. C. (1997), "Using Higher Moments to Estimate the Simple Error-in-Variables Model," *Rand Journal of Economics*, 28, S71–91.
- Crépon, B., and E. Duguet (1997a), "Research and Development, Competition and Innovation Pseudo-Maximum Likelihood and Simulated Maximum Likelihood Methods Applied to Count Data Models with Heterogeneity," *Journal of Econometrics*, 79, 355–378.
- Crépon, B., and E. Duguet (1997b), "Estimating the Innovation Function from Patent Numbers: GMM on Count Data," *Journal of Applied Econometrics*, 12, 243–264.
- Crépon, B., and B. Mairesse (1995), "The Chamberlain Approach to Panel Data: An Overview and Some Simulations," in *The Econometrics of Panel Data, A Handbook of the Theory with Applications*, L. Mátyás and P. Sevestre (Eds.), 2nd edition, Dordrecht, Kluwer.
- Dagenais, M. G., and D. L. Dagenais (1997), "Higher Moment Estimators for Linear Regression Models with Errors in Variables," *Journal of Econometrics*, 76, 193–221.
- Dagpunar, J. (1988), *Principles of Random Variate Generation*, Oxford, Oxford University Press.
- Dagsvik, J. K., and A. Karlström (2004), "Compensated Variation and Hicksian Choice Probabilities in Random Utility Models that Are Nonlinear in Income," *Review of Economic Studies*, forthcoming.
- Daly, A., and S. Zachary (1979), "Improved Multiple Choice Models," in *Identifying and Measuring the Determinants of Mode Choice*, D. Hensher and Q. Dalvi (Eds.), London, Teakfield.
- Danielsson, J., and J. -F. Richard (1993), "Accelerated Gaussian Importance Sampler with Application to Dynamic Latent Variable Models," *Journal of Applied Econometrics*, 8, S153–173.
- Das, M., W. K. Newey, and F. Vella (2003), "Nonparametric Estimation of Sample Selection Models," *Review of Economic Studies*, 70, 33–58.
- Davey, A., M. J. Shanahan, and J. L. Schafer (2001), "Correcting for Selective Nonresponse in the National Longitudinal Survey of Youth Using Multiple Imputation," *Journal of Human Resources*, 36, 500–519.
- David, H. A., and M. L. Moeschberger (1978), *The Theory of Competing Risks*, New York, Macmillan.
- Davidson, J. (1994), *Stochastic Limit Theory*, Oxford, Oxford University Press.
- Davidson, R., and J. G. MacKinnon (1984), "Model Specification Tests Based on Artificial Regressions," *International Economic Review*, 25, 485–502.
- Davidson, R., and J. G. MacKinnon (1993), *Estimation and Inference in Econometrics*, Oxford, Oxford University Press.
- Davidson, R., and J. G. MacKinnon (2004), *Econometric Theory and Methods*, Oxford, Oxford University Press.
- Davis, P. (2002), "Estimating Multi-way Error Components Models with Unbalanced Data Structures," *Journal of Econometrics*, 106, 67–95.
- Davison, A. C., D. V. Hinkley, and E. Schechtman (1986), "Efficient Bootstrap Simulation," *Biometrika*, 73, 555–566.
- Dean, C., and R. Balshaw (1997), "Efficiency Lost by Analyzing Counts Rather Than Event Times in Poisson and Overdispersed Poisson Regression Models," *Journal of the American Statistical Association*, 92, 1387–1398.

REFERENCES

- Deaton, A. (1985), "Panel Data from Time Series of Cross-Sections," *Journal of Econometrics*, 30, 109–126.
- Deaton, A. (1997), *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*, Baltimore, Johns Hopkins Press.
- Deb, P., and P. K. Trivedi (1997), "Demand for Medical Care by the Elderly: A Finite Mixture Approach," *Journal of Applied Econometrics*, 12, 313–326.
- 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 (2004), "Specification and Simulated Likelihood Estimation of a Non-normal Outcome Model with Selection: Application to Health Care Utilization," revised working paper, Department of Economics, Hunter College.
- Dehejia, R. H., (1997), *Econometric Methods for Program Evaluation*, Ph.D. Dissertation, Harvard University.
- Dehejia, R. H., and S. Wahba (1999), "Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94, 1053–1062.
- Dehejia, R. H., and S. Wahba (2002), "Propensity Score Matching Methods for Nonexperimental Causal Studies," *Review of Economics and Statistics*, 84, 151–161.
- Deistler, M., and H.-G. Seifert (1978), "Identifiability and Consistent Estimability in Econometric Models," *Econometrica*, 46, 969–980.
- Delgado, M. A. (1992), "Semiparametric Generalized Least Squares in the Multivariate Non-linear Regression Model," *Econometric Theory*, 8, 203–222.
- Delgado, M. A., and T. J. Kniesner (1997), "Count Data Models with Variance of Unknown Form: An Application to a Hedonic Model of Worker Absenteeism," *Review of Economics and Statistics*, 79, 41–49.
- Dempster, A. P., N. M. Laird, and D. B. Rubin (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, B*, 39, 1–38.
- Deng, Y., J. M. Quigley, and R. Van Order (2000), "Mortgage Terminations, Heterogeneity and the Exercise of Mortgage Options," *Econometrica*, 68, 275–307.
- Devroye, L. (1986), *Non-uniform Random Variate Generation*, New York, Springer-Verlag.
- Diggle, P. J., P. Heagerty, K.-Y. Liang, and S. L. Zeger (1994, 2002), *Analysis of Longitudinal Data*, 1st and 2nd editions, Oxford, Oxford University Press.
- DiNardo, J., and S. Pischke (1997), "The Returns to Computer Use Revisited: Have Pencils Changed the Wage Structure Too?," *Quarterly Journal of Economics*, 112, 291–303.
- Donald, S. G., and K. Lang (2001), "Inference with Differences in Differences and Other Panel Data," University of Texas Working Paper.
- Donald, S. G., and W. K. Newey (2001), "Choosing the Number of Instruments," *Econometrica*, 69, 1161–1191.
- Dorsey, R. E., and W. J. Mayer (1995), "Genetic Algorithms for Estimation Problems with Multiple Optima, Nondifferentiability, and Other Irregular Features," *Journal of Business and Economic Statistics*, 13, 53–66.
- Drukker, D. M. (2002), "Bootstrapping a Conditional Moments Test for Normality after Tobit Estimation," *The Stata Journal*, 2, 125–139.
- Duan, N., W. G. Manning, C. N. Morris, and J. P. Newhouse (1983), "A Comparison of Alternative Models for the Demand for Medical Care," *Journal of Business and Economic Statistics*, 1, 115–126.
- Dubin, J. A., and D. L. McFadden (1984), "An Econometric Analysis of Residential Electric Appliance Holdings and Consumption," *Econometrica*, 52, 345–362.
- DuMouchel, W. K., and G. J. Duncan (1983), "Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples," *Journal of the American Statistical Association*, 78, 535–543.

REFERENCES

- Durbin, J. (1954), "Errors in Variables," *Review of the International Statistical Institute*, 22, 23–32.
- Efron, B. (1979), "Bootstrapping Methods: Another Look at the Jackknife," *Annals of Statistics*, 7, 1–26.
- Efron, B. (1982), *The Bootstrap and Other Resampling Plans*, Philadelphia, Society for Industrial and Applied Mathematics.
- Efron, B. (1987), "Better Bootstrap Confidence Intervals (with Discussion)," *Journal of the American Statistical Association*, 82, 171–200.
- Efron, B., and J. Tibsharani (1993), *An Introduction to the Bootstrap*, London, Chapman and Hall.
- Eicker, F. (1963), "Asymptotic Normality and Consistency of the Least Squares Estimators for Families of Linear Regressions," *Annals of Mathematical Statistics*, 34, 447–456.
- Eicker, F. (1967), "Limit Theorems for Regressions with Unequal and Dependent Errors," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, L. LeCam and J. Neyman (Eds.), 59–82, Berkeley, CA, University of California Press.
- Elbers, C., and G. Ridder (1982), "True and Spurious Duration Dependence: The Identifiability of Proportional Hazards Model," *Review of Economic Studies*, 49, 403–410.
- Engel, E. (1857), "Die Produktions- und Consumptionsverhältnisse des Königreichs Sachsen," *Zeitschrift des Statistischen Büros des Königlich Sächsischen Ministerium des Innern*, 22 November 1857. Reprinted in 1895 as appendix to E. Engel, "Die Lebenskosten belgischer Arbeiter-Familien früher und jetzt," *Bulletin de l'Institut International de Statistique*, 9, 1–124.
- Engle, R. F., C. W. J. Granger, J. Rice, and A. Weiss (1986), "Semiparametric Estimates of the Relationship between Weather and Electricity Sales," *Journal of the American Statistical Association*, 81, 310–320.
- Engle, R. F., D. F. Hendry, and J-F. Richard (1983), "Exogeneity," *Econometrica*, 51, 277–304.
- Engle, R. F., and J. R. Russell (1998), "Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data," *Econometrica*, 66, 1127–1162.
- Fahrmeir, L., and G. Tutz (1994), *Multivariate Statistical Modelling Based on Generalized Linear Models*, New York, Springer-Verlag.
- Fan, J., and I. Gijbels (1996), *Local Polynomial Modelling and Its Applications*, London, Chapman and Hall.
- Ferall, C. (1997), "Unemployment Insurance Eligibility and School-to-Work Transition in Canada and the United States," *Journal of Business and Economic Statistics*, 15, 115–129.
- Ferguson, T. S. (1958), "A Method of Generating Best Asymptotically Normal Estimates with Application to the Estimation of Bacterial Densities," *Annals of Mathematical Statistics*, 29, 1046–1062.
- Fisher, R. A. (1922), "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society of London A*, 222, 309–368.
- Fisher, R. A. (1928), *Statistical Methods for Research Workers*, 2nd edition, London, Oliver and Boyd.
- Fleming, T. R., and D. P. Harrington (1991), *Counting Process and Survival Analysis*, New York, John Wiley.
- Flinn, C., and J. J. Heckman (1982), "New Methods for Analyzing Structural Models of Labor Force Dynamics," *Journal of Econometrics*, 18, 115–168.
- Freedman, D. A. (1984), "On Bootstrapping Two-Stage Least-Squares Estimates in Stationary Linear Models," *Annals of Statistics*, 3, 827–842.
- Freedman, D. A. (1999), "From Association to Causation: Some Remarks on the History of Causation," *Statistical Science*, 14, 243–258.

REFERENCES

- Frees, E. W., and E. A. Valdez (1998), "Understanding Relationships Using Copulas," *North American Actuarial Journal*, 1, 1–25.
- Friedlander, D., D. Greenberg, and P. Robins (1997), "Evaluating Government Training Programs for the Economically Disadvantaged," *Journal of Economic Literature*, 35, 1809–1855.
- Friedman, J. H. (1984), "A Variable Span Smoother," LCS Technical Report No. 5, Department of Statistics, Stanford University.
- Fuller, W. (1987), *Measurement Error Models*, New York, John Wiley.
- Gallant, A. R. (1981), "On the Bias in Flexible Functional Forms and an Essentially Unbiased Form: The Fourier Flexible Form," *Journal of Econometrics*, 15, 211–245.
- Gallant, A. R. (1987), *Nonlinear Statistical Models*, New York, John Wiley.
- Gallant, A. R., and D. W. Nychka (1987), "Semiparametric Maximum Likelihood Estimation," *Econometrica*, 55, 363–390.
- Gallant, A. R., and G. Tauchen (1996), "Which Moments to Match?," *Econometric Theory*, 12, 657–681.
- Gallant, A. R., and H. White (1987), *A Unified Theory of the Estimation and Inference for Nonlinear Statistical Models*, New York, Basil Blackwell.
- Gamerman, D. (1997), *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, London, Chapman and Hall.
- Geil, P., A. Melion, R. Rotte, and K. F. Zimmermann (1997), "Economic Incentives and Hospitalization," *Journal of Applied Econometrics*, 12, 295–311.
- Gelfand, A. E., and A. F. M. Smith (1990), "Sampling Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.
- Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995), *Bayesian Data Analysis*, London, Chapman and Hall.
- Gelman, A., and D. B. Rubin (1992), "Inference from Iterative Simulations Using Multiple Sequences," *Statistical Science*, 7, 457–511.
- Geman, S., and D. Geman (1984), "Stochastic Relaxation, Gibbs Distributions and Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.
- Georges, P., A.-G. Lamy, E. Nicolas, G. Quibel, and T. Roncalli (2001), "Multivariate Survival Modeling: A Unified Approach with Copulas". Preprint from Groupe de Recherche Opérationnelle, Credit Lyonnais, France.
- Geweke, J. (1988), "Antithetic Acceleration of Monte Carlo Integration in Bayesian Inference," *Journal of Econometrics*, 38, 73–89.
- Geweke, J. (1989), "Bayesian Inference in Econometric Models Using Monte Carlo Integration," *Econometrica*, 57, 1317–1339.
- Geweke, J. (1992), "Evaluating the Accuracy of Sampling-based Approaches to the Calculation of Posterior Moments (with Discussion)," in *Bayesian Statistics*, J. Bernardo, J. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), Volume 4, 169–193, Oxford, Oxford University Press.
- Geweke, J. (1995), "Monte Carlo Simulation and Numerical Integration," Federal Reserve Bank of Minneapolis, Research Department Staff Report 192.
- Geweke, J., and M. Keane (2001), "Computationally Intensive Methods for Integration in Econometrics," in *Handbook of Econometrics*, J. J. Heckman, and E. E. Leamer (Eds.), Volume 5, 3463–3567, Amsterdam, North-Holland.
- Geweke, J., M. Keane, and D. Runkle (1997), "Comparing Simulation Estimators for the Multinomial Probit Model," *Journal of Econometrics*, 80, 125–166.
- Gill, J. (2002), *Bayesian Methods: A Social and Behavioral Sciences Approach*, Boca Raton, FL, Chapman and Hall.

REFERENCES

- Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (1995), *Markov Chain Monte Carlo in Practice*, London, Chapman and Hall.
- Girma, S. (2000), “A Quasi-differencing Approach to Dynamic Modeling from a Time Series of Independent Cross-Sections,” *Journal of Econometrics*, 98, 365–383.
- Gleason, J. R. (1988), “Algorithms for Balanced Bootstrap Simulations,” *The American Statistician*, 42, 263–266.
- Godambe, V. P. (1960), “An Optimum Property of Regular Maximum Likelihood Estimation,” *Annals of Statistics*, 31, 1208–1211.
- Goffe, W. L., Ferrier, G. D., and J. Rogers (1994), “Global Optimization of Statistical Functions with Simulated Annealing,” *Journal of Econometrics*, 60, 65–99.
- Goldberg, P. K. (1995), “Product Differentiation and Oligopoly in International Markets: The Case of the U. S. Automobile Industry,” *Econometrica*, 63, 891–952.
- Goldberger, A. S. (1968), *Topics in Regression Analysis*, New York, Macmillan.
- Goldberger, A. S. (1981), “Linear Regression after Selection,” *Journal of Econometrics*, 15, 357–366.
- Goldberger, A. S. (1984), “Reverse Regression and Salary Discrimination,” *Journal of Human Resources*, 19, 293–318.
- Goldberger, A. S. (1991), *A Course in Econometrics*, Cambridge, MA, Harvard University Press.
- Götze, F., and H. R. Künsch (1996), “Second-Order Correctness of the Blockwise Bootstrap for Stationary Observations,” *Annals of Statistics*, 24, 1914–1933.
- Gouriéroux, C. (2000), *Econometrics of Qualitative Dependent Variables*, translated by Paul B. Klassen, Cambridge, UK, Cambridge University Press.
- Gouriéroux, C., A. Holly, and A. Monfort (1982), “Likelihood Ratio Test, Wald Test, and Kuhn–Tucker Test in Linear Models with Inequality Constraints on the Regression Parameters,” *Econometrica*, 50, 63–80.
- Gouriéroux, C., and A. Monfort (1981), “On the Problem of Missing Data in Linear Models,” *Review of Economic Studies*, 48, 579–586.
- Gouriéroux, C., and A. Monfort (1989), *Statistics and Econometric Models: Volume 2*, translated by Q. Vuong, Cambridge, UK, Cambridge University Press.
- Gouriéroux, C., and A. Monfort (1991), “Simulation Based Inference in Models with Heterogeneity,” *Annales d’Economie et de Statistique*, 20/21, 69–107.
- Gouriéroux, C., and A. Monfort (1996), *Simulation-Based Econometric Methods*, Oxford, Oxford University Press.
- Gouriéroux, C., A. Monfort, and E. Renault (1993), “Indirect Inference,” *Journal of Applied Econometrics*, 8, S85–118.
- Gourieroux, C., A. Monfort, E. Renault, and A. Trognon (1987), “Generalized Residuals,” *Journal of Econometrics*, 34, 5–32.
- Gourieroux, C., A. Monfort, and A. Trognon (1984a), “Pseudo Maximum Likelihood Methods: Theory,” *Econometrica*, 52, 681–700.
- Gourieroux, C., A. Monfort, and A. Trognon (1984b), “Pseudo Maximum Likelihood Methods: Applications to Poisson Models,” *Econometrica*, 52, 701–720.
- Gowrisankaran, G., and R. J. Town (1999) “Estimating the Quality of Care in Hospitals Using Instrumental Variables,” *Journal of Health Economics*, 18, 747–767.
- Grasdal, A. (2001), “The Performance of Sample Selection Estimators to Control for Attrition Bias,” *Health Economics*, 10, 385–398.
- Graubard, B. I., and E. L. Korn (1994), “Regression Analysis with Clustered Data,” *Statistics in Medicine*, 13, 509–522.
- Greene, W. H. (1981), “Sample Selection Bias as a Specification Error: Comment,” *Econometrica*, 49, 795–798.

REFERENCES

- Greene, W. H. (1983), "Reverse Regression: The Algebra of Discrimination," *Journal of Business and Economic Statistics*, 2, 117–120.
- Greene, W. H. (1997), "FIML Estimation of Sample Selection Models for Count Data," Discussion Paper EC-97-02, Department of Economics, Stern School of Business, New York University.
- Greene, W. H. (2003), *Econometric Analysis*, fifth edition, Upper Saddle River, NJ, Prentice-Hall.
- Greene, W. H. (2004a), "The Behavior of the Fixed Effects Estimator in Nonlinear Models," *The Econometrics Journal*, 7, 98–119.
- Greene, W. H. (2004b), "Estimating Econometric Models with Fixed Effects," *Econometric Reviews*, 23, 125–147.
- Greenwald, B. C. (1983), "A General Analysis of Bias in the Estimated Standard Errors of the Least Squares Coefficients," *Journal of Econometrics*, 22, 323–338.
- Gregory, A. W., and M. R. Veall (1985), "On Formulating Wald Test for Nonlinear Restrictions," *Econometrica*, 53, 1465–1468.
- Gritz, M. R. (1993), "The Impact of Training on the Frequency and Duration of Employment," *Journal of Econometrics*, 57, 21–51.
- Gronau, R. (1973), "The Effect of Children on the Housewife's Value of Time," *Journal of Political Economy*, 81, S168–199.
- Grootendorst, P. (1995), "Effects of Drug Plan Eligibility on Prescription Drug Utilization," Ph. D. Dissertation, McMaster University.
- Grosh, M. E., and P. Glewwe (1998), "The World Bank's Living Standards Measurement Study Household Surveys," *Journal of Economic Perspectives*, 12, 187–196.
- Groves, R. M. (1989), *Survey Errors and Survey Costs*, New York, John Wiley.
- Gruber, J., and M. Hanratty (1995), "The Labor-Market Effects of Introducing National Health Insurance: Evidence from Canada," *Journal of Business and Economic Statistics*, 13, 163–173.
- Guo, J. Q., and T. Li (2002), "Poisson Regression Models with Errors-in-Variables," *Journal of Statistical Planning and Inference*, 104, 391–401.
- Guo, J.-Q., and P. K. Trivedi (2002), "Flexible Parametric Distributions for Long-tailed Patent Count Distributions," *Oxford Bulletin of Economics and Statistics*, 64, 63–82.
- Gurmu, S., and P. K. Trivedi (1992), "Overdispersion in the Truncated Poisson Regression Model," *Journal of Econometrics*, 54, 347–370.
- Gurmu, S., and P. K. Trivedi (1996), "Excess Zeros in Count Models for Recreational Trips," *Journal of Business and Economic Statistics*, 14, 469–477.
- Haan, W. J., and A. T. Levin (1997), "A Practitioner's Guide to Robust Covariance Matrix Estimation," in *Handbook of Statistics, Volume 15, Robust Inference*, G. S. Maddala and C. R. Rao (Eds.), 299–342, Amsterdam, North-Holland.
- Hahn, J. (1998), "On the Role of Propensity Score in Efficient Semiparametric Estimation Average Treatment Effects," *Econometrica*, 66, 315–331.
- Hahn, J., and J. A. Hausman (2002), "A New Specification Test for the Validity of Instrumental Variables," *Econometrica*, 70, 163–189.
- Hahn, J., and J. Hausman (2003a), "IV Estimation with Valid and Invalid Instruments," manuscript, Harvard University.
- Hahn, J., and J. Hausman (2003b), "Weak Instruments: Diagnosis and Cures in Empirical Econometrics," *American Economic Review Papers and Proceedings*, 93, 118–125.
- Hahn, J., J. A. Hausman, and G. Kuersteiner (2001), "Higher Order MSE of Jackknife 2SLS," manuscript, Department of Economics, MIT.
- Hahn, J., P. Todd, and W. Van der Klaauw (2001), "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69, 201–209.

REFERENCES

- Haisken-DeNew, J., and C. M. Schmidt (1999), "Money for Nothing and Your Chips for Free? The Anatomy of the PC Wage Differential," IZA Discussion Paper No. 86.
- Hajivassiliou, V. A. (2000), "Some Practical Issues in Maximum Simulated Likelihood," in *Simulation-Based Inference in Econometrics*, R. S. Mariano, T. Schuermann, and M. J. Weeks (Eds.), 71–99, Cambridge, UK, Cambridge University Press.
- Hajivassiliou, V. A., and McFadden, D. (1994), "A Simulation Estimation Analysis of the External Debt Crises of Developing Countries," *Journal of Applied Econometrics*, 9, 109–131.
- Hajivassiliou, V. A., and D. McFadden (1998), "The Method of Simulated Scores for the Estimation of LDV Models," *Econometrica*, 66, 863–896.
- Hajivassiliou, V. A., and P. A. Ruud (1994), "Classical Estimation Methods for LDV Models Using Simulation, in *Handbook of Econometrics*, R. F. Engle and D. L. McFadden (Eds.), Volume 4, 2384–2441, Amsterdam, North-Holland.
- Hall, A. (1987), "The Information Matrix Test for the Linear Model," *Review of Economic Studies*, 54, 257–263.
- Hall, A., G. Rudebusch, and D. Wilcox (1996), "Judging Instrument Relevance in Instrumental Variables Estimation," *International Economic Review*, 37, 283–298.
- Hall, B., Z. Griliches, and J. A. Hausman (1986), "Patents and R&D: Is There a Lag?," *International Economic Review*, 27, 265–283.
- Hall, D. B., and T. A. Severini (1998), "Extended Generalized Estimating Equations for Clustered Data," *Journal of the American Statistical Association*, 93, 1365–1375.
- Hall, P. (1986), "On the Bootstrap and Confidence Intervals," *Annals of Statistics*, 14, 1431–1452.
- Hall, P. (1992), *The Bootstrap and Edgeworth Expansions*, New York, Springer-Verlag.
- Hall, P., and J. L. Horowitz (1996), "Bootstrap Critical Values for Tests Based on Generalized Method of Moments Estimators," *Econometrica*, 64, 891–916.
- Hall, P., and M. A. Martin (1988), "On Bootstrap Resampling and Iteration," *Biometrika*, 75, 661–671.
- Hamilton, J. (1994), *Time Series Analysis*, Princeton, Princeton University Press.
- Han, A. K. (1987), "Non-parametric Analysis of a Generalized Regression Model: The Maximum Rank Correlation Estimator," *Journal of Econometrics*, 35, 303–316.
- Han, A., and J. A. Hausman (1990), "Flexible Parametric Estimation of Duration and Competing Risk Models," *Journal of Applied Econometrics*, 5, 1–28.
- Hanemann, W. M. (1984), "Discrete/Continuous Models of Consumer Demand," *Econometrica*, 52, 541–562.
- Hansen, L. P. (1982), "Large Sample Properties of Generalized Methods of Moments Estimators," *Econometrica*, 1029–1054.
- Hansen, L. P., and K. J. Singleton (1982), "Generalized Instrumental Variables Estimators of Nonlinear Rational Expectations Models," *Econometrica*, 1269–1286.
- Hansen, P. R. (2003), "Asymptotic Tests of Composite Hypotheses," Working Paper 2003-09, Department of Economics, Brown University.
- Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge, UK, Cambridge University Press.
- Härdle, W., and O. Linton (1994), "Applied Nonparametric Methods," in *Handbook in Econometrics*, R. F. Engle and D. L. McFadden (Eds.), Volume 4, 2297–2339, Amsterdam, North-Holland.
- Härdle, W., and E. Mammen (1993), "Comparing Nonparametric versus Parametric Regression Fits," *Annals of Statistics*, 21, 1926–1947.
- Härdle, W., and J. S. Marron (1985), "Optimal Bandwidth Selection in Nonparametric Regression Function Estimation," *Annals of Statistics*, 13, 1465–1481.

REFERENCES

- Härdle, W., and T. M. Stoker (1989), "Investigating Smooth Multiple Regression by the Method of Average Derivatives," *Journal of the American Statistical Association*, 84, 986–995.
- Harris, R. D. F., and E. Tzavalis (1999), "Inference for Unit Roots in Dynamic Panels Where the Time Dimension is Fixed," *Journal of Econometrics*, 91, 201–226.
- Harvey, A. (1990), *The Econometric Analysis of Time Series*, 2nd edition, Cambridge, MA, MIT Press.
- Harville, D. A. (1977), "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems," *Journal of the American Statistical Association*, 72, 320–338.
- Hastie, T. J., and R. J. Tibsharani (1990), *Generalized Additivity Models*, London, Chapman and Hall.
- Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chain and Their Applications," *Biometrika*, 57, 97–109.
- Haughton, D. (1997), "Packages for Estimating Finite Mixtures," *The American Statistician*, 51, 194–205.
- Hausman, J. A. (1978), "Specification Tests in Econometrics," *Econometrica*, 46, 1251–1271.
- Hausman, J. A. (2001), "Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left," *Journal of Economic Perspectives*, 15, 57–68.
- Hausman, J. A., J. Abrevaya, and F. M. Scott-Morton (1998), "Misclassification of the Dependent Variable in a Discrete Response Setting," *Journal of Econometrics*, 87, 239–269.
- 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.
- Hausman, J. A., A. W. Lo, and A. C. MacKinlay (1992), "An Ordered Probit Analysis of Transaction Stock Prices," *Journal of Financial Economics*, 31, 319–379.
- Hausman, J. A., and D. McFadden (1984), "Specification Tests for the Multinomial Logit Model," *Econometrica*, 52, 1219–240.
- Hausman, J. A., W. K. Newey, and J. L. Powell (1995), "Nonlinear Errors in Variables: Estimation of Some Engel Curves," *Journal of Econometrics*, 65, 205–233.
- Hausman, J. A., and W. E. Taylor (1981), "Panel Data and Unobservable Individual Effects," *Econometrica*, 49, 1377–1398.
- Hausman, J. A., and D. A. Wise (1979), "Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment," *Econometrica*, 47, 455–474.
- Hausman, J. A., and D. A. Wise (1985), *Social Experimentation*, NBER Economic Research Conference Report, Chicago and London, University of Chicago Press.
- Hayashi, F. (2000), *Econometrics*, Princeton, NJ, Princeton University Press.
- Heckman, J. J. (1974), "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, 42, 679–694.
- Heckman, J. J. (1976), "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, 5, 475–492.
- Heckman, J. J. (1978), "Dummy Endogenous Variables in a Simultaneous Equations System," *Econometrica*, 46, 931–960.
- Heckman, J. J. (1979), "Sample Selection as a Specification Error," *Econometrica*, 47, 153–161.
- Heckman, J. J. (1996), "Randomization as an Instrumental Variable," *Review of Economics and Statistics*, 78, 336–341.
- Heckman, J. J. (1997), "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations," *Journal of Human Resources*, 32, 441–462.
- Heckman, J. J. (2000), "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Perspective," *Quarterly Journal of Economics*, 115, 45–98.
- Heckman, J. J. (2001), "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture," *Journal of Political Economy*, 109, 673–748.

REFERENCES

- Heckman, J. J., and G. Borjas (1980), "Does Unemployment Cause Future Unemployment? Definitions, Questions, and Answers from Continuous Time Model of Heterogeneity and State Dependence," *Economica*, 47, 247–283.
- Heckman, J. J., and B. E. Honoré (1989), "The Identifiability of the Competing Risks Model," *Biometrika*, 76, 325–330.
- Heckman, J. J., and V. Hotz (1989), "Choosing among Alternative Nonexperimental Methods for Evaluating the Impact of Social Programs," *Journal of the American Statistical Association*, 84, 862–880.
- Heckman, J. J., V. Hotz, and J. Walker (1985), "New Evidence on Timing and Spacing of Births," *American Economic Review Papers and Proceedings*, 75, 179–184.
- Heckman, J. J., H. Ichimura, J. A. Smith, and P. Todd (1998), "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017–1098.
- Heckman, J. J., H. Ichimura, and P. Todd (1997), "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program," *Review of Economic Studies*, 64, 605–654.
- Heckman, J. J., R. J. Lalonde, and J. A. Smith (1999), "The Economics and Econometrics of Active Labor Market Programs," in *Handbook of Labor Economics*, O. Ashenfelter, and D. Card (Eds.), Volume 3A, 1865–2097, Amsterdam, North-Holland.
- Heckman, J. J., and T. E. MacCurdy (1980), "A Life-Cycle Model of Female Labor Supply," *Review of Economic Studies*, 47, 47–74.
- Heckman, J. J., and R. Robb (1985), "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data*, J. J. Heckman and B. Singer (Eds.), Cambridge, UK, Cambridge University Press.
- Heckman, J. J., and B. Singer (1984a), "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models of Duration Data," *Econometrica*, 52, 271–320.
- Heckman, J. J., and B. Singer (1984b), "The Identifiability of the Proportional Hazard Model," *Review of Economic Studies*, 51, 231–241.
- Heckman, J. J., and J. A. Smith (1995), "Assessing the Case for Social Experiments," *Journal of Economic Perspectives*, 9, 85–110.
- Heckman, J. J., and J. A. Smith (1998), "Evaluating the Welfare State," in *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial Symposium*, S. Strøm, (Ed.), 241–318, Cambridge, UK, Cambridge University Press.
- Heckman, J. J., J. I. Tobias, and E. Vytlacil (2003), "Simple Estimators for Treatment Parameters in a Latent-Variable Framework," *Review of Economics and Statistics*, 85, 748–755.
- Heckman, J. J., and E. Vytlacil (1998), "Instrumental Variable Methods for the Correlated Random Coefficient Model: Estimating the Average Rate of Return to Schooling When the Return Is Correlated with Schooling," *Journal of Human Resources*, 33, 974–987.
- Heckman, J. J., and E. J. Vytlacil (2001), "Local Instrumental Variables," C. Hsiao, K. Morimune and J. L. Powell (Eds.), *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, 1–46, Cambridge, UK, Cambridge University Press.
- Hendry, D. F. (1984), "Monte Carlo Experimentation in Econometrics," in *Handbook of Econometrics*, Z. Griliches and M. Intrilligator (Eds.), Volume 2, 937–976, Amsterdam, North-Holland.
- Hendry, D. F. (1995), *Dynamic Econometrics*, Oxford, Oxford University Press.
- Hendry, D. F., and M. S. Morgan (1996), *Foundations of Econometrics*, Cambridge and New York, Cambridge University Press.
- Herriges, J. A., and C. L. Kling (1999), "Nonlinear Income Effects in Random Utility Models," *Review of Economics and Statistics*, 81, 62–72.

REFERENCES

- Heyde, C. C., and I. M. Johnstone (1979), "On Asymptotic Posterior Normality of Stochastic Processes," *Journal of the Royal Statistical Society, B*, 41, 184–189.
- Hirano, K., G. Imbens, and G. Ridder (2003), "Efficient Estimation of Average Treatment Effects Using Estimated Propensity Score," *Econometrica*, 71, 1161–1190.
- Hirano, K., and J. Porter (2003), "Asymptotic Efficiency in Parametric Structural Models with Parameter-Dependent Support," *Econometrica*, 71, 1307–1338.
- Hoch, I. (1962), "Estimation of Production Function Parameters Combining Time-Series and Cross-Section Data," *Econometrica*, 30, 34–53.
- Hoerl, A. E., and R. W. Kennard (1970), "Ridge Regression: Applications to Non-orthogonal Problems," *Technometrics*, 12, 69–82.
- Holland, P. W. (1986), "Statistics of Causal Inference," *Journal of the American Statistical Association*, 81, 945–960.
- Holly, A. (1982), "A Remark on Hausman's Specification Test," *Econometrica*, 49, 749–759.
- Holly, A. (1987), "Specification Tests: an Overview," in *Advances in Economics and Econometrics: Theory and Applications*, T. F. Bewley (Ed.), Volume 1, 59–97, Cambridge, UK, Cambridge University Press.
- Holtz-Eakin, D., W. Newey, and H. S. Rosen (1988), "Estimating Vector Autoregressions with Panel Data," *Econometrica*, 56, 1371–1395.
- Honoré, B. E. (1992), "Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects," *Econometrica*, 60, 533–565.
- Honoré, B. E. (1993), "Identification Results for Duration Models with Multiple Spells," *Review of Economic Studies*, 60, 241–246.
- Honoré, B. E., and E. Kyriazidou (2000), "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica*, 88, 839–874.
- Honoré, B. E., and J. Powell (1994), "Pairwise Difference Estimators for Censored and Truncated Models," *Journal of Econometrics*, 64, 241–278.
- Horowitz, J. L. (1992), "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica*, 60, 505–531.
- Horowitz, J. L. (1994), "Bootstrap-Based Critical Values for the Information Matrix Test," *Journal of Econometrics*, 61, 395–411.
- Horowitz, J. L. (1997), "Bootstrap Methods in Econometrics: Theory and Numerical Performance," in *Advances in Economics and Econometrics: Theory and Applications, Seventh World Congress*, D. M. Kreps and K. F. Wallis (Eds.), Volume 3, 188–222, Cambridge, UK, Cambridge University Press.
- Horowitz, J. L. (1998a), "Bootstrap Methods for Covariance Structures," *Journal of Human Resources*, 33, 39–61.
- Horowitz, J. L. (1998b), *Semiparametric Methods in Econometrics*, Lecture Notes in Statistics No. 131, New York, Springer-Verlag.
- Horowitz, J. L. (2001), "The Bootstrap," in *Handbook of Econometrics*, J. J. Heckman and E. Leamer (Eds.), Volume 5, 3159–3228, Amsterdam, North-Holland.
- Horowitz, J. L. (2002), "Bootstrap Critical Values for Tests Based on the Smoothed Maximum Score Estimator," *Journal of Econometrics*, 111, 141–167.
- Horowitz, J. L., and W. Härdle (1994), "Testing a Parametric Model against a Nonparametric Alternative," *Econometric Theory*, 10, 821–848.
- Horowitz, J. L., and W. Härdle (1996), "Direct Semiparametric Estimation of Single-Index Models with Discrete Covariates," *Journal of the American Statistical Association*, 91, 1632–1640.
- Horowitz, J. L., and S. Lee (2004), "Semiparametric Estimation of a Panel Data Proportional Hazard Model with Fixed Effects," *Journal of Econometrics*, 119, 155–198.

REFERENCES

- Horton, N. J., and S. R. Lipsitz (2001), "Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables," *The American Statistician*, 55, 244–254.
- Hosmer, D. W., and S. Lemeshow (1999), *Applied Survival Analysis: Regression Analysis of Time to Event Data*, New York, Wiley Interscience.
- Hotz, J. (1992), "Designing an Evaluation of the Job Training Partnership Act," in *Evaluating Welfare and Training Programs*, C. Manski and I. Garfinkel (Eds.), Cambridge, MA, Harvard University Press.
- Hougaard, P. (1984), "Life Table Methods for Heterogenous Populations: Distributions Describing the Heterogeneity," *Biometrika*, 71, 75–83.
- Hougaard, P. (1986), "Survival Models for Heterogeneous Populations Derived from Stable Distributions," *Biometrika*, 73, 387–396.
- Hougaard, P. (1995), "Frailty Models for Survival Data," *Lifetime Data Analysis*, 1, 255–274.
- Hougaard, P. (2000), *Analysis of Multivariate Survival Data*, New York, Springer-Verlag.
- Houthakker, H. S. (1957), "An International Comparison of Household Expenditure Patterns," *Econometrica*, 25, 532–552.
- Hsiao, C. (1986, 2003), *Analysis of Panel Data*, 1st and 2nd editions, Cambridge, UK, Cambridge University Press.
- Hsiao, C. (1989), "Consistent Estimation for Some Nonlinear Errors-in-Variables Models," *Journal of Econometrics*, 41, 159–185.
- Hsiao, C. (1992), "Nonlinear Latent Variable Models," in *The Econometrics of Panel Data*, L. Mátyás and P. Sevestre (Eds.), Dordrecht, Kluwer Academic Publishers.
- Hsiao, C. (1995), "Logit and Probit Models," in *The Econometrics of Panel Data: A Handbook of the Theory with Applications*, 3rd edition, L. Mátyás and P. Sevestre (Eds.), Dordrecht, Kluwer Academic Publishers.
- Hsiao, C., M. H. Pesaran, and A. K. Tahmisioglu (2002), "Maximum Likelihood Estimation of Fixed Effects Dynamic Panel Data Models Covering Short Periods," *Journal of Econometrics*, 109, 107–150.
- Hsiao, C., and B. H. Sun (1999), "Modeling Survey Response Bias – With an Analysis of the Demand for an Advanced Electronic Device," *Journal of Econometrics*, 89, 15–19.
- Huber, P. J. (1967), "The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions," in *Proceedings of the Fifth Berkeley Symposium*, J. Neyman (Ed.), 1, 221–233, Berkeley, CA, University of California Press.
- Huber, P. J. (1981), *Robust Statistics*, New York, John Wiley.
- Ibrahim, J. G., M.-H. Chen, and D. Sinha (2001), *Bayesian Survival Analysis*, New York, Springer-Verlag.
- Ichimura, H. (1993), "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models," *Journal of Econometrics*, 58, 71–120.
- Imbens, G. W. (1992), "An Efficient Method of Moments Estimator for Discrete Choice Models with Choice-Based Sampling," *Econometrica*, 60, 1187–1214.
- Imbens, G. W. (2002), "Generalized Method of Moments and Empirical Likelihood," *Journal of Business and Economic Statistics*, 20, 493–506.
- Imbens, G. (2005), "Semiparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *Review of Economics and Statistics*, forthcoming.
- Imbens, G. W., and J. Angrist (1994), "Identification and Estimation of Local Average Treatment Effect," *Econometrica*, 62, 467–475.
- Imbens, G. W., and T. Lancaster (1996), "Efficient Estimation and Stratified Sampling," *Journal of Econometrics*, 74, 289–318.
- Isacsson, G. (1999), *Essays on the Twins Approach in Empirical Labor Economics*, Swedish Institute for Social Research.

REFERENCES

- Jaggia, S. (1991a), "Tests of Moment Restrictions in Parametric Duration Models," *Economics Letters*, 37, 35–38.
- Jaggia, S. (1991b), "The Choice of a Mixing Distribution in Duration Models," *Economics Letters*, 37, 405–409.
- Jaggia, S. (1991c), "Specification Tests Based on the Heterogeneous Generalized Gamma Model of Duration: With an Application to Kennan's Strike Data," *Journal of Applied Econometrics*, 6, 169–180.
- Jaggia, S. (1997), "Alternative Forms of the Score Test for Heterogeneity in a Censored Exponential Model," *Review of Economics and Statistics*, 79, 340–343.
- Jaggia, S., and P. K. Trivedi (1994), "Joint and Separate Score Tests for State Dependence and Unobserved Heterogeneity," *Journal of Econometrics*, 60, 273–291.
- Jennrich, R. I. (1969), "Asymptotic Properties of Non-linear Least Squares Estimators," *Annals of Mathematical Statistics*, 40, 633–643.
- Jensen, E. R. (1999), "An Econometric Analysis of the Old-Age Security Motive for Childbearing," *International Economic Review*, 31, 953–968.
- Joe, H. (1997), *Multivariate Models and Dependence Concepts*, London, Chapman and Hall.
- Johnson, N. L., and S. Kotz (1970), *Distributions in Statistics: Continuous Univariate Distributions – I*, New York, John Wiley.
- Johnson, N. L., S. Kotz, and A. W. Kemp (1992), *Univariate Distributions*, 2nd edition, New York, John Wiley.
- Johnston, J., and J. DiNardo (1997), *Econometric Methods*, 4th edition, New York, McGraw-Hill.
- Jordan, P., D. Brubacher, S. Tsugane, Y. Tsubono, K. F. Gey, and U. Moser (1997), "Modelling of Mortality Data from a Multi-Centre Study in Japan by Means of Poisson Regression with Error in Variables," *International Journal of Epidemiology*, 26, 501–506.
- Kalbfleisch, J., and R. Prentice (1980, 2002), *The Statistical Analysis of Failure Time Data*, 1st and 2nd editions, New York, John Wiley.
- Kamakura, W. A., and M. Wedel (2004), "An Empirical Bayes Procedure for Improving Individual-Level Estimates and Predictions from Finite Mixtures of Multinomial Logit Models," *Journal of Business and Economic Statistics*, 22, 121–125.
- Kass, R. E., and A. E. Raftery (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.
- Keane, M. P. (1992), "A Note on Identification in the Multinomial Probit Model," *Journal of Business and Economic Statistics*, 10, 193–200.
- Keane, M. P. (1994), "A Computationally Practical Simulation Estimator for Panel Data," *Econometrica*, 62, 95–116.
- Keljian, H. H. (1980), "Aggregation and Disaggregation of Nonlinear Equations," in *Evaluation of Econometric Models*, J. Kmenta and J. B. Ramsey (Eds.), New York, Academic Press.
- Kenkel, D. S., and J. V. Terza (2001), "The Effect of Physician Advice on Alcohol Consumption: Count Regression with an Endogenous Treatment Effect," *Journal of Applied Econometrics*, 16, 165–184.
- Kennan, J. (1985), "The Duration of Contract Strikes in U.S. Manufacturing," *Journal of Econometrics*, 28, 5–28.
- Kennedy, W., and J. Gentle (1980), *Statistical Computing*, New York, Marcel Dekker.
- Kennickell, A. B. (1998), "Multiple Imputation in the Survey of Consumer Finances," Working Paper, Board of Governors of the Federal Reserve System, Washington, D.C.
- Kiefer, N. (1988), "Economic Duration Data and Hazard Functions," *Journal of Economic Literature*, 646–679.

REFERENCES

- Killingsworth, M., and J. Heckman (1986), "Female Labor Supply: A Survey," in *Handbook of Labor Economics*, O. Ashenfelter and R. Layard (Eds.), Volume 1, 103–144, Amsterdam, North-Holland.
- Kim, E. H., and V. Singal (1993), "Mergers and Market Power: Evidence from the Airline Industry," *American Economic Review*, 83, 549–569.
- King, G., and L. Zeng (2001), "Logistic Regression in Rare Events Data," *Political Analysis*, 9, 137–163.
- Kish, L. (1965), *Survey Sampling*, New York, John Wiley.
- Klein, J. P., and M. L. Moeschberger (1997), *Survival Analysis: Techniques for Censored and Truncated Data*, New York and Berlin, Springer-Verlag.
- Klein, R. W., and R. Sherman (1997), "Estimating New Product Demand from Biased Survey Data," *Journal of Econometrics*, 76, 53–76.
- Klein, R. W., and R. P. Sherman (2002), "Shift Restrictions and Semiparametric Estimation in Ordered Response Models," *Econometrica*, 76, 663–692.
- Klein, R. W., and R. H. Spady (1993), "An Efficient Semi-Parametric Estimator for Binary Response Models," *Econometrica*, 61, 387–423.
- Klepper, S., and E. Leamer (1984), "Consistent Set of Estimates for Regression with Errors in All Variables," *Econometrica*, 52, 163–183.
- Kling, J. R. (2001), "Interpreting Instrumental Variables Estimates of the Returns to Schooling," *Journal of Business and Economic Statistics*, 19, 358–364.
- Kloek, T. (1981), "OLS Estimation in a Model Where a Microvariable Is Explained by Aggregates and Contemporaneous Disturbances Are Equicorrelated," *Econometrica*, 49, 205–207.
- Kloek, T., and H. K. van Dijk (1978), "Bayesian Estimates of Equation System Parameters: An Application of Integration by Monte Carlo," *Econometrica*, 46, 1–19.
- Kmenta J. (1986), *Elements of Econometrics*, 2nd edition, New York, Macmillan.
- Koenker, R. (1981), "A Note on Studentizing a Test for Heteroscedasticity," *Journal of Econometrics*, 17, 107–112.
- Koenker, R., and G. Bassett (1978), "Regression Quantiles," *Econometrica*, 46, 33–50.
- Koenker, R., and G. Bassett (1982), "Robust Tests for Heteroscedasticity Based on Regression Quantiles," *Econometrica*, 50, 43–62.
- Koenker, R., and K. F. Hallock (2001), "Quantile Regression," *Journal of Economic Perspectives*, 15, 143–156.
- Koop, G. (2003), *Bayesian Econometrics*, New York, Wiley.
- Koopmans, T. C. (1949), "Identification Problems in Economic Model Construction," *Econometrica*, 17, 125–144.
- Koppelman, F., and C. Wen (1998), "Alternative Nested Logit Models: Structure, Properties and Estimation," *Transportation Research B*, 34, 75–89.
- Krashinsky, H. (2004), "Do Marital Status and Computer Usage Really Change the Wage Structure?," *Journal of Human Resources*, forthcoming.
- Krewski, D., and J. N. K. Rao (1981), "Inference from Stratified Samples: Properties of the Linearization, Jackknife and Balanced Repeated Replication Methods," *Annals of Statistics*, 9, 1010–1019.
- Krueger, A. B. (1993), "How Computers Have Changed the Wage Structure: Evidence from Microdata, 1984–1989," *Quarterly Journal of Economics*, 108, 33–60.
- Kuan, C. M., and H. White (1994), "Artificial Neural Networks: An Econometric Perspective," *Econometric Reviews*, 13, 1–91.
- Kuh, E. (1959), "The Validity of Cross-Sectionally Estimated Behavior Equations in Time-Series Applications," *Econometrica*, 27, 197–214.
- Kyriazidou, E. (1997), "Estimation of a Panel Data Sample Selection Model," *Econometrica*, 65, 1335–1364.

REFERENCES

- Lalonde, R. (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76, 604–620.
- Lancaster, T. (1979), "Econometric Models for the Duration of Unemployment," *Econometrica*, 47, 939–956.
- Lancaster, T. (1984), "The Covariance Matrix of the Information Matrix Test," *Econometrica*, 52, 1051–1054.
- Lancaster, T. (1985), "Generalized Residuals and Heterogeneous Duration Models," *Journal of Econometrics*, 28, 155–169.
- Lancaster, T. (1990), *The Econometric Analysis of Transitional Data*, Cambridge, UK, Cambridge University Press.
- Lancaster, T. (2000), "The Incidental Parameter Problem since 1948," *Journal of Econometrics*, 95, 391–413.
- Lancaster, T. (2002), "Orthogonal Parameters and Panel Data," *Review of Economic Studies*, 69, 647–666.
- Lancaster, T. (2004), *An Introduction to Modern Bayesian Econometrics*, Oxford, Blackwell.
- Lancaster, T., and G. Imbens (1996), "Case-Control with Contaminated Controls," *Journal of Econometrics*, 71, 145–160.
- Lancaster, T., and S. Nickell (1980), "The Analysis of Re-Employment Probabilities for the Unemployed," *Journal of the Royal Statistical Society, A*, 143, 141–152.
- Lawless, J. F. (1982), *Statistical Models and Methods for Lifetime Data*, New York, John Wiley.
- Leamer, E. E. (1978), *Specification Searches: Ad Hoc Inference with Nonexperimental Data*, New York, John Wiley.
- Lee, C.-I. (2001), "Finite Sample Bias in IV Estimation of Intertemporal Labor Supply Models: Is the Intertemporal Substitution Elasticity Really Small?" *Review of Economics and Statistics*, 83, 638–646.
- Lee, L.-F. (1978), "Unionism and Wage Rates: A Simultaneous Equation Model with Qualitative and Limited Dependent Variables," *International Economic Review*, 19, 415–433.
- Lee, L.-F. (1981), "Simultaneous Equations Models with Discrete and Censored Dependent Variables," in *Structural Analysis of Discrete Data with Econometric Applications*, C. F. Manski and D. McFadden (Eds.), 346–364, Cambridge, MA, MIT Press.
- Lee, L.-F. (1983), "Generalized Econometric Models with Selectivity," *Econometrica*, 51, 507–512.
- Lee, L.-F. (1995), "Semiparametric Maximum Likelihood Estimation of Polychotomous and Sequential Choice Models," *Journal of Econometrics*, 65, 381–428.
- Lee, L.-F. (2001), "Self-Selection," in *A Companion to Theoretical Econometrics*, B. Baltagi (Ed.), 383–409, Oxford, Blackwell.
- Lee, L.-F. (2004), "Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Autoregressive Models," *Econometrica*, 72, 1899–1926.
- Lee, L.-F., G. S. Maddala, and R. P. Trost (1980), "Asymptotic Covariance Matrices of Two-Stage Probit and Two-Stage Tobit Methods for Simultaneous Equations Models with Selectivity," *Econometrica*, 48, 491–504.
- Lee, L.-F., and J. H. Sepanski (1995), "Estimation of Linear and Nonlinear Errors-in-Variable Models Using Validation Data," *Journal of the American Statistical Association*, 90, 130–140.
- Lee, M.-J. (1989), "Mode Regression," *Journal of Econometrics*, 42, 337–349.
- Lee, M.-J. (1996), *Methods of Moments and Semiparametric Econometrics for Limited Dependent Variable Models*, Berlin/New York, Springer-Verlag.
- Lee, M.-J. (2002), *Panel Data Econometrics: Methods-of-Moments and Limited Dependent Variables*, San Diego, Academic Press.

REFERENCES

- Lerman, S. R., and C. F. Manski (1981), "On the Use of Simulated Frequencies to Approximate Choice Probabilities," in *Structural Analysis of Discrete Data with Econometric Applications*, C. F. Manski, and D. McFadden (Eds.), 305–319, Cambridge, MA, MIT Press.
- Leung, S. F., and S. Yu (1996), "On the Choice between Sample Selection and Two-Part Models," *Journal of Econometrics*, 72, 197–229.
- Levi, M. (1973), "Errors in the Variables Bias in the Presence of Correctly Measured Variables," *Econometrica*, 41, 985–986.
- Levin, L., and C.-F. Lin (1992), "Unit Root Tests in Panel Data: Asymptotic and Finite-Sample Properties," UCSD Department of Economics Working Paper 92-23.
- Levin, L., C.-F. Lin and S. J. C. Chu (2002), "Unit Root Tests in Panel Data: Asymptotic and Finite-Sample Properties," *Journal of Econometrics*, 108, 1–24.
- Li, H., and G. S. Maddala (1997), "Bootstrapping Cointegrating Regressions," *Journal of Econometrics*, 80, 297–318.
- Li, T. (2002), "Robust and Consistent Estimation of Non-linear Errors-in-Variables Models," *Journal of Econometrics*, 110, 1–26.
- Li, T., P. K. Trivedi, and J. Q. Guo (2003), "Modeling Response Bias in Count: A Structural Approach with an Application to the National Crime Victimization Survey Data," *Sociological Methods and Research*, 31, 514–544.
- Liang, K.-Y., and S. L. Zeger (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22.
- Lien, D., and Q. H. Vuong (1987), "Selecting the Best Linear Regression Model: A Classical Approach," *Journal of Econometrics*, 35, 3–23.
- Lillard, L., J. P. Smith, and F. Welch (1986), "What Do We Really Know about Wages? The Importance of Nonreporting and Census Imputation," *Journal of Political Economy*, 94, 489–506.
- Lindeboom, M., and G. J. Van den Berg (1994), "Heterogeneity in Bivariate Duration Models: The Importance of the Mixing Distribution," *Journal of the Royal Statistical Society, B*, 56, 49–60.
- Lindley, D. V., and Smith, A. F. M. (1972), "Bayes Estimates for the Linear Model," *Journal of the Royal Statistical Society, B*, 34, 1–41.
- Lindsey, B. G. (1995), *Mixture Models: Theory, Geometry and Applications*, Hayward, CA, Institute of Mathematical Statistics.
- Little, R. J. A. (1988), "Missing-Data Adjustment in Large Surveys" (with discussion), *Journal of Business and Economic Statistics*, 6, 287–302.
- Little, R. J. A. (1992), "Regression with Missing X's: A Review," *Journal of the American Statistical Association*, 97, 1227–1237.
- Little, R. J. A., and D. Rubin (1987), *Statistical Analysis with Missing Data*, New York, John Wiley.
- Liu, R. Y. (1988), "Bootstrap Procedures under Some Non-iid Models," *Annals of Statistics*, 16, 1696–1708.
- Loh, W.-Y. (1987), "Calibrating Confidence Coefficients," *Journal of the American Statistical Association*, 82, 155–162.
- Long, J. S., and L. H. Ervin (2000), "Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model," *The American Statistician*, 54, 217–224.
- Lovell, M. C. (1983), "Data Mining," *Review of Economics and Statistics*, 45, 1–12.
- Luce, R. D. (1959), *Individual Choice Behavior: A Theoretical Analysis*, New York, John Wiley.
- MacCurdy, T. E. (1981), "An Empirical Model of Labor Supply in a Life-Cycle Setting, *Journal of Political Economy*, 89, 1059–1085.
- MacCurdy, T. E. (1982a), "Using Information on the Moments of Disturbances to Increase the Efficiency of Estimation," NBER Technical Paper No. 22.

REFERENCES

- MacCurdy, T. E. (1982b), "The Use of Time Series Processes to Model the Error Structure of Earnings in a Longitudinal Data Analysis," *Journal of Econometrics*, 18, 83–114.
- MacCurdy, T. E. (1983), "A Simple Scheme for Estimating an Intertemporal Model of Labor Supply and Consumption in the Presence of Taxes and Uncertainty," *International Economic Review*, 24, 265–289.
- MacKinnon, J. G. (2002), "Bootstrap Inference in Econometrics," *Canadian Journal of Economics*, 35, 615–645.
- MacKinnon, J. G., and H. White (1985), "Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties," *Journal of Econometrics*, 29, 305–325.
- Madansky, A. (1959), "The Fitting of Straight Lines When Both Variables Are Subject to Error," *Journal of the American Statistical Association*, 54, 173–205.
- Maddala, G. S. (1977), *Econometrics*, New York, McGraw-Hill.
- Maddala, G. S. (1983), *Limited-Dependent and Qualitative Variables in Economics*, Cambridge, UK, Cambridge University Press.
- Magdalinos, M. A. (1988), "The Local Power of the Tests of Overidentifying Restrictions," *International Economic Review*, 29, 509–524.
- Magnac, T. (2000), "Subsidised Training and Youth Employment: Distinguishing Unobserved Heterogeneity from State Dependence in Labour Market Histories," *Economic Journal*, 110, 805–837.
- Mammen, E. (1993), "Bootstrap and Wild Bootstrap for High Dimensional Linear Models," *Annals of Statistics*, 21, 255–285.
- Manning, W. G., J. P. Newhouse, N. Duan, E. B. Keeler, A. Leibowitz, and M. S. Marquis (1987), "Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment," *American Economic Review*, 77, 251–277.
- Manski, C. F. (1975), "The Maximum Score Estimator of the Stochastic Utility Model of Choice," *Journal of Econometrics*, 3, 205–228.
- Manski, C. F. (1985), "Semiparametric Analysis of Discrete Response : Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics*, 27, 313–333.
- Manski, C. F. (1987), "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data," *Econometrica*, 55, 357–362.
- Manski, C. F. (1988a), *Analog Estimation Methods in Econometrics*, London, Chapman and Hall.
- Manski, C. F. (1988b), "Identification of Binary Response Models," *Journal of the American Statistical Association*, 43, 729–738.
- Manski, C. F. (1989), "Anatomy of the Selection Problem," *Journal of Human Resources*, 24, 343–360.
- Manski, C. F. (1991), "Regression," *Journal of Economic Literature*, 39, 34–50.
- Manski, C. F. (1995), *Identification Problems in the Social Sciences*, Cambridge, MA, Harvard University Press.
- Manski, C. F., and S. R. Lerman (1977), "The Estimation of Choice Probabilities from Choice-Based Samples," *Econometrica*, 45, 1977–1988.
- Manski, C. F., and D. McFadden (1981a), "Alternative Estimators and Sample Design for Discrete Choice Analysis," in *Structural Analysis of Discrete Data with Econometric Applications*, C. F. Manski and D. McFadden (Eds.), 2–50, Cambridge, MA, MIT Press.
- Manski, C. F., and D. McFadden Eds. (1981b), *Structural Analysis of Discrete Data with Econometric Applications*, Cambridge, MA, MIT Press.
- Manski, C. F., and T. S. Thompson (1986), "Operational Characteristics of Maximum Score Estimation," *Journal of Econometrics*, 32, 85–108.

REFERENCES

- Manton, K. G., E. Stallard, and J. W. Vaupel (1986), "Alternative Models for the Heterogeneity of Mortality Risks among the Aged," *Journal of the American Statistical Association*, 81, 635–644.
- Marschak, J. (1953), "Economic Measurement for Policy and Prediction," in *Studies in Econometric Method*, W. J. Hood and T. C. Koopmans (Eds.), Cowles Commission Monograph 14, 1–26, New York, John Wiley.
- Marschak, J., and W. H. Andrews (1944), "Random Simultaneous Equations and the Theory of Production," *Econometrica*, 12, 143–205.
- Marshall, A. W., and I. Olkin (1990), "Multivariate Distributions Generated from Mixtures of Convolution and Product Families," in *Topics in Statistical Dependence*, H. W. Block, A. R. Sampson, and T. H. Savits (Eds.), IMS Lecture Notes-Monograph Series, Volume 16, 371–393.
- Matzkin, R. L. (1994), "Restrictions of Economic Theory in Nonparametric Models," in *Handbook in Econometrics*, R. F. Engle and D. L. McFadden (eds.), Volume 4, 2523–2558, Amsterdam, North-Holland.
- Mátyás, L., and P. Sevestre, Eds. (1995), *The Econometrics of Panel Data*, 2nd edition, Dordrecht, Kluwer Academic Publishers.
- McCall, B. P. (1996), "Unemployment Insurance Rules, Joblessness, and Part-Time Work," *Econometrica*, 64, 647–682.
- McCallum, B. T. (1972), "Relative Asymptotic Bias from Errors of Omission and Measurement," *Econometrica*, 40, 757–758.
- McCullagh, P., and J. A. Nelder (1983, 1989), *Generalized Linear Models*, 1st and 2nd editions, London, Chapman and Hall.
- McCulloch R., and P. Rossi (1994), "An Exact Likelihood Analysis of the Multinomial Probit Model," *Journal of Econometrics*, 64, 207–240.
- McDonald, J., and R. Moffitt (1980), "The Uses of Tobit Analysis," *Review of Economics and Statistics*, 318–321.
- McFadden, D. (1973), "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontiers in Econometrics*, P. Zarembka (Ed.), New York, Academic Press.
- McFadden, D. (1974), "The Measurement of Urban Travel Demand," *Journal of Public Economics*, 3, 303–328.
- McFadden, D. (1978), "Modelling the Choice of Residential Location," in *Spatial Interaction Theory and Planning Models*, 75–96, A. Karlquist L. Lundquist, F. Snickars, and J. W. Weibull et al. (Eds.), Amsterdam, New York, North-Holland.
- McFadden, D. (1981), "Econometric Models of Probabilistic Choice" in *Structural Analysis of Discrete Data with Econometric Applications*, C. F. Manski and D. McFadden (Eds.), 198–272, Cambridge, MA, MIT Press.
- McFadden, D. (1984), "Econometric Analysis of Qualitative Response Models," in *Handbook of Econometrics*, Z. Griliches and M. Intriligator (Eds.), Volume 2, 1395–1457, Amsterdam, North-Holland.
- McFadden, D. (1989), "A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration," *Econometrica*, 57, 995–1026.
- McFadden, D. (1995), "Computing Willingness-to-Pay in Random Utility Models," Department of Economics, University of California, Berkeley.
- McFadden, D. (2001), "Economic Choices," *American Economic Review*, 91, 351–378.
- McFadden, D., and F. Reid (1975), "Aggregate Travel Demand Forecasting from Disaggregate Models," *Transportation Research Record*, 534, 24–37.
- McFadden, D., and P. A. Ruud (1994), "Estimation by Simulation," *Review of Economics and Statistics*, 76, 591–608.

REFERENCES

- McFadden, D., and K. Train (2000), "Mixed MNL Models for Discrete Response," *Journal of Applied Econometrics*, 15, 447–470.
- McManus, D. A. (1991), "Who Invented Local Power Analysis?," *Econometric Theory*, 265–268.
- Mealli, F., and S. Pudney (1996), "Occupational Pensions and Job Mobility in Britain: Estimation of a Random-Effects Competing Risks Model," *Journal of Applied Econometrics*, 11, 293–320.
- Melenberg, B., and A. Van Soest (1996), "Parametric and Semi-Parametric Modelling of Vacation Expenditures," *Journal of Applied Econometrics*, 11, 59–76.
- Melino, A., and G. T. Sueyoshi (1990), "A Simple Approach to the Identifiability of the Proportional Hazards Model," *Economics Letters*, 33, 63–68.
- Meng, X.-L. (2000), "Missing Data: Dial M for ???," *Journal of the American Statistical Association*, 95, 1325–1331.
- Meyer, B. D. (1990), "Unemployment Insurance and Unemployment Spells," *Econometrica*, 58, 757–782.
- Meyer, B. D. (1995), "Natural and Quasi-experiments in Economics," *Journal of Business and Economic Statistics*, 13, 151–161.
- Miller, R. G. (1974), "The Jackknife – A Review," *Biometrika*, 61, 1–17.
- Mittelhammer, R. C., Judge, G. G., and D. J. Miller (2000), *Econometric Foundations*, Cambridge, UK, Cambridge University Press.
- Mizon, G. E., and J.-F. Richard (1986), "The Encompassing Principle and Its Application to Testing Non-nested Hypotheses," *Econometrica*, 54, 657–678.
- Moffitt, R. (1992), "Incentive Effects of the U.S. Welfare System: A Review," *Journal of Economic Literature*, 30, 1–61.
- Montalvo, J. G. (1997), "GMM Estimation of Count-Panel-Data Models with Fixed Effects and Predetermined Instruments," *Journal of Business and Economic Statistics*, 15, 82–89.
- Morgan, M. S. (1990), *The History of Econometric Ideas*, Cambridge, UK, Cambridge University Press.
- Mosteller, F., and J. Tukey (1977), *Data Analysis and Regression: A Second Course in Statistics*, Reading, MA, Addison-Wesley.
- Moulton, B. R. (1986), "Random Group Effects and the Precision of Regression Estimates," *Journal of Econometrics*, 32, 385–397.
- Moulton, B. R. (1987), "Diagnostic Tests for Group Effects in Regression Analysis," *Journal of Business and Economic Statistics*, 6, 275–282.
- Moulton, B. R. (1990), "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units," *Review of Economics and Statistics*, 72, 334–338.
- Mroz, T. A. (1987), "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions," *Econometrica*, 55, 765–800.
- Mudholkar, G. S., D. K. Srivastava, and G. D. Kollia (1996), "A Generalization of the Weibull Distribution and Application to the Analysis of Survival Data," *Journal of the American Statistical Association*, 91, 1575–1583.
- Mullahy, J. (1986), "Specification and Testing of Some Modified Count Data Models," *Journal of Econometrics*, 33, 341–365.
- Mullahy, J. (1997), "Instrumental Variable Estimation of Poisson Regression Models: Application to Models of Cigarette Smoking Behavior," *Review of Economics and Statistics*, 79, 586–593.
- Mullahy, J. (1998), "Much Ado about Two: Reconsidering Retransformation and Two-Part Model in Health Econometrics," *Journal of Health Economics*, 17, 247–281.

REFERENCES

- Mundlak, Y. (1978), "On the Pooling of Time Series and Cross Section Data," *Econometrica*, 46, 69–85.
- Munkin, M., and P. K. Trivedi (1999), "Simulated Maximum Likelihood Estimation of Multivariate Mixed-Poisson Regression Models, with Application," *Econometrics Journal*, 2, 29–48.
- Murphy, K., and R. Topel (1985), "Estimation and Inference in Two-Step Econometrics Models," *Journal of Business and Economic Statistics*, 3, 370–379.
- Nadaraya, E. A. (1964), "On Estimating Regression," *Theory of Probability and Its Applications*, 9, 141–142.
- Nagar, A. L. (1959), "The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations," *Econometrica*, 27, 575–595.
- Nakamura, T. (1990), "Corrected Score Function for Errors-in-Variables Models: Methodology and Application to Generalized Linear Models," *Biometrika*, 77, 127–137.
- Narendranathan, W., S. Nickell, and J. Stern (1985), "Unemployment Benefits Revisited," *Economic Journal*, 95, 307–329.
- Nawata, K. (1993), "A Note on Estimation with Sample-Selection Bias," *Economics Letters*, 42, 15–24.
- Nawata, K., and N. Nagase (1996), "Estimation of Sample-Selection Bias Models," *Econometric Reviews*, 15, 387–400.
- Nelder, J. A., and R. W. M. Wedderburn (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society, A*, 135, 370–384.
- Nelsen, R. B. (1999), *An Introduction to Copulas*, Lecture Notes in Statistics, 139, New York, Springer-Verlag.
- Nelson, C. R., and R. Startz (1990), "The Distribution of the Instrumental Variables Estimator and Its t-Ratio When the Instrument Is a Poor One," *Journal of Business*, 63, S125–140.
- Nelson, F. D. (1977), "Censored Regression Models with Unobserved Stochastic Censoring Thresholds," *Journal of Econometrics*, 6, 309–327.
- Nelson, F. D., and L. Olson (1978), "Specification and Estimation of a Simultaneous Equations Model with Limited Dependent Variables," *International Economic Review*, 19, 695–709.
- Nerlove, M. (1963), "Returns to Scale in Electricity Supply," in *Measurement In Economics: Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld*, C. F. Christ (Ed.), 167–200, Stanford, Stanford University Press.
- Nevo, A. (2001), "Measuring Market Power in the Ready-to-Eat Cereal Industry," *Econometrica*, 69, 307–342.
- Newey, W. K. (1984), "A Methods of Moments Interpretation of Sequential Estimators," *Economics Letters*, 14, 201–206.
- Newey, W. K. (1985), "Maximum Likelihood Specification Testing and Conditional Moment Tests," *Econometrica*, 53, 1047–1070.
- Newey, W. K. (1990a), "Efficient Instrumental Variable Estimation of Nonlinear Models," *Econometrica*, 58, 809–838.
- Newey, W. K. (1990b), "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135.
- Newey, W. K. (1993), "Efficient Estimation of Models with Conditional Moment Restrictions," in *Handbook of Statistics, Volume 11, Econometrics*, G. S. Maddala, C. R. Rao and H. D. Vinod (Eds.), 419–454, Amsterdam, North-Holland.
- Newey, W. K. (1997), "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Applied Econometrics*, 5, 99–135.
- Newey, W. K., and D. McFadden (1994), "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, R. F. Engle and D. McFadden (Eds.), Volume 4, 2111–2245, Amsterdam, North-Holland.

REFERENCES

- Newey, W. K., J. L. Powell, and J. R. Walker (1990), "Semiparametric Estimation of Selection Models: Some Empirical Results," *A. E. R. Papers and Proceedings*, 324–328.
- Newey, W. K., and K. D. West (1987a), "Hypothesis Testing with Efficient Method of Moments Estimators," *International Economic Review*, 28, 777–787.
- Newey, W. K., and K. D. West (1987b), "A Simple, Positive Semi-Definite, Heteroscedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.
- Newhouse, J. P., and the Insurance Experiment Group (1993), *Free for All? Lessons from the RAND Health Insurance Experiment*, Cambridge and London, Harvard University Press.
- Newman, J. L., and C. E. McCulloch (1984), "A Hazard Rate Approach to the Timing of Births," *Econometrica*, 52, 939–962.
- (Splawa)-Neyman, J. (1923, 1990), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," translated and edited by D. M. Dabrowska and T. P. Speed from *Roczniki Rolniczych Tom X* (1923), 1–51 (Annals of Agricultural Sciences), *Statistical Science*, 5, 465–471.
- Neyman, J. (1937), "‘Smooth Test’ for Goodness of Fit," *Skandinaviske Aktuarietidskrift*, 20, 150–199.
- Neyman, J. (1959), "Optimal Asymptotic Tests of Composite Statistical Hypotheses," in *Probability and Statistics*, U. Grenander (Ed.), New York, John Wiley.
- Neyman, J., and E. S. Pearson (1933), "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society, A*, 231, 289–337.
- 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*, 1399–1416.
- Olsen, R. J. (1980), "A Least-Squares Correction for Selectivity Bias," *Econometrica*, 48, 1815–1820.
- Olsen, R. J., and G. Farkas (1989), "Endogenous Covariates in Duration Models and the Effect of Adolescent Childbirth on Schooling," *Journal of Human Resources*, 24, 39–53.
- Owen, A. B. (1988), "Empirical Likelihood Ratio Confidence Intervals for a Single Functional," *Biometrika*, 75, 237–249.
- Owen, A. B. (2001), *Empirical Likelihood*, London, Chapman and Hall.
- Paarsch, H.J. (1982), "A Monte Carlo Comparison of Estimators for Censored Regression Models," *Journal of Econometrics*, 24, 197–213.
- Pagan, A. R. (1986), "Two Stage and Related Estimators and Their Applications," *Review of Economic Studies*, 53, 517–538.
- Pagan, A. R., and A. Ullah (1999), *Nonparametric Econometrics*, Cambridge, UK, Cambridge University Press.
- Pagan, A. R., and F. Vella (1989), "Diagnostic Tests for Models Based on Individual Data: A Survey," *Journal of Applied Econometrics*, 3, S29–60.
- Pakes, A. S., and D. Pollard (1989), "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027–1057.
- Palmgren, J. (1981), "The Fisher Information Matrix for Log-Linear Models Arguing Conditionally in the Observed Explanatory Variables," *Biometrika*, 68, 563–566.
- Parzen, E. (1962), "On Estimation of a Probability Density Function and Mode," *Annals of Mathematical Statistics*, 33, 1065–1076.
- Pearl, J. (2000), *Causality: Models, Reasoning and Inference*, Cambridge, UK, Cambridge University Press.
- 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.

REFERENCES

- Pendergast, J., S. J. Gange, M. A. Newton, M. J. Lindstrom, M. Palta, and M. R. Fisher (1996), "A Survey of Methods for Analyzing Clustered Binary Response Data," *International Statistical Review*, 64, 89–118.
- Pesaran, B., and M. H. Pesaran (1995), "A Non-nested Test of Level-Differenced versus Log-Differenced Stationary Models," *Econometric Reviews*, 14, 377–392.
- Pesaran, M. H. (1974), "On the General Problem of Model Selection," *Review of Economic Studies*, 41, 153–171.
- Pesaran, M. H. (2004), "Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure," Working Paper, Cambridge University.
- Pesaran, M. H., and B. Pesaran (1993), "A Simulation Approach to the Problem of Computing Cox's Statistic for Testing Nonnested Models," *Journal of Econometrics*, 57, 377–392.
- Pesaran, M. H., and R. Smith (1995), "Estimating Long-Run Relationships from Dynamic Heterogeneous Panels," *Journal of Econometrics*, 68, 79–113.
- Phillips, G. D. A., and C. Hale (1977), "The Bias of Instrumental Variable Estimators of Simultaneous Equation Systems," *International Economic Review*, 18, 219–228.
- Phillips, P. C. B. (1983), "Exact Small Sample Theory in the Simultaneous Equations Model," in *Handbook of Econometrics*, Z. Griliches and M. D. Intrilligator (Eds.), Volume 2, 449–516, Amsterdam, North-Holland.
- 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 H. R. Moon (2000), "Nonstationary Panel Data Analysis: An Overview of Some Recent Developments," *Econometric Reviews*, 19, 263–286.
- Phillips, P. C. B., and J. Y. Park (1988), "On the Formulation of Wald Tests of Nonlinear Restrictions," *Econometrica*, 56, 1065–1083.
- Pitt, M. M., and M. R. Rosenzweig (1990), "Estimating the Intrahousehold Incidence of Illness: Child Health and Gender-Inequality in the Allocation of Time," *International Economic Review*, 31, 969–80.
- Pohlmeier, W., and V. Ulrich (1995), "An Econometric Model of the Two-Part Decision-making Process in the Demand for Health Care," *Journal of Human Resources*, 30, 339–361.
- Politis, D. N., and J. P. Romano (1994), "Large Sample Confidence Regions Based on Subsamples under Minimal Assumptions," *Annals of Statistics*, 22, 2031–2050.
- Politis, D. N., J. P. Romano, and M. Wolf (1999), *Subsampling*, New York, Springer-Verlag.
- Polivka, A. E. (1996), "Data Watch: The Redesigned Current Population Survey," *Journal of Economic Perspectives*, 10, 169–180.
- Poskitt, D. S., and C. L. Skeels (2002), "Assessing Instrumental Variable Relevance: An Alternative Measure and Some Exact Finite Sample Theory," University of Melbourne, Department of Economics Research Paper Number 862.
- Poterba, J. M., and L. H. Summers (1995), "Unemployment Benefits and Labor Market Transitions: A Multinomial Logit Model with Errors in Classification," *Review of Economics and Statistics*, 77, 207–216.
- Powell, J. L. (1984), "Least Squares Absolute Deviations Estimation for the Censored Regression Model," *Journal of Econometrics*, 25, 303–325.
- Powell, J. L. (1986a), "Censored Regression Quantiles," *Journal of Econometrics*, 32, 143–155.
- Powell, J. L. (1986b), "Symmetrically Trimmed Least Squares Estimation for Tobit Models," *Econometrica*, 54, 1435–1460.
- Powell, J. L. (1994), "Estimation of Semiparametric Methods," in *Handbook in Econometrics*, R. F. Engle and D. L. McFadden (Eds.), Volume 4, 2443–2521, Amsterdam, North-Holland.
- Powell, J. L., J. H. Stock, and T. M. Stoker (1989), "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57, 1403–1430.

REFERENCES

- Prais, S. J., and H. S. Houthakker (1955), *Analysis of Family Budgets*, Cambridge, UK, Cambridge University Press.
- Prentice, R. L., and R. Pyke (1979), "Logistic Disease Incidence Models and Case-Control Studies," *Biometrika*, 66, 403–411.
- Press, W., B. Flannery, S. Teukolsky, and W. Vetterling (1986), *Numerical Recipes: The Art of Scientific Computing*, Cambridge, UK, Cambridge University Press.
- Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling (1993), *Numerical Recipes in C: The Art of Scientific Computing*, 2nd edition, Cambridge, UK, Cambridge University Press.
- Pudney, S. (1989), *Modelling Individual Choice: The Econometrics of Corners, Kinks, and Holes*, Oxford and New York, Blackwell.
- Quandt, R. (1983), "Computational Problems and Methods," in *Handbook of Econometrics*, Z. Griliches and M. Intrilligator (Eds.), Volume 1, 699–764, Amsterdam, North-Holland.
- Quenouille, M. H. (1956), "Notes on Bias in Estimation," *Biometrika*, 43, 353–360.
- Qin, J., and J. Lawless (1994), "Empirical Likelihood and General Estimating Equations," *Annals of Statistics*, 22, 300–325.
- Radulović, D., and M. Wegkamp (undated), "An Easy Proof of Sklar's Theorem," Department of Statistics, Yale University, preprint.
- Ramsey, J. B. (1969), "Tests of Specification Errors in Classical Linear Least Squares Regression Analysis," *Journal of the Royal Statistical Society, B*, 31, 350–71.
- Rao, C. R. (1947), "Large Sample Tests of Statistical Hypotheses Concerning Several Parameters with Applications to Problems of Estimation," *Proceedings of the Cambridge Philosophical Society*, 44, 50–57.
- Rao, C. R. (1973), *Linear Statistical Inference and Its Applications*, 2nd edition, New York, John Wiley.
- Rao, J. N. K., and C. F. J. Wu (1988), "Resampling Inference with Complex Survey Data," *Journal of the American Statistical Association*, 83, 231–241.
- Reiersol, O. (1941), "Confluence Analysis by Means of Lag Moments and Other Methods of Confluence Analysis," *Econometrica*, 9, 1–24.
- Revelt, D., and K. Train (1998), "Mixed Logit with Repeated Choices: Households' Choices of Appliance Efficiency Level," *Review of Economics and Statistics*, 80, 647–657.
- Ripley, B. D. (1987), *Stochastic Simulation*, New York, John Wiley.
- Rivers, D., and Q. H. Vuong (1988), "Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models," *Journal of Econometrics*, 39, 347–366.
- Robert, C. P., and G. Casella (1999), *Monte Carlo Methods*, New York, Springer-Verlag.
- Robinson, P. M. (1987), "Asymptotically Efficient Estimation in the Presence of Heteroskedasticity of Unknown Form," *Econometrica*, 55, 875–891.
- Robinson, P. M. (1988a), "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56, 931–954.
- Robinson, P. M. (1988b), "Semiparametric Econometrics: A Survey," *Journal of Applied Econometrics*, 3, 35–51.
- Rose, N. (1990), "Profitability and Product Quality: Economic Determinants of Airline Safety Performance," *Journal of Political Economy*, 98, 944–964.
- Rosenblatt, M. (1956), "Remarks on Some Nonparametric Estimates of a Density Function," *Annals of Mathematical Statistics*, 27, 832–837.
- Rosenbaum, P., and D. B. Rubin (1983), "The Central Role of Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.
- Rosenzweig, M., and K. Wolpin (1980), "Testing the Quality-Quantity Fertility Model: The Use of Twins as a Natural Experiment," *Econometrica*, 48, 227–240.

REFERENCES

- Rosenzweig, M., and K. Wolpin (2000), “‘Natural’ Natural Experiments” in Economics,” *Journal of Economic Literature*, 38, 827–874.
- Rothenberg, T. J. (1973), *Efficient Estimation with A Priori Information*, Cowles Foundation Monograph No. 23, New Haven, Yale University Press.
- Rothenberg, T. J. (1984), “Approximating the Distributions of Econometric Estimators and Test Statistics,” in *Handbook of Econometrics*, Z. Griliches and M. D. Intrilligator (Eds.), Volume 2, 881–935, Amsterdam, North-Holland.
- Roy, A. (1951), “Some Thoughts on the Distribution of Earnings,” *Oxford Economic Papers*, 3, 135–146.
- Rubin, D. B. (1974), “Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies,” *Journal of Educational Psychology*, 66, 688–701.
- Rubin, D. B. (1976), “Inference and Missing Data,” *Biometrika*, 63, 581–592.
- Rubin, D. B. (1978), “Bayesian Inference for Causal Effects,” *Annals of Statistics*, 6, 34–58.
- Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York, John Wiley.
- Rubin, D. B. (1996), “Multiple Imputation after 18+ Years,” *Journal of the American Statistical Association*, 91, 473–489.
- Ruppert, D., M. P. Wand, and R. J. Carroll (2003), *Semiparametric Regression*, Cambridge, UK, Cambridge University Press.
- Rust, J. (1994), “Structural Estimation of Markov Decision Processes,” in *Handbook of Econometrics*, R. F. Engle and D. McFadden (Eds.), Volume 4, 3081–3143, Amsterdam, North-Holland.
- Ruud, P. A. (1984), “Tests of Specification in Econometrics,” *Econometric Reviews*, 3, 211–242.
- Ruud, P. A. (1983), “Sufficient Conditions for the Consistency of Maximum Likelihood Estimation Despite Misspecification of Distribution in Multinomial Discrete Choice Models,” *Econometrica*, 51, 225–228.
- Ruud, P. A. (1986), “Consistent Estimation of Limited Dependent Variable Models Despite Misspecification of Distribution,” *Journal of Econometrics*, 32, 157–187.
- Ruud, P. A. (2000), *An Introduction to Classical Econometric Theory*, Oxford, Oxford University Press.
- Sakata, S. (1998), “Quasi-maximum Likelihood Estimation with Complex Survey Data,” preprint.
- Salant, S. W. (1977), “Search Theory and Duration Data: A Theory of Sorts,” *Quarterly Journal of Economics*, 91, 39–57.
- Sapra, S. (1998), “Further Examples of Accelerated Failure Time Models Which Are Not Proportional Hazard Models,” Problem, *Econometric Theory*, 14, 688. Solution, *Econometric Theory* (1999), 15, 786–787.
- Sapra, S. (2000), “Simple Applications of the Cox–Tsairis Result on Unidentifiability of Dependent Competing Risks Model without Regressors,” Problem, *Econometric Theory*, 16, 619. Solution, *Econometric Theory* (2001), 17, 855–856.
- Sapra, S. (2001), “Identification of Parameters in Two Competing Risks Models,” Problem, *Econometric Theory*, 17, 1158. Solution, *Econometric Theory* (2002), 18, 6, 1463–1465.
- Sapra, S. (2002), “Effects of Heterogeneity on Unconditional (Population) Hazard Functions,” Problem SP02/3, *Statistical Papers*, 43/2, 301. Solution, *Statistical Papers*, 2003, 44/3, 445–446.
- Sargan, J. D. (1958), “The Estimation of Economic Relationships Using Instrumental Variables,” *Econometrica*, 26, 393–415.
- Sargan, J. D. (1980), “Some Approximations to the Distribution of Econometric Criteria Which Are Asymptotically Distributed as Chi-Squared,” *Econometrica*, 48, 1107–1138.

REFERENCES

- Sargan, J. D. (1988), *Lectures on Advanced Econometric Theory*, Oxford, Blackwell.
- Sargan, J. D. (2001), "Model Building and Data Mining," *Econometric Reviews*, 20, 159–170.
- Savin, N. E. (1984), "Multiple Hypothesis Testing," in *Handbook of Econometrics*, Z. Griliches and M. D. Intriligator (Eds.), Volume 2, 827–879, Amsterdam, North-Holland.
- Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, London, Chapman and Hall.
- Schmidt, P. (1976), *Econometrics*, New York, Marcel Dekker.
- Schwarz, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–464.
- Schweizer, B., and E. F. Wolff (1981), "On Nonparametric Measures of Dependence for Random Variables," *Annals of Statistics*, 9, 879–883.
- Scott, A. J., and D. Holt (1982), "The Effect of Two-Stage Sampling on Ordinary Least Squares Methods," *Journal of the American Statistical Association*, 77, 848–854.
- Serfling, R. (1980), *Approximation Theorems of Mathematical Statistics*, New York, John Wiley.
- Severini, T. A., and G. Tripathi (2001), "A Simplified Approach to Computing Efficiency Bounds in Semiparametric Models," *Journal of Econometrics*, 102, 23–66.
- Shao, J., and D. Tu (1995), *The Jackknife and Bootstrap*, New York, Springer-Verlag.
- Shaw, D. (1988), "On-Site Samples' Regression: Problems of Non-negative Integers, Truncation, and Endogenous Stratification," *Journal of Econometrics*, 37, 211–223.
- Shea, J. (1997), "Instrument Relevance in Multivariate Linear Models: A Simple Measure," *Review of Economics and Statistics*, 79, 348–352.
- Sherman, R. C. (1993), "The Limiting Distribution of the Maximum Rank Correlation Estimator," *Econometrica*, 61, 123–137.
- Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, London, Chapman and Hall.
- Silvey, S. D. (1959), "The Lagrangian Multiplier Test," *Annals of Mathematical Statistics*, 30, 389–407.
- Singh, K. (1981), "On the Asymptotic Accuracy of Efron's Bootstrap," *Annals of Statistics*, 9, 1187–1195.
- Skinner, C. J. (1989), "Introduction to Part A," in *Analysis of Complex Surveys*, C. J. Skinner, D. Holt, and T. M. F. Smith (Eds.), New York, John Wiley.
- Sklar, A. (1973), "Random Variables, Joint Distribution Functions, and Copulas," *Kybernetika*, 9, 449–460.
- Skrondal, A., and S. Rabe-Hesketh (2004), *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*, Boca Raton, FL, Chapman and Hall.
- Small, K. A., and H. S. Rosen (1981), "Applied Welfare Economics with Discrete Choice Models," *Econometrica*, 49, 105–130.
- Smith, J. A., and P. E. Todd (2004), "Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators?," *Journal of Econometrics*, forthcoming.
- Smith, M. D. (2003), "Modelling Sample Selection Using Archimedean Copulas," *Econometrics Journal*, 6, 99–123.
- Smith, R. J., and R. W. Blundell (1986), "An Exogeneity Test for a Simultaneous Equation Tobit Model with an Application to Labor Supply," *Econometrica*, 54, 679–685.
- Solari, M. E. (1969), "The Maximum Likelihood Solution to the Problem of Estimating a Linear Functional Relationship," *Journal of the Royal Statistical Society, B*, 31, 611–613.
- Solon, G. (1992), "Intergenerational Income Mobility in the United States," *American Economic Review*, 82, 393–408.
- Speckman, P. (1988), "Kernel Smoothing in Partial Linear Models," *Journal of the Royal Statistical Society, B*, 50, 413–446.
- Staiger, D., and J. Stock (1997), "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

REFERENCES

- Stefanski, L. A. (1989), "Unbiased Estimation of a Nonlinear Function of a Normal Mean with the Application to Measurement Error Models," *Communications in Statistics, A*, 18, 4335–4358.
- Stefanski, L. A. (2000), "Measurement Error Models," *Journal of the American Statistical Association*, 95, 1353–1358.
- Steinbrickner, T. R. (1999), "Estimation of a Duration Model in the Presence of Missing Data," *Review of Economics and Statistics*, 81, 529–542.
- Stern, S. (1997), "Simulation-Based Estimation," *Journal of Economic Literature*, 35, 2006–2039.
- Stevens, A. H. (1999), "Climbing Out of Poverty, Falling Back In: Measuring the Persistence of Poverty over Multiple Spells," *Journal of Human Resources*, 34, 557–588.
- Stigler, S. M. (1986), *The History of Statistics: The Measurement of Uncertainty before 1900*, Cambridge, MA, Harvard University Press.
- Stock, J. H., and F. Trebbi (2003), "Retrospectives: Who Invented Instrumental Variables Regression?," *Journal of Economic Perspectives*, 17, 177–194.
- Stock, J. H., J. H. Wright, and M. Yogo (2002), "A Survey of Weak Instruments and Weak Identification in GMM," *Journal of Business and Economic Statistics*, 20, 518–529.
- Stock, J. H., and M. Yogo (2003), "Testing for Weak Instruments in Linear IV Regression," prepared for Festschrift in honor of Thomas Rothenberg.
- Stoker, T. M. (1982), "The Use of Cross-Section Data to Characterize Macro Functions," *Journal of the American Statistical Association*, 77, 369–380.
- Stoker, T. M. (1984), "Completeness, Distribution Restrictions, and the Form of Aggregate Functions," *Econometrica*, 52, 887–907.
- Stoker, T. M. (1986), "Consistent Estimation of Scaled Coefficients," *Econometrica*, 54, 1461–1481.
- Stone, C. J. (1977), "Consistent Nonparametric Regression," *Annals of Statistics*, 5, 595–620.
- Stone, C. J. (1980), "Optimal Convergence Rates for Nonparametric Estimates," *Annals of Statistics*, 8, 1348–1360.
- Stone, R. (1953), *The Measurement of Consumers' Expenditure and Behaviour in the United Kingdom, 1920–1938*, Vol. 1, Cambridge, Cambridge University Press.
- Sueyoshi, G. T. (1992), "Semiparametric Proportional Hazards Estimation of Competing Risks Models with Time-Varying Covariates," *Journal of Econometrics*, 51, 25–58.
- Sullivan, R., A. Timmermann, and H. White (2001), "Dangers of Data Mining: The Case of Calendar Effects in Stock Returns," *Journal of Econometrics*, 105, 249–286.
- Swait, J. (2003), "Flexible Covariance Structures for Categorical Dependent Variables through Finite Mixtures of Generalized Extreme Value Models," *Journal of Business and Economic Statistics*, 21, 80–87.
- Swamy, P. A. V. B. (1970), "Efficient Inference in a Random Coefficient Regression Model," *Econometrica*, 38, 311–323.
- Szu, H., and R. Hartley (1987), "Fast Simulated Annealing," *Physics Letters A*, 122, 157–162.
- Tanner, M. A., and W. H. Wong (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528–549.
- Tauchen, G. (1985), "Diagnostic Testing and Evaluation of Maximum Likelihood Models," *Journal of Econometrics*, 30, 415–443.
- Taylor, H. M., and S. Karlin (1994), *An Introduction to Stochastic Modelling*, revised edition, San Diego and New York, Academic Press.
- Terza, J. (1998), "Estimating Count Data Models with Endogenous Switching: Sample Selection and Endogenous Switching Effects," *Journal of Econometrics*, 84, 129–139.
- Theil, H. (1953), "Repeated Least Squares Applied to Complete Equation System," mimeograph, The Hague, Central Planning Bureau.

REFERENCES

- Thistlethwaite, D., and D. Campbell (1960), "Regression Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment," *Journal of Educational Psychology*, 51, 309–317.
- Thomas, J. M. (1996), "On the Interpretation of Covariate Estimates in Independent Competing Risks Models," *Bulletin of Economic Research*, 48, 27–39.
- Thomson, C. J., and S. T. Crooke (1991), "Results of the Southern California Sportfish Economic Survey," NOAA Technical Memorandum, National Marine Fisheries Service, Southwest Fisheries Science Center.
- Tierney, L., R. E. Kass, and J. B. Kadane (1986), "Fully Exponential Laplace Approximations to Expectations and Variances of Nonpositive Functions," *Journal of the American Statistical Association*, 81, 82–86.
- Tobin, J. (1958), "Estimation of Relationships for Limited Dependent Variables," *Econometrica*, 26, 24–36.
- Train, K. E. (1986), *Qualitative Choice Analysis: Theory, Practice and an Application to Automobile Demand*, Cambridge, MA, MIT Press.
- Train, K. E. (2001), "A Comparison of Hierarchical Bayes and Maximum Simulated Likelihood for Mixed Logit," Working Paper, Department of Economics, University of California, Berkeley.
- Train, K. E. (2003), *Discrete Choice Methods with Simulation*, Cambridge, UK, Cambridge University Press.
- Trivedi, P. K., and J. N. Alexander (1989), "Reemployment Probability and Multiple Unemployment Spells: A Partial Likelihood Approach," *Journal of Business and Economic Statistics*, 7, 395–402.
- Trochim, W. (1984), *Research Design for Program Evaluation: The Regression-Discontinuity Approach*, Beverly Hills, CA, Sage Publications.
- Tsiatis, A. (1975), "A Nonidentifiability Aspect of the Problem of Competing Risks," *Proceedings of the National Academy of Sciences*, 72, 20–22.
- Tukey, J. W. (1958), "Bias and Confidence in Not Quite Large Samples" (abstract), *Annals of Mathematical Statistics*, 29, 614.
- Tunali, I., and J. B. Pritchett (1997), "Cox Regression with Alternative Concepts of Waiting Time: The New Orleans Yellow Fever Epidemic of 1853," *Journal of Applied Econometrics*, 12, 1–25.
- Ullah, A., and R. Breunig (1998), "Econometric Analysis in Complex Surveys," in *Handbook of Applied Economic Statistics*, D. Giles and A. Ullah (Eds.), New York, Marcel Dekker.
- U.S. Census Bureau (2002), "Current Population Survey: Design and Methodology," Technical Paper 63RV.
- Van den Berg, G. (1990), "Nonstationarity in Job Search," *Review of Economic Studies*, 57, 255–277.
- Van den Berg, G. (1997), "Association Measures for Durations in Bivariate Hazard Rate Models," *Journal of Econometrics*, 79, 221–245.
- Van den Berg, G. (2001), "Duration Models: Specification, Identification, and Multiple Durations," in *Handbook of Econometrics*, J. J. Heckman and E. Leamer (Eds.), Volume 5, 3381–3460, Amsterdam, North-Holland.
- Van der Klaauw, W. (2003), "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach," *International Economic Review*, 43, 1249–1287.
- Van Ophem, H. (2000), "Modeling Selectivity in Count-Data Models," *Journal of Business and Economic Statistics*, 18, 503–511.
- Vaupel, J. W., K. G. Manton, and E. Stallard (1979), "The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality," *Demography*, 16, 439–454.
- Vella, F. (1998), "Estimating Models with Sample Selection Bias: A Survey," *Journal of Human Resources*, 33, 127–169.

REFERENCES

- Vella, F., and M. Verbeek (1999), "Estimating and Interpreting Models with Endogenous Treatment Effects," *Journal of Business and Economic Statistics*, 17, 473–478.
- Verbeek, M. (1995), "Pseudo Panel Data," in *The Econometrics of Panel Data*, L. Matyas and P. Sevestre (Eds.), 2nd ed., 303–315, Norwell, MA, Kluwer Academic Publishers, and Dordrecht, Kluwer Academic.
- Verbeek, M., and T. Nijman (1992a), "Can Cohort Data Be Treated as Genuine Panel Data?," *Empirical Economics*, 17, 9–23.
- Verbeek, M., and T. Nijman (1992b), "Testing for Selectivity Bias in Panel Data Models," *International Economic Review*, 33, 681–703.
- Verdinelli, I., and L. Wasserman (1995), "Computing Bayes Factors Using a Generalization of the Savage-Dickey Density Ratio," *Journal of the American Statistical Association*, 90, 614–618.
- Vuong, Q. H. (1989), "Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses," *Econometrica*, 57, 307–333.
- Wald, A. (1940), "The Fitting of Straight Lines If Both Variables Are Subject to Error," *Annals of Mathematical Statistics*, 11, 284–300.
- Wald, A. (1943), "Test of Statistical Hypotheses Concerning Several Parameters When the Number of Observations Is Large," *Transactions of the American Mathematical Society*, 54, 426–482.
- Walker, J., and M. Ben-Akiva (2002), "Generalized Random Utility Model," *Mathematical Social Sciences*, 43, 303–343.
- Wang, P., I. M. Cockburn, and M. L. Puterman (1998), "Analysis of Patent Data – A Mixed-Poisson Regression Model Approach," *Journal of Business and Economic Statistics*, 16, 27–41.
- Wansbeek, T. J., and A. Kapteyn (1989), "Estimation of the Error Components Model with Incomplete Panels," *Journal of Econometrics*, 41, 341–361.
- Wansbeek, T., and E. Meijer (2000), *Measurement Error and Latent Variables in Econometrics*, Amsterdam, North-Holland.
- Watson, G. S. (1964), "Smooth Regression Analysis," *Sankhya A*, 26, 359–72.
- Wedel, W., W. S. DeSarbo, J. R. Bult, and V. Ramaswamy (1993), "A Latent Class Poisson Regression Model for Heterogeneous Count Data," *Journal of Applied Econometrics*, 8, 397–411.
- White, H. (1980a), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817–838.
- White, H. (1980b), "Using Least Squares to Approximate Unknown Regression Functions," *International Economic Review*, 21, 149–170.
- White, H. (1980c), "Nonlinear Regression on Cross-Section Data," *Econometrica*, 48, 721–746.
- White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25.
- White, H. (1994), *Estimation, Inference, and Specification Analysis*, Cambridge, UK, Cambridge University Press.
- White, H. (2001a), *Asymptotic Theory for Econometricians*, revised edition, San Diego, Academic Press.
- White, H. (2001b), "A Reality Check for Data Snooping," *Econometrica*, 68, 1097–1126.
- White, H., and I. Domowitz (1984), "Nonlinear Regression with Dependent Observations," *Econometrica*, 52, 143–161.
- Williams, H. (1977), "On the Formation of Travel Demand Models and Economics Evaluation Measures of User Benefit," *Environmental Planning A*, 9, 285–344.
- Willmot, G. E. (1987), "The Poisson-Inverse Gaussian Distribution as an Alternative to the Negative Binomial," *Scandinavian Actuarial Journal*, 198, 113–127.

REFERENCES

- Winkelmann, R. (1995), "Duration Dependence and Dispersion in Count-Data Models," *Journal of Business and Economic Statistics*, 13, 467–474.
- Winkelmann, R. (1997), *Econometric Analysis of Count Data*, Berlin, Springer-Verlag.
- Wolak, F. A. (1991), "The Local Nature of Hypothesis Tests Involving Inequality Constraints in Nonlinear Models," *Econometrica*, 59, 981–995.
- Wold, H., and L. Jureen (1953), *Demand Analysis*, New York, John Wiley.
- Wolter, K. M. (1985), *Introduction to Variance Estimation*, New York, Springer-Verlag.
- Wooldridge, J. M. (1991), "On the Application of Robust, Regression-Based Diagnostics to Models of Conditional Means and Variances," *Journal of Econometrics*, 47, 5–46.
- Wooldridge, J. M. (1995), "Selection Corrections for Panel Data Models under Conditional Mean Independence Assumptions," *Journal of Econometrics*, 68, 115–132.
- Wooldridge, J. M. (1997a), "Multiplicative Panel Data Models without the Strict Exogeneity Assumption," *Econometric Theory*, 13, 667–678.
- Wooldridge, J. M. (1997b), "On Two Stage Least Squares Estimation of the Average Treatment Effect in a Random Coefficient Model," *Economics Letters*, 56, 129–133.
- Wooldridge, J. M. (2001), "Asymptotic Properties of Weighted M-Estimators for Standard Stratified Samples," *Econometric Theory*, 17, 451–470.
- Wooldridge, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA, MIT Press.
- Wooldridge, J. M. (2003), "Cluster-Sample Methods in Applied Econometrics," *American Economic Review*, 93, 133–138.
- Wu, C. F. G. (1986), "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis," *Annals of Statistics*, 14, 1261–1295.
- Wu, D. (1973), "Alternative Tests of Independence between Stochastic Regressors and Disturbances," *Econometrica*, 41, 733–775.
- Yatchew, A. (1997), "An Elementary Estimator of the Partial Linear Model," *Economics Letters*, 57, 135–143.
- Yatchew, A. (1998), "Nonparametric Regression Techniques in Econometrics," *Journal of Economic Literature*, 36, 669–721.
- Yatchew, A. (2003), *Semiparametric Regression for the Applied Econometrician*, Cambridge, UK, Cambridge University Press.
- Yule, G. U. (1897), "On the Theory of Correlation," *Journal of the Royal Statistical Society*, 60, 812–854.
- Zaman, A. (1996), *Statistical Foundations for Econometric Techniques*, San Diego, CA, Academic Press.
- Zellner, A. (1962), "Estimators for Seemingly Unrelated Regression and Tests of Aggregation Bias," *Journal of the American Statistical Association*, 57, 500–509.
- Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, New York, John Wiley.
- Zellner, A. (1978), "Jeffreys–Bayes Posterior Odds Ratio and the Akaike Information Criterion for Discriminating between Models," *Economics Letters*, 1, 337–342.
- Zellner, A., and C.-K. Min (1995), "Gibbs Sampler Convergence Criteria," *Journal of the American Statistical Association*, 90, 921–927.
- Zellner, A., and H. Theil (1962), "Three Stage Least Squares: Simultaneous Estimation of Simultaneous Equations," *Econometrica*, 30, 63–68.
- Ziliak, J. P. (1997), "Efficient Estimation with Panel Data When Instruments Are Predetermined: An Empirical Comparison of Moment-Condition Estimators," *Journal of Business and Economic Statistics*, 15, 419–431.
- Zimmerman, D. J. (1992) "Regression toward Mediocrity in Economic Stature," *American Economic Review*, 82, 409–429.

Author Index

- Abe, M., 523
Abowd, J. M., 767
Abramowitz, M., 352, 390
Abrevaya, J., 914
Ahn, S. C., 766, 768, 777
Aitchison, J., 235
Aitken, M., 621, 623
Akaike, H., 278
Albert, J. H., 442, 450, 519
Alexander, J. N., 630
Allen, R. G. D., 3, 19
Allenby, G. M., 482, 513
Allison, P. D., 608, 609, 927, 940
Al-Osh, M. A., 806
Altman, N. S., 334
Altonji, J. G., 177, 192, 379, 403, 767
Alzaid, A. A., 806
Amemiya, T., 3, 4, 119, 124, 126, 128, 163–4, 192, 196, 219–20, 237, 278, 283, 293, 328, 352, 370, 472–4, 479–80, 487, 499, 528, 537, 541, 547–8, 557, 561, 567, 569, 761, 857, 932, 946, 953, 956
Amemiya, Y., 912–3, 921
Andersen, E. B., 782, 809
Andersen, P. K., 595, 609
Anderson, K. H., 663
Anderson, T. W., 191, 373, 758, 764, 765
Andrews, D. W. K., 250, 257, 266–7, 322, 325, 334, 361, 362, 566, 638
Andrews, W. H., 3
Angrist, J. D., 34, 36, 38, 62, 70, 97, 106, 112, 192, 199, 770, 858, 872, 879, 883–5, 896–7, 909, 920
Anselin, L., 845
Arellano, M., 220, 707, 723, 740, 753, 758–9, 765–7, 777, 809
Ashenfelter, O., 54, 57, 896
Athey, S., 770
Avery, R. B., 762, 789, 796
Baker, M., 104, 106–8, 113, 624, 637–8, 767
Balestra, P., 777
Balshaw, R., 682
Baltagi, B. H., 693, 722, 737–41, 760, 762, 767, 777, 801, 809
Basmann, R. L., 190
Bassett, G., 67, 87, 90, 113
Bayes, T., 458
Becker, S. O., 878, 889, 896
Beckett, S., 61
Beggs, S., 529
Begun, J. M., 330
Bekker, P. A., 191
Bellemare, C., 334
Ben-Akiva, M., 487, 516, 528
Bera, A. K., 220, 233, 630
Beran, R., 374
Berkson, J., 472, 487
Berndt, E., 138, 210, 238, 343
Berry, S. T., 82, 513
Bertrand, M., 708, 777
Bhargava, A., 764
Bhat, C. R., 409–10
Bickel, P. J., 376, 382
Bierens, H. J., 163, 268, 331, 374
Bilias, Y., 220
Binder, M., 768
Bingley, P., 888
Björklund, A., 886, 896
Bjørn, E., 920
Blake, D., 600–1, 609
Bliss, C. I., 487
Blomquist, S., 192
Blundell, R., 468, 558, 561–2, 766, 768, 805–7, 810
Böckenholz, U., 807
Bolduc, D., 528
Bollinger, C. R., 907–8, 921
Bond, S., 113, 758, 765–6, 777

AUTHOR INDEX

- Borgan, O., 595, 609
Borjas, G., 641
Borsch-Supan, A., 506, 528
Bound, J., 104, 106–8, 113, 911
Bover, O., 759, 766
Bowley, A. L., 3, 19
Bradley, P., 411–2
Brännäs, K., 691
Breitung, J., 789
Breslow, N. E., 479
Brennig, R., 820, 857
Breusch, T. S., 275, 737, 761
Bronars, S. G., 57
Brooks, J. C., 562
Brown, C., 58, 379–80, 911
Browning, M., 771
Brownstone, D., 382, 939
Brundy, J. M., 211
Bryk, A. S., 429, 847, 858
Buchinsky, M., 88, 113, 361–2
Bult, J. R., 622
Bunch, D., 517
Burkhauser, R. V., 663
Burtless, G., 49–50, 61
Buse, A., 235, 257
Butler, J. S., 663, 786, 809
- Cameron, A. C., 149, 158, 267–9, 292, 329, 347, 482, 487, 520, 560, 562, 638, 665–6, 671, 673, 675–6, 683, 685, 687, 691, 806, 809, 915
Campbell, D., 55, 62, 897
Cappè, O., 458
Card, D. E., 54, 97, 110, 112, 767, 770, 896, 900
Cardell, S., 529
Carlin, J. B., 432, 445–6, 451–2
Carrasco, R., 888
Carroll, R. J., 328, 333, 901, 921
Carter, C. A., 562
Casella, G., 409, 446, 459
Chamberlain, G., 202, 220, 334, 701, 719, 753, 766, 777, 784, 786, 796–8, 802, 809–10
Charlier, E., 808
Chay, K. Y., 565, 570, 797–8
Chen, M.-H., 609
Chernozhukov, V., 190
Chesher, A., 265, 267, 290, 292, 544, 916
Chib, S., 450, 458–9, 519, 563, 595, 609, 687
Chu, S. J. C., 767
Cincera, M., 667
Clayton, D. G., 649–50
Cleveland, W., 320
Cleves, M. R., 607–8
Cochran, W. G., 817, 857
Cockburn, I. M., 667
Collado, M. D., 771, 773
Conway, D., 920
Cornwell, C., 762
Cosslett, S. R., 479, 827
- Cox, D. R., 280, 289, 292, 574, 594–5, 608, 646
Cragg, J. G., 545, 903–4, 909
Crépon, B., 667, 687, 753, 804, 807
Crooke, S. T., 491
Cuzick, J., 649–50
- Dagenais, D. L., 909
Dagenais, M. G., 909
Dagpunar, J., 411
Dagsvik, J. K., 507, 512, 528
Dahlberg, M., 192
Daly, A., 506
Danielsson, J., 409
Das, M., 566
Davey, A., 939
David, H. A., 643
Davidson, J., 164, 956
Davidson, R., 38, 109, 112, 134, 163, 175, 180, 191, 220, 233, 241, 251, 257, 276, 283, 292, 352, 956
Davis, P., 739, 857
Davison, A. C., 374
Dean, C., 682
Deaton, A., 59–61, 771–3, 820, 857–8
Deb, P., 553, 666, 671, 679, 688, 690–1, 888
Dehejia, R. H., 873, 875–6, 889–90, 893–8
Deistler, M., 126
Delgado, M. A., 685
Dempster, A. P., 346
Deng, Y., 641
DeSarbo, W. S., 622
Devroye, L., 411
Diggle, P. J., 740, 783, 809
DiNardo, J., 112, 716
Domowitz, I., 159, 164
Donald, S. G., 103, 858
Dorsey, R. E., 486
Drukker, D. M., 379
Duan, N., 545, 552, 555, 569
Dubin, J. A., 558–9, 569
Duflo, E., 708, 777
Duguet, E., 667, 687, 807
DuMouchel, W. K., 819–21
Duncan, G. J., 58, 819–21, 911
Durbin, J., 271, 908
- Efron, B., 257, 357, 361, 364–5, 375, 382
Eicker, F., 81, 112, 137
Elbers, C., 619
Engel, E., 3
Engle, R. F., 38, 325, 654
Ervin, L. H., 75
- Fahrmeir, L., 600, 621, 809
Fan, J., 320, 333
Farkas, G., 888
Ferall, C., 609
Ferguson, T. S., 202
Ferrier, G. D., 347, 352

AUTHOR INDEX

- Fisher, R. A., 49, 139, 164, 459
Fleming, T. R., 609
Flinn, C., 663
Fox, B. L., 411–2
Freedman, D. A., 38, 321, 381, 382
Frees, E. W., 655, 663
Friedlander, D., 860
Friedman, J. H., 382
Fuller, W., 920

Gallant, A. R., 163, 321, 328–9, 404, 564–5
Gamerman, D., 433, 446, 458
Geil, P., 42, 666
Gelfand, A. E., 449
Gelman, A., 432, 445–6, 451–2, 458
Geman, D., 449
Geman, S., 449
Gentle, J., 352
George, E., 459
Georges, P., 655
Geweke, J., 403, 406–8, 444–5, 458–9, 518
Ghosh, A., 233
Gijbels, I., 320, 333
Gilks, W. R., 447, 458
Gill, J., 458
Gill, R. D., 595, 609
Girma, S., 771
Gleason, J. R., 374
Glewwe, P., 59
Godambe, V. P., 135
Goffe, W. L., 347, 352
Goldberg, P. K., 512
Goldberger, A. S., 68, 130, 543, 908, 920
Götze, F., 374, 376
Gould, W. W., 607–8
Gouriéroux, C., 38, 148, 158, 257, 290, 394, 396, 404,
 416, 569, 621, 667–8, 683, 685, 691, 940
Govrisankaran, G., 888
Granger, C. W. J., 325
Grasdal, A., 800–1
Graubard, B. I., 857
Greenberg, D., 860
Greenberg, E., 450, 459
Greene, W. H., 14, 38, 112, 119, 164, 191, 214,
 220, 252, 280, 292, 487, 511, 523, 528, 550,
 569, 609, 638, 666, 736, 740, 784–5, 800, 908,
 920, 956
Greenwald, B. C., 858
Gregory, A. W., 232
Griffith, R., 805–7
Griliches, Z., 667, 691, 792, 793, 802, 805–6, 810,
 853
Gritz, M. R., 609
Grogger, J., 57
Gronau, R., 558
Groendorst, P., 666
Grosh, M. E., 59
Groves, R. M., 857

Gruber, J., 54
Guitirrez, R. G., 607–8
Guo, J. Q., 677, 915, 921
Gurmu, S., 665–6, 691

Haan, W. J., 175
Hahn, J., 107, 113, 192, 877, 896, 920
Haisken-DeNew, J., 716
Hajivassiliou, V. A., 396, 407–8, 416, 518–9, 521
Hale, C., 192
Hall, A., 105, 267
Hall, B., 138, 343, 667, 691, 792–3, 802, 805–6, 810,
 853
Hall, D. B., 844
Hall, R., 138, 343
Hall, P., 371, 374, 379, 382
Hall, W. J., 330
Hallock, K. F., 89–90, 113
Hamilton, J., 131, 175, 220
Han, A. K., 485, 601–2, 609, 620, 637–8, 662–3
Hanemann, W. M., 560
Hanratty, M., 54
Hansen, C., 190
Hansen, L. P., 166, 171, 176, 220, 277, 789, 796
Hansen, P. R., 286
Härdle, W., 295, 299, 314–9, 321, 326, 333–4, 395,
 415
Harrington, D. P., 609
Harris, R. D. F., 768
Hartley, R., 347
Harvey, A., 352
Harville, D. A., 775–6
Hastie, T. J., 327
Hastings, W. K., 451
Haughton, D., 625
Hausman, J. A., 52, 61, 107, 113, 138, 192, 271, 276,
 292, 343, 503, 520, 529, 601–2, 609, 620, 637–8,
 662–3, 667, 691, 718, 761, 777, 792–3, 801–2,
 805–6, 810, 829, 853, 857, 904, 913–4, 918,
 920
Hayashi, F., 112, 220
Heagerty, P., 740, 783, 809
Heckman, J. J., 3, 38, 49, 61, 52–3, 543, 547, 549–50,
 558, 562, 569, 617, 619–22, 637–8, 641, 646, 655,
 663, 800, 860–1, 869, 874, 877, 886, 896–7
Hendry, D. F., 19, 38, 257, 956
Herriges, J. A., 491, 512
Heyde, C. C., 459
Hinkley, D. V., 374
Hirano, K., 143, 896
Hoch, I., 740
Hoerl, A. E., 440
Holland, P. W., 23, 32, 34, 38
Holly, A., 257, 274
Holt, D., 724, 836, 838
Holtz-Eakin, D., 758, 765
Honore, B. E., 565, 570, 646, 656, 766, 777, 798, 802,
 808–9

AUTHOR INDEX

- Horowitz, J. L., 177, 319, 326, 333, 347, 361, 368, 370, 372–3, 377, 379, 380–2, 484, 486–7, 569, 656, 808
Horton, N. J., 940
Hosmer, D. W., 638
Hotz, J., 53, 655, 789, 796, 869, 896
Hougaard, P., 615, 621, 637, 651
Houthakker, H. S., 3
Hoynes, H., 797
Hsiao, C., 36, 737–8, 740, 758, 764–6, 768, 777, 801, 809, 909, 912–4, 918
Huang, W., 330
Huber, P. J., 119, 137, 144, 146
Hyslop, D., 797–8

Ibrahim, J. G., 609
Ichimura, H., 325, 327, 485, 874, 877, 896
Ichino, A., 878, 889, 896
Imbens, G. W., 192, 206, 220, 479, 770, 822, 827, 857, 863, 883–5, 896–7
Irish, M., 62, 290, 544, 771
Isacsson, G., 62

Jaeger, D. A., 104, 106–8, 113
Jaggia, S., 574, 609, 621, 630, 632, 635–6, 638
Jeliazkov, I., 458
Jenrich, R. I., 126
Jensen, E. R., 888
Joe, H., 652–3, 654
Johansson, P., 329, 347, 675, 687, 691
Johnson, N. L., 112, 477, 541, 674
Johnston, J., 112
Johnstone, I. M., 459
Jordan, P., 915
Jorgenson, D. W., 211
Judge, G. G., 112, 220
Jureen, L., 3

Kadane, J. B., 390
Kalbfleisch, J., 576, 578, 580, 582, 576, 595, 597–8, 600, 608
Kamakura, W. A., 515
Kapteyn, A., 739
Karlström, A., 507, 512, 528
Karlin, S., 638
Kass, R. E., 390, 456, 457
Kazimi, C., 382
Keane, M. P., 403, 406, 408, 459, 517–8
Keiding, N., 595, 609
Kelijian, H. H., 487
Kemp, A. W., 674
Kinkel, D. S., 888
Kennan, J., 574
Kennard, R. W., 440
Kennedy, W., 352
Kennickell, A. B., 924, 939
Kiefer, N., 590, 600, 609, 638
Killingsworth, M., 558

Kim, E. H., 62
King, G., 479
Kish, L., 857
Klein, J. P., 588
Klein, R. W., 324, 485–6, 523, 914
Klepper, S., 906, 907, 908
Kling, C. L., 491, 512
Kling, J. R., 110–1
Kloek, T., 407, 444, 724, 836, 838, 857
Kmenta J., 723
Kniesner, T. J., 685
Koenker, R., 67, 87, 89–90, 113, 275
Kollia, G. D., 586
Koop, G., 458, 775, 809
Koopmans, T. C., 38
Koppelman, F., 511
Korn, E. L., 857
Kotz, S., 477, 541, 674
Krashinsky, H., 910
Krewski, D., 855
Krueger, A. B., 36, 38, 54, 57, 70, 97, 106, 112, 192, 196, 716, 770, 872, 920
Kuan, C. M., 322
Kuersteiner, G., 192
Kuh, E., 740
Künsch, H. R., 374
Kyriazidou, E., 798, 808–9

Laird, N. M., 346
Lalonde, R., 860, 873, 889, 896
Lancaster, T., 265, 292, 458, 479, 576, 578, 584, 586, 588, 595, 609, 616–8, 629, 638, 782, 809, 822, 827, 857
Lang, K., 858
Lavy, V., 62, 858, 879
Lawless, J. F., 220, 586, 608
Leamer, E. E., 419, 439, 458–9, 906–8, 920
Lechner, M., 789
Lee, C.-I., 113
Lee, L.-F., 523, 557–8, 561, 563, 569–70, 810, 845, 911
Lee, M.-J., 68, 302, 304, 306, 331–3, 487, 569, 701, 730, 748, 777, 798, 809
Lee, S., 656, 808
Lemeshow, S., 638
Lerman, S. R., 416, 479, 487, 528, 828, 857
Leung, S. F., 551–3
Levi, M., 904
Levin, A. T., 175
Levin, L., 767–8
Li, D., 693
Li, H., 381
Li, Q., 722
Li, T., 687, 913, 915, 918, 921
Liang, K.-Y., 796, 809, 970
Lien, D., 283
Lillard, L., 940
Lin, C.-F., 767–8

AUTHOR INDEX

- Lindeboom, M., 648, 651
 Lindley, D. V., 429, 775, 847
 Lindsey, B. G., 622
 Linton, O., 314, 333, 415
 Lipsitz, S. R., 940
 Little, R. J. A., 940
 Liu, R. Y., 377
 Lo, A. W., 520
 Loh, W.-Y., 374
 Long, J. S., 75
 Lovell, M. C., 285
 Luce, R. D., 500
 Lunde, A., 600–1, 609
 MacCurdy, T. E., 171, 729, 754, 761, 767–8, 800
 MacKinlay, A. C., 520
 MacKinnon, J. G., 38, 75, 109, 112, 134, 163, 175,
 180, 191, 220, 233, 241, 251, 257, 276, 283, 292,
 352, 362, 382, 956
 Madansky, A., 920
 Maddala, G. S., 3, 4, 6, 352, 381, 472–4, 480, 487,
 499, 509, 511, 526, 528, 537, 541, 556–7, 561, 569
 Magdalinos, M. A., 277
 Magnac, T., 798
 Mairesse, B., 753
 Mammen, E., 319, 377
 Manning, W. G., 545, 552, 555, 569, 671
 Manski, C. F., 29, 38, 44, 67–8, 85, 112, 135, 381,
 416, 479, 483–4, 487, 528, 566, 808, 828, 857
 Manton, K. G., 620, 637
 Marron, J. S., 395
 Marschak, J., 3, 28, 38
 Marshall, A. W., 649
 Martin, M. A., 374
 Mathiowetz, N., 920
 Matzkin, R. L., 322
 Mátyás, L., 740, 777, 809
 Mayer, W. J., 486
 McCall, B. P., 603, 639, 644–5, 658, 661, 663
 McCallum, B. T., 900
 McCullagh, P., 149, 164, 289, 343, 667, 683, 691, 783
 McCulloch, C. E., 655
 McCulloch, R., 519
 McDonald, J., 542
 McFadden, D., 3, 124, 126, 127, 163–4, 248, 400,
 403, 406–7, 416, 474, 479, 482, 487, 499, 503, 506,
 508–9, 510–2, 515, 518–9, 528, 558–9, 569, 953
 Mealli, F., 643, 655, 657–8, 663
 Meijer, E., 906, 909–10, 920, 921
 Melenberg, B., 334, 544, 565–6, 570, 808
 Melino, A., 619, 624, 637–8
 Melion, A., 42, 666
 Meng, X.-L., 940
 Meyer, B. D., 55, 61, 600–1, 609, 620, 637–8, 662
 Miller, D. J., 112, 220
 Miller, R. G., 375
 Milne, F., 560, 665–6, 687
 Min, C.-K., 458
 Mittelhammer, R. C., 112, 220
 Mizon, G. E., 283, 761
 Moeschberger, M. L., 643
 Moffitt, R., 542, 786, 809, 860, 886, 896
 Monfort, A., 38, 148, 158, 257, 290, 394, 396, 404,
 416, 569, 621, 667–8, 683, 685, 691, 940
 Montalvo, J. G., 807
 Moon, H. R., 767
 Morgan, M. S., 19
 Morris, C. N., 545, 552, 555, 569
 Mosteller, F., 89
 Moulton, B. R., 836, 837, 841, 857
 Mroz, T. A., 276, 548, 558, 569
 Mudholkar, G. S., 586
 Mullahy, J., 546, 555, 684, 687–8, 691
 Mullanathan, S., 708, 777
 Mundlak, Y., 719, 740, 786
 Munkin, M., 687
 Murphy, K., 200, 220
 Nadaraya, E. A., 312, 334
 Nagar, A. L., 113
 Nagase, N., 551
 Nakamura, T., 915–6, 918
 Narendranathan, W., 609, 616
 Nawata, K., 551
 Nelder, J. A., 149, 164, 289, 343, 667, 683, 691, 783
 Nelsen, R. B., 652
 Nelson, C. R., 108, 113–4
 Nelson, F. D., 547, 561, 651
 Nerlove, M., 36, 777
 Nevo, A., 482, 513
 Newey, W. K., 103, 124, 126–7, 137, 159, 163–4, 175,
 180, 200, 220, 257, 264, 292, 322, 330, 334, 380,
 566, 723, 758, 765, 913, 918, 953
 Newhouse, J. P., 50, 545, 552, 555, 569, 671
 Newman, J. L., 655
 Neyman, J., 32, 233, 237, 781
 Nickell, S., 609, 616, 618, 764
 Nychka, D. W., 328–9, 564–5
 Oakes, D., 595, 608
 Olkin, L., 561, 649
 Olsen, R. J., 551, 888
 Owen, A. B., 203, 220
 Paarsch, H. J., 563
 Pagan, A. R., 200, 220, 264, 275, 292, 302, 304, 306,
 310, 317, 319, 325, 330, 333–4, 486–7, 534, 565,
 569, 737
 Pakes, A. S., 406, 416
 Palmgren, J., 805
 Park, J. Y., 233
 Parzen, E., 334
 Pearl, J., 38
 Pearson, E. S., 237
 Pedroni, P., 767
 Pendergast, J., 841, 857

AUTHOR INDEX

- Pesaran, B., 280–1, 292, 492
Pesaran, M. H., 280–1, 292, 492, 723, 766–8
Phillips, G. D. A., 192
Phillips, P. C. B., 233, 371, 767
Piggott, J., 560, 665–6, 687
Pischke, S., 716
Pitt, M. M., 888
Pohlmeier, W., 666, 681, 691
Politis, D. N., 373
Polivka, A. E., 58
Pollard, D., 406, 416
Porter, J., 143
Poskitt, D. S., 105
Poterba, J. M., 914
Powell, J. L., 85, 88, 322, 324, 326, 334, 486, 564–6,
 570, 808, 913, 918
Prais, S. J., 3
Prentice, R. L., 479, 576, 578, 580, 582, 595, 597–8,
 600, 608
Press, W. H., 352, 390, 409, 416
Pritchett, J. B., 609
Pudney, S., 5, 643, 655, 657–8, 663, 857
Puterman, M. L., 667
Pyke, R., 479
- Qin, J., 220
Quandt, R., 352
Quenouille, M. H., 375
Quigley, J., 641
- Rabe-Hesketh, S., 786
Radulović, D., 663
Raftery, A. E., 456–7
Ramaswamy, V., 622
Ramsey, J. B., 278
Rao, C. R., 946–50, 956
Rao, J. N. K., 855
Raudenbush, S. W., 429, 847, 858
Reid, F., 482
Reiersol, O., 908
Renault, E., 290, 404
Revelt, D., 513, 528
Rice, J., 325
Richard, J.-F., 38, 283, 409
Richardson, S., 447, 458
Ridder, G., 619, 896
Ripley, B. D., 408, 411
Rivers, D., 561
Robb, R., 896
Robert, C. P., 409, 446, 458
Roberts, H. V., 920
Robins, P., 860
Robinson, P. M., 323–5, 328, 334, 565, 566
Rogers, J., 347, 352
Rogers, W. L., 911
Romano, J. P., 373
Rose, N., 665
Rosen, H. S., 528, 758, 765
- Rosenbaum, P., 865, 873, 881, 896
Rosenblatt, M., 299, 334
Rosenzweig, M., 57, 62, 888
Rossi, P. E., 482, 513, 519
Rothenberg, T. J., 202, 370–1
Rotte, R., 42, 666
Roy, A., 32, 555, 557, 569
Rubin, D. B., 32, 346, 432, 445–6, 451–2, 458, 621,
 623, 863, 865, 873, 881, 883, 885, 896, 925–6, 934,
 940
Rubin, H., 191
Rudebusch, G., 105
Runkle, D., 403
Rupert, P., 762
Ruppert, D., 333, 901, 921
Russell, J. R., 654
Rust, J., 29
Ruud, P. A., 112, 292, 403, 416, 472, 521, 956
- Sakata, S., 856
Salant, S. W., 617, 626–7, 638
Sapra, S., 609, 638, 663
Sargan, J. D., 20, 23–5, 40, 113, 220, 277, 286, 371,
 764
Savin, N. E., 210, 230, 238
Schafer, J. L., 927–8, 935, 939
Schafgens, M. M. A., 566
Schechtman, E., 374
Schmidt, C. M., 716
Schmidt, P., 25, 214, 761, 766, 768, 777
Schrage, L. E., 411–2
Schwarz, G., 279
Schweizer, B., 653
Scott, A. J., 724, 836, 838
Scott, E. L., 781
Scott-Morton, F. M., 914
Segal, L. M., 177, 192, 379, 403, 767
Seifert, H.-G., 126
Sepanski, J. H., 911
Serfling, R., 119
Severini, T. A., 330, 334, 844
Sevestre, P., 740, 777, 809
Shanahan, M. J., 939
Shao, J., 855
Shaw, D., 823
Shea, J., 105, 107, 109
Sherman, R. C., 485, 523, 914
Silverman, B. W., 334
Silvey, S. D., 235
Singal, V., 62
Singer, B., 617, 619–22, 637–8
Singh, K., 370, 382
Singleton, K. J., 171
Sinha, D., 609
Skeels, C. L., 105
Skinner, C. J., 857
Sklar, A., 651, 663
Skrondal, A., 786

AUTHOR INDEX

- Small, K. A., 528
 Smith, A. A., 404
 Smith, A. F. M., 429, 449, 775, 847
 Smith, J. A., 49, 52, 53, 61, 860–1, 874, 877, 896
 Smith, J. P., 940
 Smith, M. D., 654
 Smith, R., 767
 Smith, R. J., 561–2
 Snell, E. J., 289
 Solaro, M. E., 907
 Solon, G., 767, 921
 Spady, R. H., 324, 485–6, 523
 Speckman, P., 325
 Spiegelhalter, D. J., 447, 458
 Srivastava, D. K., 586
 Stafford, F. P., 58
 Staiger, D., 105, 109
 Stallard, E., 620, 637
 Startz, R., 108, 113–4
 Stefanski, L. A., 915–6, 921
 Stegun, I. A., 352, 390
 Steinbrickner, T. R., 939
 Stern, H. S., 432, 445–6, 451–2
 Stern, J., 609, 616
 Stern, S., 416
 Stevens, A. H., 656
 Stigler, S. M., 112, 458
 Stock, J. H., 105, 109, 112–3, 326
 Stoker, T. M., 92, 326, 487
 Stone, C. J., 322, 334
 Stone, R., 3
 Sueyoshi, G. T., 601, 619
 Sullivan, R., 286
 Summers, L. H., 914
 Sun, B. H., 914
 Swait, J., 515
 Swamy, P. A. V. B., 775
 Szu, H., 347
 Tahmisioglu, M., 766
 Tanner, M. A., 449, 455
 Tauchen, G., 264, 292, 404
 Taylor, H. M., 638
 Taylor, W. E., 761, 777
 Terza, J., 688, 691, 888
 Theil, H., 189, 209, 214
 Thistletonwaite, D., 897
 Thomas, J. M., 646
 Thompson, T. S., 484
 Thomson, C. J., 491
 Tibsharani, R. J., 327
 Tierney, L., 390
 Timmermann, A., 286, 600–1, 609
 Tobias, J. I., 897
 Tobin, J., 3, 529, 536, 569
 Todd, P. E., 874, 896
 Topel, R., 200, 220
 Town, R. J., 888
 Trebbi, F., 113
 Train, K. E., 409–10, 416, 433, 459, 487, 505, 511, 513–4, 518, 528, 798
 Tripathi, G., 330, 334
 Trivedi, P. K., 149, 158, 267–8, 289, 292, 520, 553, 560, 630, 632, 638, 655–6, 671, 673, 675–7, 683, 685, 687–8, 690–1, 806, 809, 888, 915, 921
 Trochim, W., 879
 Trognon, A., 148, 158, 290, 667–8, 683, 685, 691
 Trost, R. P., 561
 Tsatsis, A., 646
 Tu, D., 855
 Tukey, J. W., 89, 375
 Tunali, I., 609
 Tutz, G., 600, 621, 809
 Tzavalis, E., 768
 Ullah, A., 302, 304, 306, 310, 317, 319, 325, 330, 333–4, 486–7, 534, 565, 569, 820, 857
 Ulrich, V., 666, 681, 691
 U. S. Census Bureau, 855
 Valdez, E. A., 655, 663
 Valletta, R., 939
 Van den Berg, G., 609, 619, 638, 646, 648, 651, 802, 809
 Van Dyck, H., 407, 444
 Van der Klaauw, W., 879–81, 896
 Van Ophem, H., 888
 Van Order, R., 641
 Van Reenen, J., 805–7
 Van Soest, A., 334, 544, 565–6, 570, 808
 Van Zwet, W. R., 376
 Vaupel, J. W., 620, 637
 Veall, M. R., 232
 Vella, F., 264, 275, 292, 534, 566, 570, 801, 897
 Verbeek, M., 773, 801, 897
 Verdinelli, I., 457
 Vuong, Q. H., 280–3, 292, 479, 561
 Vytlacil, E., 886, 897
 Wahba, S., 873–5, 889–96, 898
 Wald, A., 99, 224
 Walker, I., 888
 Walker, J., 516
 Walker, J. R., 566
 Wand, M. P., 333
 Wang, P., 667
 Wansbeek, T. J., 739, 906, 909–10, 920–1
 Wasserman, L., 457
 Watson, G. S., 312, 334
 Wedderburn, R. W. M., 667, 683
 Wedel, M., 515, 622
 Wegkamp, M., 663
 Weiss, A., 325
 Welch, F., 940
 Wellner, J. A., 330

AUTHOR INDEX

- Wen, C., 511
West, K. D., 137, 159, 175, 257, 723
White, H., 75–6, 91, 102, 112, 131, 137, 146–7, 155,
159, 163–4, 265, 275, 286, 292, 322, 948, 951, 956
Wilcox, D., 105
Williams, H., 506
Willmot, G. E., 677
Windmeijer, F. A. G., 113, 289, 806, 810
Winkelmann, R., 665, 674, 687
Wise, D. A., 52, 61, 801, 829, 857
Wolak, F. A., 257
Wold, H., 3
Wolf, M., 373
Wolff, E. F., 653
Wolpin, K., 57, 62
Wolter, K. M., 855
Wong, W. H., 449, 455
Wooldridge, J. M., 119, 150, 217, 220, 238, 244, 263,
274–5, 292, 701, 718, 740, 784, 801, 809, 858, 863,
886–7, 953, 956
Wright, J. H., 109, 112–3
Wu, C. F. G., 377, 855
Wu, D., 271
Yatchew, A., 319, 325, 333–4, 380
Yogo, M., 109, 112–3
Yoon, M-J., 630
Yu, S., 551–3
Yule, G. U., 112
Zachary, S., 506
Zaman, A., 164
Zellner, A., 209, 214, 419, 435–6,
458–9
Zeng, L., 479
Ziliak, J. P., 192, 709, 752, 754–6
Zeger, S. L., 796, 809, 970
Zimmer, D. M., 687
Zimmerman, D. J., 921
Zimmermann, K. F., 42, 666

Subject Index

- accelerated failure time (AFT) model, 591–2
 coefficient interpretation, 606–7
 definition, 592
 leading examples, 585
accept-reject methods, 413–4, 445
ACD. *See* average completed duration
acronyms, 17
AD estimator. *See* average derivative
adaptive estimator, 323, 328, 684
adding-up constraints, 210
additive model, 323, 327, 523
additive random utility model (ARUM)
 binary outcome models, 476–8
 generalized random utility models, 515–6
 identification, 504
 multinomial outcome models, 504–7
 nested logit model, 509, 526–7
 RPL model, 513
 welfare analysis in, 506–7
admissible estimator, 435
AFT. *See* accelerated failure time
aggregated data
 binary outcomes, 480–2
 cohort-level, 772
 nonlinear models, 482, 487
 multinomial outcomes, 513
 time-aggregated durations, 578, 600–3
 see also discrete-time duration data
AIC. *See* Akaike information criterion
AID. *See* average interrupted duration
Akaike information criterion (AIC), 278–9, 284, 624
almost sure convergence, 947–8
analog estimator, 135
analogy principle, 135
 and method of moments estimators, 167
analysis of covariance, 733
analysis of variance, 733
Anscombe residual, 289
antithetic sampling, 408–9, 445
applications with data
 competing risks models, 658–62
 duration models, 603–8, 632–6
 IV estimation, 110–2
 kernel regression, 295–7, 300
 logit and probit models, 464–6, 486
 multinomial and nested logit models, 491–5, 511
 Poisson and negative binomial models, 671–4, 690
 panel fixed and random effects estimation, 708–15
 panel GMM linear estimation, 754–6
 panel nonlinear estimation, 792–5
 quantile regression, 88–90
 selection and two-part models, 553–6, 565
 survival function, 574–5, 582
 treatment evaluation estimation, 889–96
 see also data sets used in applications
Archimedean family, 654
Arellano-Bond estimator, 765–6, 777
 application, 754–6
 nonlinear models, 791
 unit roots, 768
ARMA. *See* autoregressive moving average
artificial nesting, 283
ARUM. *See* additive random utility model
asymptotic distribution, 953–4
 asymptotic efficiency, 954
 asymptotic normal distribution, 953
 definition, 74, 120, 953
 estimated asymptotic variance, 954
 of extremum estimators, 127–31
 of FGLS estimator, 82–3
 of FGNLS estimator, 156–7
 of first-differences estimator, 730–1
 of fixed effects estimator, 727–9
 of GMM estimator, 173–4, 182–3, 185–6, 194–5,
 745–6
 of Hausman test statistic, 271–4

- of kernel density estimator, 301–2, 330–1
 of kernel regression estimator, 313, 331–3
 of LM test statistic, 235, 237–8
 of LR test statistic, 235, 237
 of m-estimators, 119–21
 of MD estimator, 292
 of ML estimator, 142–3
 of MM estimator, 134, 174
 of MSL estimator, 394–5
 of MSM estimator, 400–2
 of m-test statistics, 260, 263
 of NLS estimator, 152–4
 of NL2SLS estimator, 195–6
 of OIR test statistic, 181, 183
 of OLS estimator, 73–4, 80–1
 of panel GMM estimator, 745–6
 of quasi-ML estimator, 146
 of random effects estimator, 735
 of Wald test statistic, 226–8
see also asymptotic theory
 asymptotic efficiency, 954
 of optimal GMM, 177
 asymptotic refinement, 359, 371–2
 by bootstrap, 256, 363–7, 371–2, 378–9
 definition, 359
 by Edgeworth expansion, 371–2
 by nested bootstrap, 374, 379
 asymptotic theory definitions, 943–55
 asymptotic distribution, 953
 asymptotic variance, 954
 central limit theorems, 949–52
 consistency, 945
 convergence in distribution, 948–9
 convergence in probability, 944–7
 laws of large numbers, 947–8
 limit distribution, 948
 limit variance, 952–3
 stochastic order of magnitude, 954
 summary of definitions and theorems, 944
 asymptotic variance, 74, 120, 954
 estimated asymptotic variance, 74, 954
 see also asymptotic distribution
 asymptotically pivotal statistic, 359–60, 363–4, 366,
 372, 374, 379–80
 ATE. *See* average treatment effect
 ATET. *See* average treatment effect on the treated
 attenuation bias, 903–5, 911, 915, 919–20
 attrition bias, 739, 800–1, 940
 augmented regression model, 429
 autocorrelation
 in panel model errors, 705–8, 714–5, 722–5, 745–6
 dynamic panel models, 763–8, 791–2, 797–9,
 806–7
 see also panel-robust inference
 autoregressive moving average (ARMA) errors
 definition, 159
 NLS estimator, 159
 panel data, 722–5, 729
 auxiliary model, 404
 auxiliary regression
 bootstrapping, 379, 382
 example, 241–3, 269–71
 Hausman test, 276, 718–9
 LM test, 240–1, 274
 m-test, 261–4, 544
 available case analysis. *See* pairwise deletion
 average completed duration (ACD), 626
 average derivative (AD) estimator
 definition, 326
 uses, 317, 483
 average interrupted duration (AID), 626
 average selection bias, 868
 average squared error, 315
 average treatment effect (ATE), 33–4, 866–71
 definition, 866
 difficulties estimating, 866
 local ATE, 883–6
 matching estimators, 871–8
 potential outcome model, 33–4
 selection on observables only, 868–9
 selection on unobservables, 868–71
 see also ATET; LATE; MTE
 average treatment effect on the treated (ATET),
 866–78
 application, 889–6
 definition, 866
 difficulties estimating, 866
 matching estimators, 871–8, 894–6
 selection on observables only, 868–9
 selection on unobservables, 868–71
 see also ATE; LATE; MTE
 averaged data. *See* aggregated data
 backward recurrence time, 626
 balanced bootstrap, 374
 balanced repeated replication, 855
 balancing condition, 864, 893–4
 bandwidth, 299, 307, 312
 bandwidth choice for kernel density estimator, 302–4
 cross validation, 304
 example, 296–7
 optimal, 303, 306
 Silverman's plug-in estimate, 304
 bandwidth choice for kernel regression estimator,
 314–6
 cross validation, 314–6
 example, 297, 316
 optimal, 314, 318
 plug-in estimate, 314
 baseline hazard, 591
 in AFT model, 592
 identification in mixture models, 618–20
 in multiple spells models, 655–6
 in PH model, 591, 596–7, 601–2
 Bayes factors, 456–8
 Bayes rule. *See* Bayes theorem

- Bayes theorem, 421
 example, 422–4, 435–9
- Bayesian central limit theorem, 433
- Bayesian information criterion (BIC), 278, 284
see also AIC
- Bayesian methods, 419–59
 Bayes 1764 example, 458–9
 Bayesian approach, 420–35
 binary outcome models, 475
 compared to non-Bayesian, 164, 424–5, 432–41, 439–41
 count models, 687
 data augmentation, 454–5, 932–3, 935–9
 decision analysis, 434–5
 examples, 452–4
 hierarchical linear model, 847
 importance sampling, 443–5
 linear regression, 435–43, 449–50, 452–4
 Markov chain Monte Carlo simulation, 445–54, 935–9
 measurement error model, 915
 mixed linear model, 775
 model selection, 456–8
 multinomial outcome models, 514, 519
 panel data, 775, 809
 posterior distribution, 421, 430–4
 prior distribution, 425–30
 Tobit model, 563
- BCA method. *See* bias-corrected and accelerated before-after comparison
 application, 890–1
- Berkson error model, 920
- Berkson's minimum chi-square estimator, 480–1
- Berndt, Hall, Hall, and Hausman (BHHH) estimate, 138, 241, 395
- Berndt, Hall, Hall, and Hausman (BHHH) iterative method, 343–4
- Bernoulli distribution, 140, 148, 468, 475, 483
- Bernstein-von Mises Theorem, 433, 459
- best linear unbiased predictor, 738, 776
- between estimator, 702, 736, 841
 application, 710–3
- between-group variation, 709, 733
- between model, 702
- BFGS algorithm. *See* Boyden, Fletcher, Goldfarb, and Shannon
- BHHH estimate. *See* Berndt, Hall, Hall, and Hausman
- BHHH method. *See* Berndt, Hall, Hall, and Hausman
- bias-corrected and accelerated (BCA) bootstrap method, 360
- biased sampling, 42–5, 626–7
see also sample selection; endogenous stratification
- BIC. *See* Bayesian information criterion
- binary endogenous variable, 562
- binary outcome models, 463–89
 additive random utility model, 476–8
 aggregated data, 480–2
 alternative-invariant regressors, 478
- alternative-varying regressors, 478
- choice-based samples, 478–9
- corrected score estimator, 916–8
- definition, 466
- example, 464–5
- identification, 476, 483
- index function model, 475–6
- marginal effects, 467, 470–1
- measurement error in dependent variable, 914
- measurement error in regressors, 919
- ML estimator, 468–9
- model misspecification, 472
- multiple imputation example, 937–8
- OLS estimator, 471
- panel data, 795–9
- semiparametric estimation, 482–6
see also logit models; probit models
- binding function, 404–5
- bivariate counts, 215, 685–7
- bivariate negative binomial distribution, 686–7
- bivariate ordered probit model, 523
- bivariate Poisson distribution, 686
- bivariate Poisson-lognormal mixture, 686
- bivariate probit model, 522–3
- bivariate sample selection model, 547–53
 application, 553–5
 bounds, 566
 conditional mean, 548–50
 conditional variance, 549–50
 definition, 547
 Heckman two-step estimator, 550–1
 identification, 551, 565–6
 marginal effects, 552
 ML estimator, 548
 outcome equation, 547
 participation equation, 547
 semiparametric estimator, 565–6
 versus two-part model, 546, 552–3
- Bonferroni test, 230
- bootstrap hypothesis tests
 asymptotic refinement, 363–4, 366–7, 371–2, 378–9
 bootstrap critical value, 256, 363
 bootstrap p-value, 256, 363
 example, 366–8
 nonsymmetrical test, 363, 380
 power, 372–3
 symmetrical test, 363
 without asymptotic refinement, 363, 367–8, 378
- bootstrap methods, 357–83
 asymptotic refinement, 359, 366–7
 bias estimate, 365
 bias-corrected estimator, 365, 368
 clustered data, 363, 377–8, 845
 confidence intervals, 364–5, 368
 consistency, 369–70
 critical value, 363

- examples, 254–6, 366–8
 for functions of parameters, 363
 general algorithm, 360
 for GMM, 379–80
 heteroskedastic data, 363, 376–7
 introduction, 254–6
 for nonsmooth estimators, 373, 380–1
 number of bootstrap samples, 361–2
 panel data, 363, 377–8, 708, 746, 751
 p-value, 363
 recentering, 374, 379
 rescaling, 374
 sampling methods for, 360
 smoothness requirements, 370
 standard error estimate, 362, 366
 time series data, 381
 variance estimate, 362
 without asymptotic refinement, 358, 367–8
see also bootstrap hypothesis tests
- bounds identification, 29
 in measurement error models, 906–8
- bounds in selection model, 566
- Boyden, Fletcher, Goldfarb, and Shannon (BFGS) algorithm, 344
- CAIC. *See* consistent Akaike information criterion
- calibrated bootstrap, 374
- caliper matching, 874, 895
- canonical link function, 149, 469, 783
- case-control analysis, 479, 823
- causality, 18–38
 examples, 69–70, 98
 Granger causality, 22
 identification frameworks and strategies, 35–3
 in linear regression model, 68–9
 in potential outcome models, 32–4, 862–5
 in simultaneous equations model, 26–7
 in single-equation model, 31
 and weighting, 820–1
see also endogeneity
- cdf. *See* cumulative distribution function
- censored least absolute deviations (CLAD) estimator, 564–5, 808
- censored models, 530–44, 579–80
 conditional mean, 535
 count models, 680
 definitions, 532, 579–80
 examples, 530–1, 535
 ML estimator, 533–4
 semiparametric estimation, 563–5
see also duration model; selection models; Tobit models; truncated models
- censored normal regression model. *See* Tobit model
- censoring mechanisms, 532, 579–80
 censoring from above, 532, 579
 censoring from below, 532, 579
 left censoring, 532, 579, 588
- independent censoring, 580
 interval censoring, 579, 588
 noninformative censoring, 580
 random censoring, 579
 right censoring, 532, 579, 581, 589
 sample selection, 44–5, 547
 type 1 censoring, 579
 type 2 censoring, 580
- census coefficient, 819
- central limit theorem (CLT), 949–2
 Cramer linear transformation, 952
 Cramer-Wold device, 951
 definition, 950
 examples of use, 80, 130
 Liapounov CLT, 950
 Lindeberg-Levy CLT, 950
 multivariate, 951–2
 sample average, 949
 sampling scheme, 131, 950
- CGF tests. *See* chi-square goodness-of-fit
- characteristic function, 370, 913, 950
- chatter, 394, 410
- Chebychev's inequality, 946
- chi-square goodness-of-fit (CGF) tests, 266–7, 270–1, 474
- choice-based samples, 823
 binary outcome models, 478–9
see also endogenous stratification
- Choleski decomposition, 416, 448
- CL model. *See* conditional logit
- CLAD estimator. *See* censored least absolute deviations
- Clayton copula, 654
- CLT. *See* central limit theorem
- clustered data, 829–53
 application, 848–53
 cluster bootstrap, 363, 377–8, 845
 cluster-robust inference, 707, 834, 842, 845
 cluster sampling, 41–2
 cluster-specific effects, 830–2, 837–45
 comparison to panel data, 831–2
 diagnostic tests, 841
 dummy variables model, 840
 fixed effects estimator, 840–1, 843–5
 hierarchical models, 845–8
 large clusters, 832
 nonlinear models, 841–5
 OLS estimator, 75, 833–7
 quasi-ML estimator, 150
 random effects estimator, 837–9, 843
 small clusters, 832
see also panel data
- cluster-robust standard errors
 bootstrap, 363, 377–8, 845
 clustered data, 834, 842
 panel data, 706–7, 745–6, 789
see also robust standard errors

- cluster-specific fixed effects (CSFE) estimator, 839–41, 843–4
 application, 848–53
 between estimator, 840–1
 nonlinear models, 843–4
 within estimator, 140–1
- cluster-specific fixed effects (CSFE) model, 831, 843
- cluster-specific random effects (CSRE) estimator, 837–9, 843–4
 application, 848–53
- cluster-specific random effects (CSRE) model, 831, 843–4
- cluster variable, 707
- CM tests. *See* conditional moment
- coefficient interpretation
 in binary outcome models, 467, 473
 in competing risks model, 646
 in count model, 669
 in duration models, 606–7
 in misspecified linear model, 91–2
 in multinomial outcome models, 493–4, 501–3
 in nonlinear models, 122–4, 162–3
 in Tobit model, 541–2
see also marginal effects
- coherency condition, 562
- cohort-level data. *See* pseudo panels
- cointegration, 382, 767
- common parameters, 801
- compensating variation, 500–7, 512
- competing risks model (CRM), 642–8, 658–62
 application, 658–62
 censoring, 642
 coefficient interpretation, 646
 definitions, 642–4
 dependent risks, 647–8
 exit route, 643
 identification, 646
 independent risks, 644–6
 ML estimator, 644–5
 proportional hazards, 645–6
 spell duration, 643
 with unobserved heterogeneity, 647, 659
- complementary log-log model, 466–7, 603
- complete case analysis. *See* listwise deletion
- complex surveys, 41–2, 814–6, 853–6
- composition methods, 415
- computational difficulties, 350–2
- concentration parameter, 109
- conditional analysis, 717
- conditional expectations, 955–6
- conditional independence assumption, 23, 863, 865
 definition, 863
 for participation, 863
 given propensity score, 865
 selection on observables only, 868
 unconfoundedness, 863
- conditional likelihood, 139–40, 824
- panel models, 731–2, 782–3, 796–9, 805
- conditional logit (CL) model, 500–3, 524–5
 application, 491–4
 definition, 500
 fixed effects binary logit, 797, 844
 marginal effects, 493, 501–3, 525
 ML estimator, 501
 from ARUM, 505
see also multinomial outcome models
- conditional ML estimator, 731–2, 782–3, 796–9, 805, 824
- conditional moment (CM) tests, 264–5, 267–9, 319
 consistent CM test, 268
 in duration models, 632
 example, 269–71
 in Tobit model, 544
see also m-tests
- conditional mean
 squared error loss, 67–9
- conditional mode
 step loss, 68
- condition number, 350
- conditional quantile
 asymmetric absolute loss, 68
- confidence intervals, 231–2, 316, 364–5, 368
- consistent Akaike information criterion (CAIC), 278
- consistent test statistic, 248
- consistency
 definition, 945
 of extremum estimators, 125–7, 132–3
 of GMM estimator, 173–4, 182
 of m-estimator, 132–3
 of ML estimator, 142, 146–50
 of NLS estimator, 155
 of OLS estimator, 73, 80
 strong consistency, 947
 weak consistency, 947
see also asymptotic distribution; identification;
 pseudo-true value
- constant coefficients model. *See* pooled model
- contagion, 612
- contamination bias, 903–4
- contemporaneous exogeneity assumption, 748–9, 752, 781
- continuous mapping theorem, 949
- control function approach, 37
- control function estimator, 869–70, 890
- control group, 49
- conventions, 16–17
- convergence criteria, 339–40, 458
- convergence in distribution, 948–9
 continuous mapping theorem, 949
 definition, 948
 limit distribution, 948
 transformation theorem, 949
- vector random variables, 949
see also central limit theorem
- convergence in probability, 944–7
 alternative modes of convergence, 945

- consistency, 945
 definition, 945
 probability limit, 945
 Slutsky's theorem, 945
 uniform convergence, 126, 301
 vector random variables, 945
see also law of large numbers
 copulas, 216, 651–5
 count example, 687
 definition, 651–2
 dependence parameter, 653–4
 leading examples, 654
 ML estimator, 655
 survival copulas, 652
 correlated random effects model, 719, 786
 counterfactual, 32, 555, 861, 871
 see also potential outcome model
 count data, 665
 examples, 665
 heteroskedasticity, 665
 right-skewness, 665
 see also count models
 count models, 665–93
 censored, 680
 application, 671–4, 690
 endogenous regressors, 683, 687–9
 endogenous sampling, 823
 finite mixture models, 678–9
 hurdle models, 680–1
 measurement error in dependent variable, 915
 measurement error in regressors, 915–8
 mixture models, 675–7
 multivariate, 685–7
 OLS estimator, 684
 negative binomial model, 675–7
 NLS estimator, 684
 panel data, 792–5, 802–8
 Poisson model, 666–74
 sample selection, 680
 semiparametric regression, 684–5
 truncated, 679–80
 zero-inflated, 681
 covariance matrix. *See* variance matrix
 covariance structures, 177, 379, 753, 766–7
 covariates. *See* regressors
 Cox CRM model. *See* competing risks
 Cox PH model. *See* proportional hazards
 Cox-Snell residual, 289, 631, 633–6
 CPS. *See* Current Population Survey
 Cramer linear transformation, 952
 Cramer-Rao lower bound, 143, 954
 see also semiparametric efficiency bound
 Cramer's theorem, 949
 Cramer-Wold device, 130, 951
 CRM. *See* competing risks model
 cross-equation parameter restrictions, 210
 cross-section data, 47
 cross-validation, 304, 314–6, 318, 321
 CSFE estimator. *See* cluster-specific fixed effects
 CSRE. *See* cluster-specific random effects
 cumulant, 370
 cumulative distribution function (cdf), 576
 cumulative hazard function
 definition, 577–8
 in competing risks model, 644–5
 as diagnostic tool, 631–2
 in likelihood function, 588
 Nelson-Aalen estimator, 582–4, 605–6, 662
 in proportional hazards model, 590
 Current Population Survey (CPS), 58, 814–5
 curse of dimensionality
 in Bayesian methods, 419–20
 multivariate kernel density estimator, 306
 multivariate kernel regression estimator, 319
 high-dimensional integrals, 393
 data augmentation, 454–5, 932
 imputation step, 455, 932
 for missing data, 932–8
 prediction step, 455, 933
 regression example, 933
 data-generating process (dgp), 72–3, 124
 misspecified, 90, 132
 data mining, 285–6
 data sets. *See* microdata
 data sets used in applications
 Current Population Survey Displaced Workers
 Supplement (McCall), 603–8, 632–6, 658–62
 fishing-mode choice data (Kling and Herriges),
 463–6, 486, 491–5
 National Longitudinal Survey (Kling), 110–2
 National Supported Work demonstration project
 (Dehejia and Wahba), 889–95
 Panel Survey of Income Dynamics cross-section
 sample, 295–7, 300
 Panel Survey of Income Dynamics panel sample
 (Ziliak), 708–15, 754–6
 patents-R&D panel data (Hausman, Hall, and
 Griliches), 792–5
 Rand Health Insurance Experiment expenditures,
 553–6, 565
 Rand Health Insurance Experiment medical doctor
 contacts, 671–4, 692
 strike duration data (Kennan), 574–5, 582
 Vietnam World Bank Livings Standards Survey,
 88–90, 848–53
 see also applications with data
 data structures, 39–62
 data sources, 58–9
 handling microdata, 59–61
 natural experiments, 54–8
 observational data, 40–8
 social experiments, 48–54
 data summary approach to regression, 820
 Davidon, Fletcher, and Powell (DFP) algorithm, 344,
 350–1

- decomposition of variance, 955–6
 degenerate distribution, 948
 degrees-of-freedom adjustment, 75, 102, 138, 185–6,
 278, 841
 delta method, 231–2
 bootstrap alternative, 363
 density kernel, 421
 density-weighted average derivative (DWAD)
 estimator, 326
 dependent variable, 71
 descriptive approach to regression, 820
 deviance, 149, 244
 deviance residual, 289, 291
 DFP algorithm. *See* Davidon, Fletcher, and Powell
 algorithm
 dgp. *See* data-generating process
 diagnostic tests. *See* specification tests
 DID estimator. *See* differences-in-differences
 differences-in-differences (DID) estimator, 55–7,
 768–70, 878–9
 application, 890–1
 consistency, 770
 definition, 768
 introduction, 55–7
 natural experiments, 878
 with controls, 878–9
 without controls, 878
 direct regression, 906
 disaggregated data
 contrasted with aggregated data, 5–10
 discrete factor models, 678
 see also finite mixture models
 discrete outcomes. *See* binary outcomes; counts;
 multinomial outcomes
 discrete-time duration data, 577–8, 600–3
 cumulative hazard function, 578
 discrete-time proportional hazards, 600–3
 gamma heterogeneity, 620
 hazard function, 578
 logit model, 602
 ML estimator, 601
 nonparametric estimation, 581–4
 probit model, 602
 survivor function, 578
 dissimilarity parameter, 509
 disturbance term. *See* error term
 double bootstrap, 374
 dummy endogenous variable model, 557
 dummy variable estimator, 784–5, 800, 805, 840
 see also LSDV estimator
 duration data, 573–664
 different types, 626, 641
 duration models, 573–664
 accelerated failure time, 591–2
 applications, 574–5, 583, 589, 603–8, 632–6,
 658–62
 censoring, 579–82, 587–9, 595, 642
 competing risks, 642–8, 658–62
 cumulative hazard function, 577–8
 discrete time, 577–8, 600–3
 generalized residual, 631
 hazard function, 576, 578
 key concepts, 576–8
 mixture models, 613–25
 ML estimator, 587–9
 multiple spells, 655–8
 multivariate, 648–55
 nonparametric estimators, 580–4
 OLS estimator, 590–1
 panel data, 801–2
 parametric models, 584–91
 proportional hazards, 592–7
 risk set, 581, 594
 semiparametric estimation, 594–600, 610–2
 specification tests, 628–32
 survivor function, 576, 578
 time-varying regressors, 597–600
 see also proportional hazards model
 DWAD estimator. *See* density-weighted average
 derivative
 dynamic panel models, 763–8, 791–2, 797–9,
 806–7
 Arellano-Bond estimator, 765–6
 binary outcome models, 806–7
 count models, 806–7
 covariance structures, 766–7
 inconsistency of standard estimators, 764–5
 initial conditions, 764–5
 IV estimators, 764–5
 linear models, 763–8
 MD estimator, 767
 nonlinear models, 791–2, 797–9, 806–7
 nonstationary data, 767–8
 transformed ML estimator, 766
 true state dependence, 763–4
 unobserved heterogeneity, 764
 weak exogeneity, 749
 EDF bootstrap. *See* empirical distribution function
 bootstrap
 Edgeworth expansions, 370–1
 efficient score, 141
 Eicker-White robust standard errors, 74–5, 80–1, 112,
 137, 164, 175
 see also heteroskedasticity robust-standard errors
 EM algorithm *see* expectation maximization
 empirical Bayes method, 442
 empirical distribution function (EDF) bootstrap, 360
 see also paired bootstrap
 empirical likelihood, 203–6
 empirical likelihood bootstrap, 379–80
 encompassing principle, 283
 endogeneity
 definition, 92
 due to endogenous stratification, 78, 824–5
 Hausman test for, 271–2, 275–6

- identification frameworks and strategies, 35–7
see also endogenous regressors; exogeneity
- endogenous regressors, 78
- binary, 557, 562
 - in count models, 683–4, 687–9
 - in discrete outcome models, 473
 - in duration models, 598
 - dummy, 557, 562
 - inconsistency of OLS, 95–6
 - in linear panel models, 744–63
 - in linear simultaneous equations model, 23–30
 - in nonlinear panel models, 792
 - in potential outcome model, 30–3
 - returns-to-schooling example, 69–70
 - in selection models, 559–62
 - in single-equation models, 30
 - see also* GMM estimator; IV estimator
- endogenous sampling, 42–5, 78, 822–9, 856
- consistent estimation, 827–9
 - leading examples, 823
 - see also* censored models; endogenous stratification; sample selection models
- endogenous stratification, 820, 826–7, 856
- equation-by-equation OLS, 210
- equicorrelated errors, 701, 722–4, 804
- equidispersion, 668, 670
- error components model. *See* RE model
- error components SEM, 762
- error components SUR model, 762
- error components 2SLS estimator, 760
- error components 3SLS estimator, 762
- error term, 71, 168
- additive, 168
 - nonadditive, 168
- errors-in-variables. *See* measurement error
- estimated asymptotic variance, 954
- see also* asymptotic distribution
- estimated prediction error. *See* cross-validation
- estimating equations estimator, 13–5
- asymptotic distribution, 134–5, 174
 - clustered data, 842
 - computation, 339
 - definition, 134
 - generalized, 134, 790, 794, 804
 - variance matrix estimation, 137–9
 - weighted, 829
 - see also* MM estimator
- Euler conditions, 171, 749
- exact identification. *See* just identification
- exchangeable errors, 701, 804
- exhaustive sampling, 815–6
- exogeneity, 22–3
- conditional independence, 23
 - Granger causality, 22
 - of instrument, 106
 - overidentifying restrictions test for, 277
 - panel data assumptions, 700, 748–52, 754,
- strong exogeneity, 22
 - weak exogeneity, 22
- exogenous sampling, 42–3
- exogenous stratified sampling, 42, 78, 814–5, 820, 825, 856
- exogenous regressor. *See* exogeneity
- expectation maximization (EM) algorithm, 345–7
- for data imputation, 930–2
 - E (Expectation) step, 346
 - for finite mixture model, 623–5
 - M (Maximization) step, 346
 - compared to NR algorithm, 625
- expected elapsed duration, 626
- experimental data, 48–58
- control group, 49
 - natural experiments, 54–8
 - social experiments, 48–54
 - treatment group, 49
- explanatory variables. *See* regressors
- exponential conditional mean, 124, 155, 669
- coefficient interpretation, 124, 162–3, 669
- exponential distribution, 140, 584–6
- for generalized (Cox-Snell) residual, 631
- exponential family density, 427
- conjugate prior for, 427–8
 - see also* linear exponential family
- exponential-gamma regression model, 616, 633–4
- exponential-IG regression model, 634
- exponential regression model
- application with censored data, 606–8, 633
 - example with uncensored data, 159–63
- extreme value distribution. *See* type 1 extreme value
- extremum estimator, 124–39
- asymptotic distribution, 127–31
 - consistency, 125–7
 - definition, 125
 - formal proofs, 130–2
 - informal approach, 132–3
 - statistical inference, 135–9
 - variance matrix estimation, 137–9
- factor analysis, 650
- factor loadings, 517, 650–1, 689
- factor model, 517, 648, 686
- Fairlee-Gumble-Morgenstern copula, 654
- fast simulated annealing (FSA) method, 347–8
- FD estimator. *See* first-differences
- FE estimator. *See* fixed effects
- feasible generalized least squares (FGLS) estimator, 81–3
- asymptotic distribution, 82
 - definition, 82
 - example, 84–5
 - in fixed effects model, 729
 - in mixed linear model, 775
 - nonlinear, 155–8
 - in pooled model, 720–1

- feasible generalized least squares (*cont.*)
 in random effects model, 705, 734–6, 738, 837–9,
 849–51
 as sequential two-step m-estimator, 201
 systems FGLS, 208–9
- feasible generalized nonlinear least squares (FGNLS)
 estimator, 155–8
 asymptotic distribution, 156
 definition, 156
 example, 159–63
 as optimal GMM estimator, 180–1
 systems FGNLS, 217
- FGLS estimator. *See* feasible generalized least squares
 FGNLS estimator. *See* feasible generalized nonlinear
 least squares
- FIML estimator. *See* full information maximum
 likelihood
- finite mixture models, 621–5
 counts, 678–9
 definition, 622
 EM algorithm, 623–5
 latent class interpretation, 623
 number of components, 624–5
 panel data, 786
see also mixture models
- finite-sample bias
 of GMM estimator, 177
 of IV estimator, 108–12
 of tests, 250–4, 262
- finite-sample correction term
 for sampling without replacement, 817
- first-differences (FD) estimator, 704–5, 729–31
 application, 710–11, 714
 asymptotic distribution, 730–1
 compared to FE estimator, 731
 consistency, 730, 764
 definition, 704–5, 730
 IV estimator, 758
- first-differences (FD) model, 704, 729–31, 758
- first-differences (FD) transformation, 783–4
- fixed effects (FE) estimator, 704, 726–9, 756–9,
 781–5, 791–2
 application, 710–3, 792–5
 asymptotic distribution, 727–9
 binary outcome models, 796–9
 clustered data, 839–41
 compared to DID estimator, 768
 compared to FD estimator, 729
 as conditional ML estimator, 732
 consistency, 727, 764, 781–2, 784–5
 count models, 802–8
 definition, 704, 726, 781–4
 duration models, 802
 dynamic models, 764–6, 791–2, 797–9, 806–7
 as FGLS estimator, 729
 Hausman test for, 717–9
 identification, 702
 incidental parameters, 704, 726
 inconsistency, 764, 781–2, 784–5
 IV estimators, 758
 as LSDV estimator, 733
 multinomial outcome models, 798
 selection models, 801
 Tobit model, 800
 versus random effects, 701–2, 715–9, 788
- fixed effects (FE) model, 704, 726–33, 756–9, 781–5,
 791–2
 cohort-level, 772
 clustered data, 831, 843
 definition, 700, 726
 dynamic models, 764–6, 791–2, 797–9, 806–7
 endogenous regressors, 756–9
 identification, 702
 incidental parameters, 704, 726
 marginal effects, 702
 nonlinear models, 781–5, 796–808, 791
 time-varying regressors, 702
 versus random effects, 701–2, 715–9, 788
see also fixed effects estimators
- fixed coefficient, 846
- fixed design. *See* fixed in repeated samples
- fixed in repeated samples, 76–7
 bootstrap sampling method, 360
 in kernel regression, 312
 Liapounov CLT, 951
 Markov LLN, 948
 Monte Carlo sampling method, 251
- fixed regressors. *See* fixed in repeated samples
- flexible parametric models
 count models, 674–5
 hazard models, 592
 selection models, 563
 flow sampling, 44, 626
 forward orthogonal deviations IV estimator, 759
 forward orthogonal deviations model, 759
 forward recurrence time, 626
 Fourier flexible functional form, 321
 frailty, 612, 662
see also unobserved heterogeneity
- Frank copula, 654
- Frechet bounds, 653–4
- frequentist approach, 421–2, 424, 439–40
- FSA method. *See* fast simulated annealing
- full conditional distributions, 431
see also Gibbs sampler
- full information maximum likelihood (FIML)
 estimator, 214
 nested logit model, 510–2
 nonlinear models, 219
- functional approach
 to measurement error, 901
- functional form misspecification, 91–2
 diagnostics for, 272–3, 277–8
- gamma distribution, 585–6, 614
- gamma function, 586

- Gaussian quadrature, 389–90, 393, 809
 Gauss-Hermite quadrature, 389–90
 Gauss-Laguerre quadrature, 389–90
 Gauss-Legendre quadrature, 389–90
 Gauss-Newton (GN) algorithm, 345
 example, 348
 GEE estimator. *See* generalized estimating equations
 general to specific tests, 285
 generalized additive model, 323, 327
 generalized cross-validation, 315
 generalized estimating equations (GEE) estimator, 790, 794, 804, 809
 generalized extreme value (GEV) distribution, 508
 see also nested logit model
 generalized information matrix equality, 142, 145, 264
 generalized inverse, 261
 generalized IV estimator, 187
 generalized least squares (GLS) estimator, 81–5
 asymptotic distribution, 82
 definition, 82
 as efficient GMM, 179
 example, 84–5
 nonlinear, 155–8
 generalized linear models (GLMs), 149–50, 155
 count data, 683
 conditional ML estimator, 783
 GEE estimator, 791
 quasi-ML estimator, 149–50
 see also LEF models
 generalized method of moments (GMM) estimator, 166–222
 asymptotic distribution, 173–4, 182–3
 based on additional moment restrictions, 169, 178–9
 based on moment conditions from economic theory, 171
 based on optimal conditional moment, 179–80
 bootstrap for, 379–80
 computation, 339
 definition, 173
 endogenous counts, 683–4, 687–9
 with endogenous stratification, 827
 with exogenous stratification, 823–4
 examples, 167–71, 178–9
 finite-sample bias, 177
 identification, 173, 182
 linear IV, 183–92
 linear systems, 211–2
 nonlinear IV, 192–9
 one-step GMM estimator, 187, 196, 746, 755
 optimal GMM, 176
 optimal moment condition, 179–81, 188
 optimal weighting matrix, 175–6
 panel data, 744–66, 789–90, 792
 practical considerations, 219–20
 test based on, 245
 two-step, 176, 187, 746, 755
 variance matrix estimation, 174–5
 weak instruments, 177–8
 see also panel GMM estimator
 generalized nonlinear least squares (GNLS) estimator.
 See feasible generalized nonlinear least squares
 generalized partially linear model, 323
 generalized random utility models, 515–6
 generalized residual, 289–90
 in duration models, 631
 in LM test, 239–40
 plots of, 633–6
 generalized Tobit model, 548
 generalized Weibull distribution, 584–6
 genetic algorithms, 341
 GEV distribution. *See* generalized extreme value
 Geweke, Hajivassiliou, Keane (GHK) simulator, 407–8
 for MNP model, 518
 GHK simulator. *See* Geweke, Hajivassiliou, Keane simulator
 Gibbs sampler, 448–50
 data augmentation, 454–5, 933
 example, 452–4
 in latent variable models, 514, 519, 563
 see also Markov chain Monte Carlo
 GLMs. *See* generalized linear models
 GLS estimator. *See* generalized least squares
 GMM estimator. *See* generalized method of moments
 GN algorithm. *See* Gauss-Newton
 GNLS estimator. *See* feasible generalized nonlinear least squares
 Gompertz distribution, 585–6
 Gompertz regression model, 606–8
 gradient methods, 337–48
 see also iterative methods
 Granger causality, 22
 grid search methods, 337, 351
 grouped data. *See* aggregated data
 Halton sequences, 409–10
 Hausman test, 271–4
 applications, 719, 850–1
 asymptotic distribution, 272
 auxiliary regressions, 273
 bootstrap, 378
 computation, 272–3, 378, 717–9
 definition, 271–2
 for endogeneity, 271–2, 275–6
 for fixed effects, 717–9, 737, 788, 839
 for multinomial logit model, 503
 power, 273–4
 robust versions, 273, 378, 718–9
 Hausman-Taylor IV estimator, 761
 Hausman-Taylor model, 760–2
 Hawthorne effect, 53
 hazard function
 baseline in PH model, 591
 cumulative hazard, 577–8, 582–4
 definition, 576, 578

- hazard function (*cont.*)
 in mixture models, 616–8
 multivariate, 649
 nonparametric estimator, 581, 583
 parametric examples, 585
 piecewise constant, 591
 see also duration models
- Health and Retirement Study (HRS), 58
- Heckit estimator. *See* Heckman two-step estimator
- Heckman two-step estimator
 application, 554
 in Roy model, 556
 in selection model, 550–1
 semiparametric estimator, 565–6
 in Tobit model, 543, 567–8
- Hessian matrix
 estimate, 137
 Newton-Raphson algorithm, 341–2
 singular, 350–1
- heterogeneous treatment effects, 882, 885–7
 IV estimator, 886–7
 LATE estimator, 885
 RD design, 882
- heterogeneity
 within-cell, 480
 see also unobserved heterogeneity
- heteroskedastic errors
 adaptive estimation, 323, 328
 conditional heteroskedasticity, 78
 definition, 78
 in GLMs, 149–50
 in linear model, 84–5, 94–5
 multiplicative, 84–5, 86–7
 in nonlinear model, 157–63
 residuals, 289–90
 tests for, 241, 267, 275
 Tobit MLE inconsistency, 538
 working matrix for, 82–3, 156–8
- heteroskedasticity-robust standard errors
 bootstrap, 379–80
 clustered data, 834
 example, 84–5
 for extremum estimator, 137, 164
 intuition, 81
 for NLS estimator, 155, 164
 for OLS estimator, 74–5, 80–1, 112
 panel data, 705
 for WLS estimator, 83
 see also robust standard errors
- hierarchical linear models (HLMs), 845–8
 Bayesian analysis, 847
 clustered data, 845
 coefficient types, 846–7
 individual-specific effects, 848
 mixed linear models, 774–6, 847
 panel data, 847–8
 random coefficients model, 847
 two-level model, 846
- hierarchical models, 429
 Bayesian analysis, 441–2, 447, 450, 514
 see also hierarchical linear models
- histogram, 298
 see also kernel density estimator
- HLM. *See* hierarchical linear model
- hot deck imputation, 929, 940
- HRS. *See* Health and Retirement Study
- Huber-White robust standard errors, 137, 144, 146
 see also robust standard errors
- hurdle model, 680–1, 690
 see also two-part model
- hyperparameters, 428, 847
- hypothesis tests, 223–58
 based on extremum estimator, 224–33
 based on ML estimator, 233–43
 based on GMM estimator, 245
 based on m-estimator, 244
 bootstrap, 254–6, 363–8, 372–3, 378–9
 for common misspecifications, 274–7, 670–1
 examples, 236, 241–3, 252–4, 254–6, 372–3
 induced test, 230
 joint versus separate, 230–1, 285, 629–30
 power, 247–50, 253–4
 size, 246–7, 251–3
 see also LM tests; LR test; Wald tests, m-tests
- identification
 in additive random utility models, 504
 in binary outcome models, 476, 483
 bounds identification, 29
 definitions, 29–31
 in fixed effects model, 702
 of GMM estimator, 173, 182
 just identification, 31, 214
 in linear regression model, 71–2
 in measurement error models, 905–14
 in mixture models, 618–20
 in multinomial probit model, 517
 in natural experiments, 57–8
 observational equivalence, 29
 order condition, 31, 213
 over identification, 31, 214
 rank condition, 31
 in sample selection model, 551, 565, 566
 set identification, 29
 in simultaneous equations model, 29–31, 213–4
 in single-index models, 325
 and singular Hessian, 351
 weak identification, 100
 see also identification strategies
- identification strategies, 36–7
 control function approach, 37
 exogenization, 36
 incidental parameter elimination, 36–7
 instrumental variables, 37
 matching, 37
 reweighting, 37

- identified reduced form, 36
 IG distribution. *See* inverse-Gaussian
 ignorable missingness, 927
 estimator consistency if MCAR, 927
 estimator inconsistency if MAR only, 927
 problems if nonignorable, 940
 weak exogeneity, 927
 ignorability assumption, 863
 see also conditional independence assumption
 importance sampling, 407–8, 443–5, 518
 accelerated, 409
 GHK simulator, 407–8
 importance sampling density, 444
 importance sampling estimator, 444
 importance weight, 445
 target density, 444
 imputation methods, 928–39
 data augmentation, 454–5, 932–4
 example, 936–8
 hot deck imputation, 929
 listwise deletion, 928
 mean imputation, 928–9
 multiple imputation, 934–5
 pairwise deletion, 928
 regression-based imputation, 930–2
 imputation (I) step, 455, 932
 IM test. *See* information matrix test
 IMSE. *See* integrated mean squared error
 incidental parameters, 36
 clustered data FE model, 832, 840, 844
 panel data FE model, 704, 726, 781–2, 805
 inclusive value, 510–1
 incomplete gamma function, 586
 incomplete panels. *See* unbalanced panels
 independence of irrelevant alternatives, 503, 505, 527
 independent variables. *See* regressors
 independently-weighted IV estimator, 192
 independently-weighted optimal GMM estimator, 177
 index function model
 binary outcome model, 475–6, 482–3
 bivariate probit model, 522–3
 ordered multinomial model, 519–20
 Tobit model, 536
 see also single-index model
 indicator function, 298
 indirect inference, 404–5
 individual-specific effects model
 additive, 780
 binary outcome models, 795–6
 cluster-specific effects, 830
 count models, 802–3
 definitions, 700, 780
 duration models, 802
 multiplicative, 780, 793
 one-way, 700
 parametric, 780
 selection models, 801
 single-index, 780
 Tobit models, 800–1
 two-way, 738
 see also FE models; RE models
 induced test, 230
 information criteria, 278–9, 283–4
 Akaike, 278–9, 284, 624
 Bayesian, 278, 284
 consistent Akaike, 278
 Kullback-Liebler, 147, 169, 278, 280
 Schwarz, 278, 284
 information matrix, 142
 block-diagonal, 144, 240, 329
 information matrix equality, 141–2, 145
 generalized, 142, 145
 see also BHHH estimate; OPG version
 information matrix (IM) test, 265–6
 bootstrap, 378
 computation, 261–2, 378
 definition, 265
 example, 270
 power, 267
 instrumental variables (IV) estimator
 alternative estimators, 190–2
 application, 110–2
 definition, 100–1
 example, 102–3
 finite-sample bias, 108–12, 191–2, 196
 identification, 100, 105–7
 independently-weighted IV estimator, 192
 jackknife IV estimator, 192
 LIML estimator, 191, 214
 in linear model, 98–112, 183–92, 211–2
 linear IV as GMM estimator, 170, 186
 local average treatment effects estimator, 883–9
 in measurement error models, 908–10, 912–3
 in natural experiments, 54–5
 in nonlinear models, 192–9
 in panel models, 764–5, 757–61
 quantile regression, 190
 in selection models, 559
 split-sample estimator, 191–2
 systems IV estimator, 211–2, 218–9
 in treatment effects models, 883–9
 two-stage IV estimator, 102, 187
 two-stage least squares estimator, 101–2, 187–91
 Wald estimator, 98–9
 see also GMM estimator; panel GMM estimator
 instruments
 definition, 96–7, 100
 examples, 97–8
 by exclusion restriction, 106
 by functional form restriction, 106
 invalid, 100, 105–7
 optimal, 180
 for panel data, 750–1, 754–6
 relevance, 108
 weak, 100, 104–12, 177–8, 191–2, 196, 751–2, 756
 see also instrumental variables estimator

- integrated hazard function. *See* cumulative hazard function
- integrated mean squared error (IMSE), 303
- integrated squared error (ISE), 302, 314
- interval data models
- definition, 532–3, 579
 - ML estimator, 534–5
- interruption bias, 626
- intraclass correlation, 816, 831, 835–8
- inverse-Gaussian (IG) distribution, 614–5, 677
- inverse law of probability, 421
- inverse-Mills ratio, 540–1, 553–4
- inverse transformation method, 409, 412–3
- inverse-Wishart distribution, 443, 453, 514
- irrelevant regressors, 93
- ISE. *See* integrated squared error
- iterated bootstrap, 374
- iterative methods, 337–48
- BFGS, 344
 - BHHH, 343–4
 - convergence criteria, 339–40
 - DFP, 344, 350–1
 - expectation maximization, 345–7, 623–5, 930–2
 - fast simulated annealing, 347–8
 - Gauss-Newton, 345, 348
 - line search, 338
 - Newton-Raphson, 338–9, 341–3, 348
 - numerical derivatives, 340
 - simulated annealing, 347
 - starting values, 340, 351
 - step size adjustment, 338
- IV estimator. *See* instrumental variables
- jackknife, 374–6
- bias estimate, 375
 - bias-corrected estimator, 375
 - example, 376
 - IV estimator, 192
 - standard error estimate, 375, 855
- Jensen's inequality, 956
- jittered data, 290
- joint duration distributions, 648–55
- copulas, 651–5
 - mixtures, 650–1
 - multivariate hazard function, 649
 - multivariate survivor function, 649–50
- joint limits, 767
- joint versus separate tests, 230–1, 285, 629–30
- just identification, 31, 100, 173
- Kaplan-Meier (KM) estimator, 581–3
- application, 575, 583, 604–5
 - for baseline hazard, 596–7
 - confidence bands for, 583
 - definition, 581
 - tied data, 582
- kernel density estimator, 298–306
- alternatives to, 306
- application, 296–7, 300
- asymptotic distribution, 301–2, 330–1
- bandwidth choice, 302–4
- bias, 301, 330–1
- confidence interval for, 305
- consistency, 300
- convergence rate, 302
- definition, 299
- derivative estimator, 305
- examples, 252–3, 367–8
- multivariate, 305–6
- Nadaraya-Watson kernel regression estimator, 312
- optimal bandwidth, 303
- optimal kernel, 303
- variance, 301, 331
- kernel functions, 299–300
- comparison, 300
 - definition, 299
 - higher-order, 299, 306, 313
 - leading examples, 300
 - optimal for density estimation, 303
 - properties, 299
- kernel matching, 875, 895–6
- kernel regression estimator, 311–9
- alternatives to, 319–22
 - asymptotic distribution, 313, 331–3
 - bandwidth choice, 314–6
 - bias, 313, 331–2
 - bootstrap confidence interval for, 380–1
 - boundary problems, 309, 320–1
 - conditional moment estimator, 317–8
 - confidence interval for, 316
 - consistency, 313
 - convergence rate, 314
 - definition, 312
 - derivative estimator, 317
 - introduction to nonparametric regression, 307–11
 - multivariate, 318–9
 - optimal bandwidth, 314
 - optimal kernel, 314
 - undersmoothing, 380
 - variance, 301, 331
- see also* nonparametric regression
- Khinchine's theorem, 948
- KLIC. *See* Kullback-Liebler information criterion
- KM estimator. *See* Kaplan-Meier
- k-NN estimator. *See* nearest neighbors estimator
- Kolmogorov LLN, 80, 111, 947
- Kolmogorov test, 267
- Kullback-Liebler information criterion (KLIC), 147, 169, 278, 280
- LAD estimator. *See* least absolute deviations
- Lagrange multiplier (LM) test
- asymptotic distribution, 235, 237–8
 - based on GMM-estimator, 245
 - based on m-estimator, 244
 - bootstrap, 379

- comparison with LR and Wald tests, 238–9
 computation, 239–41, 256, 274
 definition, 234–5
 examples, 236, 241–3
 for heteroskedasticity, 241, 267, 275
 in duration models, 632
 interpretation, 239–40
 for omitted variables, 274
 OPG version, 240–1
 for random effects, 737, 841
 score test, 234–5
 in Tobit model, 544
 for unobserved heterogeneity, 630, 636
see also hypothesis tests
- Laplace approximation, 390
 Laplace distribution, 178, 541
 Laplace transform, 577
 LATE estimator. *See* local average treatment effects
 latent class model, 622
see finite mixture models
 latent variable, 475, 532
 latent variable models
 additive random utility model, 476–8, 504–7
 binary outcomes, 475–8
 endogenous, 560–1
 ordered multinomial model, 519–20
see also censored models; truncated models
 law of iterated expectations, 955
 law of large numbers (LLN), 947–8
 definition, 947
 examples of use, 80, 129
 Khinchine's theorem, 948
 Kolmogorov LLN, 947
 Markov LLN, 948
 sampling schemes, 131, 948
 strong law, 947
 weak law, 947
- least absolute deviations (LAD) estimator
 application, 88–90
 asymptotic distribution, 88
 binary outcome models, 484
 bootstrap, 381
 censored LAD, 564–5, 808
 definition, 87
 two-stage LAD, 190
see also quantile regression
- least-squares dummy variable (LSDV) estimator, 704, 732–3, 840
 least-squares dummy variable (LSDV) model, 704, 732, 840
- least squares (LS) estimators
 clustered data, 833–7
 feasible generalized LS, 81–3, 155–8
 generalized LS, 81–5, 155–8
 linear, 70–85
 nonlinear LS, 150–9
 ordinary LS, 70–81
 panel data, 211, 702–3, 720–5
- systems of equations, 207–8, 211, 217
see also FGOLS; FGNLS; OLS; NLS
 leave-one-out estimate, 192, 304, 315, 375
 LEF. *See* linear exponential family
 length-biased sampling, 43–4, 626
 Liapounov CLT, 80, 131, 950
 likelihood-based hypothesis tests, 233–43
 comparisons of, 235–6, 238–9
 definitions, 234–5
 examples, 236–7, 241–3
see also LM tests; LR tests; Wald tests
- likelihood function, 139–41
 conditional likelihood function, 139, 731–2, 824
 definition, 139
 joint, 19, 824–7
 leading examples, 140–1
 marginal, 432, 595
 partial, 594–6
- likelihood principle, 139, 420, 433
- likelihood ratio (LR) test
 asymptotic distribution, 235, 237
 based on GMM-estimator, 245
 based on m-estimator, 244
 comparison with LM and Wald tests, 238–9
 definition, 234
 examples, 236, 241–3
 nonnested models, 279–83
 quasi-LR test statistic, 244
 uniformly most powerful test, 237
see also hypothesis tests
- LIML estimator. *See* limited information maximum likelihood
- limit distribution, 948
see also asymptotic distribution
- limit variance matrix, 952–3
 definition, 952
 replacement by consistent estimate, 952
 sandwich form, 953
- limited information maximum likelihood (LIML)
 estimator, 191, 214
- Lindeberg-Levy CLT, 80, 131, 950
- line search, 338
- linear exponential family (LEF) models, 147–9
 conjugate priors, 427–8
 conditional ML estimator, 782
 consistency, 148
 leading examples, 148
 pseudo-R², 288
 residuals, 289–90
 tests based on, 240, 268, 274–5
see also generalized linear models
- linear panel estimators, 695–778
 application, 708–15, 725
 Arellano-Bond estimator, 764–5
 between estimator, 703
 covariance estimator, 733
 conditional ML estimator, 731–2
 differences-in-differences estimator, 768–70

- linear panel estimators (*cont.*)
 error components 2SLS estimator, 760
 error components 3SLS estimator, 762
 first differences estimator, 704–5, 729–31
 first differences IV estimator, 758
 fixed effects estimator, 704, 726–9
 fixed effects IV estimators, 757–9
 forward orthogonal deviations IV estimator, 759
 Hausman-Taylor IV estimator, 761
 LSDV estimator, 704, 732–3
 MD estimator, 753, 76–7
 panel bootstrap, 708, 377–8, 708, 746, 751
 panel GMM estimators, 744–68
 panel-robust inference, 705–8, 722, 745–6, 751
 pooled OLS estimator, 702–3, 720–5
 random effects estimator, 705, 734–6
 random effects IV estimator, 759–60
 within estimator, 704, 726–9
 within IV estimator, 758
- linear panel models, 695–778
 analysis-of-covariance model, 733
 application, 708–15, 725
 between model, 702
 dynamic models, 763–8
 endogenous regressors, 744–63
 first differences model, 704, 730, 758
 fixed effects model, 700–2, 726–34, 757–9
 fixed versus random effects, 701–2, 715–9
 forward orthogonal deviations model, 759
 Hausman-Taylor model, 760–2
 incidental parameters problem, 704, 726
 individual dummies, 699
 individual-specific effects model, 700
 LSDV model, 704, 732
 minimum distance estimator, 753, 766–7
 mean-differenced model, 758
 measurement error, 739, 905
 mixed linear models, 774–6
 pooled model, 699, 720–5
 random effects differenced model, 760–1
 random effects model, 700–2, 734–6, 759–60
 residual analysis, 714–5
 strong exogeneity, 700, 749–50, 752
 time dummies, 699
 time-invariant regressors, 702, 749–51
 time-varying regressors, 702, 749–51
 two-way effects model, 738
 unbalanced data, 739
 weak exogeneity, 749, 752, 758
 within model, 704, 758
see also linear panel estimators
- linear probability model, 466–7
 linear programming methods, 341
 linear regression model
 definition, 16–17, 70–1
 linear systems of equations, 207–14
 panel data models as, 211
 seemingly unrelated regressions, 209–10
- simultaneous equations, 22–31, 213–4
 systems FGLS estimator, 208
 systems GLS estimator, 208
 systems GMM estimator, 208
 systems ML estimator, 214
 systems OLS estimator, 211
 systems 2SLS estimator, 212
 linearization method, 855
 link function, 149, 469, 783
 listwise deletion, 60, 928
 consistency under MCAR, 928
 example, 936–8
 inconsistency under MAR only, 928
 Living Standards Measurement Study (LSMS), 59, 88–90, 848–53
 LLN. *See* law of large numbers
 LM test. *See* Lagrange multiplier test
 local alternative hypotheses, 238, 247–8, 254
 local average treatment effects (LATE) estimator, 883–9
 assumptions, 884–5
 comparison with IV estimator, 885
 definition, 884
 heterogeneous treatment effect, 885
 monotonicity assumption, 885
 selection on unobservables, 883
 Wald estimator, 886
see also ATE; ATET; MTE
 local linear regression estimator, 320–1, 333
 local polynomial regression estimator, 320–1
 local running average estimator, 308, 320
 local weighted average estimator, 307–8
 logistic distribution, 476–7
 logistic regression. *See* logit model
 logit model, 469–70
 application, 464–5
 as ARUM, 477, 486–7
 clustered data, 844
 definition, 469
 for discrete-time duration data, 602
 GLM, 149
 imputation example, 937–9
 index function model, 476
 marginal effects, 470
 measurement error example, 919
 ML estimator, 468–9
 multinomial logit, 494–5, 500–3, 525
 nested logit, 509–12, 526–7
 ordered logit, 520
 panel data, 795–9
 probit model comparison, 471–3
 random parameters logit, 512–6
see also binary outcome models
 log-likelihood function. *See* likelihood function
 length-biased sampling, 43–4
 log-logistic distribution, 585–6, 592
 log-normal distribution, 585–6, 592
 log-normal model, 533, 545–6

- log-odds ratio, 470, 472
 log-sum, 510
 log-Weibull distribution. *See* type 1 extreme value
 long panel, 723–5, 767
 longitudinal data. *See* panel data
 loss function, 66–69
 - absolute error, 67
 - asymmetric expected error, 67
 - Bayesian decision analysis, 434–5
 - expected, 66
 - KLIC, 68, 147, 168, 278–9
 - squared error, 67–9, 156
 - step, 67–8
 Lowess regression estimator, 320–1
 - application, 297, 309–10, 712–5
 LR test. *See* likelihood ratio test
 LS estimators. *See* least squares
 LSDV. *See* least-squares dummy variable
 LSMS. *See* Living Standards Measurement Study
- MAR. *See* missing at random
 marginal analysis of panel data, 717, 787
 marginal effects, 122–4
 - in binary outcome models, 466–5, 467, 470–1
 - calculus method, 123
 - computing, 122–4
 - definition, 122
 - example, 162–3
 - finite-difference method, 123
 - in fixed effects model, 702, 788
 - in multinomial models, 493–4, 501–3, 519–23, 525
 - population-weighted, 821
 - in sample selection models, 552
 - in single-index models, 123
 - in Tobit model, 541–2
 - see also* coefficient interpretation
 marginal likelihood, 432, 595
 marginal treatment effects (MTE) estimator, 886
 market-level data, 482, 513
 Markov chain Monte Carlo (MCMC) methods, 445–54
 - convergence, 449, 458
 - in data augmentation, 933
 - examples, 452–4, 512, 687, 936–9
 - Gibbs sampler, 448–50, 514, 519, 563
 - Metropolis algorithm, 450–1
 - Metropolis-Hastings algorithm, 451–2, 512
 Markov LLN, 77, 131, 948
 Marshall-Olkin method, 649–51, 686
 matching assumption, 864
 - see also* overlap assumption
 matching estimators, 871–8, 889–96
 - application, 889–96
 - assumptions, 863–5
 - ATE matching estimator, 877
 - ATET matching estimator, 874, 877, 894–6
 - balancing condition, 893
 - caliper matching, 874
 counterfactuals, 871
 exact matching, 872, 891
 inexact matching, 873
 interval matching, 875–6
 kernel matching, 875, 895–6
 nearest-neighbor matching, 875, 894–6
 propensity score matching, 873–8, 892
 radius matching, 876, 895–6
 selection on observables only, 871
 stratification matching, 875–6, 893–6
 variance computation, 877–8, 895
- maximum empirical likelihood (MEL) estimator, 206
 maximum likelihood (ML) estimator, 139–46
 - asymptotic distribution, 142–3
 - conditional ML estimator, 731–2, 782–3, 796–9
 - consistency, 142, 824
 - definition, 141
 - endogenous stratification, 824–7
 - example, 143–4
 - exogenous stratification, 824
 - MSL estimator, 393–8
 - quasi-ML estimator, 146–50
 - regularity conditions, 141, 145–6
 - restricted, 233
 - unrestricted, 233
 - variance matrix estimation, 144
 - weighted ML estimator, 828
 - see also* quasi-ML estimator
 maximum rank correlation estimator, 485
 maximum score estimator, 341, 381, 483–4, 800
 maximum simulated likelihood (MSL) estimator, 393–8
 - asymptotic distribution, 394–5
 - bias-adjusted MSL, 396–7
 - compared to MSM, 402–3
 - count model examples, 677–8, 687, 689
 - definition, 394
 - example, 397–8
 - multinomial probit model, 518
 - number of simulations, 396
 - random parameters logit model, 522
 MCAR. *See* missing completely at random
 MD estimator. *See* minimum distance estimator
 mean-differenced estimator, 783, 805–6
 mean-differenced model, 758, 783
 mean imputation, 928, 936–8
 mean integrated squared error (MISE), 303, 314
 mean-scaling estimator, 783, 805–6
 mean-square convergence, 946
 mean substitution. *See* mean imputation
 measurement error
 - in cohort-level data, 772–3
 - in dependent variable, 913–4
 - in microdata, 46, 60
 - in panel data, 739, 905
 - in regressors, 899–922
 - see also* measurement error model estimators; measurement error models

- measurement error model estimators, 899–922
 attenuation bias, 903–5, 911, 915, 919–20
 bounds identification, 906–8
 corrected score estimator, 916–8
 IV estimator, 908–10, 912–3
 linear models, 900–11
 nonlinear models, 911–20
 OLS estimator inconsistency, 902–4
 using additional moment restrictions, 909–10
 using instruments, 908–9
 using known measurement error variance, 902–3,
 910
 using replicated data, 910–1, 913
 using validation sample, 911
- measurement error models, 899–922
 attenuation bias, 903–5, 911, 915, 919–20
 classical measurement error model, 901–2
 dependent variable measured with error, 913–4
 examples, 919–20
 identification, 905–14
 linear models, 900–11
 multiple regressors, 904
 nonclassical measurement error, 904, 920
 nonlinear models, 911–20
 panel models, 905
 scalar regressor, 903
 serial correlation, 909
 variance inflation, 904, 916
see also measurement error model estimators
- median regression. *See* LAD estimator
- MEL. *See* maximum empirical likelihood
- m-estimator, 118–22
 asymptotic distribution, 120
 clustered data, 842–3
 definition, 118–9
 sequential two-step, 200–2
 simulated m-estimator, 398–9
 tests based on, 244, 263–4
 weighted m-estimator, 829, 856
see also extremum estimators
- method of moments (MM) estimator
 asymptotic distribution, 134, 174
 definition, 172
 examples, 167
see also estimating equations estimator; GMM estimator
- method of scoring, 343, 348
- method of simulated moments (MSM) estimator, 399–404
 asymptotic distribution, 400–2
 compared to MSL, 402–3
 definition, 400
 example, 403
 MNP model, 497, 518
 number of simulations, 399
- method of simulated scores (MSS) estimator
 for MNP model, 519
- method of steepest ascent, 344
- Metropolis algorithm, 450–1
- Metropolis-Hastings algorithm, 451–2, 512
- microdata sets, 58–61
 handling, 59–61
 leading examples, 58–9
- microeconomics overview, 1–17
- midpoint rule, 388, 391–2
- minimum chi-square estimator, 203
see also Berkson's minimum chi-square estimator
- minimum distance (MD) estimator, 202–3, 753, 766–7
 asymptotic distribution, 202
 bootstrap for, 379–80
 covariance structures, 766–7
 definition, 202
 equally-weighted, 202
 generalized, 222
 indirect inference, 404–5
 OIR test, 203
 optimal, 202, 753
 panel data, 753, 766–7
 relation to GMM, 203, 753
- misclassification, 914
- MISE. *See* mean integrated squared error
- missing at random (MAR), 926–7
 definition, 926
 and ignorable missingness, 927, 932
 relation to MCAR, 927
- missing completely at random (MCAR), 926–7
 definition, 927
 and ignorable missingness, 927
 relation to MCAR, 927
- missing data, 923–41
 deletion methods, 928
 examples, 924
 ignorable assumption, 927
 imputation with models, 929–41
 imputation without models, 928–9
 MAR assumption, 926–7
 MCAR assumption, 927
 nonignorable missingness, 927, 940
see also imputation methods
- misspecification tests. *See* specification tests
- mixed estimator, 439–41
- mixed linear model, 774–6
 Bayesian methods, 775
 FGLS estimator, 775
 fixed parameters, 774
 ML estimator, 776
 random parameters, 774
 restricted ML estimator, 776
 nonstationary panel data, 767–8
 prediction, 776
see also hierarchical linear model
- mixed logit model, 500–3
 example, 495
 definition, 500
see also RPL model

- mixed proportional hazards (MPH) model, 611–25
Weibull-gamma mixture, 615
see also mixture models
- mixture hazard function, 616–8
- mixture models, 611–25
 application, 623–6
 counts, 675–9
 durations, 611–25
 identification, 618–20
MSL estimator, 393–8, 687
 multinomial outcomes, 515–6
 multiplicative heterogeneity, 613
 specification tests, 628–32
see also finite mixture models; unobserved heterogeneity
- ML estimator. *See* maximum likelihood
- MM estimator. *See* method of moments
- MNL estimator. *See* multinomial logit
- MNP estimator. *See* multinomial probit
- model diagnostics, 287–91
 binary outcome models, 473–4
 duration models, 628–32
 example, 290–1
 multinomial outcome models, 499
 pseudo-R² measures, 287–9, 291
 residual analysis, 289–91
see also model selection methods
- model misspecification, 90–4
see also endogeneity; functional form
 misspecification; heterogeneity; omitted values;
 pseudo-true value
- model selection methods
 Bayesian, 456–8
 nested models, 278–81
 nonnested models, 278–84
 order of testing, 285
see also model diagnostics; specification tests
- moment-based simulation estimators, 398–404
see also MSL estimator; MSM estimator
- moment-based tests. *See* m-tests
- moment matching. *See* indirect inference
- Monte Carlo integration, 391–2
 direct, 391
 example, 392
 importance sampling, 407, 443–5
 simulators, 393–4, 406–10
see also quadrature
- Monte Carlo studies, 250–4
 example, 251–4
- moving average estimator, 308
- moving blocks bootstrap, 373, 381
- MPH model. *See* mixed proportional hazards
- MSL estimator. *See* maximum simulated likelihood
- MSM estimator. *See* method of simulated moments
- MSS estimator. *See* method of simulated scores
- MTE. *See* marginal treatment effects
- m-tests, 260–71
 asymptotic distribution, 260, 263
 auxiliary regressions, 261–3
 bootstrap, 261, 379
 chi-square goodness of fit, 266–7, 270–1, 474
 conditional moment test, 264–5, 267–9, 319
 CM test interpretation, 268
 computation, 261–3
 definition, 260
Hausman test, 271–4, 717–9
 information matrix tests, 265–6, 270
 outer-product-of-the-gradient form, 262
 overidentifying restrictions test, 181, 183, 267, 747
 power, 268
 rank, 261
- multicollinearity, 350–1
 in multinomial probit model, 517
 in panel model, 752
 in sample selection model, 542, 551
- multilevel models. *See* hierarchical models
- multinomial logit (MNL) model, 500–3, 525
 application, 494–5
 as additive random utility model, 505
 definition, 500
 marginal effects, 494, 501–3, 525
 ML estimator, 501
 panel data, 798
see also multinomial outcome models
- multinomial outcome models, 490–528
 application, 491–5
 alternative-invariant regressors, 498
 alternative-varying regressors, 497
 conditional logit, 500–3, 524–5
 definition, 496–7
 identification, 504
 index function model, 519–20
 marginal effects, 501–3, 524–5
 mixed logit, 500–3
 ML estimator, 496, 501
 multinomial logit, 500–3, 525
 multinomial probit, 516–9
 ordered models, 519–20
 OLS estimator, 471
 panel data, 798
 random parameters logit, 512–6
 random utility model, 504–7
 semiparametric estimation, 523–4
- multinomial probit (MNP) model, 516–9
 Bayesian Methods, 519
 definition, 516–7
 identification, 517
 ML estimator, 518
MSL estimator, 518
MSM estimator, 518
MSS estimator, 518
see also multinomial outcome models

- multiple duration spells, 655–8
 fixed effects, 656
 lagged duration dependence, 657
 ML estimator, 658
 random effects, 657
 recurrent spells, 655
 multiple imputation, 934–9
 estimator, 934
 examples, 935–9
 relative efficiency, 935
 variance of estimator, 934–5
 multiple treatments, 860
 multiplicative errors
 multistage surveys, 41–2, 814–6, 853–6
 variance estimation, 853
 multivariate data
 binary outcomes, 521–3
 counts, 685–7
 durations, 640–64
see also systems of equations
 multivariate-*t* distribution, 442
- NA estimator. *See* Nelson-Aalen
 National Longitudinal Survey (NLS), 58, 110–2
 National Longitudinal Survey of Youth (NLSY), 58–9
 National Supported Work (NSW) demonstration project, 889–95
 natural conjugate pair, 427–8
 natural experiments, 32, 54–8
 definition, 54
 differences-in-differences estimator, 55–7, 768–70, 878–9
 examples, 54
 exogenous variation, 54–5
 identification, 57–8
 instrumental variables, 54–5
 regression discontinuity design, 879–83
 ncp. *See* noncentrality parameter
 nearest neighbors (k-NN) estimator, 319–20
 definition, 319
 example, 308–9
 symmetrized, 308, 320
see also nonparametric regression
 nearest-neighbor matching, 875, 894–6
 negative binomial distribution, 675
 negative binomial model, 675–7
 application, 690
 bivariate, 215, 686–7
 hurdle model, 681
 ML estimator, 677
 MSL estimator, 677–8
 NB1 variant, 676
 NB2 variant, 676
 panel data, 804, 806
 negative hypergeometric distribution, 806
 neglected heterogeneity. *See* unobserved heterogeneity
 Nelson-Aalen (NA) estimator, 582–4
 application, 605–6, 662
 confidence bands for, 584
 definition, 582
 tied data, 582
 nested bootstrap, 374, 379
 nested logit model, 507–12, 526–7
 from ARUM, 526–7
 definition 510–1
 different versions of, 511–2
 example, 511
 GEV model, 508, 526
 ML estimator, 510
 sequential estimator, 510
 welfare analysis, 510
see also multinomial models
 nested models 278, 281
see also nonnested models
 neural network models, 322
 Newey-West robust standard errors, 137, 175, 723
 definition, 175
see also robust standard errors
 Newton-Raphson (NR) method, 341–3
 examples, 338–9, 348
 NLFIML estimator. *See* nonlinear full-information maximum likelihood
 NLS estimator. *See* nonlinear least squares
 NLSY. *See* National Longitudinal Survey of Youth
 NL2SLS estimator. *See* nonlinear two-stage least squares
 NL3SLS estimator. *See* nonlinear three-stage least squares
 noise-to-signal ratio, 903
 noncentral chi-square distribution, 248
 noncentrality parameter (ncp), 248
 nonclassical measurement error, 904, 920
 nongradient methods, 337, 341, 347–8
 nonignorable missingness, 927, 940
 attrition bias due to, 940
 selection bias due to, 927, 932, 940
 nonlinear estimators
 coefficient interpretation, 122–4
 extremum estimator
 m-estimator, 118–22
 GMM estimator, 166–222
 ML estimator, 139–46
 NLS estimator, 150–9
 overview, 117–22
 panel models, 779–810
 nonlinear full-information maximum likelihood (NLFIML) estimator, 219
 nonlinear GMM estimator, 192–9
 asymptotic distribution, 194–5
 definition, 194–5
 example, 197–8, 199, 688
 instrument choice, 196
 NL2SLS estimator, 196

- optimal, 195
 panel data, 789–90
 nonlinear in parameters, 27
 nonlinear in variables, 27
 nonlinear IV estimator. *See* nonlinear GMM
 nonlinear least squares (NLS) estimator, 150–9
 asymptotic distribution, 152–4
 consistency, 152–3
 definition, 151
 example, 155, 159–64
 time series, 158–9
 variance matrix estimation, 154–5
 nonlinear panel estimators, 779–810
 application, 792–5
 conditional ML estimator, 781–2, 805
 dummy variable estimator, 784–5, 800, 805
 first-differences estimator, 783–4
 fixed effects estimator, 783–5, 794, 796–802, 805–8
 GEE estimator, 790, 794, 804
 mean-differenced estimator, 783, 805–6
 mean-scaling estimator, 783, 805–6
 ML estimator, 785–6
 NLS estimator, 787, 794
 panel GMM estimator, 789–90
 panel-robust inference, 788–91
 quadrature, 785–6, 796, 800
 quasi-differenced estimator, 783–4
 quasi-ML estimator, 791
 random effects estimator, 785–6, 794–6, 800–1,
 803–4
 selection models, 801
 semiparametric, 808
 nonlinear panel models, 779–810
 application, 792–5
 binary outcome models, 795–6
 conditional mean models, 780–1
 count models, 792–5, 802–6
 dynamic models, 791–2, 797–9, 806–7
 endogenous regressors, 792
 exogeneity assumptions, 781
 finite mixture models, 786
 fixed effects models, 781–5, 791–2
 fixed versus random effects, 788
 incidental parameters problem, 781–2, 805
 individual-specific effects models, 780–1
 parametric models, 780, 782–3, 785–7, 792
 pooled models, 787, 794
 random effects models, 785–6, 792
 selection models, 801
 semiparametric, 808
 Tobit models, 800–1
 transition models, 801–2
 nonlinear regression model, 151
 additive error, 168, 193, 217
 nonadditive error, 168, 193, 218
 nonlinear systems of equations, 214–9
 additive errors, 217
 copulas, 651–5
 mixtures, 650–1
 ML estimator, 215–6
 NLFIML estimator, 219
 NL3SLS estimator, 219
 nonadditive errors, 217–8
 nonlinear panel model, 216
 nonlinear SUR model, 216
 quasi-ML estimator, 150
 seemingly unrelated regressions, 216
 simultaneous equations, 219
 systems FGNLS estimator, 217
 systems GMM estimator, 219
 systems IV estimator, 218–9
 systems MM estimator, 218
 systems NLS estimator, 217
 nonlinear three-stage least squares (NL3SLS)
 estimator, 219
 nonlinear two-stage least squares (NL2SLS) estimator
 asymptotic distribution, 195–6
 definition, 195–6
 example, 199
 see also nonlinear GMM estimator
 nonnested models
 Cox LR test, 279–80
 definition, 278
 example, 283–4
 information criteria comparison, 278–9
 overlapping, 281
 strictly nonnested, 281
 Vuong LR test, 280–3
 nonparametric bootstrap. *See* paired bootstrap
 nonparametric density estimation. *See* kernel density
 estimator
 nonparametric maximum likelihood (NPML)
 estimator, 622
 nonparametric regression, 307–22
 convergence rate, 311, 314
 kernel, 311–9
 local linear, 320
 local weighted average, 307–8
 Lowess, 320
 nearest-neighbors, 308–9, 319–20
 series, 321
 statistical inference intuition, 309–11
 test against parametric model, 319
 see also semiparametric regression
 nonrandomly varying coefficient, 846
 normal copula, 654
 normal distribution, 140
 truncated moments, 540, 566–7
 normal limit product rule. *See* Cramer linear
 transformation
 NPML estimator. *See* nonparametric maximum
 likelihood
 NR method. *See* Newton-Raphson method
 NSW demonstration project. *See* National Supported
 Work
 nuisance parameters. *See* incidental parameters

- numerical derivatives, 340, 350
 numerical integration. *See quadrature*
- observational data, 40–8, 814–7
 biased samples, 42–5
 clustering, 42
 identification strategies, 36–7
 measurement error, 46
 missing data, 46
 population, 40
 sample attrition, 47
 sampling methods, 40–4, 815–7
 sampling units, 41, 815
 sampling without replacement, 816–7
 survey methods, 41–2, 814–7
 survey nonresponse, 45–6
 types of data, 47–8
- observational equivalence, 29
- odds ratio, 470
see also posterior odds ratio
- OIR test. *See* overidentifying restrictions test
- OLS estimator. *See* ordinary least squares
- omitted variables bias, 92–3, 700, 716
 LM tests for, 274
- one-step GMM estimator, 187, 196
 panel, 746, 755
see also two-stage least squares
- one-way individual-specific effects model. *See*
 individual-specific effects model
- on-site sampling, 43, 823
- optimal Bayesian estimator, 434
- optimal GMM estimator, 176, 179–81, 187, 195
 compared to 2SLS, 187–8
- optimal MD estimator, 202, 753
- OPG. *See* outer-product of the gradient
- Orbit model, 914
- order of magnitude, 954
- ordered logit model, 520, 682
- ordered multinomial models, 519–20
- ordered probit model, 520, 535
- ordinary least squares (OLS) estimator, 70–81
 asymptotic distribution, 73–4, 80–1
 bias in standard errors with clustering, 836–7
 binary data, 471
 clustered data, 833–7
 coefficient interpretation in misspecified model,
 91–2
 consistency 72, 80
 definition, 71
 example, 84–5
 finite-sample distribution, 79
 heteroskedasticity-robust standard errors, 74–5, 81
 identification, 71–2
 inconsistency, 91, 95–6
 inefficiency, 80
 nonlinear, 150–9
 panel data, 702–3, 720–5
see also least squares estimators
- orthogonal polynomials, 321, 329, 390
 definition 390
- orthogonal regression approach, 920
- orthonormal polynomials, 321, 329, 390
- outcome equation, 547, 867
- outer product (OP) estimate, 138, 241, 395
 outer-product of the gradient (OPG) version
 LM test, 240–1
 m-test, 262–4
 small-sample performance, 262
- overdispersion, 670–1, 674–6, 690
 measurement error, 915–6
 panel data, 794, 806
 tests for, 671
- overidentification, 31, 100, 173, 176, 379–80, 747
see also GMM estimator
- overidentifying restrictions (OIR) test
 asymptotic distribution, 181, 183
 bootstrap, 379–80
 definition, 181, 267, 277
 panel data, 747, 756
- overlap assumption, 864, 871
 in RD design, 881
- oversampling, 41, 478–9, 814, 872
- paired bootstrap, 360, 366–8, 376, 378
- pairwise deletion, 928
 biased standard errors, 928
- panel attrition, 739, 801
- panel bootstrap, 377, 707, 746, 751, 789
- panel data, 47
- panel data models and estimators, 695–810
 comparison to clustered data, 831–2
see also linear panel; nonlinear panel
- panel GMM estimators, 744–68, 789–90
 application, 754–6
 Arellano-Bond estimator, 765–6
 asymptotic distribution, 745–6
 bootstrap, 389–90
 compared to MD estimator, 753
 computation, 751–2
 definition, 745
 efficiency, 747, 756
 exogeneity assumptions, 748–52
 instruments, 744, 747–51
 IV estimators for FE model, 757–9
 IV estimators for RE model, 759–60
 just-identified, 745
 nonlinear, 789–90
 OIR test, 747, 756
 one-step GMM estimator, 746, 755
 overidentified, 745
 2SLS estimator, 746, 755
 two-step GMM estimator, 746, 755
 variance matrix estimation, 751
- panel GMM model, 744–66
 application, 754–6
 dynamic, 763–6

- with individual-specific effects, 750–62
 without individual-specific effects, 744–53
see also panel GMM estimators
- panel IV estimators. *See* panel GMM estimators
- panel-robust statistical inference, 377, 705–7, 722, 746, 751, 788–90
 for Hausman test, 718
- Panel Study in Income Dynamics (PSID), 58, 889
- parametric bootstrap, 360
- Pareto distribution
 of the first kind, 609
 of the second kind, 616
- partial additive model, 323
- partial equilibrium analysis, 53, 862, 972
see also SUTVA
- partial F-statistic, 105, 109, 111
- partial likelihood estimator, 594–6
- partial ML estimator, 140
- partial R-squared, 104–5, 111
- partially linear model, 323–5, 327, 565, 684
- participation equation, 547, 551
- Pearson chi-square goodness-of-fit test, 266
- Pearson residual, 289, 291
- peer-effects model, 832
- percentile, 86
- percentile method, 364–5, 367–8
- percentile-*t* method, 364, 366–7
- PH model. *See* proportional hazards
- piecewise constant hazard model, 591
- Pitman drift, 248
- PML estimator. *See* pseudo-ML estimator
- Poisson distribution, 668
- Poisson-gamma mixture, 675
- Poisson-IG mixture, 677
- Poisson regression model, 666–74
 application, 671–4, 690, 792–5, 850–3
 asymptotic distribution of estimators, 668–9
 bivariate, 686
 censored MLE, 535
 with clustered data, 844, 850–3
 coefficient interpretation, 669
 definition, 668
 equidispersion, 668
 example, 117–8, 121–2
 LEF density, 148
 measurement error, 915–8
 mixtures, 675–9
 ML estimator, 668
 overdispersion, 670–1
 panel data, 792–5, 802–6
 quasi-ML estimator, 668–9, 682–3
 truncated MLE, 535
 underdispersion, 671
 zero-truncated, 680
see also count models
- polynomial baseline hazard, 591, 636
- pooled cross-section time series model. *See* pooled model
- pooled estimators, 702–3, 720–5
 application, 710–2, 725
 FGLS estimator, 720–1
 GEE estimator, 790, 794
 NLS estimator, 794
 OLS estimator, 211, 702–3, 720–5
 WLS estimator, 702–3, 721
- pooled model, 699, 720–5, 787–8
- pooling tests, 737
- population-averaged model. *See* pooled model
- population moment conditions
 for estimation, 172
 for testing, 260
see also GMM estimator; MM estimator; m-tests
- posterior distribution, 421, 430–4
 asymptotic behavior, 432–4
 conditional posterior, 431
 definition, 421
 expected posterior loss, 434
 expected posterior risk, 434
 full conditional distribution, 431
 highest posterior density interval, 431
 highest posterior density region, 431
 marginal posterior, 430
 observed-data posterior, 930
 posterior density interval, 431
 posterior mean, 423, 434
 posterior mode, 433
 posterior moments, 430
 posterior precision, 423
see also Bayesian methods
- posterior odds ratio, 456
- posterior (P) step, 455, 933
- potential outcome model, 30–4, 861–5
see also treatment effects; treatment evaluation
- power of tests, 247–50, 253–4
 bootstrapped tests, 372–3
 conditional moment test, 267–9
 example, 253–4
 Hausman test, 273–4
 local alternative hypotheses, 247–8
 uniformly most powerful test, 237
 Wald tests, 248–50
- precision parameter, 423
- predetermined instruments. *See* weak exogeneity
- prediction, 66–70
 best linear, 70
 conditional, 66
 error, 66–70
 in linear panel models, 738
 in mixed linear model, 774–6
 optimal, 66–70
 rotation groups, 814
 in structural model, 28
 weighted, 821
- pretest estimator, 285
- primary sampling units (PSUs), 41, 815, 845–55

- prior distribution, 425–30
 conjugate prior, 427
 definition, 420
 Dickey's prior, 439
 diffuse prior, 426
 flat prior, 426
 hierarchical priors, 428–9, 441–2
 improper prior, 426
 informative prior, 437–9
 Jeffreys' prior, 426
 noninformative prior, 425, 435–7
 normal-gamma prior, 437
 sensitivity analysis for, 429–30
see also Bayesian methods
- probit model, 470–71
 application, 465–6
 as additive random utility model, 477
 bivariate probit, 522–3
 bootstrap example, 254–6
 definition, 470
 discrete-time duration data, 602
 as GLM, 149
 index function model, 476
 logit model comparison, 471–3
 marginal effects, 467, 471
 ML estimator, 470
 Monte Carlo study example, 251–4
 multinomial probit, 516–9
 ordered probit, 520, 535
 panel data, 795–6
 simultaneous equations probit, 523, 560–1
see also binary outcome models
- probit selection equation, 548
- product copula, 654
- product integral, 578
- product rule, 949
see also Cramer linear transformation
- program evaluation. *See* treatment evaluation
- projection pursuit model, 323
- propensity score, 864–5
 application, 893–4
 balancing condition, 864, 893–4
 conditional independence assumption, 865
 definition, 864
 matching, 873–8, 892
see also treatment evaluation
- proportional hazards (PH) model, 592–7
 application, 605–7
 baseline survivor function estimator, 596–7
 coefficient interpretation, 606–7
 competing risks model, 645–6
 definition, 591
 discrete-time model, 600–3
 leading examples, 585
 mixed PH, 611–25
 panel data, 802
 partial likelihood estimator, 594–6
- pseudo-ML estimator (PML). *See* quasi-ML estimator
- pseudo panels, 771–3
 cohort, 771
 cohort fixed effects, 772–3
 measurement error, 772–3
- pseudo-random number generators, 410–6, 957–9
 accept-reject methods, 413–4
 composition methods, 415
 inverse transformation method, 413
 leading distributions, 957–9
 multivariate normal, 416
 transformation method, 413
 uniform variates, 412
see also MCMC methods
- pseudo R-squared measures
 for binary outcome models, 473–4
 definitions, 287–9
 example, 290–1
 for multinomial outcome models, 499
- pseudo-true value, 94, 132, 146, 281
- PSID. *See* Panel Study in Income Dynamics
- PSUs. *See* primary sampling units
- pure exogenous sampling, 825
- p-value, 226, 229, 234, 286, 363
- quadrature, 388–90
 Gaussian, 389–90
 multidimensional, 393
 in nonlinear panel models, 785–6, 796, 800
see also Monte Carlo integration
- qualitative response models. *See* binary outcomes, multinomial outcomes
- quantile, 86–7
- quantile regression, 85–90
 application, 88–90
 asymmetric absolute loss, 68, 85
 asymptotic distribution, 88
 bootstrap, 381
 computation, 341
 definition, 87
 IV estimator, 190
 multiplicative heteroskedasticity, 86–7
- quasi-difference, 783–4
- quasi-experiment. *See* natural experiment
- quasi-maximum likelihood (QML) estimator, 146–50
 asymptotic distribution, 146
 in binary outcome models, 469
 in clustered models, 842–3
 definition, 146
 in LEF, 147–9
 with multivariate dependent variable, 150
 in nonlinear systems, 216
 in panel models, 768, 786
 in Poisson model, 668–9, 682–3
- quasi-random numbers. *See* pseudo-random numbers
- QML estimator. *See* quasi-ML estimator
- random assignment, 49–50, 862
see also sampling schemes

- random coefficients model, 94, 385, 774–6, 786
see also hierarchical models
- random effects (RE) estimator, 705, 734–6, 759–62, 785–6
- application, 710–1, 725
 - asymptotic distribution, 735
 - clustered data, 837–9, 843–4
 - consistency, 699, 764
 - definition, 705, 734
 - error components 2SLS estimator, 760
 - error components 3SLS estimator, 762
 - FGLS estimator, 734–6
 - GEE estimator, 790, 794, 804
 - Hausman test, 717–9
 - incidental parameters, 704, 726
 - IV estimators, 759–60
 - ML estimator, 736, 785–6, 794–7, 800–1, 803–4
 - NLS estimator, 787, 794
 - quasi-ML estimator, 791
 - two-way effects model, 738
 - versus fixed effects, 701–2, 715–9
- random effects (RE) model, 700–2, 734–6, 759–62, 785–6
- binary outcome models, 795–6
 - Chamberlain model, 719, 786
 - clustered data, 831, 843–4
 - count models, 794, 803–4
 - definition, 700, 734
 - dynamic models, 792
 - duration models, 801–2
 - endogenous regressors, 756–7, 759–62
 - Mundlak model, 719
 - nonlinear models, 785–6
 - selection models, 801
 - Tobit model, 800–1
 - two-way effects model, 738
 - versus random effects, 701–2, 715–9
 - see also* hierarchical models; random effects estimator
- random number generators. *See* pseudo-random numbers
- random parameters logit (RPL) model, 512–6
- Bayesian methods, 514
 - definition, 513
 - ML estimator, 513–4
- random parameters model. *See* random coefficients model
- random utility models. *See* ARUM
- randomization bias, 53, 867
- randomized experiment, 50–3
- National Supported Work demonstration project, 889
- randomized trials, 49–53
- randomly varying coefficient, 847–8
- rank condition for identification, 31, 182, 214
- rank-ordered logit model, 521
- rank-ordered probit model, 521
- raw residual, 289, 291
- RD design. *See* regression discontinuity design
- receiver operators characteristics (ROC) curve, 474
- reduced form, 21, 25, 213
- see also* structural model
- RE estimator. *See* random effects
- regression-based imputation, 930–2
- EM algorithm, 932
 - nonignorable missingness, 932
- regression discontinuity (RD) design, 879–83
- fuzzy RD design, 882
 - heterogeneous treatment effects, 882
 - RD estimator, 882–3
 - sharp RD design, 880–1
 - treatment assignment mechanism, 879–81
- regressors, 71
- alternative-varying, 478, 497–8
 - endogenous, 23–33
 - fixed, 76–7
 - irrelevant, 93
 - omitted, 92–3
 - stochastic, 77
 - time-varying, 597–600, 702, 749–51
 - see also* endogenous regressors
- regularity conditions for ML, 141–2, 151–6
- relative risk, 470, 503
- reliability ratio, 903
- renewal function, 626
- renewal process, 626, 638
- repeated cross section data, 47, 770–3
- see also* differences-in-differences
- repeated measures. *See* panel data
- replicated data, 910–1, 913
- RESET test, 277–8
- residual analysis
- definitions, 289–90
 - duration data, 633–6
 - example, 290–1
 - panel data, 714–5
 - small-sample correction, 289
- residual bootstrap, 361
- response-based sampling, 43
- restricted ML estimator, 233, 776
- revealed preference data, 498, 516
- ridge regression estimator, 440
- Robinson difference estimator, 324–5, 565
- robust sandwich variance matrix estimate. *See* sandwich variance matrix
- robust standard errors
- bootstrap, 362–3, 376–8
 - Eicker-White, 74–5, 80–1, 112, 137
 - for extremum estimator, 137–9
 - Huber-White, 137, 144, 146
 - Newey-West, 137, 175, 723
 - see also* cluster-robust; heteroskedasticity-robust; panel-robust; systems-robust
- ROC curve. *See* receiver operators characteristics curve
- rotating panels, 739

- Roy model, 555–7, 562
 definition, 556
 dummy endogenous variable, 557
 Heckman two-step estimator, 556
 ML estimator, 556
 panel semiparametric estimation, 808
 as treatment effects model, 867
- RPL model. *See* random parameters logit
- R-squared, 287
 pseudo, 287–9
 uncentered, 241, 263
- running mean estimator, 308
- SA method. *See* simulated annealing
- sample attrition, 47
- sample moment conditions
see population moment conditions
- sample selection bias, 44–5
- sample weights, 817–21, 853–6
see also weighting
- sampling schemes
 assumptions for OLS, 76–78
 case-control, 479, 823
 choice-based sampling, 43, 478–9, 823
 endogenous sampling, 42–5, 78, 822–9, 856
 endogenous stratified sampling, 78, 820, 825–6, 856
 exogenous stratified sampling, 42, 78, 814–5, 820, 825, 856
 fixed in repeated samples, 76–7
 flow sampling, 44, 626
 multi-stage surveys, 41–2, 814–6, 853–6
 on-site sampling, 43, 823
 simple random sampling, 41, 76–7, 816
 stock sampling, 44, 626–7
 with replacement, 816
 without replacement, 816–7
- sandwich variance matrix
 clustered data, 834, 842
 extremum estimator, 132, 137–9
 GMM estimator, 175
 ML estimator, 144, 148
 NLS estimator, 150
 OLS estimator, 74
 panel data, 705–7, 722, 746, 751
 for Wald test, 277
see also robust standard errors
- Sargan test, 277
see also overidentifying restrictions test
- scale parameter, 509
- scanner data, 499
- Schwarz criterion. *See* BIC
- SCLS estimator. *See* symmetrically censored least squares
- score test, *see* Lagrange multiplier test
- score vector, 141
- secondary sampling units (SSUs), 41, 815, 854
- seed, 411
- seemingly unrelated regressions (SUR) model, 209–10, 216
- Bayesian MCMC example, 452–4
- count data, 685
- error components, 762
- nonlinear, 216
- selection bias, 445
 nonignorable missingness, 927, 932, 940
 treatment effects models, 867–71
see also selection models
- selection models, 546–62
 bivariate sample selection model, 547–53
 count models, 680
 example, 553–5
 panel data, 801
 Roy model, 555–7, 867
 sample selection, 546
 self selection, 546
 semiparametric estimation, 565–6
 structural models, 558–62
 treatment effects model, 862–4
 versus selection on observables only, 552–3, 864, 868–71
 versus two-part models, 546, 552–3
see also Tobit models
- selection on observables only, 552–3, 862–4, 868–9, 878–3, 889–96
 compared to selection models, 552–3, 864, 871
 conditional independence assumption, 868
 control function estimator, 869
 definition, 868–9
 DID estimator, 878–9
 RD design estimator, 879–83
 treatment effects model, 862–4, 889–96
- selection on unobservables, 552–3, 865–71, 883–9
 definition, 868
 in treatment effects model, 862–4
 IV estimators, 883–9
 Roy model, 867
 selection bias, 867–71
 selection model, 552–3
- self-weighting sample, 818
- SEM. *See* simultaneous equations model
- seminonparametric ML estimator, 328–9, 485
- semiparametric efficiency bounds, 323, 329–30, 485
- semiparametric estimators, 322–30
 adaptive, 323
 application, 565
 average derivative estimator, 326
 efficiency bounds, 323, 329–30
 nonparametric FGLS, 328
 Robinson difference estimator, 324–5, 565
 semiparametric least squares, 327, 483
 seminonparametric ML estimator, 328–9, 485
see also semiparametric models
- semiparametric heterogeneity model, 622
see also finite mixture models
- semiparametric least squares, 327, 483

- semiparametric ML estimator, 328–9, 485
 semiparametric models, 322–30
 additive models, 327
 binary outcome models, 482–6
 censored models, 563–5
 count models, 684–5
 definition, 322
 duration models, 594–600, 601–2
 flexible parametric models, 563
 heteroskedastic linear model, 323, 328
 identification, 325–6
 leading examples, 322
 multinomial outcome models, 523–4
 panel data models, 808
 partially linear model, 324–5
 selection models, 565–6
 single-index models, 325–7
 see also semiparametric estimators
 sequential limits, 767
 sequential multinomial models, 520–1
 sequential two-step m-estimator, 200–2
 bootstrap for, 362
 sequence of random variables, 943, 945
 serial correlation. *See* autocorrelation
 set identification, 29
 series estimator, 321
 for binary outcomes, 483
 shared frailty model, 662
 short panel
 definition, 700
 statistical inference in, 705–8, 721–2, 746, 751, 768
 shrinkage estimator, 440
 Silverman's plug-in estimate, 304
 simple random sampling (SRS), 41, 76–7, 816
 simple stratified sampling, 818
 Simpson's rule, 388–9
 simulated annealing (SA) method, 347
 simulated m-estimator, 398–9
 simulation-based estimation methods, 364–418
 motivating examples, 385–6
 see also MSL, MSM, indirect inference, simulators
 simulators, 393–4, 406–10
 antithetic sampling, 408–9
 direct, 393
 frequency, 406
 GHK, 407–8
 Halton sequences, 409–10
 importance sampling, 407
 smooth, 407
 subsimulator, 394
 unbiased, 394, 400
 see also quadrature
 simultaneous equations model (SEM), 22–31, 213–4,
 219
 causal interpretation, 26
 error components, 762
 extension to nonlinear models, 27
 FIML estimator, 214
 identification, 29–31, 213–4
 JML estimator, 214
 nonlinear, 219
 order condition, 213
 rank condition, 214
 reduced form, 25, 213
 single-equation models, 31
 structural form, 25, 213
 structural model, 24
 2SLS estimator, 214
 3SLS estimator, 214
 simultaneous equations probit, 523, 560–1
 simultaneous equations Tobit, 560–1
 single-index models, 123, 323, 325–7
 definition, 123
 identification, 325
 marginal effects, 123
 nonlinear panel model, 780
 semiparametric estimators, 325–7
 SIPP. *See* Survey of Income and Program Participation
 size of test, 246–7, 251–3
 nominal size, 251
 size-corrected test, 251
 true size, 251–3
 Sklar's theorem, 652
 Slutsky's Theorem, 945–6
 alternative version, 949
 small-sample bias. *See* finite-sample bias
 smooth maximum score estimator, 484
 smoothing parameters, 307
 smoothing spline estimator, 321
 social experiments, 32, 48–54
 advantages, 50–2
 examples, 51, 889
 limitations, 52–4
 randomization, 49–50
 span, 320
 specific to general test, 285
 specification tests, 259–78
 for clustered data, 840
 for duration models, 628–32
 for endogeneity, 275–6
 for exogeneity, 277
 for heteroskedasticity, 275
 for individual-specific effects, 737
 for omitted variables, 274
 for overdispersion, 670–1
 for pooling, 737
 for unobserved heterogeneity, 628–32
 for Tobit model, 543–4
 see also m-tests; model diagnostics
 spherical errors, 78
 split-sample IV estimator, 191–2
 SRS. *See* simple random sampling
 SSUs. *See* secondary sampling units
 stable family of distributions, 621
 stable unit treatment value assumption (SUTVA), 872
 standard errors. *See* robust standard errors

- starting values, 340, 351
 state dependence. *See* true state dependence
 stated preference data, 498, 516
 stationary population, 40
 statistical packages, 349
 step size adjustment, 338
 stochastic order of magnitude, 954–5
 stock sampling, 44, 626–7
 strata, 41, 815
see also sampling schemes; weighting
 stratification matching, 875–6, 893–6
 stratified random sampling, 76–7, 814–5
 use of Liapounov CLT, 951
 use of Markov LLN, 948
see also sampling schemes; weighting
 strict exogeneity. *See* strong exogeneity
 strong consistency, 947
 strong exogeneity, 22
 in panel models, 700, 749–50, 752, 781
 structural approach
 to measurement error, 901
 to weighting, 820–1
 structural economic models, 28, 171
 with selection, 558–60
 structural form, 20, 25, 223
 structural model, 20–31, 35–6
 based on economic model, 28
 exogeneity, 22–3
 full information, 35
 limited information, 35
 reduced form, 21, 25, 223
 structural form, 20, 25, 223
 structure, 20
see also simultaneous equations model
 structural selection models, 558–62
 based on utility maximization, 558–60
 endogenous regressors, 561–2
 simultaneous equations Tobit, 560–1
 studentized statistic, 359
 subsampling method, 373
 substitution bias, 53, 867
 sufficient statistic, 732, 782, 799, 805
 definition, 782
 summation assumption, 748, 752
 superpopulation, 40, 816
 supersmooth, 321
 SUR model. *See* seemingly unrelated regressions
 survey methods, 41–2, 84–7, 814–8, 853–6
 survey nonresponse, 45–6, 60, 739
see also attrition bias; imputation methods
 Survey of Income and Program Participation (SIPP), 59
 survival analysis. *See* duration models
 survival function. *See* survivor function
 survivor function
 aggregate survivor function, 619
 definition, 576–8
 estimator in PH model, 596–7
 Kaplan-Meier estimator, 581–2, 604–5
 in mixture models, 615–6
 multivariate, 649–50
 parametric examples, 585
 SUTVA. *See* stable unit treatment value assumption
 switching regressions model. *See* Roy model
 symmetrically censored least squares (SCLS)
 estimator, 565
 synthetic panels. *See* pseudo panels
 systems of equations, 206–19
 linear systems, 206–14
 nonlinear systems, 214–9
 seemingly unrelated regression, 209–10, 216
 simultaneous equations model, 22–31, 213–4, 219
 systems-robust standard errors, 208–9, 212, 219
 target density, 444
 tests. *See* hypothesis tests, m-tests, specification tests
 three-stage least squares (3SLS) estimator, 214
 3SLS estimator. *See* three-stage least squares
 time series data
 bootstrap, 381
 NLS estimator, 158–9
 Newey-West standard errors, 137, 175, 727
 time-varying regressors
 in duration models, 597–9
 in panel data models, 702, 749–51
 Tobit model, 536–44
 Bayesian methods, 563
 censored mean, 538–41
 censoring mechanism, 532, 579
 consistency of MLE, 538
 definition, 536
 example, 530–1
 generalized, 548
 Heckman two-step estimator, 543, 567–8
 identification, 536
 as imputation method, 932
 inverse-Mills ratio, 540–1
 marginal effects, 541–2
 measurement error in dependent variable, 914
 ML estimator, 537–8
 NLS estimator, 542
 OLS estimator, 543
 panel data, 800–1
 simultaneous equations, 560–1
 specification tests, 543–4
 with stochastic thresholds, 547
 with truncated data, 538
 truncated mean, 538–41, 566–7
 two-limit, 536
 type 2, 547
 type 5, 557
see also selection models
 top-coded data, 532–3, 541, 563
 transformation methods, 413
 transformation theorem, 949
 transformed ML estimator, 766

- transition data. *See* duration models
 trapezoidal rule, 388
 treatment-control comparison
 application, 890–1
 treatment effects framework, 862–5, 871–8, 889–96
 balancing condition, 864, 893–4
 binary treatment variable, 862
 conditional independence assumption, 863, 865
 conditional mean independence assumption, 864
 heterogeneous treatment effects, 882, 885
 multiple treatments, 860
 overlap assumption, 864, 871
 propensity score, 864–5
 Roy model, 867
 stable unit treatment value assumption, 872
 see also treatment evaluation
 treatment evaluation, 860–98
 application, 889–96
 IV estimators, 883–9
 matching estimators, 871–8
 DID estimators, 878–9
 selection bias, 865–71
 selection on observables, 862–4, 878–3, 889–96
 selection on unobservables, 865–71, 883–9
 regression discontinuity design, 879–83
 see also treatment effects framework
 treatment group, 49, 862
 trimming, 316, 333
 trivariate reduction, 686
 true state dependence
 duration models, 612, 630, 636
 dynamic panel models, 763–4, 798, 802
 see also unobserved heterogeneity
 truncated models, 530–44
 conditional mean, 535
 count models, 679–80
 definition, 532
 examples, 530–1, 535
 ML estimator, 534
 see also Tobit model; selection models
 truncated moments of standard normal, 540, 566–7
 truncation mechanisms, 532
 truncation from above, 532
 truncation from below, 532
 2SLS estimator. *See* two-stage least squares
 two-limit Tobit model, 536
 two-part model, 544–6
 application, 553–5
 compared to selection models, 546, 552–3
 definition, 545
 example, 545–6
 see also hurdle model
 two-stage IV estimator, 187
 two-stage least squares (2SLS) estimator, 101–2,
 187–91
 alternatives to, 190–2
 Basmann's approach, 190–1
 compared to optimal GMM, 187–8
 as GLS in transformed model, 188–9
 as GMM estimator, 187
 nonlinear, 195–6, 199
 panel data, 746, 755
 in SEM, 214
 Theil's interpretation, 189–90
 two-stage sampling, 41, 818
 two-step estimators
 GMM, 176, 187
 Heckman, 543, 550–1, 556, 567–8
 sequential m-estimator, 200–2
 two-step GMM estimator, 176, 187
 panel, 746, 755
 two-way effects model, 738
 type I error, 246–7
 type II error, 246–7
 type 1 extreme value distribution, 477, 486–7
 duration model error, 590
 multinomial logit model, 505
 type 2 Tobit. *See* bivariate sample selection model
 type 5 Tobit. *See* Roy model
 ultimate sampling units (USUs), 41, 815
 unbalanced panels, 739
 uncentered explained sum of squares (ESS), 241
 uncentered R-squared, 241, 263
 unconfoundedness assumption. *See* conditional
 independence assumption
 underrecording, 915
 undersmoothing, 305, 333, 380
 uniform convergence in probability, 126, 301
 uniform number generators, 412
 uniformly most powerful (UMP) test, 247
 unit roots, 382, 767–8
 universal logit model, 500
 unobserved heterogeneity
 application, 632–6
 in competing risks model, 647
 in count models, 675–7, 686
 distributions for, 614–5, 620–1
 in duration models, 611–25
 finite mixture models for, 621–5
 identification, 618–20
 IM test for, 267
 individual-specific effects, 700, 764
 mixture models for, 613–21
 MSL example, 397–8
 MSM example, 403
 multiplicative, 613, 686
 in nonlinear systems, 215
 specification tests for, 629–32
 variance inflation, 614
 versus true state dependence, 612, 630, 636, 763–4,
 798, 802
 USUs. *See* ultimate sampling units
 validation sample, 911
 variance components, 735, 845

- variance matrix estimation
 BHHH estimate, 138
 degrees-of-freedom adjustment, 75, 102, 138,
 185–6, 278, 841
 expected Hessian estimate, 138
 for extremum estimator, 137–9
 for GMM estimator, 174–5
 Hessian estimate, 138
 for NLS estimator, 154–5
 OPG estimate, 138
 robust estimate, 137
 sandwich estimate, 137, 144
 for weighted estimators, 854–6
see also robust standard errors
 variance reduction for simulation, 478
- Wald estimator
 in treatment effects models, 886
 Wald test, 136–7, 224–33
 asymptotic distribution, 226–8
 comparison with LM and, LR tests, 238–9
 definition, 136
 examples, 236, 241–3
 exclusion restrictions, 227
 F-test version, 226
 introduction, 136–7
 lack of invariance, 232–3
 likelihood based, 234, 241–3
 linear models, 224–5
 linear restrictions, 136–7
 in misspecified models, 229–30
 nonlinear restrictions, 224, 229
 power, 248–50
 of statistical significance, 228
 t-test version, 226–8
see also hypothesis tests
 weak consistency, 947
 weak exogeneity, 22
 in panel data, 749, 752, 758
 weak instruments, 100, 104–12
 application, 110–2
 definition, 104
 finite sample bias, 108–12, 177–8, 191–2, 196
 GMM estimator, 177–8
 inconsistency, 105–7
 indicators 104–5, 756
 panel data, 751–2, 756
 Weibull distribution, 584–6
 Weibull-gamma regression model, 615
 Weibull regression model, 143–4, 589, 606–8, 635
 weighted estimation
 endogenous stratification, 828–9
 exogenous stratification, 818–20
- weighted exogenous sampling ML (WESML)
 estimator, 828
 weighted least squares (WLS) estimator, 81–5
 asymptotic distribution, 83
 contrasted with GLS, 83
 definition, 83
 example, 84–5
 in pooled model, 702–3, 721
see also FGLS estimator
 weighted maximum likelihood (WML) estimator,
 828
 weighted semiparametric least squares (WSWL)
 estimator, 327
 for binary outcome models, 485
 weighting, 817–21, 827–9, 853–6
 descriptive versus structural approach, 820
 with endogenous stratification, 827–9
 sample weights, 817–8
 variance estimation, 853–6
 weighted prediction, 821
 weighted regression, 818–20
 whether to weight, 820–1
 welfare analysis
 with ARUM, 506–7
 with nested logit model, 512
 WESML estimator. *See* weighted exogenous sampling
 ML
 White standard errors. *See* robust standard errors
 wild bootstrap, 377–8
 window width, 299, 307, 312
 Wishart distribution, 443
see also inverse-Wishart distribution
 within estimator. *See* fixed effects estimator
 within model. *See* fixed effects model
 within-group variation, 709, 733
 with-zeros model, 681
 WLS estimator. *See* weighted least squares
 WML estimator. *See* weighted maximum likelihood
 WNLS estimator, 156–7
 asymptotic distribution, 156
 definition, 156
 example, 159–63
 as GLM, 158
 working matrix
 definition, 82
 for GLM estimator, 158
 for pooled GEE estimator, 794
 for pooled WLS estimator, 721
 for WLS estimator, 82–3
 WSLS estimator. *See* weighted semiparametric least
 squares
 zero-inflated count model, 680–1