Longitudinal and Panel Data:
Analysis and Applications for the Social Sciences

Brief Table of Contents

Chapter 1. Introduction

PART I - LINEAR MODELS

Chapter 2. Fixed Effects Models
Chapter 3. Models with Random Effects
Chapter 4. Prediction and Bayesian Inference
Chapter 5. Multilevel Models
Chapter 6. Random Regressors
Chapter 7. Modeling Issues
Chapter 8. Dynamic Models

PART II - NONLINEAR MODELS

Chapter 9. Binary Dependent Variables
Chapter 10. Generalized Linear Models
Chapter 11. Categorical Dependent Variables and Survival Models

Appendix A. Elements of Matrix Algebra
Appendix B. Normal Distribution
Appendix C. Likelihood-Based Inference
Appendix D. Kalman Filter
Appendix E. Symbols and Notation
Appendix F. Selected Longitudinal and Panel Data Sets
Appendix G. References

This draft is partially funded by the Fortis Health Insurance Professorship of Actuarial Science.

© 2003 by Edward W. Frees. All rights reserved

# Longitudinal and Panel Data: Analysis and Applications for the Social Sciences

## Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table of Contents</td>
<td>i</td>
</tr>
<tr>
<td>Preface</td>
<td>vi</td>
</tr>
<tr>
<td><strong>1. Introduction</strong></td>
<td></td>
</tr>
<tr>
<td>1.1 What are longitudinal and panel data?</td>
<td>1-1</td>
</tr>
<tr>
<td>1.2 Benefits and drawbacks of longitudinal data</td>
<td>1-4</td>
</tr>
<tr>
<td>1.3 Longitudinal data models</td>
<td>1-9</td>
</tr>
<tr>
<td>1.4 Historical notes</td>
<td>1-13</td>
</tr>
<tr>
<td><strong>PART I - LINEAR MODELS</strong></td>
<td></td>
</tr>
<tr>
<td><strong>2. Fixed Effects Models</strong></td>
<td></td>
</tr>
<tr>
<td>2.1 Basic fixed effects model</td>
<td>2-1</td>
</tr>
<tr>
<td>2.2 Exploring longitudinal data</td>
<td>2-5</td>
</tr>
<tr>
<td>2.3 Estimation and inference</td>
<td>2-10</td>
</tr>
<tr>
<td>2.4 Model specification and diagnostics</td>
<td>2-14</td>
</tr>
<tr>
<td>2.4.1 Pooling test</td>
<td>2-14</td>
</tr>
<tr>
<td>2.4.2 Added variable plots</td>
<td>2-15</td>
</tr>
<tr>
<td>2.4.3 Influence diagnostics</td>
<td>2-16</td>
</tr>
<tr>
<td>2.4.4 Cross-sectional correlation</td>
<td>2-17</td>
</tr>
<tr>
<td>2.4.5 Heteroscedasticity</td>
<td>2-18</td>
</tr>
<tr>
<td>2.5 Model extensions</td>
<td>2-19</td>
</tr>
<tr>
<td>2.5.1 Serial correlation</td>
<td>2-20</td>
</tr>
<tr>
<td>2.5.2 Subject-specific slopes</td>
<td>2-21</td>
</tr>
<tr>
<td>2.5.3 Robust estimation of standard errors</td>
<td>2-22</td>
</tr>
<tr>
<td>Further reading</td>
<td>2-23</td>
</tr>
<tr>
<td>Appendix 2A - Least squares estimation</td>
<td>2-24</td>
</tr>
<tr>
<td>2A.1 Basic Fixed Effects Model – Ordinary Least Squares Estimation</td>
<td>2-24</td>
</tr>
<tr>
<td>2A.2 Fixed Effects Model – Generalized Least Squares Estimation</td>
<td>2-24</td>
</tr>
<tr>
<td>2A.3 Diagnostic Statistics</td>
<td>2-25</td>
</tr>
<tr>
<td>2A.4 Cross-sectional Correlation</td>
<td>2-26</td>
</tr>
<tr>
<td>2. Exercises and Extensions</td>
<td>2-27</td>
</tr>
<tr>
<td><strong>3. Models with Random Effects</strong></td>
<td></td>
</tr>
<tr>
<td>3.1 Error components / random intercepts model</td>
<td>3-1</td>
</tr>
<tr>
<td>3.2 Example: Income tax payments</td>
<td>3-7</td>
</tr>
<tr>
<td>3.3 Mixed effects models</td>
<td>3-12</td>
</tr>
<tr>
<td>3.3.1 Linear mixed effects model</td>
<td>3-12</td>
</tr>
<tr>
<td>3.3.2 Mixed linear model</td>
<td>3-15</td>
</tr>
<tr>
<td>3.4 Inference for regression coefficients</td>
<td>3-16</td>
</tr>
</tbody>
</table>
3.5 Variance component estimation 3-20
   3.5.1 Maximum likelihood estimation 3-20
   3.5.2 Restricted maximum likelihood 3-21
   3.5.3 MIVQUE estimators 3-23

Further reading 3-25

Appendix 3A – REML calculations 3-26
   3A.1 Independence Of Residuals And Least Squares Estimators 3-26
   3A.2 Restricted Likelihoods 3-26
   3A.3 Likelihood Ratio Tests And REML 3-28

3. Exercises and Extensions 3-30

4. Prediction and Bayesian Inference 4-2
   4.1 Prediction for one-way ANOVA models 4-2
   4.2 Best linear unbiased predictors (BLUP) 4-4
   4.3 Mixed model predictors 4-7
      4.3.1 Linear mixed effects model 4-7
      4.3.2 Linear combinations of global parameters and subject-specific effects 4-7
      4.3.3 BLUP residuals 4-8
      4.3.4 Predicting future observations 4-9
   4.4 Example: Forecasting Wisconsin lottery sales 4-10
      4.4.1 Sources and characteristics of data 4-11
      4.4.2 In-sample model specification 4-13
      4.4.3 Out-of-sample model specification 4-15
      4.4.4 Forecasts 4-16
   4.5 Bayesian inference 4-17
   4.6 Credibility theory 4-20
      4.6.1 Credibility theory models 4-21
      4.6.2 Credibility theory ratemaking 4-22

Further reading 4-25

Appendix 4A – Linear unbiased prediction 4-26
   4A.1 Minimum Mean Square Predictor 4-26
   4A.2 Best Linear Unbiased Predictor 4-26
   4A.3 BLUP Variance 4-27

4. Exercises and Extensions 4-28

5. Multilevel Models 5-1
   5.1 Cross-sectional multilevel models 5-1
      5.1.1 Two-level models 5-1
      5.1.2 Multiple level models 5-5
      5.1.3 Multiple level modeling in other fields 5-6
   5.2 Longitudinal multilevel models 5-6
      5.2.1 Two-level models 5-6
      5.2.2 Multiple level models 5-10
   5.3 Prediction 5-11
   5.4 Testing variance components 5-13

Further reading 5-15

Appendix 5A – High order multilevel models 5-16
5. Exercises and Extensions 5-19
6. **Random Regressors**

6.1 Stochastic regressors in non-longitudinal settings
   6.1.1 Endogenous stochastic regressors 6-1
   6.1.2 Weak and strong exogeneity 6-3
   6.1.3 Causal effects 6-4
   6.1.4 Instrumental variable estimation 6-5

6.2 Stochastic regressors in longitudinal settings
   6.2.1 Longitudinal data models without heterogeneity terms 6-7
   6.2.2 Longitudinal data models with heterogeneity terms and strictly exogenous regressors 6-8

6.3 Longitudinal data models with heterogeneity terms and sequentially exogenous regressors

6.4 Multivariate responses
   6.4.1 Multivariate regressions 6-16
   6.4.2 Seemingly unrelated regressions 6-17
   6.4.3 Simultaneous equations models 6-18
   6.4.4 Systems of equations with error components 6-20

6.5 Simultaneous equation models with latent variables
   6.5.1 Cross-sectional models 6-23
   6.5.2 Longitudinal data applications 6-26

Further reading 6-29
Appendix 6A – Linear projections 6-30

7. **Modeling Issues**

7.1 Heterogeneity 7-1
7.2 Comparing fixed and random effects estimators
   7.2.1 A special case 7-7
   7.2.2 General case 7-9

7.3 Omitted variables
   7.3.1 Models of omitted variables 7-13
   7.3.2 Augmented regression estimation 7-14

7.4 Sampling, selectivity bias, attrition
   7.4.1 Incomplete and rotating panels 7-17
   7.4.2 Unplanned nonresponse 7-18
   7.4.3 Non-ignorable missing data 7-21

7. Exercises and Extensions 7-24

8. **Dynamic Models**

8.1 Introduction 8-1
8.2 Serial correlation models
   8.2.1 Covariance structures 8-3
   8.2.2 Nonstationary structures 8-4
   8.2.3 Continuous time correlation models 8-5

8.3 Cross-sectional correlations and time-series cross-section models

8.4 Time-varying coefficients
   8.4.1 The model 8-9
   8.4.2 Estimation 8-10
8.4.4 Forecasting 8-11
8.5 Kalman filter approach 8-13
  8.5.1 Transition equations 8-14
  8.5.2 Observation set 8-15
  8.5.3 Measurement equations 8-15
  8.5.4 Initial conditions 8-16
  8.5.5 Kalman filter algorithm 8-16
8.6 Example: Capital asset pricing model 8-18
Appendix 8A – Inference for the time-varying coefficient model 8-23
  8A.1 The Model 8-23
  8A.2 Estimation 8-23
  8A.3 Prediction 8-25

PART II - NONLINEAR MODELS

9. Binary Dependent Variables 9-1
  9.1 Homogeneous models
    9.1.1 Logistic and probit regression models 9-2
    9.1.2 Inference for logistic and probit regression models 9-5
    9.1.3 Example: Income tax payments and tax preparers 9-7
  9.2 Random effects models 9-9
  9.3 Fixed effects models 9-13
  9.4 Marginal models and GEE 9-16
Further reading 9-20
Appendix 9A – Likelihood calculations 9-20
  9A.1 Consistency Of Likelihood Estimators 9-21
  9A.2 Computing Conditional Maximum Likelihood Estimators 9-22
9. Exercises and Extensions 9-23

10. Generalized Linear Models 10-1
  10.1 Homogeneous models
    10.1.1 Linear exponential families of distributions 10-1
    10.1.2 Link functions 10-2
    10.1.3 Estimation 10-3
  10.2 Example: Tort filings 10-5
  10.3 Marginal models and GEE 10-8
  10.4 Random effects models 10-12
  10.5 Fixed effects models
    10.5.1 Maximum likelihood estimation for canonical links 10-16
    10.5.2 Conditional maximum likelihood estimation for canonical links 10-17
    10.5.3 Poisson distribution 10-18
  10.6 Bayesian inference 10-19
Further reading 10-22
Appendix 10A – Exponential families of distributions 10-23
  10A.1 Moment Generating Functions 10-23
  10A.2 Sufficiency 10-24
  10A.3 Conjugate Distributions 10-24
10A.4 Marginal Distributions 10-25
10. Exercises and Extensions 10-27

11. Categorical Dependent Variables and Survival Models
11.1 Homogeneous models 11-1
   11.1.1 Statistical inference 11-2
   11.1.2 Generalized logit 11-2
   11.1.3 Multinomial (conditional) logit 11-4
   11.1.4 Random utility interpretation 11-6
   11.1.5 Nested logit 11-7
   11.1.6 Generalized extreme value distribution 11-8
11.2 Multinomial logit models with random effects 11-8
11.3 Transition (Markov) models 11-10
11.4 Survival models 11-18
Appendix 11A. Conditional likelihood estimation for multinomial logit models with random effects 11-21

APPENDICES
Appendix A. Elements of Matrix Algebra A-1
   A.1 Basic Definitions A-1
   A.2 Basic Operations A-1
   A.3 Further Definitions A-2
   A.4 Matrix Decompositions A-2
   A.5 Partitioned Matrices A-3
   A.6 Kronecker (Direct) Products A-4
Appendix B. Normal Distribution A-5
Appendix C. Likelihood-Based Inference A-6
   C.1 Characteristics of Likelihood Functions A-6
   C.2 Maximum Likelihood Estimators A-6
   C.3 Iterated Reweighted Least Squares A-8
   C.4 Profile Likelihood A-8
   C.5 Quasi-Likelihood A-8
   C.6 Estimating Equations A-9
   C.7 Hypothesis Tests A-11
   C.8 Information Criteria A-12
   C.9 Goodness of Fit Statistics A-13
Appendix D. Kalman Filter A-14
   D.1 Basic State Space Model A-14
   D.2 Kalman Filter Algorithm A-14
   D.3 Likelihood Equations A-15
   D.4 Extended State Space Model and Mixed Linear Models A-15
   D.5 Likelihood Equations for Mixed Linear Models A-16
Appendix E. Symbols and Notation A-18
Appendix F. Selected Longitudinal and Panel Data Sets A-24
Appendix G. References A-28
Index A-40
Preface

Intended Audience and Level

This text focuses on models and data that arise from repeated measurements taken from a cross-section of subjects. These models and data have found substantive applications in many disciplines within the biological and social sciences. The breadth and scope of applications appears to be increasing over time. However, this widespread interest has spawned a hodgepodge of terms; many different terms are used to describe the same concept. To illustrate, even the subject title takes on different meanings in different literatures; sometimes this topic is referred to as “longitudinal data” and sometimes as “panel data.” To welcome readers from a variety of disciplines, I use the cumbersome yet more inclusive descriptor “longitudinal and panel data.”

This text is primarily oriented to applications in the social sciences. Thus, the data sets considered here are from different areas of social science including business, economics, education and sociology. The methods introduced into text are oriented towards handling observational data, in contrast to data arising from experimental situations, that are the norm in the biological sciences.

Even with this social science orientation, one of my goals in writing this text is to introduce methodology that has been developed in the statistical and biological sciences, as well as the social sciences. That is, important methodological contributions have been made in each of these areas; my goal is to synthesize the results that are important for analyzing social science data, regardless of their origins. Because many terms and notations that appear in this book are also found in the biological sciences (where panel data analysis is known as longitudinal data analysis), this book may also appeal to researchers interested in the biological sciences.

Despite its forty-year history and widespread usage, a survey of the literature shows that the quality of applications is uneven. Perhaps this is because longitudinal and panel data analysis has developed in separate fields of inquiry; what is widely known and accepted in one field is given little prominence in a related field. To provide a treatment that is accessible to researchers from a variety of disciplines, this text introduces the subject using relatively sophisticated quantitative tools, including regression and linear model theory. Knowledge of calculus, as well as matrix algebra, is also assumed. For Chapter 8 on dynamic models, a time series course would also be useful.

With this level of prerequisite mathematics and statistics, I hope that the text is accessible to quantitatively oriented graduate social science students who are my primary audience. To help students work through the material, the text features several analytical and empirical exercises. Moreover, detailed appendices on different mathematical and statistical supporting topics should help students develop their knowledge of the topic as they work the exercises. I also hope that the textbook style, such as the boxed procedures and an organized set of symbols and notation, will appeal to applied researchers that would like a reference text on longitudinal and panel data modeling.

Organization

The beginning chapter sets the stage for the book. Chapter 1 introduces longitudinal and panel data as repeated observations from a subject and cites examples from many disciplines in which longitudinal data analysis is used. This chapter outlines important benefits of longitudinal data analysis, including the ability to handle the heterogeneity and dynamic features of the data. The chapter also acknowledges some important drawbacks of this scientific methodology, particularly the problem of attrition. Furthermore, Chapter 1 provides an overview of the several types of models used to handle longitudinal data; these models are considered in greater detail in
When discussing heterogeneity in the context of longitudinal data analysis, we mean that observations from different subjects tend to be dissimilar when compared to observations from the same subject that tend to be similar. One way of modeling heterogeneity is to use fixed parameters that vary by individual; this formulation is known as a fixed effects model and is described in Chapter 2. A useful pedagogic feature of fixed effects models is that they can be introduced using standard linear model theory. Linear model and regression theory is widely known among research analysts; with this solid foundation, fixed effects models provide a desirable foundation for introducing longitudinal data models. This text is written assuming that readers are familiar with linear model and regression theory at the level of, for example, Draper and Smith (1995) or Greene (1993). Chapter 2 provides an overview of linear models with a heavy emphasis on analysis of covariance techniques that are useful for longitudinal and panel data analysis. Moreover, the Chapter 2 fixed effects models provide a solid framework for introducing many graphical and diagnostic techniques.

Another way of modeling heterogeneity is to use parameters that vary by individual yet that are represented as random quantities; these quantities are known as random effects and are described in Chapter 3. Because models with random effects generally include fixed effects to account for the mean, models that incorporate both fixed and random quantities are known as linear mixed effects models. Just as a fixed effects model can be thought of in the linear model context, a linear mixed effects model can be expressed as a special case of the mixed linear model. Because mixed linear model theory is not as widely known as regression, Chapter 3 provides more details on the estimation and other inferential aspects than the corresponding development in Chapter 2. Still, the good news for applied researchers is that, by writing linear mixed effects models as mixed linear models, widely available statistical software can be used to analyze linear mixed effects models.

By appealing to linear model and mixed linear model theory in Chapters 2 and 3, we will be able to handle many applications of longitudinal and panel data models. Still, the special structure of longitudinal data raises additional inference questions and issues that are not commonly addressed in the standard introductions to linear model and mixed linear model theory. One such set of questions deals with the problem of “estimating” random quantities, known as prediction. Chapter 4 introduces the prediction problem in the longitudinal data context and shows how to “estimate” residuals, conditional means and future values of a process. Chapter 4 also shows how to use Bayesian inference as an alternative method for prediction.

To provide additional motivation and intuition for Chapters 3 and 4, Chapter 5 introduces multilevel modeling. Multilevel models are widely used in educational sciences and developmental psychology where one assumes that complex systems can be modeled hierarchically; that is, modeling one level at a time, each level conditional on lower levels. Many multilevel models can be written as linear mixed effects models; thus, the inference properties of estimation and prediction that we develop in Chapters 3 and 4 can be applied directly to the Chapter 5 multilevel models.

Chapter 6 returns to the basic linear mixed effects model but now adopts an econometric perspective. In particular, this chapter considers situations where the explanatory variables are stochastic and may be influenced by the response variable. In such circumstances, the explanatory variables are known as endogenous. Difficulties associated with endogenous explanatory variables, and methods for addressing these difficulties, are well known for cross-sectional data. Because not all readers will be familiar with the relevant econometric literature, Chapter 6 reviews these difficulties and methods. Moreover, Chapter 6 describes the more recent literature on similar situations for longitudinal data.

Chapter 7 analyzes several issues that are specific to a longitudinal or panel data study. One issue is the choice of the representation to model heterogeneity. The many choices include...
fixed effects, random effects and serial correlation models. Chapter 7.1 reviews important identification issues when trying to decide upon the appropriate model for heterogeneity. One issue is the comparison of fixed and random effects models that has received substantial attention in the econometrics literature. As described in Chapter 7, this comparison involves interesting discussions of the omitted variables problem. Briefly, we will see that time-invariant omitted variables can be captured through the parameters used to represent heterogeneity, thus handling two problems at the same time. Chapter 7 concludes with a discussion of sampling and selectivity bias. Panel data surveys, with repeated observations on a subject, are particularly susceptible to a type of selectivity problem known as attrition, where individuals leave a panel survey.

Longitudinal and panel data applications are typically “long” in the cross-section and “short” in the time dimension. Hence, the development of these methods stem primarily from regression-type methodologies such as linear model and mixed linear model theory. Chapters 2 and 3 introduce some dynamic aspects, such as serial correlation, where the primary motivation is to provide improved parameter estimators. For many important applications, the dynamic aspect is the primary focus, not an ancillary consideration. Further, for some data sets, the temporal dimension is “long,” thus providing opportunities to model the dynamic aspect in detail. For these situations, longitudinal data methods are closer in spirit to multivariate time series analysis than to cross-sectional regression analysis. Chapter 8 introduces dynamic models, where the time dimension is of primary importance.

Chapters 2 through 8 are devoted to analyzing data that may be represented using models that are linear in the parameters, including linear and mixed linear models. In contrast, Chapters 9 through 11 are devoted to analyzing data that can be represented using nonlinear models. The collection of nonlinear models is vast. To provide a concentrated discussion that relates to the applications orientation of this book, we focus on models where the distribution of the response cannot be reasonably approximated by a normal distribution and alternative distributions must be considered.

We begin in Chapter 9 with a discussion of modeling responses that are dichotomous; we call these binary dependent variable models. Because not all readers with a background in regression theory have been exposed to binary dependent models such as logistic regression, Chapter 9 begins with an introductory section under the heading of “homogeneous” models; these are simply the usual cross-sectional models without heterogeneity parameters. Then, Chapter 9 introduces the issues associated with random and fixed effects models to accommodate the heterogeneity. Unfortunately, random effects model estimators are difficult to compute and the usual fixed effects model estimators have undesirable properties. Thus, Chapter 9 introduces an alternative modeling strategy that is widely used in biological sciences based on a so-called “marginal” model. This model employs generalized estimating equation (GEE), or generalized method of moments (GMM), estimators that are simple to compute and have desirable properties.

Chapter 10 extends that Chapter 9 discussion to generalized linear models (GLMs). This class of models handles the normal-based models of Chapter 2 through 8, the binary models of Chapter 9 as well as additional important applied models. Chapter 10 focuses on count data through the Poisson distribution although the general arguments can also be used for other distributions. Like Chapter 9, we begin with the homogeneous case to provide a review for readers that have not been introduced to GLM. The next section is on marginal models that are particularly useful for applications. Chapter 10 follows with an introduction to random and fixed effects models.

Using the Poisson distribution as a basis, Chapter 11 extends the discussion to multinomial models. These models are particularly useful in economic “choice” models that have seen broad applications in the marketing research literatures. Chapter 11 provides a brief overview of the economic basis for these choice models and then shows how to apply these to random effects multinomial models.
**Statistical Software**

My goal in writing this text is to reach a broad group of researchers. Thus, to avoid excluding large segments of individuals, I have chosen not to integrate any specific statistical software package into the text. Nonetheless, because of the applications orientation, it is critical that the methodology presented be easily accomplished using readily available packages. For the course taught at the University of Wisconsin, I use the statistical package SAS. (Although many of my students opt to use alternative packages such as STATA and “R,” I encourage free choice!) In my mind, this is the analog of an “existence theorem.” If a procedure is important and can be readily accomplished by one package, then it is (or will soon be) available through its competitors. On the book web site, users will find routines written in SAS for the methods advocated in the text, thus “proving” that they are readily available to applied researchers. Routines written for STATA and R are also available on the web site. For more information on SAS, STATA and R, visit their web sites:

- [http://www.sas.com](http://www.sas.com)
- [http://www.stata.com](http://www.stata.com)
- [http://www.r-project.org](http://www.r-project.org)

**References Codes**

In keeping with my goal of reaching a broad group of researchers, I have attempted to integrate contributions from different fields that regularly study longitudinal and panel data techniques. To this end, Appendix G contains the references that are subdivided into six sections. This subdivision is maintained to emphasize the breadth of longitudinal and panel data analysis and the impact that it has made on several scientific fields. I refer to these sections using the following coding scheme:

- **B** Biological Sciences Longitudinal Data
- **E** Econometrics Panel Data
- **EP** Educational Science and Psychology
- **O** Other Social Sciences
- **S** Statistical Longitudinal Data
- **G** General Statistics

For example, I use “Neyman and Scott (1948E)” to refer to an article written by Neyman and Scott, published in 1948, that appears in the “Econometrics Panel Data” portion of the references.

**Approach**

This book grew out of lecture notes for a course offered at the University of Wisconsin. The pedagogic approach of the manuscript evolved from the course. Each chapter consists of an introduction to the main ideas in words and then as mathematical expressions. The concepts underlying the mathematical expressions are then reinforced with empirical examples; these data are available to the reader at the Wisconsin book web site. Most chapters conclude with exercises that are primarily analytic; some are designed to reinforce basic concepts for (mathematically) novice readers. Others are designed for (mathematically) sophisticated readers and constitute extensions of the theory presented in the main body of the text. The beginning chapters (2-5) also include empirical exercises that allow readers to develop their data analysis skills in a longitudinal data context. Selected solutions to the exercises are also available from the author.

Readers will find that the text becomes more mathematically challenging as it progresses. Chapters 1–3 describe the fundamentals of longitudinal data analysis and are prerequisites for the remainder of the text. Chapter 4 is prerequisite reading for Chapters 5 and 8. Chapter 6 contains
important elements necessary for reading Chapter 7. As described above, a time series analysis course would also be useful for mastering Chapter 8, particularly the Section 8.5 Kalman filter approach.

Chapter 9 begins the section on nonlinear modeling. Only Chapters 1-3 are necessary background for the section. However, because it deals with nonlinear models, the requisite level of mathematical statistics is higher than Chapters 1-3. Chapters 10 and 11 continue the development of these models. I do not assume prior background on nonlinear models. Thus, in Chapters 9-11, the first section introduces the chapter topic in a non-longitudinal context that I call a homogeneous model.

Despite the emphasis placed on applications and interpretations, I have not shied from using mathematics to express the details of longitudinal and panel data models. There are many students with excellent training in mathematics and statistics that need to see the foundations of longitudinal and panel data models. Further, there are now a number of texts and summary articles that are now available (and cited throughout the text) that place a heavier emphasis on applications. However, applications-oriented texts tend to be field-specific; studying only from such a source can mean that an economics student will be unaware of important developments in educational sciences (and vice versa). My hope is that many instructors will choose to use this text as a technical supplement to an applications-oriented text from their own field.

The students in my course come from the wide variety of backgrounds in mathematical statistics. To develop longitudinal and panel data analysis tools and achieve a common set of notation, most chapters contain a short appendix that develops mathematical results cited in the chapter. Further, there are four appendices at the end of the text that expand mathematical developments used throughout the text. A fifth appendix, on symbols and notation, further summarizes the set of notation used throughout the text. The sixth appendix provides a brief description of selected longitudinal and panel data sets that are used in several disciplines throughout the world.

Acknowledgements

This text was reviewed by several generations of longitudinal and panel data classes here at the University of Wisconsin. The students in my classes contributed a tremendous amount of input into the text; their input drove the text’s development far more than they realize.

I have enjoyed working with several colleagues on longitudinal and panel data problems over the years. Their contributions are reflected indirectly throughout the text. Moreover, I have benefited from detailed reviews by: Anocha Ariborg, Mousumi Banerjee, Jee-Seon Kim, Yueh-Chuan Kung, and Georgios Pitelis.

Saving the most important for last, I thank my family for their support. Ten thousand thanks to my mother Mary, my wife Deirdre, our sons Nathan and Adam, and our source of amusement, Lucky (our dog).
Chapter 1. Introduction

Abstract. This chapter introduces the many key features of the data and models used in the analysis of longitudinal and panel data. Here, longitudinal and panel data are defined and an indication of their widespread usage is given. The chapter discusses the benefits of these data; these include opportunities to study dynamic relationships while understanding, or at least accounting for, cross-sectional heterogeneity. Designing a longitudinal study does not come without a price; in particular, longitudinal data studies are sensitive to the problem of attrition, that is, unplanned exit from a study. This book focuses on models that are appropriate for the analysis of longitudinal and panel data; this introductory chapter outlines the set of models that will be considered in subsequent chapters.

1.1 What are longitudinal and panel data?

Statistical modeling

Statistics is about data. It is the discipline concerned with the collection, summarization and analysis of data to make statements about our world. When analysts collect data, they are really collecting information that is quantified, that is, transformed to a numerical scale. There are many well-understood rules for reducing data, using either numerical or graphical summary measures. These summary measures can then be linked to a theoretical representation, or model, of the data. With a model that is calibrated by data, statements about the world can be made.

As users, we identify a basic entity that we measure by collecting information on a numerical scale. This basic entity is our unit of analysis, also known as the research unit or observational unit. In the social sciences, the unit of analysis is typically a person, firm or governmental unit, although other applications can and do arise. Other terms used for the observational unit include individual, from the econometrics literature, as well as subject, from the biostatistics literature.

Regression analysis and time series analysis are two important applied statistical methods used to analyze data. Regression analysis is a special type of multivariate analysis, where several measurements are taken from each subject. We identify one measurement as a response, or dependent variable; the interest is in making statements about this measurement, controlling for the other variables.

With regression analysis, it is customary to analyze data from a cross-section of subjects. In contrast, with time series analysis, we identify one or more subjects and observe them over time. This allows us to study relationships over time, the so-called dynamic aspect of a problem. To employ time series methods, we generally restrict ourselves to a limited number of subjects that have many observations over time.

Defining longitudinal and panel data

Longitudinal data analysis represents a marriage of regression and time series analysis. As with many regression data sets, longitudinal data are composed of a cross-section of subjects. Unlike regression data, with longitudinal data we observe subjects over time. Unlike time series
data, with longitudinal data we observe many subjects. Observing a broad cross-section of subjects over time allows us to study dynamic, as well as cross-sectional, aspects of a problem.

The descriptor \textit{panel data} comes from surveys of individuals. In this context, a “panel” is a group of individuals surveyed repeatedly over time. Historically, panel data methodology within economics had been largely developed through labor economics applications. Now, economic applications of panel data methods are not confined to survey or labor economics problems and the interpretation of the descriptor “panel analysis” is much broader. Hence, we will use the terms “longitudinal data” and “panel data” interchangeably although, for simplicity, we often use only the former term.

\textbf{Example 1.1 - Divorce rates}

Figure 1.1 shows the 1965 divorce rates versus AFDC (Aid to Families with Dependent Children) payments for the fifty states. For this example, each state represents an observational unit, the divorce rate is the response of interest and the level of AFDC payment represents a variable that may contribute information to our understanding of divorce rates.

The data are observational; thus, it is not appropriate to argue for a causal relationship between welfare payments (AFDC) and divorce rates without additional economic or sociological theory. Nonetheless, their relation is important to labor economists and policymakers.

Figure 1.1 shows a negative relation; the corresponding correlation coefficient is -0.37. Some argue that this negative relation is counter-intuitive in that one would expect a positive relation between welfare payments and divorce rates; states with desirable economic climates enjoy both a low divorce rate and low welfare payments. Others argue that this negative relationship is intuitively plausible; wealthy states can afford high welfare payments and produce a cultural and economic climate conducive to low divorce rates.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{divorce_vs_afdc.png}
\caption{Plot of 1965 Divorce versus AFDC Payments}
\end{figure}

\textit{Source: US Statistical Abstracts}

Another plot, not displayed here, shows a similar negative relation for 1975; the corresponding correlation is -0.425. Further, a plot with both the 1965 and 1975 data displays a negative relation between divorce rates and AFDC payments.
Figure 1.2. Plot of Divorce versus AFDC Payments – 1965 and 1975

Figure 1.2 shows both the 1965 and 1975 data; a line connects the two observations within each state. The line represents a change over time (dynamic), not a cross-sectional relationship. Each line displays a positive relationship, that is, as welfare payments increase so do divorce rates for each state. Again, we do not infer directions of causality from this display. The point is that the dynamic relation between divorce and welfare payments within a state differs dramatically from the cross-sectional relationship between states.

Some notation

Models of longitudinal data are sometimes differentiated from regression and time series through their “double subscripts.” With this notation, we may distinguish among responses by subject and time. To this end, define $y_{it}$ to be the response for the $i$th subject during the $t$th time period. A longitudinal data set consists of observations of the $i$th subject over $t=1, \ldots, T_i$ time periods, for each of $i=1, \ldots, n$ subjects. Thus, we observe:

- first subject - $\{y_{11}, y_{12}, \ldots, y_{1T_i}\}$
- second subject - $\{y_{21}, y_{22}, \ldots, y_{2T_i}\}$
- \ldots
- $n$th subject - $\{y_{n1}, y_{n2}, \ldots, y_{nT_i}\}$.

In Example 1.1, most states have $T_i = 2$ observations and are depicted graphically in Figure 1.2 by a line connecting the two observations. Some states have only $T_i = 1$ observation and are depicted graphically by an open circle plotting symbol. For many data sets, it is useful to let the number of observations depend on the subject; $T_i$ denotes the number of observations for the $i$th subject. This situation is known as the unbalanced data case. In other data sets, each subject has the same number of observations; this is known as the balanced data case. Traditionally, much of the econometrics literature has focused on the balanced data case. We will consider the more broadly applicable unbalanced data case.
Prevalence of longitudinal and panel data analysis

Longitudinal and panel databases and models have taken an important role in the literature. They are widely used in the social science literature, where panel data are also known as pooled cross-sectional time series and in the natural sciences, where panel data are referred to as longitudinal data. To illustrate, an index of business and economic journals, ABI/INFORM, lists 270 articles in 2001 and 2002 that use panel data methods. Another index of scientific journals, the ISI Web of Science, lists 811 articles in 2001 and 2002 that use longitudinal data methods. And these are only the applications that were considered innovative enough to be published in scholarly reviews!

Longitudinal data methods have also developed because important databases have become available to empirical researchers. Within economics, two important surveys that track individuals over repeated surveys include the Panel Survey of Income Dynamics (PSID) and the National Longitudinal Survey of Labor Market Experience (NLS). In contrast, the Consumer Price Survey (CPS) is another survey conducted repeatedly over time. However, the CPS is generally not regarded as a panel survey because individuals are not tracked over time. For studying firm-level behavior, databases such as Compustat and CRSP (University of Chicago’s Center for Research on Security Prices) have been available for over thirty years. More recently, the National Association of Insurance Commissioners (NAIC) has made insurance company financial statements available electronically. With the rapid pace of software development within the database industry, it is easy to anticipate the development of many more databases that would benefit from longitudinal data analysis. To illustrate, within the marketing area, product codes are scanned in when customers check out of a store and are transferred to a central database. These so-called scanner data represent yet another source of data information that may tell marketing researchers about purchasing decisions of buyers over time or the efficiency of a store’s promotional efforts. Appendix F summarizes longitudinal and panel data sets used worldwide.

1.2 Benefits and drawbacks of longitudinal data

There are several advantages of longitudinal data compared with either purely cross-sectional or purely time series data. In this introductory chapter, we focus on two important advantages: the ability to study dynamic relationships and to model the differences, or heterogeneity, among subjects. Of course, longitudinal data are more complex than purely cross-sectional or times series data and so there is a price in working with them. The most important drawback is the difficulty in designing the sampling scheme to reduce the problem of subjects leaving the study prior to its completion, known as attrition.

Dynamic relationships

Figure 1.1 shows the 1965 divorce rate versus welfare payments. Because these are data from a single point in time, they are said to represent a static relationship. To illustrate, we might summarize the data by fitting a line using the method of least squares. Interpreting the slope of this line, we estimate a decrease of 0.95% in divorce rates for each $100 increase in AFDC payments.

In contrast, Figure 1.2 shows changes in divorce rates for each state based on changes in welfare payments from 1965 to 1975. Using least squares, the overall slope represents an increase of 2.9% in divorce rates for each $100 increase in AFDC payments. From 1965 to 1975, welfare payments increased an average of $59 (in nominal terms) and divorce rates increased 2.5%. Now the slope represents a typical time change in divorce rates per $100 unit time change in welfare payments; hence, it represents a dynamic relationship.

Perhaps the example might be more economically meaningful if welfare payments were in real dollars, and perhaps not (for example, deflated by the Consumer Price Index).
Nonetheless, the data strongly reinforce the notion that dynamic relations can provide a very different message than cross-sectional relations.

Dynamic relationships can only be studied with repeated observations and we have to think carefully about how we define our “subject” when considering dynamics. To illustrate, suppose that we are looking at the event of divorce on individuals. By looking at a cross-section of individuals, we can estimate divorce rates. By looking at cross-sections repeated over time (without tracking individuals), we can estimate divorce rates over time and thus study this type of dynamic movement. However, only by tracking repeated observations on a sample of individuals can we study the duration of marriage, or time until divorce, another dynamic event of interest.

**Historical approach**

Early panel data studies used the following strategy to analyze pooled cross-sectional data:

- Estimate cross-sectional parameters using regression.
- Use time series methods to model the regression parameter estimators, treating estimators as known with certainty.

Although useful in some contexts, this approach is inadequate in others, such as Example 1.1. Here, the slope estimated from 1965 data is –0.95%. Similarly, the slope estimated from 1975 data turns out to be –1.0%. Extrapolating these negative estimators from different cross-sections yields very different results from the dynamic estimate, a positive 2.9%. Theil and Goldberger (1961E) provide an early discussion of the advantages of estimating the cross-sectional and time series aspects simultaneously.

**Dynamic relationships and time series analysis**

When studying dynamic relationships, univariate time series analysis is a well-developed methodology. However, this methodology does not account for relationships among different subjects. In contrast, multivariate time series analysis does account for relationships among a limited number of different subjects. Whether univariate or multivariate, an important limitation of time series analysis is that it requires several (generally, at least thirty) observations to make reliable inferences. For an annual economic series with thirty observations, using time series analysis means that we are using the same model to represent an economic system over a period of thirty years. Many problems of interest lack this degree of stability; we would like alternative statistical methodologies that do not impose such strong assumptions.

**Longitudinal data as repeated time series**

With longitudinal data, we use several (repeated) observations of many subjects, over different time periods. Repeated observations from the same subject tend to be correlated. One way to represent this correlation is through dynamic patterns. A model that we use is:

\[ y_{it} = E y_{it} + \epsilon_{it}, \quad t=1, ..., T_i, \quad i=1, ..., n. \]  

(1.1)

Here, \( \epsilon_{it} \) represents the deviation of the response from its mean; this deviation may include dynamic patterns. Intuitively, if there is a dynamic pattern that is common among subjects, then by observing this pattern over many subjects, we hope to estimate the pattern with fewer time series observations than required of conventional time series methods.

For many data sets of interest, subjects do not have identical means. As a first order approximation, a linear combination of known, explanatory variables such as

\[ E y_{it} = \alpha + x_{it}' \beta \]

serves as a useful specification of the mean function. Here, \( x_{it} \) is a vector of explanatory, or independent, variables.
Longitudinal data as repeated cross-sectional studies

Longitudinal data may be treated as a repeated cross-section by ignoring the information about individuals that is tracked over time. As mentioned above, there are many important repeated surveys such as the CPS where subjects are not tracked over time. Such surveys are useful for understanding aggregate changes in a variable, such as the divorce rate, over time. However, if the interest is in studying the time-varying effects of economic, demographic or sociological characteristics of an individual on divorce, then tracking individuals over time is much more informative than a repeated cross-section.

Heterogeneity

By tracking subjects over time, we may model subject behavior. In many data sets of interest, subjects are unlike one another, that is, they are heterogeneous. In (repeated) cross-sectional regression analysis, we use models such as $y_{it} = \alpha + x_{it}' \beta + \epsilon_{it}$ and ascribe the uniqueness of subjects to the disturbance term $\epsilon_{it}$. In contrast, with longitudinal data we have an opportunity to model this uniqueness. A basic longitudinal data model that incorporates heterogeneity among subjects is based on

$$E y_{it} = \alpha_i + x_{it}' \beta, \quad t=1, \ldots, T; \quad i=1, \ldots, n. \quad (1.2)$$

In cross-sectional studies where $T_i = 1$, the parameters of this model are unidentifiable. However, in longitudinal data, we have a sufficient number of observations to estimate $\beta$ and $\alpha_i, \ldots, \alpha_n$. Allowing for subject-specific parameters, such as $\alpha_i$, provides an important mechanism for controlling heterogeneity of individuals. Models that incorporate heterogeneity terms such as in equation (1.2) will be called heterogeneous models. Models without such terms will be called homogeneous models.

We may also interpret heterogeneity to mean that observations from the same subject tend to be similar compared to observations from different subjects. Based on this interpretation, heterogeneity can be modeled by examining the sources of correlation among repeated observations from a subject. That is, for many data sets, we anticipate finding a positive correlation when examining $\{y_{i1}, y_{i2}, \ldots, y_{iT_i}\}$. As noted above, one possible explanation is the dynamic pattern among the observations. Another possible explanation is that the response shares a common, yet unobserved, subject-specific parameter that induces a positive correlation.

There are two distinct approaches for modeling the quantities that represent heterogeneity among subjects, $\{\alpha_i\}$. Chapter 2 explores one approach, where $\{\alpha_i\}$ are treated as fixed, yet unknown, parameters to be estimated. In this case, equation (1.2) is known as a fixed effects model. Chapter 3 introduces the second approach, where $\{\alpha_i\}$ are treated as (ex-ante) draws from an unknown population and thus are random variables. In this case, equation (1.2) may be expressed as

$$E (y_{it} | \alpha_i) = \alpha_i + x_{it}' \beta.$$

This is known as a random effects formulation.

Heterogeneity bias

Failure to include heterogeneity quantities in the model may introduce serious bias into the model estimators. To illustrate, suppose that a data analyst mistakenly uses the function

$$E y_{it} = \alpha + x_{it}' \beta,$$

when equation (1.2) is the true function. This is an example of heterogeneity bias, or a problem with aggregation with data.

Similarly, one could have different (heterogeneous) slopes
\[ \text{E} \, y_{it} = \alpha + x_{it}' \beta_i \]

or different intercepts and slopes

\[ \text{E} \, y_{it} = \alpha_i + x_{it}' \beta_i. \]

**Omitted variables**

Incorporating heterogeneity quantities into longitudinal data models are often motivated by the concern that important variables have been omitted from the model. To illustrate, consider the true model

\[ y_{it} = \alpha_i + x_{it}' \beta + z_i' \gamma + \epsilon_{it}. \]

Assume that we do not have available the variables represented by the vector \( z_i \); these omitted variables are also said to be lurking. If these omitted variables do not depend on time, then it is still possible to get reliable estimators of other model parameters, such as those included in the vector \( \beta \). One strategy is to consider the deviations of a response from its time series average. This yields the derived model:

\[ y_{it}^* = y_{it} - \bar{y}_i = (\alpha_i + x_{it}' \beta + z_i' \gamma + \epsilon_{it}) - (\alpha_i + x_{it}' \beta + z_i' \gamma + \bar{\epsilon}_i) \]

\[ = (x_{it} - \bar{x}_i)' \beta + \epsilon_{it} - \bar{\epsilon}_i = x_{it}^* \beta + \epsilon_{it}^*. \]

Here, we use the response time series average, \( \bar{y}_i = T^{-1} \sum_{t=1}^{T} y_{it} \), and similarly for \( \bar{x}_i \) and \( \bar{\epsilon}_i \). Thus, ordinary least square estimators based on regressing the deviations in \( y \) on the deviations in \( x \) yields a desirable estimator of \( \beta \).

This strategy demonstrates how longitudinal data can mitigate the problem of omitted variable bias. For strategies that rely on purely cross-sectional data, it is well known that correlations of lurking variables, \( z \), with the model explanatory variables, \( x \), induce bias when estimating \( \beta \). If the lurking variable is time-invariant, then it is perfectly collinear with the subject-specific variables \( \alpha_i \). Thus, estimation strategies that account for subjects-specific parameters also account for time-invariant omitted variables. Further, because of the collinearity between subject-specific variables and time-invariant omitted variables, we may interpret the subject-specific quantities \( \alpha_i \) as proxies for omitted variables. Chapter 7 describes strategies for dealing with omitted variable bias.

**Efficiency of estimators**

A longitudinal data design may yield more efficient estimators than estimators based on a comparable amount of data from alternative designs. To illustrate, suppose that the interest is in assessing the average change in a response over time, such as the divorce rate. Thus, let \( \bar{y}_{s1} - \bar{y}_{s2} \) denote the difference between divorce rates between two time periods. In a repeated cross-sectional study such as the CPS, we would calculate the reliability of this statistic assuming independence among cross-sections to get

\[ \text{Var}(\bar{y}_{s1} - \bar{y}_{s2}) = \text{Var} \, \bar{y}_{s1} + \text{Var} \, \bar{y}_{s2}. \]

However, in a panel survey that tracks individuals over time, we have

\[ \text{Var}(\bar{y}_{s1} - \bar{y}_{s2}) = \text{Var} \, \bar{y}_{s1} + \text{Var} \, \bar{y}_{s2} - 2 \text{Cov}(\bar{y}_{s1}, \bar{y}_{s2}). \]

The covariance term is generally positive because observations from the same subject tend to be positively correlated. Thus, other things being equal, a panel survey design yields more efficient estimators than a repeated cross-section design.

One method of accounting for this positive correlation among same-subject observations is through the heterogeneity terms, \( \alpha_i \). In many data sets, introducing subject-specific variables \( \alpha_i \)
also accounts for a large portion of the variability. Accounting for this variation reduces the mean square error and standard errors associated with parameter estimators. Thus, we are more efficient in parameter estimation than the case without subject-specific variables $\alpha_i$.

It is also possible to incorporate subject-invariant parameters, often denoted by $\lambda_t$, to account for period (temporal) variation. For many data sets, this does not account for the same amount of variability as $\{\alpha_i\}$. With “small” numbers of time periods, it is straightforward to use time dummy (binary) variables to incorporate subject-invariant parameters.

Other things equal, standard errors become smaller and efficiency improves as the number of observations increases. For some situations, a researcher may obtain more information by sampling each subject repeatedly. Thus, some advocate that an advantage of longitudinal data is that we generally have more observations, due to the repeated sampling, and greater efficiency of estimators compared to a purely cross-sectional regression design. The danger of this philosophy is that generally observations from the same subject are related. Thus, although more information is obtained by repeated sampling, researchers need to be cautious in assessing the amount of additional information gained.

**Correlation and causation**

For many statistical studies, analysts are happy to describe associations among variables. This is particularly true of forecasting studies where the goal is to predict the future. However, for other analyses, researchers are interested in assessing causal relationships among variables.

Longitudinal and panel data are sometimes touted as providing “evidence” of causal effects. Just as with any statistical methodology, longitudinal data models in and of themselves are not enough to establish causal relationships among variables. However, longitudinal data can be more useful than purely cross-sectional data in establishing causality. To illustrate, consider the three ingredients necessary for establishing causality, taken from the sociology literature (see, for example, Toon, 2000):

- A statistically significant relationship is required.
- The association between two variables must not be due to another, omitted, variable.
- The “causal” variable must precede the other variable in time.

Longitudinal data are based on measurements taken over time and thus address the third requirement of a temporal ordering of events. Moreover, as described above, longitudinal data models provide additional strategies for accommodating omitted variables that are not available in purely cross-sectional data.

Observational data are not from carefully controlled experiments where random allocations are made among groups. Causal inference is not directly accomplished when using observational data and only statistical models. Rather, one thinks about the data and statistical models as providing relevant empirical evidence in a chain of reasoning about causal mechanisms. Although longitudinal data provide stronger evidence than purely cross-sectional data, most of the work in establishing causal statements should be based on the theory of the substantive field from which the data are derived. Chapter 6 discusses this issue in greater detail.

**Drawbacks: Attrition**

Longitudinal data sampling design offers many benefits compared to purely cross-sectional or purely time-series designs. However, because the sampling structure is more complex, it can also fail in subtle ways. The most common failure of longitudinal data sets to meet standard sampling design assumptions is through difficulties that result from attrition. In this context, attrition refers to a gradual erosion of responses by subjects. Because we follow the same subjects over time, nonresponse typically increases through time. To illustrate, consider the
US Panel Study of Income Dynamics (PSID). In the first year (1968), the nonresponse rate was 24%. However, by 1985, the nonresponse rate grew to about 50%.

Attrition can be a problem because it may result in a selection bias. Selection bias potentially occurs when a rule other than simple random (or stratified) sampling is used to select observational units. Examples of selection bias often concern endogenous decisions by agents to join a labor pool or participate in a social program. To illustrate, suppose that we are studying a solvency measure of a sample of insurance firms. If the firm becomes bankrupt or evolves into another type of financial distress, then we may not be able to examine financial statistics associated with the firm. Nonetheless, this is exactly the situation in which we would anticipate observing low values of the solvency measure. The response of interest is related to our opportunity to observe the subject, a type of selection bias. Chapter 7 discusses the attrition problem in greater detail.

1.3 Longitudinal data models

When examining the benefits and drawbacks of longitudinal data modeling, it is also useful to consider the types of inference that are based on longitudinal data models, as well as the variety of modeling approaches. The type of application under consideration influences the choice of inference and modeling approaches.

Types of inference

For many longitudinal data applications, the primary motivation for the analysis is to learn about the effect that an (exogenous) explanatory variable has on a response, controlling for other variables, including omitted variables. Users are interested in whether estimators of parameter coefficients, contained in the vector $\beta$, differ in a statistically significant fashion from zero. This is also the primary motivation for most studies that involve regression analysis; this is not surprising given that many models of longitudinal data are special cases of regression models.

Because longitudinal data are collected over time, they also provide us with an ability to predict future values of a response for a specific subject. Chapter 4 considers this type of inference, known as forecasting.

The focus of Chapter 4 is on the “estimation” of random variables, known as prediction. Because future values of a response are, to the analyst, random variables, forecasting is a special case of prediction. Another special case involves situations where we would like to predict the expected value of a future response from a specific subject, conditional on latent (unobserved) characteristics associated with the subject. For example, this conditional expected value is known in insurance theory as a credibility premium, a quantity that is useful in pricing of insurance contracts.

Social science statistical modeling

Statistical models are mathematical idealizations constructed to represent the behavior of data. When a statistical model is constructed (designed) to represent a data set with little regard to the underlying functional field from which the data emanates, we may think of the model as essentially data driven. For example, we might examine a data set of the form $(x_1, y_1), \ldots, (x_n, y_n)$ and posit a regression model to capture the association between $x$ and $y$. We will call this type of model a sampling based model, or following the econometrics literature, say that the model arises from the data generating process.

In most cases, however, we will know something about the units of measurement of $x$ and $y$ and anticipate a type of relationship between $x$ and $y$ based on knowledge of the functional field from which these variables arise. To continue our example in a finance context, suppose that $x$ represents a return from a market index and that $y$ represents a stock return from an individual
security. In this case, financial economics theory suggests a linear regression relationship of \( y \) on \( x \). In the economics literature, Goldberger (1972) defines a *structural model* to be a statistical model that represents causal relationships, as opposed to relationships that simply capture statistical associations. Chapter 6 further develops the idea of causal inference.

If a sampling based model adequately represents statistical associations in our data, then why bother with an extra layer of theory when considering statistical models? In the context of binary dependent variables, Manski (1992) offers three motivations: interpretation, precision and extrapolation.

Interpretation is important because the primary purpose of many statistical analyses is to assess relationships generated by theory from a scientific field. A sampling based model may not have sufficient structure to make this assessment, thus failing the primary motivation for the analysis.

Structural models utilize additional information from an underlying functional field. If this information is utilized correctly, then in some sense the structural model should provide a better representation than a model without this information. With a properly utilized structural model, we anticipate getting more precise estimates of model parameters and other characteristics. In practical terms, this improved precision can be measured in terms of smaller standard errors.

At least in the context of binary dependent variables, Manski (1992) feels that extrapolation is the most compelling motivation for combining theory from a functional field with a sampling based model. In a time series context, extrapolation means forecasting; this is generally the main impetus for an analysis. In a regression context, extrapolation means inference about responses for sets of predictor variables “outside” of those realized in the sample. Particularly for public policy analysis, the goal of a statistical analysis is to infer the likely behavior of data outside of those realized.

### Modeling issues

This chapter has portrayed longitudinal data modeling as a special type of regression modeling. However, in the biometrics literature, longitudinal data models have their roots in multivariate analysis. Under this framework, we view the responses from an individual as a vector of responses, that is, \( y_i = (y_{i1}, y_{i2}, \ldots, y_{iT}) \). Within the biometrics framework, the first applications are referred to as *growth curve* models. These classic examples use the height of children as the response to examine the changes in height and growth, over time; see Chapter 5. Within the econometrics literature, Chamberlain (1982, 1984) exploited the multivariate structure. The multivariate analysis approach is most effective with balanced data at equally spaced in time points. However, compared to the regression approach, there are several limitations of the multivariate approach. These include:

- It is harder to analyze “missing” data, attrition, and different accrual patterns.
- Because there is no explicit allowance for time, it is harder to forecast and predict at time points between those collected (interpolation).

Even within the regression approach for longitudinal data modeling, there are still a number of issues that need to be resolved in choosing a model. We have already introduced the issue of modeling heterogeneity. Recall that there are two important types of models of heterogeneity, fixed and random effects models, the subjects of Chapters 2 and 3.

Another important issue is the structure for modeling the dynamics; this is the subject of Chapter 8. We have described imposing a serial correlation on the disturbance terms. Another approach, described in Section 8.2, involves using lagged (endogenous) responses to account for temporal patterns. These models are important in econometrics because they are more suitable for structural modeling where there is a greater tie between economic theory and statistical modeling.
than models that are based exclusively on features of the data. When the number of (time) observations per subject, \( T \), is small, then simple correlation structures of the disturbances terms provide an adequate fit for many data sets. However, as \( T \) increases, we have greater opportunities to model the dynamic structure. The Kalman filter, described in Section 8.5, provides a computational technique that allows the analyst to handle a broad variety of complex dynamic patterns.

Many of the longitudinal data applications that appear in the literature are based on linear model theory. Hence, this text is predominantly (Chapters 1 through 8) devoted to developing linear longitudinal data models. However, nonlinear models represent an area of recent development where examples of their importance to statistical practice appear with greater frequency. The phrase “nonlinear models” in this context refers to instances where the distribution of the response cannot be reasonably approximated using a normal curve. Some examples of this occur when the response is binary or other types of count data, such as the number of accidents in a state, and when the response is from a very heavy tailed distribution, such as with insurance claims. Chapters 9 through 11 introduce techniques from this budding literature to handle these types of nonlinear models.

**Types of applications**

A statistical model is ultimately useful only if it provides a useful approximation to real data. Table 1.1 outlines the data sets used in this text to underscore the importance of longitudinal data modeling.
<table>
<thead>
<tr>
<th>Data Title</th>
<th>Subject Area</th>
<th>File Name</th>
<th>Unit of Analysis</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airline</td>
<td>Finance</td>
<td>Airline</td>
<td>Subjects are (n=19) airlines over (T=11) years: 1970-1980. (N=187) observations.</td>
<td>Examine characteristics of airlines to determine total operating costs.</td>
</tr>
<tr>
<td>Bond Maturity</td>
<td>Finance</td>
<td>Bondmat</td>
<td>Subjects are (n=328) firms over (T=10) years: 1980-1989. (N=3,280) observations.</td>
<td>Examine the maturity of debt structure in terms of corporate financial characteristics.</td>
</tr>
<tr>
<td>Capital Structure</td>
<td>Finance</td>
<td>Capital</td>
<td>Subjects are (n=361) Japanese firms over (T=15) years: 1984-1998. (N=5,415) observations.</td>
<td>Examine changes capital structure before and after the market crash for different types of cross holding structures.</td>
</tr>
<tr>
<td>Charitable Contributions</td>
<td>Accounting</td>
<td>Charity</td>
<td>Subjects are (n=47) taxpayers over (T=10) years; 1979-1988. (N=470) observations.</td>
<td>Examine characteristics of taxpayers to determine factors that influence the amount of charitable giving.</td>
</tr>
<tr>
<td>Divorce</td>
<td>Sociology</td>
<td>Divorce</td>
<td>Subjects are (n=51) states over (T=4) years: 1965, 1975, 1985 and 1995. (N=204) observations.</td>
<td>Assess socioeconomic variables that affect the divorce rate.</td>
</tr>
<tr>
<td>Electric Utilities</td>
<td>Economics</td>
<td>Electric</td>
<td>Subjects are (n=68) (electric) utilities over (T=12) months. (N=816) observations.</td>
<td>Examine the average cost of utilities in terms of the price of labor, fuel and capital.</td>
</tr>
<tr>
<td>Group Term Life Data</td>
<td>Insurance</td>
<td>Glife</td>
<td>Subjects are (n=106) credit unions over (T=7) years. (N=742) observations.</td>
<td>Forecast group term life insurance claims of Florida credit unions.</td>
</tr>
<tr>
<td>Housing Prices</td>
<td>Real estate</td>
<td>Hprice</td>
<td>Subjects are (n=36) metropolitan statistical areas (MSAs) over (T=9) years: 1986-1994. (N=324) observations.</td>
<td>Examine annual housing prices in terms of MSA demographic and economic indices.</td>
</tr>
<tr>
<td>Lottery Sales</td>
<td>Marketing</td>
<td>Lottery</td>
<td>Subjects are (n=50) postal code areas over (T=40) weeks.</td>
<td>Examine effects of area economic and demographic characteristics on lottery sales.</td>
</tr>
<tr>
<td>Medicare Hospital Costs</td>
<td>Social Insurance</td>
<td>Medicare</td>
<td>Subjects are (n=54) states over (T=6) years: 1990-1995. (N=324) observations.</td>
<td>Forecast Medicare hospital costs by state based on utilization rates and past history.</td>
</tr>
<tr>
<td>Property and Liability Insurance</td>
<td>Insurance</td>
<td>Pdemand</td>
<td>Subjects are (n=22) countries over (T=7) years: 1987-1993. (N=154) observations.</td>
<td>Examine the demand for property and liability insurance in terms of national economic and risk aversion characteristics.</td>
</tr>
<tr>
<td>Student Achievement</td>
<td>Education</td>
<td>Student</td>
<td>Subjects are (n=400) students from 20 schools are observed over (T=4) grades (3-6). (N=1,012) observations.</td>
<td>Examine student math achievement based on student and school demographic and socioeconomic characteristics.</td>
</tr>
<tr>
<td>Tax Preparers</td>
<td>Accounting</td>
<td>Taxprep</td>
<td>Subjects are (n=243) taxpayers over (T=5) years: 1982, 1984-1988. (N=1,215) observations.</td>
<td>Examine characteristics of taxpayers to determine the demand for a professional tax preparer.</td>
</tr>
<tr>
<td>Tort Filings</td>
<td>Insurance</td>
<td>Tiling</td>
<td>Subjects are (n=19) states over (T=6) years: 1984-1989. (N=114) observations.</td>
<td>Examine demographic and legal characteristics of states that influence the number of tort filings.</td>
</tr>
<tr>
<td>Worker’s Compensation</td>
<td>Insurance</td>
<td>Workerc</td>
<td>Subjects are (n=121) occupation classes over (T=7) years. (N=847) observations.</td>
<td>Forecast worker’s compensation claims by occupation class.</td>
</tr>
</tbody>
</table>
1.4 Historical notes

The term ‘panel study’ was coined in a marketing context when Lazarsfeld and Fiske (1938O) considered the effect of radio advertising on product sales. Traditionally, hearing radio advertisements was thought to increase the likelihood of purchasing a product. Lazarsfeld and Fiske considered whether those that bought the product would be more likely to hear the advertisement, thus positing a reverse in the direction of causality. They proposed repeatedly interviewing a set of people (the ‘panel’) to clarify the issue.

Baltes and Nesselroade (1979EP) trace the history of longitudinal data and methods with an emphasis on childhood development and psychology. They describe longitudinal research as consisting of “a variety of methods connected by the idea that the entity under investigation is observed repeatedly as it exists and evolves over time.” Moreover, they trace the need for longitudinal research to at least as early as the nineteenth century.

Toon (2000EP) cites Engel’s 1857 budget survey, examining how the amount of money spent on food changes as a function of income, as perhaps the earliest example of a study involving repeated measurements from the same set of subjects.

As noted in Section 1.2, in early panel data studies, pooled cross-sectional data were analyzed by estimating cross-sectional parameters using regression and using time series methods to model the regression parameter estimates, treating the estimates as known with certainty. Dielman (1989O) discusses this approach in more detail and provides examples. Early applications in economics of the basic fixed effects model include Kuh (1959E), Johnson (1960E), Mundlak (1961E) and Hoch (1962E). Chapter 2 introduces this and related models in detail.

Balestra and Nerlove (1966E) and Wallace and Hussain (1969E) introduced the (random effects) error components model, the model with \( \{ \alpha_i \} \) as random variables. Chapter 3 introduces this and related models in detail.

Wishart (1938B), Rao (1959S, 1965B), Potthoff and Roy (1964B) were among the first contributions in the biometrics literature to use multivariate analysis for analyzing growth curves. Specifically, they considered the problem of fitting polynomial growth curves of serial measurements from a group of subjects. Chapter 5 contains examples of growth curve analysis.

This approach to analyzing longitudinal data was extended by Grizzle and Allen (1969B), who introduced covariates, or explanatory variables, into the analysis. Laird and Ware (1982B) made the other important transition from multivariate analysis to regression modeling. They introduce the two-stage model that allows for both fixed and random effects. Chapter 3 considers this modeling approach.
Chapter 2. Fixed Effects Models

Abstract. This chapter introduces the analysis of longitudinal and panel data using the general linear model framework. Here, longitudinal data modeling is cast as a regression problem by using fixed parameters to represent the heterogeneity; nonrandom quantities that account for the heterogeneity are known as fixed effects. In this way, ideas of model representation and data exploration are introduced using regression analysis, a toolkit that is widely known. Analysis of covariance, from the general linear model, easily handles the many parameters that are needed to represent the heterogeneity.

Although longitudinal and panel data can be analyzed using regression techniques, it is also important to emphasize the special features of these data. Specifically, the chapter emphasizes the wide cross-section and the short time-series of many longitudinal and panel data sets, as well as the special model specification and diagnostic tools needed to handle these features.

2.1 Basic fixed effects model

Data

Suppose that we are interested in explaining hospital costs for each state in terms of measures of utilization, such as the number of discharged patients and the average hospital stay per discharge. Here, we consider the state to be the unit of observation, or subject. We differentiate among states with the index \( i \), where \( i \) may range from 1 to \( n \), and \( n \) is the number of subjects. Each state is observed \( T_i \) times and we use the index \( t \) to differentiate the observation times. With these indices, let \( y_{it} \) denote the response of the \( i \)th subject at the \( t \)th time point. Associated with each response \( y_{it} \) is a set of explanatory variables, or covariates. For example, for state hospital costs, these explanatory variables include the number of discharged patients and the average hospital stay per discharge. In general, we assume there are \( K \) explanatory variables \( x_{it,1}, x_{it,2}, \ldots, x_{it,K} \) that may vary by subject \( i \) and time \( t \). We achieve a more compact notational form by expressing the \( K \) explanatory variables as a \( K \times 1 \) column vector

\[
\mathbf{x}_{it} = \begin{pmatrix} x_{it,1} \\ x_{it,2} \\ \vdots \\ x_{it,K} \end{pmatrix}.
\]

To save space, it is customary to use the alternate expression \( \mathbf{x}_{it} = (x_{it,1}, x_{it,2}, \ldots, x_{it,K})' \), where the prime “’” means transpose. (You will find that some sources prefer to use a superscript “\( T \)” for transpose. Here, \( T \) will refer to the number of time replications.) Thus, the data for the \( i \)th subject consists of:

\[
\{x_{i1,1}, \ldots, x_{i1,K}, y_{i1}\} \\
\{x_{i2,1}, \ldots, x_{i2,K}, y_{i2}\} \\
\vdots \\
\{x_{iT,1}, \ldots, x_{iT,K}, y_{iT}\}.
\]
that can be expressed more compactly as
\[
\{x_{i1}', y_{i1} \},
\ldots,
\{x_{iT}', y_{iT} \}.
\]

Unless specified otherwise, we allow the number of responses to vary by subject, indicated with the notation \(T_i\). This is known as the *unbalanced* case. We use the notation \(T = \max \{T_1, T_2, \ldots, T_n\}\) to be the maximal number of responses for a subject. Recall from Section 1.1 that the case \(T_i = T\) for each \(i\) is called the *balanced* case.

**Basic models**

To analyze relationships among variables, the relationships between the response and the explanatory variables are summarized through the *regression function*
\[
E y_{it} = \alpha + \beta_1 x_{it, 1} + \beta_2 x_{it, 2} + \ldots + \beta_K x_{it, K},
\]  
(2.1)

that is linear in the parameters \(\alpha, \beta_1, \ldots, \beta_K\). For applications where the explanatory variables are nonrandom, the only restriction of equation (2.1) is that we believe that the variables enter linearly. As we will see in Chapter 6, for applications where the explanatory variables are random, we may interpret the expectation in equation (2.1) as conditional on the observed explanatory variables.

We focus attention on assumptions that concern the observable variables, \(\{x_{it, 1}, \ldots, x_{it, K}, y_{it}\}\).

**Assumptions of the Observables Representation of the Linear Regression Model**

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1.</td>
<td>(E y_{it} = \alpha + \beta_1 x_{it, 1} + \beta_2 x_{it, 2} + \ldots + \beta_K x_{it, K}).</td>
</tr>
<tr>
<td>F2.</td>
<td>({x_{it, 1}, \ldots, x_{it, K}}) are nonstochastic variables.</td>
</tr>
<tr>
<td>F3.</td>
<td>(\text{Var } y_{it} = \sigma^2).</td>
</tr>
<tr>
<td>F4.</td>
<td>({y_{it}}) are independent random variables.</td>
</tr>
</tbody>
</table>

The “observables representation” is based on the idea of conditional linear expectations (see Goldberger, 1991, for additional background). One can motivate assumption F1 by thinking of \(\{x_{it, 1}, \ldots, x_{it, K}, y_{it}\}\) as a draw from a population, where the mean of the conditional distribution of \(y_{it}\) given \(\{x_{it, 1}, \ldots, x_{it, K}\}\) is linear in the explanatory variables. Inference about the distribution of \(y\) is conditional on the observed explanatory variables, so that we may treat \(\{x_{it, 1}, \ldots, x_{it, K}\}\) as nonstochastic variables. When considering types of sampling mechanisms for thinking of \(\{x_{it, 1}, \ldots, x_{it, K}, y_{it}\}\) as a draw from a population, it is convenient to think of a stratified random sampling scheme, where values of \(\{x_{it, 1}, \ldots, x_{it, K}\}\) are treated as the strata. That is, for each value of \(\{x_{it, 1}, \ldots, x_{it, K}\}\), we draw a random sample of responses from a population. This sampling scheme also provides motivation for assumption F4, the independence among responses. To illustrate, when drawing from a database of firms to understand stock return performance \(y\), one can choose large firms, measured by asset size, focus on an industry, measured by standard industrial classification and so forth. You may not select firms with the largest stock return performance because this is stratifying based on the response, not the explanatory variables.

A fifth assumption that is often implicitly required in the linear regression model is:

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>F5.</td>
<td>({y_{it}}) is normally distributed.</td>
</tr>
</tbody>
</table>

This assumption is not required for all statistical inference procedures because central limit theorems provide approximate normality for many statistics of interest. However, formal justification for some, such as \(t\)-statistics, do require this additional assumption.
Chapter 2. Fixed Effects Models / 2-3

In contrast to the observables representation, the classical formulation of the linear regression model focuses attention on the “errors” in the regression, defined as \( e_{it} = y_{it} - (\alpha + \beta_1 x_{it,1} + \beta_2 x_{it,2} + \cdots + \beta_K x_{it,K}) \).

### Assumptions of the Error Representation of the Linear Regression Model

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>( y_{it} = \alpha + \beta_1 x_{it,1} + \beta_2 x_{it,2} + \cdots + \beta_K x_{it,K} + \epsilon_{it} ) where ( E \epsilon_{it} = 0 ).</td>
</tr>
<tr>
<td>E2</td>
<td>( {x_{it,1}, \ldots, x_{it,K}} ) are nonstochastic variables.</td>
</tr>
<tr>
<td>E3</td>
<td>( \text{Var} \epsilon_{it} = \sigma^2 ).</td>
</tr>
<tr>
<td>E4</td>
<td>( { \epsilon_{it} } ) are independent random variables.</td>
</tr>
</tbody>
</table>

The “error representation” is based on the Gaussian theory of errors (see Stigler, 1986, for a historical background). As described above, the linear regression function incorporates the additional knowledge from independent variables through the relation \( E y_{it} = \alpha + \beta_1 x_{it,1} + \beta_2 x_{it,2} + \cdots + \beta_K x_{it,K} \). Other unobserved variables that influence the measurement of \( y \) are encapsulated in the “error” term \( \epsilon_{it} \), which is also known as the “disturbance” term. The independence of errors, F4, can be motivated by assuming that \( \{ \epsilon_{it} \} \) are realized through a simple random sample from an unknown population of errors.

Assumptions E1-E4 are equivalent to assumptions F1-F4. The error representation provides a useful springboard for motivating goodness of fit measures. However, a drawback of the error representation is that it draws the attention from the observable quantities \( (x_{it,1}, \ldots, x_{it,K}, y_{it}) \) to an unobservable quantity, \( \{ e_{it} \} \). To illustrate, the sampling basis, viewing \( \{ e_{it} \} \) as a simple random sample, is not directly verifiable because one cannot directly observe the sample \( \{ e_{it} \} \). Moreover, the assumption of additive errors in E1 will be troublesome when we consider nonlinear regression models in Part II. Our treatment focuses on the observable representation in Assumptions F1-F4.

In assumption F1, the slope parameters \( \beta_1, \beta_2, \ldots, \beta_K \) are associated with the \( K \) explanatory variables. For a more compact expression, we summarize the parameters as a column vector of dimension \( K \times 1 \), denoted by

\[
\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \end{bmatrix}.
\]

With this notation, we may re-write assumption F1 as

\[
E y_{it} = \alpha + x_{it}' \beta,
\]

because of the relation \( x_{it}' \beta = \beta_1 x_{it,1} + \beta_2 x_{it,2} + \cdots + \beta_K x_{it,K} \). We call the representation in equation (2.2) cross-sectional because, although it relates the explanatory variables to the response, it does not use the information in the repeated measurements on a subject. Because it also does not include (subject-specific) heterogeneous terms, we also refer to the equation (2.2) representation as part of a homogeneous model.

Our first representation that uses the information in the repeated measurements on a subject is

\[
E y_{it} = \alpha_i + x_{it}' \beta.
\]

Equation (2.3) and assumptions F2-F4 comprise the basic fixed effects model. Unlike equation (2.2), in equation (2.3) the intercept terms, \( \alpha_i \), are allowed to vary by subject.

### Parameters of interest

The parameters \( \{ \beta \} \) are common to each subject and are called global, or population, parameters. The parameters \( \{ \alpha_i \} \) vary by subject and are known as individual, or subject-specific,
parameters. In many applications, we will see that population parameters capture broad relationships of interest and hence are the parameters of interest. The subject-specific parameters account for the different features of subjects, not broad population patterns. Hence, they are often of secondary interest and are called nuisance parameters.

As we saw in Section 1.3, the subject-specific parameters represent our first device that helps control for the heterogeneity among subjects. We will see that estimators of these parameters use information in the repeated measurements on a subject. Conversely, the parameters \{α_i\} are non-estimable in cross-sectional regression models without repeated observations. That is, with \(T_i = 1\), the model
\[
y_{it} = \alpha_i + \beta_1 x_{i1,t} + \beta_2 x_{i2,t} + \ldots + \beta_K x_{iK,t} + \epsilon_{it}
\]
has more parameters \((n+K)\) than observations \((n)\) and thus, we cannot identify all the parameters. Typically, the disturbance term \(\epsilon_{it}\) includes the information in \(\alpha_i\) in cross-sectional regression models. An important advantage of longitudinal data models when compared to cross-sectional regression models is the ability to separate the effects of \{\alpha_i\} from the disturbance terms \{\epsilon_{it}\}. By separating out subject-specific effects, our estimates of the variability become more precise and we achieve more accurate inferences.

**Subject and time heterogeneity**

We will argue that the subject-specific parameter \(\alpha_i\) captures much of the time-constant information in the responses. However, the basic fixed effects model assumes that \{\(y_{it}\)\} are independent terms and, in particular, that there is:

- no serial correlation (correlation over time) and
- no contemporaneous correlation (correlation across subjects).

Thus, no special relationships between subjects and time periods are assumed. By interchanging the roles of “\(i\)” and “\(t\)”, we may consider the function
\[
E y_{it} = \lambda_t + x_{it}' \beta.
\]  
(2.4)

Here, the parameter \(\lambda_t\) is a time-specific variable that does not depend on subjects.

For most longitudinal data applications, the number of subjects, \(n\), substantially exceeds the maximal number of time periods, \(T\). Further, generally the heterogeneity among subjects explains a greater proportion of variability than the heterogeneity among time periods. Thus, we begin with the “basic” function \(E y_{it} = \alpha_i + x_{it}' \beta\). This model allows explicit parameterization of the subject-specific heterogeneity.

Both functions in equations (2.3) and (2.4) are based on traditional one-way analysis of covariance models. For this reason, the basic fixed effects model is also called the one-way fixed effects model. By using binary (dummy) variables for the time dimension, we can incorporate time-specific parameters into the population parameters. In this way, it is straightforward to consider the function
\[
E y_{it} = \alpha_i + \lambda_t + x_{it}' \beta.
\]  
(2.5)

Equation (2.5) with assumptions F2-F4 is called the two-way fixed effects model.

**Example 2.1 – Urban wages**

Glaeser and Maré (2001) investigated the effects of determinants on wages, with the goal of understanding why workers in cities earn more than their non-urban counterparts. They examined two-way fixed effects models using data from the National Longitudinal Survey of Youth (NLSY); they also used data from the Panel Study of Income Dynamics (PSID) to assess
the robustness of their results to another sample. For the NLSY data, they examined \( n = 5,405 \) male heads of households over the years 1983-1993, consisting of a total of \( N = 40,194 \) observations. The dependent variable was logarithmic hourly wage. The primary explanatory variable of interest was a 3-level categorical variable that measures the city size in which workers reside. To capture this variable, two binary (dummy) variables were used: (1) a variable to indicate whether the worker resides in a large city (with more than one-half million residents), a “dense metropolitan area,” and (2) a variable to indicate whether the worker resides in a metropolitan area that does not contain a large city, a “non-dense metropolitan area.” The omitted category is non-metropolitan area. Several other control variables were included to capture effects of a worker’s experience, occupation, education and race. When including time dummy variables, there were \( K = 30 \) explanatory variables in the reported regressions.

### 2.2 Exploring longitudinal data

**Why explore?**

The models that we use to represent reality are simplified approximations. As stated by George Box (1979G), “All models are wrong, but some are useful.” The inferences that we draw by examining a model calibrated with a data set depends on the data characteristics; we expect a reasonable proximity between the model assumptions and the data. To assess this proximity, we explore the many important features of the data. By data exploration, we mean summarizing the data, either numerically or graphically, without reference to a model.

Data exploration provides hints of the appropriate model. To draw reliable inferences from the modeling procedure, it is important that the data be congruent with the model. Further, exploring the data also alerts us to any unusual observations or subjects. Because standard inference techniques described are generally non-robust to unusual features, it is important to identify these features early in the modeling process.

Data exploration also provides an important communication device. Because data exploration techniques are not model dependent, they may be better understood than model dependent inference techniques. Thus, they can be used to communicate features of a data set, often supplementing model based inferences.

**Data exploration techniques**

Longitudinal data analysis is closely linked to multivariate analysis and regression analysis. Thus, the data exploration techniques developed in these fields are applicable to longitudinal data and will not be developed here. The reader may consult Tukey (1977G) for the original source on exploratory data analysis. To summarize, the following is a list of commonly used data exploration techniques that will be demonstrated throughout this book:

- Examine graphically the distribution of \( y \) and each \( x \) through histograms, density estimates, boxplots and so on.
- Examine numerically the distribution of \( y \) and each \( x \) through statistics such as means, medians, standard deviations, minimums, maximums and so on.
- Examine the relationship between \( y \) and each \( x \) through correlations and scatter plots.

Further, summary statistics and graphs by time period may be useful for detecting temporal patterns.

Three data exploration techniques that are specific to longitudinal data are (1) multiple time series plots, (2) scatter plots with symbols and (3) added variable plots. Because these techniques are specific to longitudinal data analysis, they are less widely known and described below. Another way to examine data is through diagnostic techniques, described in Section 2.4.
In contrast to data exploration techniques, diagnostic techniques are performed after the fit of a preliminary model.

**Multiple time series plots**
A multiple time series plot is a plot of a variable, generally the response, $y_{it}$, versus time $t$. Within the context of longitudinal data, we serially (over time) connect observations over a common subject. This graph helps (1) detect patterns in the response, by subject and over time, (2) identify unusual observations and/or subjects and (3) visualize the heterogeneity.

**Scatter plots with symbols**
In the context of regression, a plot of the response, $y_{it}$, versus an explanatory variable, $x_{ijt}$, helps us to assess the relationship between these variables. In the context of longitudinal data, it is often useful to add a plotting symbol to the scatter plot to identify the subject. This allows us to see the relationship between the response and explanatory variable yet account for the varying intercepts. Further, if there is a separation in the explanatory variable, such as increasing over time, then we can serially connect the observations. In this case, we may not require a separate plotting symbol for each subject.

**Basic added variable plot**
A basic added variable plot is a scatter plot of $\{y_{it} - \bar{y}_i\}$ versus $\{x_{ijt} - \bar{x}_{ij}\}$. Here, $\bar{y}_i$ and $\bar{x}_{ij}$ are averages of $\{y_{it}\}$ and $\{x_{ijt}\}$ over time. An added variable plot is a standard regression diagnostic technique that is described in further detail in Section 2.4. Although the basic added variable plot can be viewed as a special case of the more general diagnostic technique, it can also be motivated without reference to a model. That is, in many longitudinal data sets, the subject-specific parameters account for a large portion of the variability. This plot allows us to visualize the relationship between $y$ and each $x$, without forcing our eye to adjust for the heterogeneity of the subject-specific intercepts.

**Example: Medicare hospital costs**
We consider $T=6$ years, 1990-1995, of data for inpatient hospital charges that are covered by the Medicare program. The data were obtained from the Health Care Financing Administration. To illustrate, in 1995 the total covered charges were $157.8 billions for twelve million discharges. For this analysis, we use state as the subject, or risk class. Thus, we consider $n=54$ states that include the 50 states in the Union, the District of Columbia, Virgin Islands, Puerto Rico and an unspecified “other” category.

The response variable of interest is the severity component, covered claims per discharge, which we label as CCPD. The variable CCPD is of interest to actuaries because the Medicare program reimburses hospitals on a per-stay basis. Also, many managed care plans reimburse hospitals on a per-stay basis. Because CCPD varies over state and time, both the state and time (YEAR=1, ..., 6) are potentially important explanatory variables. We do not assume a priori that frequency is independent of severity. Thus, number of discharges, NUM_DSCHG, is another potential explanatory variable. We also investigate the importance of another component of hospital utilization, AVE_DAYS, defined to be the average hospital stay per discharge in days.

Table 2.1 summarizes these basic variables, by year. Here, we see that both claims and number of discharges increase over time whereas the average hospital stay decreases. The standard deviations and extreme values indicate that there is substantial variability among states.
TABLE 2.1. Summary Statistics of Covered Claims Per Discharge, Number of Discharges and Average Hospital Stay, by Year.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Time Period</th>
<th>Mean</th>
<th>Median</th>
<th>Standard Deviation</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covered Claims per Discharge (CCPD)</td>
<td>1990</td>
<td>8,503</td>
<td>7,992</td>
<td>2,467</td>
<td>3,229</td>
<td>16,485</td>
</tr>
<tr>
<td></td>
<td>1991</td>
<td>9,473</td>
<td>9,113</td>
<td>2,712</td>
<td>2,966</td>
<td>17,637</td>
</tr>
<tr>
<td></td>
<td>1992</td>
<td>10,443</td>
<td>10,055</td>
<td>3,041</td>
<td>3,324</td>
<td>19,814</td>
</tr>
<tr>
<td></td>
<td>1993</td>
<td>11,160</td>
<td>10,667</td>
<td>3,260</td>
<td>4,138</td>
<td>21,122</td>
</tr>
<tr>
<td></td>
<td>1994</td>
<td>11,523</td>
<td>10,955</td>
<td>3,346</td>
<td>4,355</td>
<td>21,500</td>
</tr>
<tr>
<td></td>
<td>1995</td>
<td>11,797</td>
<td>11,171</td>
<td>3,278</td>
<td>5,058</td>
<td>21,032</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>10,483</td>
<td>10,072</td>
<td>3,231</td>
<td>2,966</td>
<td>21,500</td>
</tr>
<tr>
<td>Number of Discharges (NUM_DSCHG)</td>
<td>1990</td>
<td>197.73</td>
<td>142.59</td>
<td>202.99</td>
<td>0.53</td>
<td>849.37</td>
</tr>
<tr>
<td></td>
<td>1991</td>
<td>203.14</td>
<td>142.69</td>
<td>210.38</td>
<td>0.52</td>
<td>885.92</td>
</tr>
<tr>
<td></td>
<td>1992</td>
<td>210.89</td>
<td>143.25</td>
<td>218.92</td>
<td>0.65</td>
<td>908.59</td>
</tr>
<tr>
<td></td>
<td>1993</td>
<td>211.25</td>
<td>143.67</td>
<td>219.82</td>
<td>0.97</td>
<td>894.22</td>
</tr>
<tr>
<td></td>
<td>1994</td>
<td>218.87</td>
<td>150.08</td>
<td>226.78</td>
<td>1.16</td>
<td>905.62</td>
</tr>
<tr>
<td></td>
<td>1995</td>
<td>222.51</td>
<td>152.70</td>
<td>229.46</td>
<td>1.06</td>
<td>902.48</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>210.73</td>
<td>144.28</td>
<td>216.72</td>
<td>0.52</td>
<td>908.59</td>
</tr>
<tr>
<td>Average Hospital Stay (AVE_DAYS)</td>
<td>1990</td>
<td>9.05</td>
<td>8.53</td>
<td>2.08</td>
<td>6.33</td>
<td>17.48</td>
</tr>
<tr>
<td></td>
<td>1991</td>
<td>9.82</td>
<td>8.57</td>
<td>7.23</td>
<td>6.14</td>
<td>60.25</td>
</tr>
<tr>
<td></td>
<td>1992</td>
<td>8.62</td>
<td>8.36</td>
<td>1.86</td>
<td>5.83</td>
<td>16.35</td>
</tr>
<tr>
<td></td>
<td>1993</td>
<td>8.52</td>
<td>8.11</td>
<td>2.11</td>
<td>5.83</td>
<td>17.14</td>
</tr>
<tr>
<td></td>
<td>1994</td>
<td>7.90</td>
<td>7.56</td>
<td>1.73</td>
<td>5.38</td>
<td>14.39</td>
</tr>
<tr>
<td></td>
<td>1995</td>
<td>7.34</td>
<td>7.14</td>
<td>1.44</td>
<td>5.12</td>
<td>12.80</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>8.54</td>
<td>8.07</td>
<td>3.47</td>
<td>5.12</td>
<td>60.25</td>
</tr>
</tbody>
</table>

Notes: The variable CCPD is in dollars of claim per discharge. Each year summarizes \( n=54 \) states. The total summarizes (6*54=) 324 observations.

Source: Center for Medicare and Medicaid Services

Figure 2.1 illustrates the multiple time series plot. Here, we see that not only are overall claims increasing but also that claims increase for each state. Different levels of hospitals costs among states are also apparent; we call this feature heterogeneity. Further, Figure 2.1 indicates that there is greater variability among states than over time.

Figure 2.2 illustrates the scatter plot with symbols. This is a plot of CCPD versus number of discharges, connecting observations over time. This plot shows a positive overall relationship between CCPD and the number of discharges. Like CCPD, we see a substantial state variation of different numbers of discharges. Also like CCPD, the number of discharges increases over time, so that, for each state, there is a positive relationship between CCPD and number of discharges. The slope is higher for those states with smaller number of discharges. This plot also suggests that the number of discharges lagged by one year is an important predictor of CCPD.

Figure 2.3 is a scatter plot of CCPD versus average total days, connecting observations over time. This plot demonstrates the unusual nature of the second observation for the 54th state. We also see evidence of this point through the maximum statistic of the average hospital stay in Table 2.1. This point does not appear to follow the same pattern as the rest of our data and turns out to have a large impact on our fitted models.

Figure 2.4 illustrates the basic added variable plot. This plot portrays CCPD versus year, after excluding the second observation for the 54th state. In Figure 2.4 we have controlled for the state factor that we observed to be an important source of variation. Figure 2.4 shows that the rate of increase of CCPD over time is approximately consistent among states, yet there exists important variations. The rate of increase is substantially larger for the 31st state (New Jersey).
Figure 2.1 Multiple Time Series Plot of CCPD. Covered claims per discharge (CCPD) are plotted over $T=6$ years, 1990-1995. The line segments connect states; thus, we see that CCPD increases for almost every state over time.

Figure 2.2 Scatter Plot of CCPD versus Number of Discharges. The line segments connect observations within a state over 1990-1995. We see a substantial state variation of numbers of discharges. There is a positive relationship between CCPD and number of discharges for each state. Slopes are higher for those states with smaller number of discharges.
Figure 2.3 Scatter Plot of CCPD versus Average Hospital Stay. The line segments connect states over 1990-1995. This figure demonstrates that the second observation for the 54th state is unusual.

Figure 2.4 Added Variable Plot of CCPD versus Year. Here, we have controlled for the state factor. In this figure, the second observation for the 54th state has been excluded. We see that the rate of increase of CCPD over time is approximately consistent among states, yet there exists important variations. The rate of increase is substantially large for the 31st state (New Jersey).
Trellis plot

A technique for graphical display that has recently become popular in the statistical literature is a *trellis plot*. This graphical technique takes its name from a trellis which is a structure of open latticework. When viewing a house or garden, one typically thinks of a trellis as being used to support creeping plants such as vines. We will use this lattice structure and refer to a trellis plot as consisting of one or more panels arranged in a rectangular array. Graphs that contain multiple versions of a basic graphical form, each version portraying a variation of the basic theme, promote comparisons and assessments of change. By repeating a basic graphical form, we promote the process of communication. Trellis plots have been advocated by Cleveland (1993G), Becker, Cleveland and Shyu (1996G), Venables and Ripley (1999G) and by Pinheiro and Bates (2000S).

Tufte (1997G) states that using small multiples in graphical displays achieves the same desirable effects as using parallel structure in writing. Parallel structure in writing is successful because it allows readers to identify a sentence relationship only once and then focus on the meaning of each individual sentence element, such as a word, phrase or clause. Parallel structure helps achieve economy of expression and draw together related ideas for comparison and contrast. Similarly, small multiples in graphs allow us to visualize complex relationships across different groups and over time.

Figure 2.5 illustrates the use of small multiples. In each panel, the plot portrayed is identical except that it is based on a different state; this use of parallel structure allows us to demonstrate the increasing covered claims per discharge (CCPD) for each state. Moreover, by organizing the states by average CCPD, we can see the overall level of CCPD for each state as well as variations in the slope (rate of increase). This plot was produced using the statistical package “R.”

![Trellis Plot of CCPD versus Year](image_url)

**Figure 2.5 Trellis Plot of CCPD versus Year.** Each of the 54 panels represents a plot of CCPD versus YEAR, 1990-1995 (the horizontal axis is suppressed). State 31 corresponds to New Jersey.
2.3 Estimation and inference

Least squares estimation

Returning to our model in equation (2.3), we now consider estimation of the regression coefficients $\beta$ and $\alpha_i$ and then the variance parameter $\sigma^2$. By the Gauss-Markov theorem, the best linear unbiased estimators of $\beta$ and $\alpha_i$ are the ordinary least squares (OLS) estimators. These are given by

$$
\mathbf{b} = \left( \sum_{i=1}^{n} \sum_{t=1}^{T_i} (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)' \right)^{-1} \left( \sum_{i=1}^{n} \sum_{t=1}^{T_i} (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\mathbf{y}_{it} - \bar{\mathbf{y}}_i) \right),
$$

where $\mathbf{b} = (b_1, b_2, \ldots, b_K)'$, and

$$
\alpha_i = \bar{y}_i - \bar{\mathbf{x}}_i' \mathbf{b}.
$$

The derivations of these estimators are in Appendix 2A.1.

Note to Reader: We now begin to use matrix notation extensively. You may wish to review this set of notation in Appendix A, focusing on the definitions and basic operations in A.1-3, before proceeding.

Statistical and econometric packages are widely available and thus users will rarely have to code the least squares estimator expressions. Nonetheless, the expressions in equations (2.6) and (2.7) offer several valuable insights.

First, we note that there are $n+K$ unknown regression coefficients in equation (2.3), $n$ for the $\{\alpha_i\}$ parameters and $K$ for the $\beta$ parameters. Using standard regression routines, this calls for the inversion of a $(n+K) \times (n+K)$ matrix. However, the calculation of the ordinary least squares estimators in equation (2.6) requires inversion of only a $K \times K$ matrix. This is a standard feature of analysis of covariance models, treating the subject identifier as a categorical explanatory variable known as a factor.

Second, the OLS estimator of $\beta$ can also be expressed as a weighted average of subject-specific estimators. Specifically, suppose that all parameters are subject-specific, so that the regression function is $\mathbb{E} y_{it} = \alpha_i + \mathbf{x}_i' \beta_i$. Then, routine calculations show that the ordinary least squares estimator of $\beta_i$ turns out to be

$$
\mathbf{b}_i = \left( \sum_{t=1}^{T_i} (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)' \right)^{-1} \left( \sum_{t=1}^{T_i} (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\mathbf{y}_{it} - \bar{\mathbf{y}}_i) \right).
$$

Now, define a weight matrix

$$
\mathbf{W}_i = \sum_{t=1}^{T_i} (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'
$$

so that a simpler expression for $\mathbf{b}_i$ is $\mathbf{b}_i = (\mathbf{W}_i^{-1}) \sum_{t=1}^{T_i} (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\mathbf{y}_{it} - \bar{\mathbf{y}}_i)$. With this weight, we can express the estimator of $\beta$ as

$$
\mathbf{b} = \left( \sum_{i=1}^{n} \mathbf{W}_i \right)^{-1} \sum_{i=1}^{n} \mathbf{W}_i \mathbf{b}_i,
$$

a (matrix) weighted average of subject-specific parameter estimates. To help interpret equation (2.8), consider Figure 2.2. Here, we see that the response (CCPD) is positively related to number of discharges for each state. Thus, because each subject-specific coefficient is positive, we expect the weighted average of coefficients to also be positive.
For a third insight from equations (2.6) and (2.7), consider another weighting vector

\[ W_{it,1} = \left( \sum_{i=1}^{n} W_i \right)^{-1} \left( x_{it} - \bar{x}_i \right). \]

With this vector, another expression for equation (2.6) is

\[ b = \sum_{i=1}^{n} \sum_{t=1}^{T} W_{it,1} y_{it}. \] (2.9)

From this, we see that the regression coefficients in \( b \) are linear combinations of the responses. By the linearity, if the responses are normally distributed (assumption F5), then so are the regression coefficients in \( b \).

Fourth, regression coefficients associated with time-constant variables cannot be estimated using equation (2.6). Specifically, suppose that the \( j \)th variable does not depend on time, so that \( x_{it,j} = \bar{x}_{i,j} \). Then, elements in the \( j \)th row and column of

\[ \sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' \]

are identically zero, so that the matrix is not invertible. Thus, regression coefficients cannot be calculated using equation (2.6) and, in fact, are not estimable when one of the explanatory variables is time-constant.

Other properties of estimators

Both \( a_i \) and \( b \) have the usual (finite sample) properties of ordinary least squares regression estimators. In particular, they are unbiased estimators. Further, by the Gauss-Markov theorem, they are minimum variance among the class of unbiased estimators. If the responses are normally distributed (assumption F5), then so are \( a_i \) and \( b \). Further, using equation (2.9), it is easy to check that the variance of \( b \) turns out to be

\[ \text{Var} b = \sigma^2 \left( \sum_{i=1}^{n} W_i \right)^{-1}. \] (2.10)

ANOVA table and standard errors

The estimator of the variance parameter, \( \sigma^2 \), follows from the customary regression set-up. That is, it is convenient to first define residuals and the analysis of variance (ANOVA) table. From this, we get an estimator of \( \sigma^2 \), as well as standard errors for the regression coefficient estimators.

To this end, define the residuals as \( e_{it} = y_{it} - (a_i + x_{it}' b) \), the difference between the observed and fitted values. In ANOVA terminology, the sum of squared residuals is called the error sum of squares and denoted by \( \text{Error SS} = \sum_{it} e_{it}^2 \). The mean square error is our estimator of \( \sigma^2 \), denoted by

\[ s^2 = \frac{\text{Error SS}}{N - (n + K)} = \text{Error MS}. \] (2.11)

The corresponding positive square root is the residual standard deviation, denoted by \( s \). Here, recall that \( T_1 + T_2 + \ldots + T_n = N \) is the total number of observations. These calculations are summarized in Table 2.2.
Table 2.2. ANOVA Table

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>Regression SS</td>
<td>n-1+K</td>
<td>Regression MS</td>
</tr>
<tr>
<td>Error</td>
<td>Error SS</td>
<td>N - (n+K)</td>
<td>Error MS</td>
</tr>
<tr>
<td>Total</td>
<td>Total SS</td>
<td>N-1</td>
<td></td>
</tr>
</tbody>
</table>

To complete the definitions of the expressions in Table 2.2, we have

\[ Total \ SS = \sum_{i=1}^{n} (y_{it} - \bar{y})^2 \quad \text{and} \quad Regression \ SS = \sum_{i=1}^{n} (a_i + \mathbf{x}_i' \mathbf{b} - \bar{y})^2. \]

Further, the mean square quantities are the sum of square quantities divided by their respective degrees of freedom (df). The ANOVA table calculations are often reported through the goodness of fit statistic called the coefficient of determination,

\[ R^2 = \frac{Regression \ SS}{Total \ SS}, \]

or the version adjusted for degrees of freedom,

\[ R^2_a = 1 - \frac{(Error \ SS)/(N - (n + K))}{(Total \ SS)/(N - 1)}. \]

An important function of the residual standard deviation is to estimate standard errors associated with parameter estimators. Using the ANOVA table and equation (2.10), the estimated variance matrix of the vector of regression coefficients is 

\[ \text{Var} \mathbf{b} = s^2 \left( \sum_{i=1}^{n} W_i \right)^{-1}. \]

Thus, the standard error for the jth regression coefficient \( b_j \) is

\[ se(b_j) = s \sqrt{\text{jth diagonal element of } \left( \sum_{i=1}^{n} W_i \right)^{-1}}. \]

Standard errors are the basis for the t-ratios, arguably the most important (or at least most widely cited) statistics in applied regression analysis. To illustrate, the t-ratio for the jth regression coefficient \( b_j \) is

\[ t(b_j) = \frac{b_j}{se(b_j)} = \frac{b_j}{s \sqrt{\text{jth diagonal element of } \left( \sum_{i=1}^{n} W_i \right)^{-1}}}. \]

Assuming the responses are normally distributed, \( t(b_j) \) has a t-distribution with \( N-(n+K) \) degrees of freedom.

**Example – Medicare hospital costs, Continued**

To illustrate, we return to the Medicare example. Figures 2.1-2.4 suggested that the state categorical variable is important. Further, NUM_DSCH, AVE_DAYS and YEAR are also potentially important. From Figure 2.4, we noted that the increase in CCPD is higher for New Jersey than in other states. Thus, we also included a special interaction variable \( YEAR*(STATE=31) \) that allowed us to represent the unusually large time slope for the 31st state, New Jersey. Thus, we estimate the function

\[ E \ CCPD_{it} = \alpha_i + \beta_1 YEAR_t + \beta_2 AVE\_DAYS_{it} \]
\[ + \beta_3 NUM\_DSCH_{it} + \beta_4 YEAR_t*(STATE=31). \]

(2.12)
The fitted model appears in Display 2.1, using the statistical package SAS.

**Example 2.1 – Urban wages, Continued**

To illustrate, Table 2.3 summarizes three regression models reported by Glaeser and Maré (2001) in their investigation of determinants of hourly wages. The two homogeneous models do not include worker-specific intercepts whereas these are included in the fixed effects model. For the homogeneous model without controls, the only two explanatory variables are the binary variables for indicating whether a worker resides in a (dense or non-dense) metropolitan area. The omitted category in this regression is non-metropolitan area, so we interpret the 0.263 coefficient to mean that workers in dense metropolitan areas on average earn 0.263 log dollars, or 26.3%, more than their non-metropolitan area counterparts. Similarly, those in non-dense metropolitan areas earn 17.5% more than their non-metropolitan area counterparts.

Wages may also be influenced by a worker’s experience, occupation, education and race, and there is no guarantee that these characteristics are distributed uniformly over different city sizes. Thus, a regression model with these controls is also reported in Table 2.3. Table 2.3 shows that workers in cities, particularly dense cities, still receive more than workers in non-urban areas, even when controlling for a worker’s experience, occupation, education and race.

Glaeser and Maré offer additional explanations as to why workers in cities earn more than their non-urban counterparts, including higher cost of living and urban “disamenities” (they also discount these explanations as of less importance). They do posit an “omitted ability bias,” that is, suggesting that the ability of a worker is an important wage determinant that should be controlled for. They suggest that higher ability workers may flock to cities because if cities speed the flow of information, then this might be more valuable to workers who have high human capital. Further, cities may be centers of consumption that cater to the rich. Ability is a difficult attribute to measure (they examine a proxy, the Armed Forces Qualification Test, and find that it is not useful). However, if one treats ability as time-constant, then its effects on wages will be captured in the time-constant worker-specific intercept $\alpha_i$. Table 2.3 reports on the fixed effects regression that includes a worker-specific intercept. Here, we see that the parameter estimates for city premiums have been substantially reduced, although they are still statistically significantly. One interpretation is that a worker will receive a 10.9% (7%) higher wage for working in a dense (non-dense) city when compared to a non-metropolitan worker, even when controlling for a worker’s experience, occupation, education and race and omitted time-constant attributes such as.

---

**Display 2.1 SAS OUTPUT**

```
General Linear Models Procedure
Dependent Variable: CCPD

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>57</td>
<td>3258506185.0</td>
<td>57166775.2</td>
<td>203.94</td>
<td>0.0001</td>
</tr>
<tr>
<td>Error</td>
<td>265</td>
<td>74284379.1</td>
<td>280318.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>322</td>
<td>3332790564.1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Square 0.977711  C.V. 5.041266  Root MSE 529.45105  CCPD Mean 10502.344

| Parameter       | Estimate | T for H0: Parameter=0 | Pr > |T| | Std Error of Estimate |
|-----------------|----------|------------------------|------|---|----------------------|
| YEAR            | 710.884203 | 26.51 | 0.0001 | 26.8123882 |
| AVE_DAYS        | 361.290071 | 6.23  | 0.0001 | 57.9789849 |
| NUM_DCHG        | 10.754717  | 4.18  | 0.0001 | 2.5726139  |
| YR_31           | 1262.456077 | 9.82  | 0.0001 | 128.6088909 |
```
“ability.” Section 7.2 will present a much more detailed discussion of this omitted variable interpretation.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Homogenous model without controls</th>
<th>Homogeneous model with controls</th>
<th>Two-way fixed effects model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dense metropolitan premium</td>
<td>0.263 (0.01)</td>
<td>0.245 (0.01)</td>
<td>0.109 (0.01)</td>
</tr>
<tr>
<td>Non-dense metropolitan premium</td>
<td>0.175 (0.01)</td>
<td>0.147 (0.01)</td>
<td>0.070 (0.01)</td>
</tr>
<tr>
<td>Coefficient of determination $R^2$</td>
<td>1.4</td>
<td>33.1</td>
<td>38.1</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>1.4</td>
<td>33.0</td>
<td>28.4</td>
</tr>
</tbody>
</table>

Source: Glaeser and Maré (2001). Standard errors are in parentheses.

Large sample properties of estimators

In typical regression situations, responses are at best only approximately normally distributed. Nonetheless, hypothesis tests and predictions are based on regression coefficients that are approximately normally distributed. This premise is reasonable because of the Central Limit Theorem, which roughly states that weighted sums of independent random variables are approximately normally distributed and that the approximation improves as the number of random variables increases. Regression coefficients can be expressed as weighted sums of responses. Thus, if sample sizes are large, then we may assume approximate normality of regression coefficient estimators.

In the longitudinal data context, “large samples” means that either the number of subjects, $n$, and/or the number of observations per subject, $T$, becomes large. In this chapter, we discuss the case where $n$ becomes large although $T$ remains fixed. This can be motivated by the fact that in many data applications, the number of subjects is large relative to the number of time periods observed. Chapter 8 will discuss the problem of large $T$.

As the number of subjects, $n$, becomes large yet $T$ remains fixed, most of the properties of $b$ are retained from the standard regression situations. To illustrate, we have that $b$ is a weakly consistent estimate of $\beta$. Specifically, weak consistency means approaching (convergence) in probability. This is a direct result of the unbiasedness and an assumption that $\sum W_i$ grows without bound. Further, under mild regularity conditions, we have a central limit theorem for the slope estimator. That is, $b$ is approximately normally distributed even though the responses may not be.

The situation for the estimators of subject-specific intercepts $\alpha_i$ is dramatically different. To illustrate, the least squares estimator of $\alpha_i$ is not consistent as $n$ becomes large. Further, if the responses are not normally distributed, then $\alpha_i$ is not even approximately normal. Intuitively, this is because we assume that the number of observations per subject, $T_i$, is a bounded number.

As $n$ grows, the number of parameters grows, a situation called “infinite dimensional nuisance parameters” in the literature; see for example, Neyman and Scott (1948E) for a classic treatment. When the number of parameters grows with sample size, the usual large sample properties of estimators may not be valid. Section 7.1 and Chapter 9 will discuss this issue further.
2.4 Model specification and diagnostics

Inference based on a fitted statistical model often may be criticized because the features of the data are not in congruence with the model assumptions. Diagnostic techniques are procedures that examine this congruence. Because we sometimes use discoveries about model inadequacies to improve the model specification, this group of procedures is also called model specification, or mis-specification, tests or procedures. The broad distinction between diagnostics and the Section 2.2 data exploration techniques is that the former are performed after a preliminary model fit whereas the latter are done before fitting models with data.

When an analyst fits a great number of models to a data set, this leads to difficulties known as data snooping. That is, with several explanatory variables, one can generate a large number of linear models and an infinite number of nonlinear models. By searching over many models, it is possible to “overfit” a model, so that standard errors are smaller than they should be and insignificant relationships appear significant.

There are widely different philosophies espoused in the literature for model specification. One end of the spectrum believes that data snooping is a problem endemic to all data analysis. Proponents of this philosophy believe that a model should be fully specified before examining the data; in this way, inferences drawn from the data are not mitigated from data snooping. The other end of the spectrum argues that inferences from a model are unreliable if the data are not in accordance with model assumptions. Proponents of this philosophy argue that a model summarizes important characteristics of the data and that the best model should be sought through a series of specification tests.

These distinctions are widely discussed in the applied statistical modeling literature. We present here several specification tests and procedures that can be used to describe how well the data fits the model. Results from the specification tests and procedures can then be used to either re-specify the model or interpret model results, according to one’s beliefs in model fitting.

2.4.1. Pooling test

A pooling test, also known as a test for heterogeneity, examines whether or not the intercepts take on a common value, say \( \alpha \). An important advantage of longitudinal data models, compared to cross-sectional regression models, is that we can allow for heterogeneity among subjects, generally through subject-specific parameters. Thus, an important first procedure is to justify the need for subject-specific effects.

The null hypothesis of homogeneity can be expressed as \( H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_n = \alpha \). Testing this null hypothesis is simply a special case of the general linear hypothesis and can be handled directly as such. Here is one way to perform a partial \( F \)- (Chow) test.

**Procedure for the pooling test**

1. Run the “full model” with \( E \ y_{it} = \alpha_i + x_{it}' \beta \) to get Error SS and \( s^2 \).
2. Run the “reduced model” with \( E \ y_{it} = \alpha + x_{it}' \beta \) to get \((\text{Error SS})_{\text{reduced}}\).
3. Compute the partial \( F \)-statistic, \( F \text{-ratio} = \frac{(\text{Error SS})_{\text{reduced}} - \text{Error SS}}{(n-1)s^2} \).
4. Reject \( H_0 \) if \( F \text{-ratio} \) exceeds a percentile from an \( F \)-distribution with numerator degrees of freedom \( df_1 = n-1 \) and denominator degrees of freedom \( df_2 = N - (n+K) \). The percentile is one minus the significance level of the test.

This is an “exact” test in the sense that it does not require large sample sizes yet does require normality of the responses (assumption F5). Studies have shown that the \( F \)-test is not sensitive to departures from normality (see, for example, Layard, 1973G). Further, note that if the
denominator degrees of freedom, \( df_2 \), is large, then we may approximate the distribution by a chi-square distribution with \( n-1 \) degrees of freedom.

**Example – Medicare hospital costs, Continued**

To test for heterogeneity in Medicare Hospital cost, the interest is in testing the null hypothesis \( H_0: \alpha_1 = \alpha_2 = ... = \alpha_{51} \). From Display 2.1, we have \( \text{Error SS} = 74,284,379.1 \) and \( s^2 = 280,318.4 \). Fitting the pooled cross-sectional regression function (with common effects \( \alpha \))

\[
E_{CCPD} = \alpha + \beta_1 \text{YEAR}_i + \beta_2 \text{AVE\_DAYS}_{it} + \beta_3 \text{NUM\_DSCH}_{it} + \beta_4 \text{YEAR}_i \times \text{STATE=31}
\]

yields \( \text{(Error SS)}_{\text{reduced}} = 2,373,115,932.9 \). Thus, the test statistic is:

\[
F\text{-ratio} = \frac{2,373,115,932.9 - 74,284,379.1}{(54 - 1) \cdot 280,318.4} = 154.7.
\]

For an \( F \)-distribution with \( df_1 = 53 \) and \( df_2 = 323-(54+4)=265 \), the associated \( p \)-value is less than 0.0001. This provides strong evidence for the case for retaining subject-specific parameters \( \alpha_i \) in the model specification.

**2.4.2. Added variable plots**

An added variable plot, also known as a partial regression plot, is a standard graphical device used in regression analysis; see, for example, Cook and Weisberg (1982G). It allows one to view the relationship between a response and an explanatory variable, after controlling for the linear effects of other explanatory variables. Thus, added variable plots allow analysts to visualize the relationship between \( y \) and each \( x \), without forcing the eye to adjust for the differences induced by the other explanatory variables. The Section 2.2 basic added variable plot is a special case of the following procedure that can be used for additional regression variables.

To produce an added variable plot, one first selects an explanatory variable, say \( x_j \), and then follows the procedure below.

**Procedure for producing an added variable plot**

1. Run a regression of \( y \) on the other explanatory variables (omitting \( x_j \)) and calculate the residuals from this regression. Call these residuals \( e_1 \).
2. Run a regression of \( x_j \) on the other explanatory variables (omitting \( x_j \)) and calculate the residuals from this regression. Call these residuals \( e_2 \).
3. Produce a plot of \( e_1 \) versus \( e_2 \). This is an added variable plot.

**Correlations and added variable plots**

To help interpret added variable plots, use equation (2.1) to express the disturbance term as

\[
\epsilon_{it} = y_{it} - (\alpha_t + \beta_1 x_{it,1} + ... + \beta_K x_{it,K}).
\]

That is, we may think of the error as the response, after controlling for the linear effects of the explanatory variables. The residual \( e_1 \) is an approximation of the error, interpreted to be the response after controlling for the effects of explanatory variables. Similarly, we may interpret \( e_2 \) to be the \( j \)th explanatory variable, after controlling for the effects of other explanatory variables. Thus, we interpret the added variable plot as a graph of the relationship between \( y \) and \( x_j \), after controlling for the effects of other explanatory variables.
As with all scatter plots, the added variable plot can be summarized numerically through a correlation coefficient that we will denote by \( \text{corr}(e_1, e_2) \). It is related to the \( t \)-statistic of \( x_j, t(b_j) \), from the full regression equation (including \( x_j \)) through the following expression:

\[
\text{corr}(e_1, e_2) = \frac{t(b_j)}{\sqrt{t(b_j)^2 + N - (n + K)}}.
\]

Here, \( n + K \) is the number of regression coefficients in the full regression equation and \( N \) is the number of observations. Thus, the \( t \)-statistic from the full regression equation can be used to determine the correlation coefficient of the added variable plot without running the three-step procedure. However, unlike correlation coefficients, the added variable plot allows us to visualize potential nonlinear relationships between \( y \) and \( x_j \).

### 2.4.3. Influence diagnostics

Traditional influence diagnostics are important because they allow an analyst to understand the impact of individual observations on the estimated model. That is, influence statistics allow analysts to perform a type of “sensitivity analysis;” one can calibrate the effect of individual observations on regression coefficients. Cook’s distance is a diagnostic statistic that is widely used in regression analysis and is reviewed in Appendix 2A.3. For the Chapter 2 fixed effects longitudinal data models, observation level diagnostic statistics are of less interest because the effect of unusual observations is absorbed by subject-specific parameters. Of greater interest is the impact that an entire subject has on the population parameters.

To assess the impact that a subject has on estimated regression coefficients, we use the statistic

\[
B_i(b) = (b - b_{(i)})' \left( \sum_{i=1}^{n} W_i \right)^{-1} (b - b_{(i)}) / K.
\]

Here, \( b_{(i)} \) is the ordinary least squares estimator \( b \) calculated with the \( i \)-th subject omitted. Thus, \( B_i(b) \) measures the distance between regression coefficients calculated with and without the \( i \)-th subject. In this way, we can assess the effect of the \( i \)-th subject. The longitudinal data influence diagnostic is similar to Cook’s distance for regression. However, Cook’s distance is calculated at the observation level whereas \( B_i(b) \) is at the subject level.

Observations with a “large” value of \( B_i(b) \) may be influential on the parameter estimates. Banerjee and Frees (1997S) showed that the statistic \( B_i(b) \) has an approximate \( \chi^2 \) (chi-square) distribution with \( K \) degrees of freedom. Thus, we may use quantiles of the \( \chi^2 \) to quantify the adjective “large.” Influential observations warrant further investigation; they may require correction for coding errors, additional variable specification to accommodate the patterns they emphasize or deletion from the data set.

From the definition of \( B_i(b) \), it appears that the calculation of the influence statistic is computationally intensive. This is because the definition requires \( n+1 \) regression computations, one for \( b \) and one for each \( b_{(i)} \). However, as with Cook’s distance at the observation level, shortcut calculation procedures are available. The details are in Appendix 2A.3.

#### Example – Medicare hospital costs, Continued

Figure 2.3 alerted us to the unusual value of AVE_DAYS that occurred in the \( i=54 \)-th subject at the \( t=2 \)-nd time point. It turns out that this observation has a substantial impact on the fitted regression model. Fortunately, the graphical procedure in Figure 2.3 and the summary statistics in Table 2.1 were sufficient to detect this unusual point. Influence diagnostic statistics provide another tool for detecting unusual observations and subjects. Suppose that the model in equation (2.12) was fit using the full data set of \( N = 324 \) observations. It turns out that Cook’s distance was \( D_{54,2} = 17.06 \) for the \( (i=54, t=2) \) point, strongly indicating an influential
observation. The corresponding subject-level statistic was \( B_{54} = 244.3 \). Compared to a chi-square distribution with \( K = 4 \) degrees of freedom, this indicates that something about the 54th subject was unusual. For comparison, the diagnostic statistics were calculated under a fitted regression model after removing the \((i=54, t=2)\) point. The largest value of Cook’s distance was 0.0907 and the largest value of the subject-level statistic was 0.495. Neither value indicates substantial influential behavior after the unusual \((i=54, t=2)\) point was removed.

### 2.4.4. Cross-sectional correlation

Our basic model relies on assumption F4, the independence among observations. In traditional cross-sectional regression models, this assumption is untestable without a parametric assumption. However, with repeated measurements on a subject, it is possible to examine this assumption. As is traditional in the statistics literature, when testing for independence, we are really testing for zero correlation. That is, we are interested in the null hypothesis, \( H_0: \rho_{ij} = \text{Corr}(y_{it}, y_{jt}) = 0 \) for \( i \neq j \).

To understand how violations of this assumption may arise in practice, suppose that the true model is \( y_{it} = \lambda_t + x_{it}' \beta + \epsilon_{it} \). Here, we use \( \lambda_t \) for a random temporal effect that is common to all subjects. Because it is common, it induces correlation among subjects, as follows. We first note that the variance of a response is \( \text{Var} y_{it} = \sigma^2_{\lambda} + \sigma^2 \), where \( \text{Var} \epsilon_{it} = \sigma^2 \) and \( \text{Var} \lambda_t = \sigma^2_{\lambda} \).

From here, basic calculations show that the covariance between observations at the same time but from different subjects is \( \text{Cov}(y_{it}, y_{jt}) = \sigma^2_{\lambda} \), for \( i \neq j \). Thus, the cross-sectional correlation is \( \text{Corr}(y_{it}, y_{jt}) = \frac{\sigma^2_{\lambda}}{\sigma^2_{\lambda} + \sigma^2} \). Hence, a positive cross-sectional correlation may be due to unobserved temporal effects that are common among subjects.

#### Testing for non-zero cross-sectional correlation

To test \( H_0: \rho_{ij} = 0 \) for all \( i \neq j \), we use a procedure developed in Frees (1995E) where balanced data were assumed so that \( T_i = T \).

**Procedure for computing cross-sectional correlation statistics**

1. Fit a regression model and calculate the model residuals, \( \{e_{it}\} \).
2. For each subject \( i \), calculate the ranks of each residual. That is, define \( \{r_{i1}, \ldots, r_{iT}\} \) to be the ranks of \( \{e_{i1}, \ldots, e_{iT}\} \). These ranks will vary from 1 to \( T \), so that the average rank is \((T+1)/2\).
3. For the \( i \)th and \( j \)th subject, calculate the rank correlation coefficient (Spearman’s correlation)
   \[
   sr_{ij} = \frac{\sum_{t=1}^{T} \left( r_{it} - (T+1)/2 \right) \left( r_{jt} - (T+1)/2 \right)}{\sum_{t=1}^{T} \left( r_{jt} - (T+1)/2 \right)^2}.
   \]
4. Calculate the average Spearman’s correlation and the average squared Spearman’s correlation
   \[
   R_{AVE} = \frac{1}{n(n-1)/2} \sum_{i<j} sr_{ij} \quad \text{and} \quad R_{AVE}^2 = \frac{1}{n(n-1)/2} \sum_{i<j} (sr_{ij})^2.
   \]
   Here, \( \Sigma_{i<j} \) means sum over \( j = 2, \ldots, n \) and \( i = 1, \ldots, j-1 \).

**Calibration of cross-sectional correlation test statistics**

Large values of the statistics \( R_{AVE} \) and \( R_{AVE}^2 \) indicate the presence of non-zero cross-sectional correlations. In applications where either positive or negative cross-sectional
correlations prevail, one should consider the \( R_{AVE} \) statistic. Friedman (1937G) showed that \( FR = (T-1)(n-1)R_{AVE} + 1 \) follows a chi-square distribution (with \( T-1 \) degrees of freedom) asymptotically, as \( n \) becomes large. Friedman devised the test statistic \( FR \) to determine the equality of treatment means in a two-way analysis of variance. The statistic \( FR \) is also used in the problem of “\( n \)-rankings.” In this context, \( n \) “judges” are asked to rank \( T \) items and the data are arranged in a two-way analysis of variance layout. The statistic \( R_{AVE} \) is interpreted to be the average agreement among judges.

The statistic \( R_{AVE}^2 \) is useful for detecting a broader range of alternatives than the statistic \( R_{AVE} \). For hypothesis testing purposes, we compare \( R_{AVE}^2 \) to a distribution that is a weighted sum of chi-square random variables. Specifically, define

\[
Q = a(T) \left( X_1^2 - (T - 1) \right) + b(T) \left( X_2^2 - T (T-3)/2 \right).
\]

Here, \( X_1^2 \) and \( X_2^2 \) are independent chi-square random variables with \( T-1 \) and \( T (T-3)/2 \) degrees of freedom, respectively. The constants are

\[
a(T) = \frac{4(T+2)}{5(T-1)^2(T+1)} \quad \text{and} \quad b(T) = \frac{2(5T + 6)}{5T(T-1)(T+1)}.
\]

Frees (1995E) showed that \( n(R_{AVE}^2 - (T - 1)^1) \) follows a \( Q \) distribution asymptotically, as \( n \) becomes large. Thus, one rejects \( H_0 \) if \( R_{AVE}^2 \) exceeds \((T-1)^1 + Q_q/n \), where \( Q_q \) is an appropriate quantile from the \( Q \) distribution.

Because the \( Q \) distribution is a weighted sum of chi-square random variables, computing quantiles may be tedious. For an approximation, it is much faster to compute the variance of \( Q \) and use a normal approximation. Exercise 2.13 illustrates the use of this approximation.

The statistics \( R_{AVE} \) and \( R_{AVE}^2 \) are averages over \( n(n-1)/2 \) correlations, which may be computationally intensive for large values of \( n \). Appendix 2A.4 describes some short-cut calculation procedures.

**Example – Medicare hospital costs, Continued**

The main drawback of the \( R_{AVE} \) and \( R_{AVE}^2 \) statistics is that the asymptotic distributions are only available for balanced data. To achieve balanced data for the Medicare hospital costs data, we omit the 54th state. The model in equation (2.12) was fit to the remaining 53 states and residuals calculated. The values of the correlation statistics turned out to be \( R_{AVE} = 0.281 \) and \( R_{AVE}^2 = 0.388 \). Both statistics are statistically significant with p-values less than 0.001. This result indicates substantial cross-sectional correlation, indicating some type of “co-movement” among states over time that is not captured by our simple model.

For comparison, the model was re-fit using YEAR as a categorical variable in lieu of a continuous one. This is equivalent to including six indicator variables, one for each year. The values of the correlation statistics turned out to be \( R_{AVE} = 0.020 \) and \( R_{AVE}^2 = 0.419 \). Thus, we have captured some of the positive “co-movement” among state Medicare hospital costs with time indicator variables.

### 2.4.5. Heteroscedasticity

When fitting regression models to data, an important assumption is that the variability is common among all observations. This assumption of common variability is called *homoscedasticity*; this means “same scatter.” Indeed, the least squares procedure assumes that the expected variability of each observation is constant; it gives the same weight to each observation when minimizing the sum of squared deviations. When the scatter varies by observation, the data are said to be *heteroscedastic*. Heteroscedasticity affects the efficiency of the regression coefficient estimators although these estimators remain unbiased even in the presence of heteroscedasticity.
In the longitudinal data context, the variability \( \text{Var} \, y_{it} \) may depend on the subject through \( i \), or the time period through \( t \), or both. Several techniques are available for handling heteroscedasticity. First, heteroscedasticity may be mitigated through a transformation of the response variable. See Carroll and Ruppert (1988G) for a broad treatment of this approach. Second, heteroscedasticity may be explained by weighting variables, denoted by \( w_{it} \). Third, the heteroscedasticity may be ignored in the estimation of the regression coefficients yet accounted for in the estimation of the standard errors. Section 2.5.3 expands on this approach. Further, as we will see in Chapter 3, the variability of \( y_{it} \) may vary over \( i \) and \( t \) through a “random effects” specification.

One method for detecting heteroscedasticity is to perform a preliminary regression fit of the data and plot the residuals versus the fitted values. The preliminary regression fit removes many of the major patterns in the data and leaves the eye free to concentrate on other patterns that may influence the fit. We plot residuals versus fitted values because the fitted values are an approximation of the expected value of the response and, in many situations, the variability grows with the expected response.

More formal tests of heteroscedasticity are also available in the regression literature. For an overview, see Judge et al. (1985E) or Greene (2002E). To illustrate, let us consider a test due to Breusch and Pagan (1980E). Specifically, this test examines the alternative hypothesis \( H_a: \text{Var} \, y_{it} = \sigma^2 + \gamma' \, w_{it} \), where \( w_{it} \) is a known vector of weighting variables and \( \gamma \) is a \( p \)-dimensional vector of parameters. Thus, the null hypothesis is \( H_0: \text{Var} \, y_{it} = \sigma^2 \).

**Procedure for testing for heteroscedasticity**

1. Fit a regression model and calculate the model residuals, \( \{ e_{it} \} \).
2. Calculate squared standardized residuals, \( e_{it}^2 = e_{it}^2 / (\text{Error SS} / N) \).
3. Fit a regression model of \( e_{it}^2 \) on \( w_{it} \).
4. The test statistic is \( LM = (\text{Regress SS}_w) / 2 \), where \( \text{Regress SS}_w \) is the regression sum of squares from the model fit in step 3.
5. Reject the null hypothesis if \( LM \) exceeds a percentile from a chi-square distribution with \( p \) degrees of freedom. The percentile is one minus the significance level of the test.

Here, we use \( LM \) to denote the test statistic because Breusch and Pagan derived it as a Lagrange multiplier statistic; see Breusch and Pagan (1980E) for more details. Appendix C.7 reviews Lagrange multiplier statistics and related hypothesis tests.

A common approach for handling heteroscedasticity involves computing standard errors that are robust to the homoscedasticity specification. This is the topic of Section 2.5.3.

**2.5 Model extensions**

To introduce extensions of the basic model, we first provide a more compact representation using matrix notation. A matrix form of equation (2.2) function is

\[
E \, y_i = \alpha_i \, 1_i + X_i \, \beta ,
\]  

(2.13)

Here, \( y_i \) is the \( T_i \times 1 \) vector of responses for the \( i \)th subject, \( y_i = (y_{i1}, y_{i2}, ..., y_{iT_i})' \). Further, \( X_i \) is a \( T_i \times K \) matrix of explanatory variables,
$$X_i = \begin{pmatrix} x_{i1,1} & x_{i1,2} & \cdots & x_{i1,K} \\ x_{i2,1} & x_{i2,2} & \cdots & x_{i2,K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{iT_i,1} & x_{iT_i,2} & \cdots & x_{iT_i,K} \end{pmatrix} = \begin{pmatrix} x'_{i1} \\ x'_{i2} \\ \vdots \\ x'_{iT_i} \end{pmatrix}.$$  \hspace{1cm} (2.14)

This can be expressed more compactly as $X_i = (x_{i1}, x_{i2}, \ldots, x_{iT_i})'$. Finally, $I_i$ is the $T_i \times 1$ vector of ones.

### 2.5.1 Serial correlation

In longitudinal data, subjects are measured repeatedly over time and repeated measurements of a subject tend to be related to one another. As we have seen, one way to capture this relationship is through common subject-specific parameters. Alternatively, this relationship can be captured through the correlation among observations within a subject. Because this correlation is among observations taken over time, we refer to it as a **serial correlation**.

As with time series analysis, it is useful to measure tendencies in time patterns through a correlation structure. Nonetheless, it is also important to note that time patterns can be handled through the use of time-varying explanatory variables. As a special case, temporal indicator (dummy) variables can be used for time patterns in the data. Although it is difficult to isolate when examining data, time-varying explanatory variables account for time patterns in the mean response, whereas serial correlations are used to account for time patterns in second moment of the response. Further, in Chapters 6 and 8 we will explore other methods for modeling time patterns, such as using lagged dependent variables.

### Timing of observations

The actual times that observations are taken are important when examining serial correlations. This section assumes that observations are taken equally spaced in time, such as quarterly or annually. The only degree of imbalance that we explicitly allow for is the number of observations per subject, denoted by $T_i$. Chapter 7 introduces issues of missing data, attrition and other forms of unbalance. Chapter 8 introduces techniques for handling data that are not equally spaced in time.

### Temporal covariance matrix

For a full set of observations, we use $R$ to denote the $T \times T$ temporal variance-covariance matrix. This is defined by $R = \text{Var } y$, where $R_{rs} = \text{Cov } (y_r, y_s)$ is the element in the $r$th row and $s$th column of $R$. There are at most $T(T+1)/2$ unknown elements of $R$. We denote this dependence of $R$ on parameters using the notation $R(\tau)$, where $\tau$ is a vector of parameters.

For less than a full set of observations, consider the $i$th subject that has $T_i$ observations. Here, we define $\text{Var } y_i = R_i(\tau)$, a $T_i \times T_i$ matrix. The matrix $R_i(\tau)$ can be determined by removing certain rows and columns of the matrix $R(\tau)$. We assume that $R_i(\tau)$ is positive-definite and only depends on “$i$” through its dimension.

The matrix $R(\tau)$ depends on $\tau$, a vector of unknown parameters called **variance components**. Let us examine several important special cases of $R$. Table 2.4 summarizes these examples.

- $R = \sigma^2 I$, where $I$ is a $T \times T$ identity matrix. This is the case of no serial correlation.
- $R = \sigma^2 \begin{pmatrix} 1 - \rho \\ \rho \end{pmatrix} I + \rho J$, where $J$ is a $T \times T$ matrix of ones. This is the **compound symmetry** model, also known as the **uniform correlation** model.
- $R_{rs} = \sigma^2 \rho^{|r-s|}$. This is the **autoregressive** of order one model, denoted by $AR(1)$.
- Make no additional assumptions on $R$. 

Table 2.4. Covariance Structure Examples

<table>
<thead>
<tr>
<th>Structure</th>
<th>Example</th>
<th>Variance Comp (τ)</th>
<th>Structure</th>
<th>Example</th>
<th>Variance Comp (τ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independent</td>
<td>( \begin{pmatrix} \sigma^2 &amp; 0 &amp; 0 &amp; 0 \ 0 &amp; \sigma^2 &amp; 0 &amp; 0 \ 0 &amp; 0 &amp; \sigma^2 &amp; 0 \ 0 &amp; 0 &amp; 0 &amp; \sigma^2 \end{pmatrix} )</td>
<td>( \sigma^2 )</td>
<td>Autoregressive of Order 1,</td>
<td>( \begin{pmatrix} 1 &amp; \rho &amp; \rho^2 &amp; \rho^3 \ \rho &amp; 1 &amp; \rho &amp; \rho^2 \ \rho^2 &amp; \rho &amp; 1 &amp; \rho \ \rho^3 &amp; \rho^2 &amp; \rho &amp; 1 \end{pmatrix} )</td>
<td>( \sigma^2, \rho )</td>
</tr>
<tr>
<td>Compound Symmetry</td>
<td>( \sigma^2 \begin{pmatrix} 1 &amp; \rho &amp; \rho &amp; \rho \ \rho &amp; 1 &amp; \rho &amp; \rho \ \rho &amp; \rho &amp; 1 &amp; \rho \ \rho &amp; \rho &amp; \rho &amp; 1 \end{pmatrix} )</td>
<td>( \sigma^2, \rho )</td>
<td>No Structure</td>
<td>( \begin{pmatrix} \sigma_1^2 &amp; \sigma_{12} &amp; \sigma_{13} &amp; \sigma_{14} \ \sigma_{12} &amp; \sigma_2^2 &amp; \sigma_{23} &amp; \sigma_{24} \ \sigma_{13} &amp; \sigma_{23} &amp; \sigma_3^2 &amp; \sigma_{34} \ \sigma_{14} &amp; \sigma_{24} &amp; \sigma_{34} &amp; \sigma_4^2 \end{pmatrix} )</td>
<td>( \sigma_{12}, \sigma_{13}, \sigma_{14}, \sigma_{23}, \sigma_{24}, \sigma_{34}, \sigma_{4} )</td>
</tr>
</tbody>
</table>

To see how the compound symmetry model may occur, consider the model \( y_{it} = \alpha_i + \varepsilon_{it} \), where \( \alpha_i \) is a random cross-sectional effect. This yields \( \text{Var}(y_{it}) = \sigma^2 + \sigma^2 = \sigma^2 \). Similarly, for \( r \neq s \), we have \( \text{Cov}(y_{ir}, y_{is}) = \sigma^2 \cdot \rho \). To write this in terms of \( \sigma^2 \), note that the correlation is \( \text{Corr}(y_{ir}, y_{is}) = \frac{\sigma^2 \cdot \rho}{\sigma^2 + \sigma^2} = \rho \). Thus, \( \text{Cov}(y_{ir}, y_{is}) = \sigma^2 \cdot \rho \) and \( R = \sigma^2 \cdot (1 - \rho) I + \rho J \).

The autoregressive model is a standard representation used in time series analysis. Suppose that \( u_{it} \) are independent errors and the disturbance terms are determined sequentially through the relation \( \varepsilon_{it} = \rho \varepsilon_{i,t-1} + u_{it} \). This implies the relation \( \text{Cov}(y_{ir}, y_{is}) = \sigma^2 \cdot \rho |r-s| \), with \( \sigma^2 = \frac{\text{Var}(u_{it})}{(1-\rho^2)} \). This model may also be extended to the case where time observations are not equally spaced in time; Section 8.2 provides further details. For the unstructured model, there are \( T(T+1)/2 \) unknown elements of \( R \). It turns out that this choice is nonestimable for a fixed effects model with individual-specific intercepts. Chapter 7 provides additional details.

### 2.5.2. Subject-specific slopes

In the Medicare hospital costs example, we found that a desirable model of covered claims per discharge, CCPD, was of the form

\[
\text{E CCPD}_{it} = \alpha_i + \beta_1 \text{YEAR}_i + x_{it}' \beta. 
\]

Thus, one could interpret \( \beta_1 \) as the expected annual change in CCPD; this may be due, for example, to a medical inflation component. Suppose that the analyst anticipates that the medical inflation component will vary by state (as suggested by Figure 2.5). To address this concern, we consider instead the model

\[
\text{E CCPD}_{it} = \alpha_{i1} + \alpha_{i2} \text{YEAR}_i + x_{it}' \beta. 
\]

Here, subject-specific intercepts are denoted by \( \{ \alpha_{i1} \} \) and we allow for subject-specific slopes associated with year through the notation \( \{ \alpha_{i2} \} \).

Thus, in addition to letting intercepts vary by subject, it is also useful to let one or more slopes vary by subject. We will consider regression functions of the form

\[
\text{E } y_{it} = z_{it}' \alpha_i + x_{it}' \beta. 
\] (2.15)
Here, the subject-specific parameters are $\mathbf{a}_i = (\alpha_{i1}, ..., \alpha_{iq})'$ and the $q$ explanatory variables are $\mathbf{z}_{it} = (z_{it1}, z_{it2}, ..., z_{itq})'$; both column vectors are of dimension $q \times 1$. Equation (2.15) is shorthand notation for the function

$$E y_{it} = \alpha_{i1} z_{it1} + \alpha_{i2} z_{it2} + \ldots + \alpha_{iq} z_{itq} + \beta_1 x_{it1} + \beta_2 x_{it2} + \ldots + \beta_K x_{itK}.$$ 

To provide a more compact representation using matrix notation, we define $\mathbf{Z}_i = (\mathbf{z}_{i1}, \mathbf{z}_{i2}, \ldots, \mathbf{z}_{iT})'$, a $T_i \times q$ matrix of explanatory variables. With this notation, as in equation (2.13), a matrix form of equation (2.15) is

$$E \mathbf{y}_i = \mathbf{Z}_i \mathbf{a}_i + \mathbf{X}_i \mathbf{\beta}. \quad (2.16)$$

The responses between subjects are independent, yet we allow for temporal correlation and heteroscedasticity through the assumption that $\text{Var} \mathbf{y}_i = \mathbf{R}_i(\tau) = \mathbf{R}_i$.

Taken together, these assumptions comprise what we term the \textit{fixed effects linear longitudinal data model}.

<table>
<thead>
<tr>
<th>Assumptions of the Fixed Effects Linear Longitudinal Data Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1. $E \mathbf{y}_i = \mathbf{Z}_i \mathbf{a}_i + \mathbf{X}_i \mathbf{\beta}$.</td>
</tr>
<tr>
<td>F2. ${x_{it1}, ..., x_{itK}}$ and ${z_{it1}, ..., z_{itq}}$ are nonstochastic variables.</td>
</tr>
<tr>
<td>F3. $\text{Var} \mathbf{y}_i = \mathbf{R}_i(\tau) = \mathbf{R}_i$.</td>
</tr>
<tr>
<td>F4. ${y_{it}}$ are independent random vectors.</td>
</tr>
<tr>
<td>F5. ${y_{it}}$ are normally distributed.</td>
</tr>
</tbody>
</table>

Note that we use the same letters, F1-F5, to denote the assumptions of the fixed linear longitudinal data model and the linear regression model. This is because the models differ only through their mean and variance functions.

**Sampling and model assumptions**

We can use a model that represents how the sample is drawn from the population to motivate the assumptions of the fixed effects linear longitudinal data model. Specifically, assume the data arise as a \textit{stratified sample}, in which the subjects are the strata. For example, in Section 2.2 example we would identify each state as a stratum. Under stratified sampling, one assumes independence among different subjects (assumption F4). For observations within a stratum, unlike stratified sampling in a survey context, we allow for serial dependence to arise in a time series pattern through the assumption F3. In general, selecting subjects based on exogenous characteristics suggests stratifying the population and using a fixed effects model. To illustrate, many panel data studies have analyzed selected large countries, firms or important CEOs (chief executive officers). When a sample is selected based on exogenous explanatory variables and these explanatory variables are treated as fixed, yet variable, we treat the subject-specific terms as fixed, yet variable.

**Least squares estimators**

The estimators are derived in Appendix 2A.2 assuming that the temporal correlation matrix $\mathbf{R}_i$ is known. Section 3.5 will address the problems of estimating the parameters that determine this matrix. Moreover, Section 7.1 will emphasize some of the special problems of estimating these parameters in the presence of fixed effects heterogeneity. With known $\mathbf{R}_i$, the regression coefficient estimators are given by

$$\mathbf{b}_{FE} = \left( \sum_{i=1}^n \mathbf{X}_i' \mathbf{R}_i^{-1/2} \mathbf{Q}_{Z_i} \mathbf{R}_i^{-1/2} \mathbf{X}_i \right)^{-1} \sum_{i=1}^n \mathbf{X}_i' \mathbf{R}_i^{-1/2} \mathbf{Q}_{Z_i} \mathbf{R}_i^{-1/2} \mathbf{y}_i \quad (2.17)$$
and
\[ a_{FE,i} = (Z_i' R_i^{-1} Z_i)^{-1} Z_i' R_i^{-1} (y_i - X_i b_{FE}) . \] (2.18)

Here, \( Q_{Z,j} = I_j - R_i^{-1/2} Z_i (Z_i' R_i^{-1} Z_i)^{-1} Z_i' R_i^{-1/2} . \)

### 2.5.3. Robust estimation of standard errors

Equations (2.17) and (2.18) show that the regression estimators are linear combinations of the responses and thus it is straightforward to determine the variance of these estimations. To illustrate, we have
\[ \text{Var} b_{FE} = \left( \sum_{i=1}^{n} X_i' Q_i R_i^{-1/2} X_i \right)^{-1} \left( \sum_{i=1}^{n} X_i' Q_i R_i^{-1/2} Q_i R_i^{-1/2} X_i \right) \left( \sum_{i=1}^{n} X_i' Q_i R_i^{-1/2} X_i \right)^{-1} \].

Thus, standard errors for the components of \( b_{FE} \) are readily determined by using estimates of \( R_i \) and taking square roots of diagonal elements of \( \text{Var} b_{FE} \).

It is common practice to ignore serial correlation and heteroscedasticity initially when estimating \( \beta \) so that one can assume \( R_i = \sigma_i^2 I_i \). With this assumption, the least squares estimator of \( \beta \) is
\[ b = \left( \sum_{i=1}^{n} X_i' Q_i X_i \right)^{-1} \left( \sum_{i=1}^{n} X_i' Q_i y_i \right) , \]
with \( Q_i = I_j - Z_i (Z_i' Z_i)^{-1} Z_i' \). This is an unbiased and asymptotically normal estimator of \( \beta \), although it is less efficient than \( b_{FE} \). Basic calculations show that it has variance
\[ \text{Var} b = \left( \sum_{i=1}^{n} X_i' Q_i X_i \right)^{-1} \left[ \sum_{i=1}^{n} X_i' Q_i R_i Q_i R_i^{-1/2} X_i \right] \left( \sum_{i=1}^{n} X_i' Q_i X_i \right)^{-1} . \]

To estimate this, Huber (1967G), White (1980E) and Liang and Zeger (1986B) suggested replacing \( R_i \) by \( e_i e_i' \) to get an estimate that is robust to unsuspected serial correlation and heteroscedasticity. Here, \( e_i \) is the vector of residuals. Thus, a robust standard error of \( b_j \) is
\[ \text{se}(b_j) = \sqrt{\text{diagonal element of } \left( \sum_{i=1}^{n} X_i' Q_i X_i \right)^{-1} \left[ \sum_{i=1}^{n} X_i' Q_i e_i e_i' Q_i X_i \right] \left( \sum_{i=1}^{n} X_i' Q_i X_i \right)^{-1} } \]. (2.19)

For contrast, consider pooled cross-sectional regression model (based on equation (2.1)) so that \( Q_i = I_i \) and assume no serial correlation. Then, the ordinary least squares estimator of \( \beta \) has variance
\[ \text{Var} b = \left( \sum_{i=1}^{n} X_i' X_i \right)^{-1} \left[ \sum_{i=1}^{n} X_i' R_i X_i \right] \left( \sum_{i=1}^{n} X_i' X_i \right)^{-1} , \]
where \( R_i = \sigma_i^2 I_i \) for heteroscedasticity. Further, using the estimator \( s_i^2 = e_i' e_i / T_i \) for \( \sigma_i^2 \) yields the usual (White’s) robust standard errors. By way of comparison, the robust standard error in equation (2.19) accommodates heterogeneity (through the \( Q_i \) matrix) and also accounts for unsuspected serial correlation by using the \( T_i \times T_i \) matrix \( e_i e_i' \) in lieu of the scalar estimate \( s_i^2 = e_i' e_i / T_i \).
Further reading

Fixed effects modeling can be best understood based on a solid foundation in regression analysis and analysis using the general linear model. Draper and Smith (1981G) and Seber (1977G) are two classic references that introduce regression using matrix algebra. Treatments that emphasize categorical covariates in the general linear model context include Searle (1987G) and Hocking (1985G). Alternatively, most introductory graduate econometric textbooks cover this material; see, for example, Greene (2002E) or Hayashi (2000E).

This book actively uses matrix algebra concepts to develop the subtleties and nuances of longitudinal data analysis. Appendix A provides a brief overview of the key results. Graybill (1969G) provides additional background.

Early applications of basic fixed effects panel data models are by Kuh (1959E), Johnson (1960E), Mundlak (1961E) and Hoch (1962E).

Kiefer (1980E) discussed the basic fixed effects model in the presence of an unstructured serial covariance matrix. He showed how to construct two-stage generalized least squares (GLS) estimators of global parameters. Further, he gave a conditional maximum likelihood interpretation of the GLS estimator. Extensions of this idea and additional references are in Kung (1996O); see also Section 7.1.

Empirical work on estimating subject-specific slope models has been limited in a fixed effects context. An example is provided by Polachek and Kim (1994E); they used subject-specific slopes in fixed effects models when examining gaps in earnings between males and females. Mundlak (1978bE) provided some basic motivation that will be described in Section 7.3.
Appendix 2A - Least Squares Estimation

2A.1 Basic Fixed Effects Model - Ordinary Least Squares Estimation

We first calculate the ordinary least squares estimators of the linear parameters in the function

\[ E(y_{it}) = \alpha_i + \mathbf{x}_i' \beta, \quad i = 1, \ldots, n, \quad t = 1, \ldots, T_i. \]

To this end, let \( a_i^*, b_1^*, b_2^*, \ldots, b_k^* \) be “candidate” estimators of the parameters \( \alpha_i, \beta_1, \beta_2, \ldots, \beta_k \). For these candidates, define the sum of squares

\[ \text{SS}(a^*, b^*) = \sum_{i=1}^{n} \sum_{t=1}^{T_i} (y_{it} - (a_i^* + \mathbf{x}_i' b^*))^2, \]

where \( a^* = (a_1^*, \ldots, a_n^*)' \) and \( b^* = (b_1^*, \ldots, b_k^*)' \). Specifically, \( a^* \) and \( b^* \) are arguments of the sum of squares function \( \text{SS} \). To minimize this quantity, first examine partial derivatives with respect to \( a_i^* \) to get

\[ \frac{\partial}{\partial a_i^*} \text{SS}(a^*, b^*) = (-2) \sum_{t=1}^{T_i} \left[ y_{it} - (a_i^* + \mathbf{x}_i' b^*) \right]. \]

Setting these partial derivatives to zero yields the least squares estimators of \( \alpha_i \),

\[ a_i^*(b^*) = \bar{y}_i - \bar{x}_i' b^*, \]

where \( \bar{x}_i = (\sum_{t=1}^{T_i} \mathbf{x}_{it}) / T_i \). The sum of squares evaluated at this value of intercept is

\[ \text{SS}(a^*(b^*), b^*) = \sum_{i=1}^{n} \sum_{t=1}^{T_i} \left( y_{it} - \bar{y}_i - (\mathbf{x}_i' - \bar{x}_i) b^* \right)^2. \]

To minimize this sum of squares, take a partial derivative with respect to each component of \( b^* \). For \( j \)th component, we have

\[ \frac{\partial}{\partial b_j} \text{SS}(a^*(b^*), b^*) = (-2) \sum_{i=1}^{n} \sum_{t=1}^{T_i} \left( x_{ij} - \bar{x}_i \right) \left( y_{it} - \bar{y}_i - (\mathbf{x}_i' - \bar{x}_i) b^* \right). \]

Setting this equal to zero, for each component, yields the “normal equations”

\[ \sum_{i=1}^{n} \sum_{t=1}^{T_i} \left( x_{ij} - \bar{x}_i \right) \left( x_{ij} - \bar{x}_i \right)' b^* = \sum_{i=1}^{n} \sum_{t=1}^{T_i} \left( x_{ij} - \bar{x}_i \right) \left( y_{it} - \bar{y}_i \right). \]

These normal equations yields the ordinary least squares estimators

\[ b = \left( \sum_{i=1}^{n} \sum_{t=1}^{T_i} \left( x_{ij} - \bar{x}_i \right) \left( x_{ij} - \bar{x}_i \right)' \right)^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T_i} \left( x_{ij} - \bar{x}_i \right) \left( y_{it} - \bar{y}_i \right) \]

and

\[ a_i^* = \bar{y}_i - \bar{x}_i' b. \]


We consider linear longitudinal data models with regression functions

\[ E(y_i) = \mathbf{Z}_i a_i + \mathbf{X}_i \beta, \quad i = 1, \ldots, n, \]

where the variance-covariance matrix of \( y_i, \mathbf{R}_i \), is assumed known.

The generalized least squares sum of squares is
Here, \(a^*\) and \(b^*\) are “candidate” estimators of \(\alpha = (\alpha_1, \ldots, \alpha_n)\) and \(\beta = (\beta_1, \ldots, \beta_K)\); they are arguments of the function SS.

Following the same strategy as in Appendix 2A.1, begin by taking partial derivatives of SS with respect to each subject-specific term. That is,

\[
\frac{\partial}{\partial a_i} \text{SS}(a^*, b^*) = (-2)Z_i' R^{-1}_i \left( y_i - (Z_i a_i^* + X_i b^*) \right).
\]

Setting this equal to zero yields

\[
a_i(b^*) = \left( Z_i' R^{-1}_i \right)^{-1} Z_i' R^{-1}_i \left( y_i - X_i b^* \right).
\]

We work with the projection

\[
P_{Z,i} = R^{-1/2}_i Z_i \left( R^{-1/2}_i \right)^{-1} Z_i R^{-1/2}_i,
\]

that is symmetric and idempotent \((P_{Z,i} P_{Z,i} = P_{Z,i})\). With this notation, we have

\[
R^{-1/2}_i Z_i a_i(b^*) = P_{Z,i} R^{-1/2}_i \left( y_i - X_i b^* \right)
\]

and

\[
R^{-1/2}_i \left( y_i - (Z_i a_i^* + X_i b^*) \right) = R^{-1/2}_i \left( y_i - X_i b^* \right) - P_{Z,i} R^{-1/2}_i \left( y_i - X_i b^* \right)
\]

\[
= (I - P_{Z,i}) R^{-1/2}_i \left( y_i - X_i b^* \right).
\]

Now, define the projection, \(Q_{Z,i} = I - P_{Z,i}\), also symmetric and idempotent. With this notation, the sum of squares is

\[
\text{SS}(a^*, b^*) = \sum_{i=1}^n \left( y_i - X_i b^* \right)' R^{-1/2}_i Q_{Z,i} R^{-1/2}_i \left( y_i - X_i b^* \right).
\]

To minimize the sum of squares, take a partial derivative with respect to \(b^*\). Setting this equal to zero yields the generalized least squares estimators:

\[
b_{FE} = \left( \sum_{i=1}^n X_i' R^{-1/2}_i Q_{Z,i} R^{-1/2}_i X_i \right)^{-1} \sum_{i=1}^n X_i' R^{-1/2}_i Q_{Z,i} R^{-1/2}_i y_i
\]

and

\[
a_{FE,i} = \left( Z_i' R^{-1}_i Z_i \right)^{-1} Z_i' R^{-1}_i \left( y_i - X_i b_{FE} \right).
\]

### 2A.3. Diagnostic Statistics

**Observation level diagnostic statistic**

We use Cook’s distance to diagnose unusual observations. For brevity, we assume \(R_i = \sigma^2 I_i\). To define Cook’s distance, first let \(a_{i,(0)}\) and \(b_{i,(0)}\) denote OLS fixed effects estimators of \(a_i\) and \(\beta\), calculate without the observation from the \(i\)th subject at the \(r\)th time point. Without this observation, the fitted value for the \(j\)th subject at the \(r\)th time point is

\[
\hat{y}_{jr(i)} = z_{jr} a_{j,(i)} + x_{jr} b_{(i)}.
\]

We define Cook’s distance to be

\[
D_{ij} = \frac{\sum_{j=1}^n \sum_{r=1}^{T_i} (\hat{y}_{jr} - \hat{y}_{jr(i)})^2}{(nq + K)s^2},
\]
where the fitted value is calculated as $\hat{y}_i = z'_i a_{j,FE} + x'_i b_{FE}$. We calibrate $D_{it}$ using an $F$-distribution with numerator $df_1 = nq + K$ degrees of freedom and denominator $df_1 = N - (nq + K)$ degrees of freedom. The short-cut calculation form is:
1. Calculate the leverage for the $i$th subject and $r$th time point as
$$h_{it} = z'_i (Z'_i Z_i)^{-1} z_i + x'_i \left( \sum_{i=1}^{n} X'_i X_i \right)^{-1} x_i .$$
2. Residuals are calculated as $e_{it} = y_{it} - (a_{FE}^t z_{it} + b_{FE}^t x_{it})$. The mean square error is
$$s^2 = \frac{1}{N - (nq + K)} \sum_{i=1}^{n} \sum_{r=1}^{T} e_{it}^2 .$$
3. Cook’s distance is calculated as $D_{it} = \frac{e_{it}^2}{(1 - h_{it})^2 (nq + K)s^2}$.

**Subject level diagnostic statistic**

From Banerjee and Frees 1997S), the generalization of Section 2.4.3 to the fixed effects longitudinal data model defined in Section 2.5 is:
$$B_i(b) = (b_{FE} - b_{FE(i)})^t \left( \sum_{i=1}^{n} X'_i R_i^{-1/2} Q_{Z,i} R_i^{-1/2} X_i \right)^{-1} (b_{FE} - b_{FE(i)}) / K ,$$
where $Q_{Z,i} = I_i - R_i^{-1/2} Z_i (Z'_i R_i^{-1} Z_i)^{-1} Z'_i R_i^{-1/2}$. The short-cut calculation form is:
$$B_i(b) = e_i^t Q_{Z,i} R_i^{-1/2} (I_i - H_i)^{-1} H_i (I_i - H_i)^{-1} R_i^{-1/2} Q_{Z,i} e_i / K ,$$
where
$$H_i = R_i^{-1/2} Q_{Z,i} X_i \left( \sum_{i=1}^{n} X'_i R_i^{-1/2} Q_{Z,i} R_i^{-1/2} X_i \right)^{-1} X'_i Q_{Z,i} R_i^{-1/2}$$
is the leverage matrix and $e_i = y_i - X_i b_{FE}$. We calibrate $B_i$ using the chi-square distribution with $K$ degrees of freedom.

**2A.4. Cross-sectional Correlation - Short-cut calculations**

The statistics $R_{AVE}$ and $R^2_{AVE}$ are averages over $n(n-1)/2$ correlations, which may be computationally intensive for large values of $n$. For a short-cut calculation for $R_{AVE}$, we compute Friedman’s statistic directly,
$$FR = \frac{12}{nT(T+1)} \sum_{i=1}^{T} \left( \sum_{i=1}^{n} r_{i,t} \right)^2 - 3n(T+1) ,$$
and then use the relation $R_{AVE} = (FR - (T-1)) / ((n-1)(T-1))$.

For a short-cut calculation for $R^2_{AVE}$, first define the quantity
$$Z_{i,t,u} = \frac{1}{T^3 - T} 12 (r_{i,t} - (T + 1)/2)(r_{i,u} - (T + 1)/2) .$$
With this quantity, an alternative expression for $R^2_{AVE}$ is
$$R^2_{AVE} = \frac{1}{n(n-1)} \sum_{t,u} \left( \sum_{i} Z_{i,t,u} \right)^2 - \sum_i Z_{i,t,u}^2 .$$
Here, $\Sigma_{(t,u)}$ means sum over $t=1, ..., T$ and $u=1, ..., T$. Although more complex in appearance, this is a much faster computational form for $R_{AVE}$.
2. Exercises and Extensions

Section 2.1
2.1. Estimate longitudinal data models using regression routines
Consider a fictitious data set with \( x_{it} = i \times t \), for \( i = 1, 2, 3, 4 \) and \( t = 1, 2 \). That is, we have:

\[
\begin{array}{ccccccccc}
  i & 1 & 1 & 2 & 2 & 3 & 3 & 4 & 4 \\
  t & 1 & 2 & 1 & 2 & 1 & 2 & 1 & 2 \\
  x_{it} & 1 & 2 & 2 & 4 & 3 & 6 & 4 & 8 \\
\end{array}
\]

Consider the usual regression model of the form \( y = X \beta + \varepsilon \), where the matrix of explanatory variables is

\[
X = \begin{pmatrix}
  x_{11} & x_{12} & \cdots & x_{1K} \\
  x_{21} & x_{22} & \cdots & x_{2K} \\
  \vdots & \vdots & \ddots & \vdots \\
  x_{n1} & x_{n2} & \cdots & x_{nK}
\end{pmatrix}
\]

You wish to express your longitudinal data model in terms of the usual regression model.

a. Provide an expression for the matrix \( X \) for the regression model in equation (2.1). Specify the dimension of the matrix as well as each entry of the matrix in terms of the data provided above.

b. Consider the basic fixed effects model in equation (2.2). Express this in terms of the usual regression model by using binary (dummy) variables. Provide an expression for the matrix \( X \).

c. Provide an expression for the matrix \( X \) for the fixed effects model in equation (2.4).

d. Provide an expression for the matrix \( X \) for the fixed effects model in equation (2.5).

e. Suppose now that you have \( n = 400 \) instead of 4 subjects and \( T = 10 \) observations per subject instead of 2. What is the dimension of your design matrices in parts (a)-(d)? What is the dimension of the matrix, \( X'X \), that regression routines need to invert?

Section 2.3
2.2. Standard errors for regression coefficients
Consider the basic fixed effects model in equation (2.3), with \( \{\varepsilon_{it}\} \) identically and independently distributed with mean zero and variance \( \sigma^2 \).

a. Check equation (2.10), that is, prove that

\[
\text{Var}(b) = \sigma^2 \left( \sum_{i=1}^{n} W_i \right)^{-1}, \quad \text{where } W_i = \sum_{t=1}^{T} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)'
\]

b. Determine the variance of the \( i \)th intercept, \( \text{Var}(a_i) \).

c. Determine the covariance among intercepts, that is, determine \( \text{Cov}(a_i, a_j) \), for \( i \neq j \).

d. Determine the covariance between an intercept and the slope estimator, that is, determine \( \text{Cov}(a_i, b) \).

e. Determine \( \text{Var}(a_i + x^*b) \), where \( x^* \) is a known vector of explanatory variables. For what value of \( x^* \) is this a minimum?

2.3. Least squares

a. Suppose that the regression function is \( E(y_{it}) = \alpha_i \). Determine the ordinary least squares estimator for \( \alpha_i \).

b. Suppose that the regression function is \( E(y_{it}) = \alpha_i x_{it} \), where \( x_{it} \) is a scalar. Determine the ordinary least squares estimator for \( \alpha_i \).
c. Suppose that the regression function is $E y_{it} = \alpha_i x_{it} + \beta$. Determine the ordinary least squares estimator for $\alpha_i$.

2.4. Two population slope interpretations

Consider the basic fixed effects model in equation (2.3) and suppose that $K = 1$ and that $x$ is a binary variable. Specifically, let $n_{1,i} = \sum_{t=1}^{T_i} x_{it}$ be the number of ones for the $i$th subject and let $n_{2,i} = T_i - n_{1,i}$ be the number of zeroes. Further, define $\bar{y}_{1,i} = \frac{\sum_{t=1}^{T_i} x_{it} y_{it}}{n_{1,i}}$ to be the average $y$ when $x=1$, for the $i$th subject and similarly $\bar{y}_{2,i} = \frac{\sum_{t=1}^{T_i} (1-x_{it}) y_{it}}{n_{2,i}}$.

a. Show that we may write the fixed effects slope, given in equation (2.6), as

$$ b = \frac{\sum_{i=1}^{n} w_i (\bar{y}_{1,i} - \bar{y}_{2,i})}{\sum_{i=1}^{n} w_i}, $$

with weights $w_i = n_{1,i} n_{2,i} / T_i$.

b. Interpret this slope coefficient $b$.

c. Show that $\text{Var} b = \sigma^2 / \left( \sum_{i=1}^{n} w_i \right)$.

d. Suppose that you would like to minimize $\text{Var} b$ and that the set of observations numbers $\{T_1, \ldots, T_n\}$ is fixed. How could you design the binary variables $x$ (and thus, $n_{1,i}$ and $n_{2,i}$) to minimize $\text{Var} b$?

e. Suppose that $\bar{x}_i = 0$ for half the subjects and $\bar{x}_i = 1$ for the other half. What is $\text{Var} b$? Interpret this result.

f. Suppose that the $i$th subject is designed so that $\bar{x}_i = 0$. What is the contribution of this subject to $\sum_{i=1}^{n} w_i$?

2.5. Least squares bias

Suppose that the analyst should use the heterogeneous model in equation (2.2) but instead decides to use a simpler, homogeneous model of the form $E y_{it} = \alpha + x_{it}' \beta$.

a. Call the least square slope estimator $b_H$ ($H$ for homogeneous). Show that the slope estimator is

$$ b_H = \left( \sum_{i=1}^{n} \sum_{t=1}^{T_i} (x_{it} - \bar{x})(x_{it} - \bar{x}) \right)^{-1} \left( \sum_{i=1}^{n} \sum_{t=1}^{T_i} (x_{it} - \bar{x})(y_{it} - \bar{y}) \right). $$

b. Show that the deviation of $b_H$ from the slope $\beta$ is

$$ b_H - \beta = \left( \sum_{i=1}^{n} \sum_{t=1}^{T_i} (x_{it} - \bar{x})(x_{it} - \bar{x}) \right)^{-1} \left( \sum_{i=1}^{n} \sum_{t=1}^{T_i} (x_{it} - \bar{x})(\alpha_i - \bar{\alpha} + \varepsilon_{it} - \bar{\varepsilon}) \right). $$

c. Assume that $K=1$. Show that the bias in using $b_H$ can be expressed as

$$ \text{E} b_H - \beta = \frac{1}{(N-1)s_x^2} \sum_{i=1}^{n} T_i \alpha_i (\bar{x}_i - \bar{x}), $$

where $s_x^2 = \frac{1}{N-1} \sum_{i=1}^{n} \sum_{t=1}^{T_i} (x_{it} - \bar{x})^2$ is the sample variance of $x$. 
2.6. Residuals
Consider the basic fixed effects model in equation (2.3) and suppose that \( K = 1 \). Define the residuals of the ordinary least squares fit as \( e_{it} = y_{it} - (a_i + x_{it} \beta) \).

a. Show that the average residual is zero, that is, show \( \bar{e} = 0 \).

b. Show that the average residual for the \( i \)th subject is zero, that is, show \( \bar{e}_i = 0 \).

c. Show that \( \sum_{i=1}^{n} \sum_{t=1}^{T_i} e_{it} x_{it,j} = 0 \).

d. Why does (c) imply that the estimated correlation between the residuals and the \( j \)th explanatory variable is zero?

e. Show that the estimated correlation between the residuals and the fitted values is zero.

f. Show that the estimated correlation between the residuals and the observed dependent variables is, in general, not equal to zero.

g. What are the implications of parts (e) and (f) for residual analysis?

2.7. Group interpretation
Consider the basic fixed effects model in equation (2.3) and suppose that \( K = 1 \). Suppose that we are considering \( n = 5 \) groups. Each group was analyzed separately, with standard deviations and regression slope coefficients given below. For group \( i \), the sample standard deviation of the explanatory variable is given by \( s_{x,i}^2 = \frac{1}{T_i - 1} \sum_{t=1}^{T_i} (x_{it} - \bar{x}_i)^2 \).

<table>
<thead>
<tr>
<th>Group ((i))</th>
<th>Observations per group ((T_i))</th>
<th>Sample standard deviation (s_{x,i})</th>
<th>Slope ((b_i))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>11</td>
<td>9</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>4</td>
<td>-3</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

a. Use equation (2.8) to determine the overall slope estimator, \( b \).

b. Discuss the influence of the group sample standard deviations and size on \( b \).

2.8. Consistency
Consider the basic fixed effects model in equation (2.3) and suppose that \( K = 1 \). A sufficient condition for (weak) consistency of \( b \) is the mean square error tends to zero, that is, \( E (b - \beta)^2 \to 0 \) as \( n \to \infty \) (and \( T_i \) remains bounded).

a. Show that we require a sufficient amount of variability in the set of explanatory variables \( \{x_{it}\} \) in order to ensure consistency of \( b \). Explicitly, what does the phrase “a sufficient amount of variability” mean in this context?

b. Suppose that \( x_{it} = (-2)^i \) for all \( i \) and \( t \). Does this set of explanatory variables meet our sufficient condition to ensure consistency of \( b \)?

c. Suppose that \( x_{it} = t (-2)^i \) for all \( i \) and \( t \). Does this set of explanatory variables meet our sufficient condition to ensure consistency of \( b \)?

d. Suppose that \( x_{it} = t (-1/2)^i \) for all \( i \) and \( t \). Does this set of explanatory variables meet our sufficient condition to ensure consistency of \( b \)?

2.9. Least Squares
For the \( i \)th subject, consider the regression function \( E y_{it} = \alpha_i + x_{it}^t \beta_i \), \( t=1, \ldots, T_i \).

a. Write this as a regression function of the form \( E y_i = X_i \beta_i \) by giving appropriate definitions for \( X_i \) and \( \beta_i \).
b. Use a result on partitioned matrices, equation (A.1) of Appendix A5, to show that the least squares estimator of $\beta_i$ is

$$b_i = \left( \sum_{t=1}^{T} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' \right)^{-1} \left( \sum_{t=1}^{T} (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i) \right).$$

**Section 2.4**

2.10. **Pooling test**

a. Assume balanced data with $T = 5$ and $K = 5$. Use a statistical software package to show that the 95th percentile of the $F$-distribution with $df_1 = n - 1$ and $df_2 = N - (n + K) = 4 n - 5$ behaves as follows.

<table>
<thead>
<tr>
<th>$n$</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>50</th>
<th>100</th>
<th>250</th>
<th>500</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>95th percentile</td>
<td>2.1608</td>
<td>1.8760</td>
<td>1.7269</td>
<td>1.6325</td>
<td>1.4197</td>
<td>1.2847</td>
<td>1.1739</td>
<td>1.1209</td>
<td>1.0845</td>
</tr>
</tbody>
</table>

b. For the pooling test statistic defined in Section 2.4.1, show that $F$-ratio $\to 1$ as $n \to \infty$ (use weak or strong consistency). Interpret the results of part (a) in terms of this result.

2.11. **Added variable plot**

Consider the basic fixed effects model in equation (2.3) and suppose that $K = 1$.

a. Begin with the model without any explanatory variables, $y_{it} = \alpha_i + \epsilon_{it}$. Determine the residuals for this model, denoted by $e_{i1}$.

b. Now consider the “model” $x_{it} = \alpha_i + \epsilon_{it}$. Determine the residuals for this representation, denoted by $e_{i2}$.

c. Explain why a plot of $\{e_{i1}\}$ versus $\{e_{i2}\}$ is a special case of added variable plots.

d. Determine the sample correlation between $\{e_{i1}\}$ and $\{e_{i2}\}$. Denote this correlation as $\text{corr}(e_1, e_2)$.

e. For the basic fixed effects model in equation (2.3) with $K = 1$, show that

$$\left( n - (n + 1) \right) s^2 = \left( \sum_{t=1}^{T} \sum_{i=1}^{n} e_{i1}^2 \right) - b \left( \sum_{t=1}^{T} \sum_{i=1}^{n} e_{i1} e_{i2} \right).$$

f. For the basic fixed effects model in equation (2.2) with $K = 1$, show that

$$\left( N - (n + 1) \right) s^2 = \left( \sum_{t=1}^{T} \sum_{i=1}^{n} e_{i1}^2 \right) - b \left( \sum_{t=1}^{T} \sum_{i=1}^{n} e_{i1} e_{i2} \right).$$

g. For the basic fixed effects model in equation (2.2) with $K = 1$, establish the relationship described in Section 2.4.2 between the partial correlation coefficient and the $t$-statistic. That is, use parts (d)-(f) to show

$$\text{corr}(e_1, e_2) = \frac{t(b)}{\sqrt{t(b)^2 + N - (n + 1)}}.$$

2.12. **Observation level diagnostics**

We now establish a short-cut formula for calculating the usual Cook’s distance, in linear regression models. To this end, we consider the linear regression function $E y = X \beta$, that consists of $N$ rows and use the subscript “o” for a generic observation. Thus, let $x_{o}'$ be the $o$th row, or observation. Further, define $X_{(o)}$ to be the matrix of explanatory variables without the $o$th observation, and similarly for $y_{(o)}$. The ordinary least squares estimator of $\beta$ with all observations is $b = (X'X)^{-1} X' y$.

a. Use the equation just below equation (A.3) in Appendix A.5 to show that
\[
\left(X'_oX_o\right)^{-1} = \left(X'X - x'_ox_o\right)^{-1} = \left(X'X\right)^{-1} + \frac{\left(X'X\right)^{-1}x'_ox'_o\left(X'X\right)^{-1}}{1 - h_{oo}},
\]

where \(h_{oo} = x'_o\left(X'X\right)^{-1}x_o\) is the leverage for the \(o\)th observation.

b. The estimator of \(\beta\) without the \(o\)th observation is \(b_{(o)} = \left(X_{(o)}'X_{(o)}\right)^{-1}X_{(o)}'y_{(o)}\). Use part (a) to show that
\[
b_{(o)} = b - \frac{\left(X'X\right)^{-1}x_o e_o}{1 - h_{oo}}
\]

where \(e_o = y_o - x'_o b\) is the \(o\)th residual.

c. Cook’s distance is defined to be
\[
D_o = \frac{\left(\hat{y} - \hat{y}_{(o)}\right)'\left(\hat{y} - \hat{y}_{(o)}\right)}{\text{ncol}(X)s^2},
\]
where \(\hat{y} = Xb\) is the vector of fitted values and \(\hat{y}_{(o)} = Xb_{(o)}\) is the vector of fitted values without the \(o\)th observation. Show that
\[
D_o = \left(\frac{e_o}{s\sqrt{1 - h_{oo}}}\right)^2 \frac{h_{oo}}{\text{ncol}(X)(1 - h_{oo})}.
\]
d. Use the expression in part (c) to verify the formula for Cook’s Distance given in Appendix 2A.3.

2.13. Cross-sectional correlation test statistic

a. Calculate the variance of the random variable \(Q\) in Section 2.4.4.

b. The following table provides the 95th percentile of the \(Q\) random variable as a function of \(T\).

<table>
<thead>
<tr>
<th>(T)</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>12</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>95th percentile</td>
<td>0.832</td>
<td>0.683</td>
<td>0.571</td>
<td>0.495</td>
<td>0.431</td>
<td>0.382</td>
<td>0.344</td>
<td>0.286</td>
<td>0.227</td>
</tr>
</tbody>
</table>

Compute the corresponding cut-offs using the normal approximation and your answer in part (a). Discuss the performance of the approximation as \(T\) increases.

**Section 2.5**

2.14. Serial correlations

Consider the compound symmetry model, where the error variance-covariance matrix is given by \(R = \sigma^2\left((1-\rho)I + \rho J\right)\).

a. Check that the inverse of \(R\) is \(R^{-1} = \sigma^{-2}\left((1-\rho)I - (1-\rho)/(T(1-\rho)+1)J\right)\). Do this by showing that \(R R^{-1} = I\), the identity matrix.

b. Use this form of \(R\) in equation (2.17) to show that the fixed effects estimator of \(\beta\), \(b_{FE}\), equals the ordinary least squares estimator, \(b\), given in Section 2.5.3.

2.15. Regression model

Consider the general fixed effects longitudinal data model given in equation (2.16). Write this model as a regression function in the form \(\text{E}y = X^*\beta^*\). When doing this, be sure to:

a. Describe specifically how to use the matrices of explanatory variables \(\{Z_i\}\) and \(\{X_i\}\) to form \(X^*\).

b. Describe specifically how to use the vectors of parameters \(\{a_i\}\) and \(\beta\) to form \(\beta^*\).

c. Identify the dimensions of \(y, X^*\) and \(\beta^*\).
2.16. Interpreting the slope as a weighted average of subject-specific slopes
Consider the general fixed effects longitudinal data model given in equation (2.16) and define the weight matrix
\[ W_i = X'_i R_i^{-1/2} Q_{Z,i} R_i^{-1/2} X_i. \]

a. Based on data from only the \( i \)th subject, show that the least squares slope estimator is
\[ b_{FE,i} = W_i^{-1} X'_i R_i^{-1/2} Q_{Z,i} R_i^{-1/2} y_i. \]
(Hint: consider equation (2.17) with \( n=1 \)).

b. Show that the fixed effects estimator of \( \beta \) can be expressed as a matrix weighted average of the form
\[ b_{FE} = \left( \sum_{i=1}^{n} W_i \right)^{-1} \left( \sum_{i=1}^{n} W_i b_{FE,i} \right). \]

2.17 Fixed effects linear longitudinal data estimators
Consider the regression coefficient estimators of the fixed effects linear longitudinal data model in equations (2.17) and (2.18). Show that if we assume no serial correlation, \( q=1 \) and \( z_{it,1} = 1 \), then these expressions reduce to the estimators given in equations (2.6) and (2.7).

2.18 Ordinary least squares based on differenced data
Consider the based fixed effects model in equation (2.3) and use ordinary least squares based on differencing data. That is, data will be differenced over time, so that the response is \( \Delta y_{it} = y_{it} - y_{i,t-1} \) and the vector of covariates is \( \Delta x_{it} = x_{it} - x_{i,t-1} \).

a. Show that the ordinary least squares estimator of \( \beta \) based on differenced data is
\[ b_{\Delta} = \left( \sum_{i=1}^{n} \sum_{t=2}^{T_i} \Delta x_{it} \Delta x'_it \right)^{-1} \left( \sum_{i=1}^{n} \sum_{t=2}^{T_i} \Delta x_{it} \Delta y_{it} \right). \]

b. Now compute the variance of this estimator. To this end, define the vector of differenced responses \( \Delta y_i = (\Delta y_{i1}, \ldots, \Delta y_{iT_i})' \) and the corresponding matrix of differenced covariates \( \Delta X_i = (\Delta x_{i1}, \ldots, \Delta x_{iT_i})' \). With this notation, show that
\[ \text{Var} \, \, b_{\Delta} = \left( \sum_{i=1}^{n} \Delta X'_i \Delta X_i \right)^{-1} \left( \sum_{i=1}^{n} \Delta X'_i R^*_i \Delta X_i \right), \]
where
\[ R^*_i = \text{Var} \, \, \Delta y_i = \sigma^2 \]
\[ = \left( \begin{array}{cccccc}
2 & -1 & 0 & \cdots & 0 & 0 \\
-1 & 2 & -1 & \cdots & 0 & 0 \\
0 & -1 & 2 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 2 & -1 \\
0 & 0 & 0 & \cdots & -1 & 2 \\
\end{array} \right). \]

c. Now assume balanced data so that \( T_i = T \). Further assume that \( \{x_{it}\} \) are identically and independently distributed with mean \( E \, x_{it} = \mu_x \) and variance \( \text{Var} \, x_{it} = \Sigma_x \). Using equation (2.10), show that
\[ \lim_{n \to \infty} n(\text{Var}(b | X_1, \ldots, X_n)) = \sigma^2 \left( \frac{1}{(T-1)} \right) \Sigma_x^{-1}, \]
with probability one.

d. Use the assumptions of part (c). Using part (b), show that
\[
\lim_{n \to \infty} n (\text{Var}(b_{\Delta} \mid X_1, \ldots, X_n)) = \sigma^2 \frac{3T - 4}{2(T - 1)^2} \Sigma^{-1}, \text{ with probability one.}
\]

e. From the results of parts (c) and (d), argue that the equation (2.6) fixed effects estimator is more efficient than the least squares estimator based on differenced data. What is the limiting (as \( T \to \infty \)) efficiency ratio?

### Empirical Exercises

#### 2.19. Charitable Contributions

We analyze individual income tax returns data from the 1979-1988 Statistics of Income (SOI) Panel of Individual Returns. The SOI Panel is a subset of the IRS Individual Tax Model File and represents a simple random sample of individual income tax returns filed each year. Based on the individual returns data, the goal is to investigate whether a taxpayer's marginal tax rate affects private charitable contributions, and secondly, if the tax revenue losses due to charitable contributions deductions is less than the gain of charitable organizations. To address these issues, we consider a price and income model of charitable contributions, considered by Banerjee and Frees (1997).

The latter define price as the complement of an individual's federal marginal tax rate, using taxable income prior to contributions. Income of an individual is defined as the adjusted gross income. The dependent variable is total charitable contributions, which is measured as the sum of cash and other property contributions, excluding carry overs from previous years. Other covariates included in the model are age, marital status and the number of dependents of an individual taxpayer. Age is a dichotomous variable representing whether a taxpayer is over sixty four years or not. Similarly, marital status represents if an individual is married or single.

The population consists of all U.S. taxpayers who itemize their deductions. Specifically, these are the individuals who are likely to have and to record charitable contribution deductions in a given year. Among the 1,413 taxpayers in our subset of the SOI Panel, approximately 22% itemized their deductions each year during the period 1979-1988. A random sample of 47 individuals was selected from the latter group. These data are analyzed in Banerjee and Frees (1997).

#### Table 2E.1. Taxpayer Characteristics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUBJECT</td>
<td>Subject identifier, 1-47.</td>
</tr>
<tr>
<td>TIME</td>
<td>Time identifier, 1-10.</td>
</tr>
<tr>
<td>CHARITY</td>
<td>The sum of cash and other property contributions, excluding carry overs from previous years.</td>
</tr>
<tr>
<td>INCOME</td>
<td>Adjusted gross income.</td>
</tr>
<tr>
<td>PRICE</td>
<td>One minus the marginal tax rate. Here, the marginal tax rate is defined on income prior to contributions.</td>
</tr>
<tr>
<td>AGE</td>
<td>A binary variable that equal one if a taxpayer is over sixty four years and equals zero otherwise.</td>
</tr>
<tr>
<td>MS</td>
<td>A binary variable that equal one if a taxpayer is married and equals zero otherwise.</td>
</tr>
<tr>
<td>DEPS</td>
<td>Number of dependents claimed on the taxpayer’s form.</td>
</tr>
</tbody>
</table>

**Basic summary statistics**

Summarize each variable. For the binary variables, AGE and MS, provide only averages. For the other variables, CHARITY, INCOME, PRICE and DEPS, provide the mean, median, standard deviation, minimum and maximum. Further, summarize the average response variable CHARITY over TIME.
ii Create a multiple time series plot of CHARITY versus TIME.
iii Summarize the relationship among CHARITY, INCOME, PRICE, DEPS and TIME. Do this by calculating correlations and scatter plots for each pair.

b Basic fixed effects model
i Run a one-way fixed effects model of CHARITY on INCOME, PRICE, DEPS, AGE and MS. State which variables are statistically significant and justify your conclusions.
ii Produce an added variable plot of CHARITY versus INCOME, controlling for the effects of PRICE, DEPS, AGE and MS. Interpret this plot.

c Incorporating temporal effects. Is there an important time pattern?
i Re-run the model in b(i) and include TIME as an additional explanatory (continuous) variable.
ii Re-run the model in b(i) and include TIME through dummy variables, one for each year.
iii Re-run the model in b(i) and include an AR(1) component for the error.
iv Which of the three methods for incorporating temporal effects do you prefer? Be sure to justify your conclusion.

d Unusual observations
i Re-run the model in b(i) and calculate Cook’s distance to identify unusual observations.
ii Re-run the model in b(i) and calculate the influence statistic for each subject. Identify the subject with the largest influence statistic. Re-run your model by omitting the subject that you have identified. Summarize the effects on the global parameter estimates.

2.20. Tort Filings
The response that we consider is FILINGS, the number of tort actions against insurance companies per one hundred thousand population (y). For each of six years, 1984-1989, the data were obtained from 19 states. Thus, there are $6 \times 19 = 114$ observations available. The issue is to try to understand how state legal, economic and demographic characteristics affect FILINGS. Table 2E.2 describes these characteristics. More extensive motivation is provided in Section 10.2.

<table>
<thead>
<tr>
<th>Table 2E.2, State Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent Variable</strong></td>
</tr>
<tr>
<td>FILINGS</td>
</tr>
<tr>
<td><strong>State Legal Characteristics</strong></td>
</tr>
<tr>
<td>JSLIAB</td>
</tr>
<tr>
<td>COLLRULE</td>
</tr>
<tr>
<td>CAPS</td>
</tr>
<tr>
<td>PUNITIVE</td>
</tr>
<tr>
<td><strong>State Economic and Demographic Characteristics</strong></td>
</tr>
<tr>
<td>POPLAWYR</td>
</tr>
<tr>
<td>VEHCMILE</td>
</tr>
<tr>
<td>GSTATEP</td>
</tr>
<tr>
<td>POPDENS</td>
</tr>
<tr>
<td>WCMPMAX</td>
</tr>
<tr>
<td>URBAN</td>
</tr>
<tr>
<td>UNEMPLOY</td>
</tr>
</tbody>
</table>

2.21. Housing Prices

In this problem, we will examine models of housing prices in US metropolitan areas. Many studies have addressed the housing market, see, for example, Green and Malpezzi (2003O) for an introduction. The prices of houses are influenced by demand-side factors such as income and demographic variables. Supply-side factors, such as the regulatory environment of a metropolitan area, may also be important.

The data consists of annual observations from 36 metropolitan statistical areas (MSAs) over the nine-year period 1986-1994. The response variable is NARSP, an MSA’s average sale price based on transactions reported through the Multiple Listing Service, National Association of Realtors. As part of a preliminary analysis, the response variable has been transformed using a natural logarithm. For this problem, the demand-side variables are time varying yet the supply-side variables do not vary with time.
<table>
<thead>
<tr>
<th><strong>Table 2E.3. MSA Characteristics</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Response variable</strong></td>
</tr>
<tr>
<td>NARSP</td>
</tr>
<tr>
<td>an MSA's average sale price, in logarithmic units. It is based on transactions reported through the Multiple Listing Service.</td>
</tr>
<tr>
<td><strong>Demand side explanatory variables</strong></td>
</tr>
<tr>
<td>YPC</td>
</tr>
<tr>
<td>Annual Per Capita income, from the Bureau of Economic Analysis</td>
</tr>
<tr>
<td>POP</td>
</tr>
<tr>
<td>Population, from the Bureau of Economic Analysis</td>
</tr>
<tr>
<td>PERYPC</td>
</tr>
<tr>
<td>Annual percentage growth of per capita income</td>
</tr>
<tr>
<td>PERPOP</td>
</tr>
<tr>
<td>Annual percentage growth of population</td>
</tr>
<tr>
<td><strong>Supply side explanatory variables</strong></td>
</tr>
<tr>
<td>REGTEST</td>
</tr>
<tr>
<td>Regulatory index from Malpezzi (1996O).</td>
</tr>
<tr>
<td>RCDUM</td>
</tr>
<tr>
<td>Rent control dummy variable</td>
</tr>
<tr>
<td>SREG1</td>
</tr>
<tr>
<td>Sum of American Institute of Planners state regulatory questions regarding use of environmental planning and management.</td>
</tr>
<tr>
<td>AJWTR</td>
</tr>
<tr>
<td>Indicates whether the MSA is adjacent to a coastline</td>
</tr>
<tr>
<td>AJPARK</td>
</tr>
<tr>
<td>Indicates whether the MSA is adjacent to one or more large parks, military bases or reservations.</td>
</tr>
<tr>
<td><strong>Additional Variables</strong></td>
</tr>
<tr>
<td>MSA</td>
</tr>
<tr>
<td>Subject (MSA) identifier, 1-36.</td>
</tr>
<tr>
<td>TIME</td>
</tr>
<tr>
<td>Time identifier, 1-9.</td>
</tr>
</tbody>
</table>

a  **Basic summary statistics**

i  Begin by summarizing the time-constant supply side explanatory variables. Provide means for the binary variables and the mean, median, standard deviation, minimum and maximum for the other variables.

ii Assess the strength of the relationship among the supply side explanatory variables by calculating correlations.

iii Summarize the additional variables. Provide mean, median, standard deviation, minimum and maximum for these variables.

iv Examine trends over time producing the means over time for the non-supply side variables.

v Produce a multivariate time series plot of NARSP.

vi Calculate correlations among NARSP, PERYPC, PERPOP and YEAR.

vii Plot PERYPC versus YEAR. Comment on the unusual behavior in 1993 and 1994.

viii Produce an added variable plot of NARSP versus PERYPC, controlling for the effects of PERPOP and YEAR. Interpret this plot.

b  **Basic fixed effects model**

i Fit a homogeneous model, as in equation (2.1), using PERYPC, PERPOP and YEAR as explanatory variables. Comment on the statistical significance of each variable and the overall model fit.

ii Run a one-way fixed effects model of NARSP on PERYPC, PERPOP and YEAR. State which variables are statistically significant and justify your conclusions. Comment on the overall model fit.

iii Compare the models in parts b(i) and b(ii) using a partial $F$-test. State which model you prefer based on this test.

iv Re-run the step in b(ii) by excluding YEAR as an additional explanatory (continuous) variable yet including an AR(1) component for the error. State whether or not YEAR should be included in the model.
v For the model in b(ii) calculate the influence statistic for each MSA. Identify the MSA with the largest influence statistic. Re-run your model by omitting the MSA that you have identified. Summarize the effects on the global parameter estimates.

c Additional analyses
i We have not yet tried to fit any supply-side variables. Re-do the model fit part b(i), yet including supply side variables REGTEST, RCDUM, SREG1, AJPARK and AJWTR. Comment on the statistical significance of each variable and the overall model fit.
ii Re-run the model in part c(i) and include a dummy variable for each MSA, resulting a one-way fixed effects models. Comment on the difficulty of achieving unique parameter estimates with this procedure.

2.22 Bond Maturity – Unstructured problem
These data consist of observations of 328 non-regulated firms over the period 1980-1989. The goal is to assess the debt maturity structure of a firm.

For this exercise, develop a fixed effects model. For one approach, see Stohs and Mauer (1996O).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEBTMAT</td>
<td>The book value-weighted average of the firm’s debt. This is the response variable.</td>
</tr>
<tr>
<td>SIC</td>
<td>Standard Industrial Classification (SIC) of the firm.</td>
</tr>
<tr>
<td>FIRMID</td>
<td>Subject (firm) identifier, 1-328</td>
</tr>
<tr>
<td>TIME</td>
<td>Time identifier, 1-10</td>
</tr>
<tr>
<td>MVBV</td>
<td>The market value of the firm (proxied by the sum of the book value of assets and the market value of equity less the book value of equity) scaled by the book value of assets.</td>
</tr>
<tr>
<td>SIZE</td>
<td>The natural logarithm of the estimate of firm value measured in 1982 dollars using the PPI deflator.</td>
</tr>
<tr>
<td>CHANGEEPS</td>
<td>The difference between next year’s earnings per share and this year’s earnings per share scaled by this year’s common stock price per share.</td>
</tr>
<tr>
<td>ASSETMAT</td>
<td>The book value-weighted average of the maturities of current assets and net property plant and equipment.</td>
</tr>
<tr>
<td>VAR</td>
<td>Ratio of the standard deviation of the first difference in earnings before interest, depreciation and taxes to the average of assets over the period 1980-1989.</td>
</tr>
<tr>
<td>TERM</td>
<td>The difference between the long-term and short-term yields on government bonds.</td>
</tr>
<tr>
<td>BONDRATE</td>
<td>The firm’s S&amp;P bond rating.</td>
</tr>
<tr>
<td>TAXRATE</td>
<td>Ratio of income taxes paid to pretax income.</td>
</tr>
<tr>
<td>LEVERAGE</td>
<td>Ratio of total debt to the market value of the firm.</td>
</tr>
</tbody>
</table>
Chapter 3. Models with Random Effects

Abstract. This chapter considers the Chapter 2 data structure but here the heterogeneity is modeled using random quantities in lieu of fixed parameters; these random quantities are known as random effects. By introducing random quantities, the analysis of longitudinal and panel data can now be cast in the mixed linear model framework.

Although mixed linear models are an established part of statistical methodology, their use is not as widespread as regression. Thus, the chapter introduces this modeling framework, beginning with the special case of a single random intercept known as the error components model and then focusing on the linear mixed effects model that is particularly important for longitudinal data. After introducing the models, this chapter describes estimation of regression coefficients and variance components, as well as hypothesis testing for regression coefficients.

3.1 Error components / random intercepts model

Sampling and inference

Suppose that you are interested in studying the behavior of individuals that are randomly selected from a population. For example, in Section 3.2 we will study the effects that an individual’s economic and demographic characteristics have on the amount of income tax paid. Here, the set of subjects that we will study is randomly selected from a larger database that is itself a random sample of the US taxpayers. In contrast, the Chapter 2 Medicare example dealt with a fixed set of subjects. That is, it is difficult to think of the 54 states as a subset from some “super-population” of states. For both situations, it is natural to use subject-specific parameters, \{\alpha_i\}, to represent the heterogeneity among subjects. Unlike Chapter 2, Chapter 3 discusses situations in which it is more reasonable to represent \{\alpha_i\} as random variables instead of fixed, yet unknown, parameters. By arguing that \{\alpha_i\} are draws from a distribution, we will have the ability to make inferences about subjects in a population that are not included in the sample.

Basic model and assumptions

The error components model equation is

\[ y_{it} = \alpha_i + x_{it}' \beta + \epsilon_{it}. \] (3.1)

This portion of the notation is the same as the error representation of the basic fixed effects model. However, now the term \( \alpha_i \) is assumed to be a random variable, not a fixed, unknown parameter. The term \( \alpha_i \) is known as a random effect. Mixed effects models are ones that include random as well as fixed effects. Because equation (3.1) includes random effects (\( \alpha_i \)) and fixed effects (\( \beta \)), the error components model is a special case of the mixed linear model.

To complete the specification of the error components model, we assume that \{\alpha_i\} are identically and independently distributed with mean zero and variance \( \sigma_{\alpha}^2 \). Further, we assume that \{\alpha_i\} are independent of the error random variables, \{\epsilon_{it}\}. For completeness, we still assume that \( x_{it} \) is a vector of covariates, or explanatory variables, and that \( \beta \) is a vector of fixed, yet unknown, population parameters. Note that because \( E \alpha_i = 0 \), it is customary to include a constant
within the vector \( x_{it} \). This was not true of the fixed effects models in Chapter 2 where we did not center the subject-specific terms about 0.

Linear combinations of the form \( x_{it}' \beta \) quantify the effect of known variables that may affect the response. Additional variables, that are either unimportant or unobservable, comprise the “error term.” In the error components model, we may think of a regression model \( y_{it} = x_{it}' \beta + \eta_{it} \), where the error term \( \eta_{it} \) is decomposed into two components so that \( \eta_{it} = \alpha_i + \varepsilon_{it} \). The term \( \alpha_i \) represents the time-constant portion whereas \( \varepsilon_{it} \) represents the remaining portion. To identify the model parameters, we assume that the two terms are independent. In the biological sciences, the error components model is known as the *random intercepts model*; this descriptor is used because the intercept \( \alpha_i \) is a random variable. We will use the descriptors “error components” and “random intercepts” interchangeably although, for simplicity, we often use only the former term.

**Traditional ANOVA set-up**

In the error components model, the terms \( \{\alpha_i\} \) account for the heterogeneity among subjects. To help interpret this feature, consider the special case where \( K = 1, x_{it} = 1 \) and denote \( \mu = \beta_1 \). In this case, equation (3.1) contains no explanatory variables and reduces to

\[
y_{it} = \mu + \alpha_i + \varepsilon_{it},
\]

the traditional random effects, one-way ANOVA model. Neter and Wasserman (1974G) describe this classic model. This model can be interpreted as arising from a two-stage sampling scheme:

**Stage 1.** Draw a random sample of \( n \) subjects from a population. The subject-specific parameter \( \alpha_i \) is associated with the \( i \)th subject.

**Stage 2.** Conditional on \( \alpha_i \), draw realizations of \( \{y_{it}\} \), for \( t = 1, \ldots, T_i \) for the \( i \)th subject.

That is, in the first stage, we draw a sample from a population of subjects. In the second stage, we observe each subject over time. Because the first stage is considered a random draw from a population of subjects, we represent characteristics that do not depend on time through the random quantity \( \alpha_i \). Figure 3.1 illustrates the two-stage sampling scheme.

**Stage 1**

**Stage 2**

*Figure 3.1. Two-stage random effects sampling.* In the left panel, unobserved subject-specific components are drawn from an unobserved population. In the right panel, several observations are drawn for each subject. These observations are centered about the unobserved subject-specific components from the first stage. Different plotting symbols represent draws for different subjects.
Within this traditional model, interest generally centers about the distribution of the population of subjects. For example, the parameter $\text{Var } \alpha_i = \sigma^2_\alpha$ summarizes the heterogeneity among subjects. In Chapter 2 on fixed effects models, we examined the heterogeneity issue through a test of the null hypothesis $H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_n$. In contrast, under the random effects model, we examine the null hypothesis $H_0: \sigma^2_\alpha = 0$. Furthermore, estimates of $\sigma^2_\alpha$ are of interest but require scaling to interpret. A more useful quantity to report is $\sigma^2_\alpha / (\sigma^2_\alpha + \sigma^2)$, the intra-class correlation. As we saw in Section 2.5.1, this quantity can be interpreted as the correlation between observations within a subject. The correlation is constrained to lie between 0 and 1 and does not depend on the units of measurement for the response. Further, it can also be interpreted as the proportion of variability of a response that is due to heterogeneity among subjects.

**Sampling and model assumptions**

The Section 2.1 basic fixed effects and error components models are similar in appearance yet, as will be discussed in Section 7.2, can lead to different substantive conclusions in the context of a specific application. As we have described, the choice between these two models is dictated primarily by the method in which the sample is drawn. On the one hand, selecting subjects based on a two-stage, or cluster, sample implies use of the random effects model. On the other hand, selecting subjects based on exogenous characteristics suggests a stratified sample and thus using a fixed effects model.

The sampling basis allows us to restate the error components model, as follows.

**Error Components Model Assumptions**

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1.</td>
<td>$E(y_{it}</td>
</tr>
<tr>
<td>R2.</td>
<td>${x_{it,1}, \ldots, x_{it,K}}$ are nonstochastic variables.</td>
</tr>
<tr>
<td>R3.</td>
<td>$\text{Var}(y_{it}</td>
</tr>
<tr>
<td>R4.</td>
<td>${y_{it}}$ are independent random variables, conditional on ${\alpha_1, \ldots, \alpha_n}$.</td>
</tr>
<tr>
<td>R5.</td>
<td>$y_{it}$ is normally distributed, conditional on ${\alpha_1, \ldots, \alpha_n}$.</td>
</tr>
<tr>
<td>R6.</td>
<td>$E \alpha_i = 0$, $\text{Var } \alpha_i = \sigma^2_\alpha$ and ${\alpha_1, \ldots, \alpha_n}$ are mutually independent.</td>
</tr>
<tr>
<td>R7.</td>
<td>${\alpha_i}$ is normally distributed.</td>
</tr>
</tbody>
</table>

Assumptions R1-R5 are similar to the fixed effects models assumptions F1-F5; the main difference is that we now condition on random subject-specific terms, $\{\alpha_1, \ldots, \alpha_n\}$. Assumptions R6 and R7 summarize the sampling basis of the subject-specific terms. Taken together, these assumptions comprise our error components model.

However, assumptions R1-R7 do not provide an “observables” representation of the model because they are based on unobservable quantities, $\{\alpha_1, \ldots, \alpha_n\}$. We summarize the effects of Assumptions R1-R7 on the observable variables, $\{x_{it,1}, \ldots, x_{it,K}, y_{it}\}$.

**Observables Representation of the Error Components Model**

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RO1.</td>
<td>$E y_{it} = x_{it}' \beta$.</td>
</tr>
<tr>
<td>RO2.</td>
<td>${x_{it,1}, \ldots, x_{it,K}}$ are nonstochastic variables.</td>
</tr>
<tr>
<td>RO3.</td>
<td>$\text{Var } y_{it} = \sigma^2 + \sigma^2_\alpha$ and $\text{Cov}(y_{ir}, y_{is}) = \sigma^2_\alpha$, for $r \neq s$.</td>
</tr>
<tr>
<td>RO4.</td>
<td>${y_{it}}$ are independent random vectors.</td>
</tr>
<tr>
<td>RO5.</td>
<td>${y_i}$ is normally distributed.</td>
</tr>
</tbody>
</table>

To reiterate, the properties RO1-5 are a consequence of R1-R7. As we progress into more complex situations, our strategy will consist of using sampling bases to suggest basic assumptions, such as R1-R7, and then convert them into testable properties such as RO1-5. Inference about the testable properties then provides information about the more basic assumptions. When considering nonlinear models beginning in Chapter 9, this conversion will not
be as direct. In some instances, we will focus on the observable representation directly and refer to it as a *marginal or population-averaged model*. The marginal version emphasizes the assumption that observations are correlated within subjects (Assumption RO3), not the random effects mechanism for inducing the correlation.

For more complex situations, it will be useful to describe these assumptions in matrix notation. As in equation (2.13), the regression function can be expressed more compactly as

$$E(y_i | \alpha_i) = \alpha_i \mathbf{1}_i + X_i \beta$$

and thus,

$$E y_i = X_i \beta.$$  \hspace{1cm} (3.2)

Recall that $\mathbf{1}_i$ is a $T_i \times 1$ vector of ones and, from equation (2.14), that $X_i$ is a $T_i \times K$ matrix of explanatory variables, $X_i = \left( x_{i1}, x_{i2}, \ldots, x_{iT_i} \right)$. The expression for $E (y_i | \alpha_i)$ is a restatement of Assumption R1 in matrix notation. Equation (3.2) is a restatement of Assumption RO1. Alternatively, equation (3.2) is due to the law of iterated expectations and assumptions R1 and R6, because $E y_i = E E (y_i | \alpha_i) = E \alpha_i \mathbf{1}_i + X_i \beta = X_i \beta$. For Assumption RO3, we have

$$\text{Var} y_i = V_i = \sigma^2 \mathbf{J}_i + \sigma^2 \mathbf{I}_i.$$  \hspace{1cm} (3.3)

Here, recall that $\mathbf{J}_i$ is a $T_i \times T_i$ matrix of ones, and $\mathbf{I}_i$ is a $T_i \times T_i$ identity matrix.

**Structural models**

Model assumptions are often dictated by sampling procedures. However, we also wish to consider stochastic models that represent causal relationships suggested by a substantive field, known as *structural models*. Section 6.1 describes structural modeling in longitudinal and panel data analysis. To illustrate, in models of economic applications, it is important to consider carefully what one means by the “population of interest.” Specifically, when considering choices of economic entities, a standard defense for a probabilistic approach to analyzing economic decisions is that, although there may be a finite number of economic entities, there is an infinite range of economic decisions. For example, in the Chapter 2 Medicare hospital cost example, one may argue that each state faces a distribution of infinitely many economic outcomes and that this is the population of interest. This viewpoint argues that one should use an error components model. Here, we interpret $\{\alpha_i\}$ to represent those aspects of the economic outcome that are unobservable yet constant over time. In contrast, in Chapter 2 we implicitly used the sampling based model to interpret $\{\alpha_i\}$ as fixed effects.

This viewpoint is the standard rationale for studying stochastic economics. To illustrate, a quote from Haavelmo (1944E) is related to this point:

“… the class of populations we are dealing with does not consist of an infinity of different individuals, it consists of an infinity of possible decisions which might be taken with respect to the value of $y$.”

This defense is well summarized by Nerlove and Balestra in a monograph edited by Mátyás and Sevestre (1996E, Chapter 1) in the context of panel data modeling.

**Inference**

When designing a longitudinal study and considering whether to use a fixed or random effects model, keep in mind the purposes of the study. If you would like to make statements about a population larger than the sample, then use the random effects model. Conversely, if you are simply interested in controlling for subject-specific effects (treating them as nuisance parameters) or in making predictions for a specific subject, then use the fixed effects model.
Time-constant variables

When designing a longitudinal study and considering whether to use a fixed or random effects model, also keep in mind the variables of interest. Often, the primary interest is in testing for the effect of a time-constant variable. To illustrate, in our taxpayer example, we may be interested in the effects that gender may have on an individual’s tax liability (we assume that this variable does not change for an individual over the course of our study). Another important example of a time-constant variable is a variable that classifies subjects into groups. Often, we wish to compare the performance of different groups, for example, a “treatment group” and a “control group.”

In Section 2.3, we saw that time-constant variables are perfectly collinear with subject-specific intercepts and hence are inestimable. In contrast, it will turn out that coefficients associated with time-constant variables are estimable in a random effects model. Hence, if a time-constant variable such as gender or treatment group is the primary variable of interest, one should design the longitudinal study so that a random effects model can be used.

Degrees of freedom

When designing a longitudinal study and considering whether to use a fixed or random effects model, also keep in mind the size of the data set necessary for inference. In most longitudinal data studies, inference about the population parameters \( \beta \) is the primary goal whereas the terms \( \{ \alpha_i \} \) are included to control for the heterogeneity. In the basic fixed effects model, we have seen that there are \( n+K \) linear regression parameters plus 1 variance parameter. This is compared to only \( 1+K \) regression plus 2 variance parameters in the basic random effects model. Particularly in studies where the time dimension is small (such as \( T = 2 \) or 3), a design suggesting a random effects model may be preferable because fewer degrees of freedom are necessary to account for the subject-specific parameters.

GLS estimation

Equations (3.2) and (3.3) summarize the mean and variance of the vector of responses. To estimate regression coefficients, this chapter uses generalized least squares (GLS) equations of the form:

\[
\left( \sum_{i=1}^{n} X_i' V_i^{-1} X_i \right) \beta = \sum_{i=1}^{n} X_i' V_i^{-1} y_i.
\]

The solution of these equations yields generalized least square estimators that, in this context, we call the error components estimator of \( \beta \). Additional algebra (Exercise 3.1) shows that this estimator can be expressed as

\[
b_{EC} = \left( \sum_{i=1}^{n} X_i' \left( I_i - \frac{\zeta_i}{T_i} J_i \right) X_i \right)^{-1} \sum_{i=1}^{n} X_i' \left( I_i - \frac{\zeta_i}{T_i} J_i \right) y_i, \tag{3.4}
\]

Here, the quantity \( \zeta_i = \frac{T_i \sigma^2}{T_i \sigma^2 + \sigma^2} \) is a function of the variance components \( \sigma^2_{\alpha} \) and \( \sigma^2 \). In Chapter 4, we will refer to this quantity as the credibility factor. Further, the variance of the error components estimator turns out to be

\[
\text{Var} b_{EC} = \sigma^2 \left( \sum_{i=1}^{n} X_i' \left( I_i - \frac{\zeta_i}{T_i} J_i \right) X_i \right)^{-1}.
\]
To interpret $b_{EC}$, we give an alternative form for the corresponding Chapter 2 fixed effects estimator. That is, from equation (2.6) and some algebra, we have

$$b = \left( \sum_{i=1}^{n} X_i (I_i - T_i^{-1} J_i) X_i \right)^{-1} \sum_{i=1}^{n} X_i (I_i - T_i^{-1} J_i) y_i.$$ 

Thus, we see that the random effects $b_{EC}$ and fixed effects $b$ are approximately equal when the credibility factors are close to one. This occurs when $\sigma_2$ is large relative to $\sigma$. Intuitively, when there is substantial separation among the intercept terms, relative to the uncertainty in the observations, we anticipate that the fixed and random effect estimators will behave similarly. Conversely, equation (3.3) shows that $b_{EC}$ is approximately equal to an ordinary least squares estimator when $\sigma_2$ is large relative to $\sigma_2^2$ (so that the credibility factors are close to zero). Section 7.2 further develops the comparison among these alternative estimators.

**Feasible generalized least squares estimator**

The calculation of the GLS estimator in equation (3.4) assumes that the variance components $\sigma_2$ and $\sigma$ are known.

**Procedure for computing a “feasible” generalized least squares estimator**

1. First run a regression assuming $\sigma_2 = 0$, resulting in an ordinary least squares estimate of $\beta$.
2. Use the residuals from Step 1 to determine estimates of $\sigma_2$ and $\sigma$.
3. Using the estimates of $\sigma_2$ and $\sigma$ from Step 2, determine $b_{EC}$ using equation (3.4).

For Step 2, there are many ways of estimating the variance components. Section 3.5 provides details. This procedure could be iterated. However, studies have shown that iterated versions do not improve the performance of the one-step estimators. See, for example, Carroll and Rupert (1988G).

To illustrate, we consider some simple moment-based estimators of $\sigma_2$ and $\sigma$ due to Baltagi and Chang (1994E). Define the residuals $e_{it} = y_{it} - (a_i + x_{it}' b)$ using $a_i$ and $b$ according to the Chapter 2 fixed effects estimators in equations (2.6) and (2.7). Then, the estimator of $\sigma_2$ is $s_2$ as given in equation (2.11). The estimator of $\sigma_2$ is:

$$s_2 = \frac{\sum_{i=1}^{n} T_i (a_i - \bar{a}_w)^2 - s_2 c_n}{N - \sum_{i=1}^{n} T_i^2 / N},$$

where $\bar{a}_w = N^{-1} \sum_{i=1}^{n} T_i a_i$ and

$$c_n = n - 1 + \text{trace} \left( \sum_{i=1}^{n} \sum_{t=1}^{T_i} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' \right)^{-1} \sum_{i=1}^{n} T_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x}').$$

A potential drawback is that a particular realization of $s_2$ may be negative; this feature is undesirable for a variance estimator.

**Pooling test**

As with the traditional random effects ANOVA model, the test for heterogeneity, or pooling test, is written as a test of the null hypothesis $H_0$: $\sigma_2 = 0$. That is, under the null hypothesis, we do not have to account for subject-specific effects. Although this is a difficult issue for the general case, in the special case of error components, desirable test procedures have been developed. We discuss here a test that extends a Lagrange multiplier test statistic, due to
Breusch and Pagan (1980E), to the unbalanced data case. (See Appendix C.7 for an introduction to Lagrange multiplier statistics.) This test is a simpler version of one developed by Baltagi and Li (1990E) for a more complex model (specifically, a two-way error component model that we will introduce in Chapter 6).

### Pooling test procedure

1. Run the pooled cross-sectional regression model 
   \[ y_{it} = x_{it}' \beta + \varepsilon_{it}, \]
   to get residuals \( e_{it}. \)

2. For each subject, compute an estimator of \( \sigma^2_\alpha, s_i = \frac{1}{T_i(T_i - 1)} \left( T_i \bar{\varepsilon}_i^2 - \sum_{t=1}^{T_i} \varepsilon_{it}^2 \right), \)
   where
   \[ \bar{\varepsilon}_i = T_i^{-1} \sum_{t=1}^{T_i} e_{it}. \]

3. Compute the test statistic, \( TS = \frac{1}{2n} \left( \frac{\sum_{i=1}^{n} s_i \sqrt{T_i(T_i - 1)}}{N^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T_i} \varepsilon_{it}^2} \right)^2. \)

4. Reject \( H_0 \) if \( TS \) exceeds a percentile from a \( \chi^2 \) (chi-square) distribution with one degree of freedom. The percentile is one minus the significance level of the test.

Note that the pooling test procedure uses estimators of \( \sigma^2_\alpha, s_i, \) that may be negative with positive probability. Section 5.4 discusses alternative procedures where we restrict variance estimators to be nonnegative.

### 3.2 Example: Income tax payments

In this section, we study the effects that an individual’s economic and demographic characteristics have on the amount of income tax paid. Specifically, the response of interest is \( \text{LNTAX}, \) defined as the natural logarithm of the liability on the tax return. Table 3.1 describes several taxpayer characteristics that may affect tax liability.

The data for this study are from the Statistics of Income (SOI) Panel of Individual Returns, a part of the Ernst and Young/University of Michigan Tax Research Database. The SOI Panel represents a simple random sample of unaudited individual income tax returns filed for tax years 1979-1990. The data are compiled from a stratified probability sample of unaudited individual income tax returns, Forms 1040, 1040A and 1040EZ, filed by U.S. taxpayers. The estimates that are obtained from these data are intended to represent all returns filed for the income tax years under review. All returns processed are subjected to sampling except tentative and amended returns.

We examine a balanced panel from 1982-1984 and 1986-1987 taxpayers included in the SOI panel; a four percent sample of this comprises our sample of 258 taxpayers. These years are chosen because they contain the interesting information on paid preparer usage. Specifically, these data include line item tax return data plus a binary variable noting the presence of a paid tax preparer for years 1982-1984 and 1986-1987. These data are also analyzed in Frischmann and Frees (1999O).

The primary goal of this analysis is to determine whether tax preparers significantly affect tax liability. To motivate this question, we note that preparers have the opportunity to impact virtually every line item on a tax return. Our variables are selected because they appear consistently in prior research and are largely outside the influence of tax preparers (that is, they are “exogenous”). Briefly, our explanatory variables are as follows: MS, HH, AGE, EMP, and PREP are binary variables coded for married, head-of-household, at least 65 years of age, self-
employed, and paid preparer, respectively. Further, DEPEND is the number of dependents and MR is the marginal tax rate measure. Finally, LNTPI and LNTAX are the total positive income and tax liability as stated on the return in 1983 dollars, in logarithmic units.

| Demographic Characteristics | | | | | |
|----------------------------|-------------------------------|------------------|------------------|------------------|
| MS                         | is a binary variable, one if the taxpayer is married and zero otherwise. | | | | |
| HH                         | is a binary variable, one if the taxpayer is the head of household and zero otherwise. | | | | |
| DEPEND                     | is the number of dependents claimed by the taxpayer. | | | | |
| AGE                        | is a binary variable, one if the taxpayer is age 65 or over and zero otherwise. | | | | |

| Economic Characteristics | | | | | |
|----------------------------|-------------------------------|------------------|------------------|------------------|
| LNTPI                      | is the natural logarithm of the sum of all positive income line items on the return, in 1983 dollars. | | | | |
| MR                         | is the marginal tax rate. It is computed on total personal income less exemptions and the standard deduction. | | | | |
| EMP                        | is a binary variable, one if Schedule C or F is present and zero otherwise. Self-employed taxpayers have greater need for professional assistance to reduce the reporting risks of doing business. | | | | |
| PREP                       | is a variable indicating the presence of a paid preparer. | | | | |
| LNTAX                      | is the natural logarithm of the tax liability, in 1983 dollars. This is the response variable of interest. | | | | |

Tables 3.2 and 3.3 describe the basic taxpayer characteristics used in our analysis. The binary variables in Table 3.2 indicate that over half the sample is married (MS) and approximately half the sample uses a paid preparer (PREP). Preparer use appears highest in 1986 and 1987, years straddling significant tax law change. Slightly less than ten percent of the sample is 65 or older (AGE) in 1982. The presence of self-employment income (EMP) also varies over time.

<table>
<thead>
<tr>
<th>YEAR</th>
<th>MS</th>
<th>HH</th>
<th>AGE</th>
<th>EMP</th>
<th>PREP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1982</td>
<td>0.597</td>
<td>0.081</td>
<td>0.085</td>
<td>0.140</td>
<td>0.450</td>
</tr>
<tr>
<td>1983</td>
<td>0.597</td>
<td>0.093</td>
<td>0.105</td>
<td>0.159</td>
<td>0.442</td>
</tr>
<tr>
<td>1984</td>
<td>0.624</td>
<td>0.085</td>
<td>0.112</td>
<td>0.155</td>
<td>0.484</td>
</tr>
<tr>
<td>1986</td>
<td>0.647</td>
<td>0.081</td>
<td>0.132</td>
<td>0.147</td>
<td>0.508</td>
</tr>
<tr>
<td>1987</td>
<td>0.647</td>
<td>0.093</td>
<td>0.147</td>
<td>0.147</td>
<td>0.516</td>
</tr>
</tbody>
</table>

The summary statistics for the other non-binary variables are in Table 3.3. Further analyses indicate an increasing income trend, even after adjusting for inflation, as measured by total positive income (LNTPI). Moreover, both the mean and median marginal tax rates (MR) are decreasing, although mean and median tax liabilities (LNTAX) are stable (see Figure 3.2). These results are consistent with congressional efforts to reduce rates and expand the tax base through broadening the definition of income and eliminating deductions.
### Table 3.3 Summary Statistics for Other Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Median</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEPEND</td>
<td>2.419</td>
<td>2.000</td>
<td>0.000</td>
<td>6.000</td>
<td>1.338</td>
</tr>
<tr>
<td>LNTPI</td>
<td>9.889</td>
<td>10.051</td>
<td>-0.128</td>
<td>13.222</td>
<td>1.165</td>
</tr>
<tr>
<td>MR</td>
<td>23.523</td>
<td>22.000</td>
<td>0.000</td>
<td>50.000</td>
<td>11.454</td>
</tr>
<tr>
<td>LNTAX</td>
<td>6.880</td>
<td>7.701</td>
<td>0.000</td>
<td>11.860</td>
<td>2.695</td>
</tr>
</tbody>
</table>

**Figure 3.2. Boxplot of LNTAX versus YEAR.** Logarithmic tax liability (in real dollars) is stable over the years 1982-1987.

To explore the relationship between each indicator variable and logarithmic tax, Table 3.4 presents the average logarithmic tax liability by level of indicator variable. This table shows that married filers pay greater tax, head of household filers pay less tax, taxpayers 65 or over pay less, taxpayers with self-employed income pay less and taxpayers that use a professional tax preparer pay more.

### Table 3.4 Averages of Logarithmic Tax by Level of Explanatory Variable

<table>
<thead>
<tr>
<th>Explanatory Variable</th>
<th>MS</th>
<th>HH</th>
<th>AGE</th>
<th>EMP</th>
<th>PREP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level of Explanatory Variable</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5.973</td>
<td>7.013</td>
<td>6.939</td>
<td>6.983</td>
<td>6.624</td>
</tr>
<tr>
<td></td>
<td>7.430</td>
<td>5.480</td>
<td>6.431</td>
<td>6.297</td>
<td>7.158</td>
</tr>
</tbody>
</table>
Table 3.5 summarizes basic relations among logarithmic tax and the other non-binary explanatory variables. Both LNTPI and MR are strongly correlated with logarithmic tax whereas the relationship between DEPEND and logarithmic tax is positive, yet weaker. Further, Table 3.5 shows that LNTPI and MR are strongly positively correlated.

<table>
<thead>
<tr>
<th>TABLE 3.5 Correlation Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>LNTPI</td>
</tr>
<tr>
<td>MR</td>
</tr>
<tr>
<td>LNTAX</td>
</tr>
</tbody>
</table>

Although not presented in detail here, exploration of the data revealed several other interesting relationships among the variables. To illustrate, a basic added variable plot in Figure 3.3 shows the strong relation between logarithmic tax liability and total income, even after controlling for subject-specific time-constant effects.

The error components model described in Section 3.1 was fit, using the explanatory variables described in Table 3.1. The estimated model appears in Display 3.1, from a fit using the statistical package SAS. Display 3.1 shows that HH, EMP, LNTPI and MR are statistically significant variables that affect LNTAX. Somewhat surprisingly, the PREP variable was not statistically significant.

To test for the importance of heterogeneity, the Section 3.1 pooling test was performed. A fit of the pooled cross-sectional model, with the same explanatory variables, produced residuals and an error sum of squares equal to $Error SS = 3599.73$. Thus, with $T = 5$ years and $n = 258$ subjects, the test statistic is $TS = 273.5$. Comparing this test statistic to a chi-square distribution with one degree of freedom indicates that the null hypothesis of homogeneity is rejected. As we
will see in Chapter 7, there are some unusual features of this data set that cause this test statistic to be large.

**Display 3.1 Selected SAS Output**

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Evaluations</th>
<th>-2 Log Like</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>4984.68064143</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>4791.25465804</td>
<td>0.00000001</td>
</tr>
</tbody>
</table>

Convergence criteria met.

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Subject</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>SUBJECT</td>
<td>0.9217</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>1.8740</td>
</tr>
</tbody>
</table>

**Fit Statistics**

-2 Log Likelihood 4791.3
AIC (smaller is better) 4813.3
AICC (smaller is better) 4813.5
BIC (smaller is better) 4852.3

**Solution for Fixed Effects**

<table>
<thead>
<tr>
<th>Effect</th>
<th>Estimate</th>
<th>Error</th>
<th>DF</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>t</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-2.9604</td>
<td>0.5686</td>
<td>257</td>
<td>-5.21</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MS</td>
<td>0.03730</td>
<td>0.1818</td>
<td>1024</td>
<td>0.21</td>
<td>0.8375</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HH</td>
<td>-0.6890</td>
<td>0.2312</td>
<td>1024</td>
<td>-2.98</td>
<td>0.0029</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE</td>
<td>0.02074</td>
<td>0.1993</td>
<td>1024</td>
<td>0.10</td>
<td>0.9171</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMP</td>
<td>-0.5048</td>
<td>0.1674</td>
<td>1024</td>
<td>-3.02</td>
<td>0.0026</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PREP</td>
<td>-0.02170</td>
<td>0.1171</td>
<td>1024</td>
<td>-0.19</td>
<td>0.8530</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LNTPI</td>
<td>0.7604</td>
<td>0.06972</td>
<td>1024</td>
<td>10.91</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DEPEND</td>
<td>-0.1128</td>
<td>0.05907</td>
<td>1024</td>
<td>-1.91</td>
<td>0.0566</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MR</td>
<td>0.1154</td>
<td>0.007288</td>
<td>1024</td>
<td>15.83</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3.3 Mixed effects models

Similar to the extensions for the fixed effects model described in Section 2.5, we now extend the error components model to allow for variable slopes, serial correlation and heteroscedasticity.

3.3.1 Linear mixed effects model

We now consider conditional regression functions of the form

\[ E( y_{it} | \alpha_i ) = z_{it}' \alpha_i + x_{it}' \beta. \]  

(3.5)

Here, the term \( z_{it}' \alpha_i \) comprises the random effects portion of the model. The term \( x_{it}' \beta \) comprises the fixed effects portion. As with equation (2.15) for fixed effects, equation (3.5) is short-hand notation for

\[ E( y_{it} | \alpha_i ) = \alpha_{i1} z_{it1} + \alpha_{i2} z_{it2} + ... + \alpha_{iq} z_{itq} + \beta_1 x_{it1} + \beta_2 x_{it2} + ... + \beta_K x_{itK}. \]

As in equation (2.16), a matrix form of equation (3.5) is

\[ E( y_i | \alpha_i ) = Z_i \alpha_i + X_i \beta. \]

(3.6)

We also wish to allow for serial correlation and heteroscedasticity. Similar to Section 2.5.1 for fixed effects, we can incorporate these extensions through the notation \( \text{Var}( y_i | \alpha_i ) = R_i \). We maintain the assumption that the responses between subjects are independent.

Further, we assume that the subject-specific effects \{\alpha_i\} are independent with mean \( E \alpha_i = 0 \) and variance-covariance matrix \( \text{Var} \alpha_i = D \), a \( q \times q \) positive definite matrix. By assumption, the random effects are mean zero; thus, any nonzero mean for a random effect must be expressed as part of the fixed effects terms. The columns of \( Z \) are usually a subset of the columns of \( X \).

Taken together, these assumptions comprise what we term the linear mixed effects model.

Linear Mixed Effects Model Assumptions

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1.</td>
<td>( E( y_i</td>
</tr>
<tr>
<td>R2.</td>
<td>{( x_{it1}, ..., x_{itK})} and {( z_{it1}, ..., z_{itq})} are nonstochastic variables.</td>
</tr>
<tr>
<td>R3.</td>
<td>( \text{Var}( y_i</td>
</tr>
<tr>
<td>R4.</td>
<td>{( y_i )} are independent random vectors, conditional on {( \alpha_i, ..., \alpha_n )}.</td>
</tr>
<tr>
<td>R5.</td>
<td>{( y_i )} is normally distributed, conditional on {( \alpha_i, ..., \alpha_n )}.</td>
</tr>
<tr>
<td>R6.</td>
<td>( E \alpha_i = 0 ), ( \text{Var} \alpha_i = D ) and {( \alpha_1, ..., \alpha_n )} are mutually independent.</td>
</tr>
<tr>
<td>R7.</td>
<td>{( \alpha_i )} is normally distributed.</td>
</tr>
</tbody>
</table>

With assumptions R3 and R6, the variance of each subject can be expressed as

\[ \text{Var}( y_i ) = Z_i D Z_i' + R_i = V_i(\tau) = V_i. \]

(3.7)

The notation \( V_i(\tau) \) means that the variance-covariance matrix of \( y_i \) depends on variance components \( \tau \). Section 2.5.1 provided several examples that illustrate how \( R_i \) may depend on \( \tau \); we will give special cases to show how \( V_i \) may depend on \( \tau \).

With this, we may summarize the effects of Assumptions R1-R7 on the observables variables, \{\( x_{it1}, ..., x_{itK}, z_{it1}, ..., z_{itq}, y_{it} \)\}. 

Observables Representation of the Linear Mixed Effects Model

RO1. \( E y_i = X_\beta \).

RO2. \( \{x_{it,1}, \ldots, x_{it,k}\} \) and \( \{z_{it,1}, \ldots, z_{it,q}\} \) are nonstochastic variables.

RO3. \( \text{Var} y_i = Z_i D Z_i^\prime + R_i = V_i(\tau) = V_i \).

RO4. \( \{y_i\} \) are independent random vectors.

RO5. \( \{y_i\} \) is normally distributed.

As in Chapter 2 and Section 3.1, the properties RO1-5 are a consequence of R1-R7. We focus on these properties because they are the basis for testing our specification of the model. The observable representation is also known as a marginal or population-averaged model.

Example – Trade localization

Feinberg, Keane and Bognano (1998E) studied \( n = 701 \) U.S. based multinational corporations over the period 1983-1992. Using firm-level data available from the Bureau of Economic Analysis of the U.S. Department of Commerce, they documented how large corporation’s allocation of employment and durable assets (property, plant and equipment) of Canadian affiliates changed in response to changes in Canadian and U.S. tariffs. Specifically, their model was

\[
\ln y_{it} = \beta_1 CT_{it} + \beta_2 UT_{it} + \beta_3 \text{Trend}_t + x_{it}^* \beta + \epsilon_{it}
\]

\[
= (\beta_1 + \alpha_1) CT_{it} + (\beta_2 + \alpha_2) UT_{it} + (\beta_3 + \alpha_3) \text{Trend}_t + x_{it}^* \beta + \epsilon_{it}
\]

\[
= \alpha_1 CT_{it} + \alpha_2 UT_{it} + \alpha_3 \text{Trend}_t + x_{it}^* \beta + \epsilon_{it}.
\]

Here, \( CT_{it} \) is the sum over all industry Canadian tariffs in which firm \( i \) belongs, and similarly for \( UT_{it} \). The vector \( x_{it} \) includes \( CT_{it}, UT_{it} \) and \( \text{Trend}_t \) (for the mean effects), as well as real U.S. and Canadian wages, gross domestic product, price-earnings ratio, real U.S. interest rates and a measure of transportation costs. For the response, they used both Canadian employment and durable assets.

The first equation emphasizes that response to changes in Canadian and U.S. tariffs, as well as time trends, is firm-specific. The second equation provides the link to the third expression that is in terms of the linear mixed effects model form. Here, we have included \( CT_{it}, UT_{it} \) and \( \text{Trend}_t \) in \( x_{it}^* \) to get \( x_{it} \). With this reformulation, the mean of each random slope is zero, that is, \( E \alpha_{1i} = E \alpha_{2i} = E \alpha_{3i} = 0 \). In the first specification, the means are \( E \beta_{1i} = \beta_1, E \beta_{2i} = \beta_2 \) and \( E \beta_{3i} = \beta_3 \). Feinberg, Keane and Bognano found that a significant portion of the variation was due to firm-specific slopes; they attribute this variation to idiosyncratic firm differences such as technology and organization. They also allowed for heterogeneity in the time trend. This allows for unobserved time-varying factors (such as technology and demand) that affect individual firms differently.

A major finding of this paper is that Canadian tariff levels were negatively related to assets and employment in Canada; this finding contradicts the hypothesis that lower tariffs would undermine Canadian manufacturing.
Special cases
To help interpret linear mixed effects models, we consider several important special cases. We begin by emphasizing the case where \( q = 1 \) and \( z_{it} = 1 \). In this case, the linear mixed effects model reduces to the error components model, introduced in Section 3.1. For this model, we have only subject-specific intercepts, no subject-specific slopes and no serial correlation.

Repeated measures design
Another classic model is the so-called repeated measures design. Here, several measurements are collected on a subject over a relatively short period of time, under controlled experimental conditions. Each measurement is subject to a different treatment but the order of treatments is randomized so that no serial correlation is assumed.

Specifically, we consider \( i = 1, \ldots, n \) subjects. A response for each subject is measured based on each of \( T \) treatments, where the order of treatments is randomized. The mathematical model is:

\[
Y_{it} = \alpha_i + \beta_i + \epsilon_{it}.
\]

The main research question of interest is \( H_0: \beta_1 = \beta_2 = \ldots = \beta_T \), that is, the null hypothesis is no treatment differences.

The repeated measures design is a special case of equation (3.5), taking \( q = 1, z_{it} = 1, T_i = T, K = T \) and using the \( t \)th explanatory variable, \( x_{it,t} \), to indicate whether the \( t \)th treatment has been applied to the response.

Random coefficients model
We now return to the linear mixed effects model and suppose that \( q = K \) and \( z_{it} = x_{it} \). In this case the linear mixed effects model reduces to a random coefficients model, of the form

\[
E(y_{it} | \alpha_i) = X_{it}' \beta_i.
\]

Here, \( \{ \beta_i \} \) are random vectors with mean \( \beta \). The random coefficients model can be easily interpreted as a two-stage sampling model. In the first stage, one draws the \( i \)th subject from a population that yields a vector of parameters \( \beta_i \). From the population, this vector has mean \( E \beta_i = \beta \) and variance \( \text{Var} \beta_i = D \). At the second stage, one draws \( T_i \) observations for the \( i \)th observation, conditional on having observed \( \beta_i \). The mean and variance of the observations are \( E(y_i | \beta_i) = X_i \beta_i \) and \( \text{Var}(y_i | \beta_i) = R_i \). Putting these two stages together yields

\[
E y_i = X_i E \beta_i = X_i \beta
\]

and

\[
\text{Var} y_i = E(\text{Var}(y_i | \beta_i)) + \text{Var}(E(y_i | \beta_i)) = R_i + \text{Var}(X_i \beta_i) = R_i + X_i' D X_i = V_i.
\]

Example – Taxpayer study – Continued
The random coefficients model was fit using the Taxpayer data with \( K = 8 \) variables. The model fitting was done using the statistical package SAS, with the MIVQUE(0) variance components estimation techniques, described in Section 3.5. The resulting fitting \( D \) matrix appears in Table 3.6. Display 3.2 provides additional details of the model fit.
### Table 3.6 Values of the Estimated D Matrix

<table>
<thead>
<tr>
<th></th>
<th>INTERCEPT</th>
<th>MS</th>
<th>HH</th>
<th>AGE</th>
<th>EMP</th>
<th>PREP</th>
<th>LNTPI</th>
<th>MR</th>
<th>DEPEND</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERCEPT</td>
<td>47.86</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MS</td>
<td>-0.40</td>
<td>20.64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HH</td>
<td>-1.26</td>
<td>1.25</td>
<td>23.46</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE</td>
<td>18.48</td>
<td>2.61</td>
<td>-0.79</td>
<td>22.33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMP</td>
<td>-8.53</td>
<td>0.92</td>
<td>0.12</td>
<td>0.21</td>
<td>20.60</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PREP</td>
<td>4.22</td>
<td>-0.50</td>
<td>-1.85</td>
<td>-0.21</td>
<td>-0.50</td>
<td>21.35</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LNTPI</td>
<td>-4.54</td>
<td>-0.17</td>
<td>0.15</td>
<td>-2.38</td>
<td>1.18</td>
<td>-0.38</td>
<td>21.44</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MR</td>
<td>0.48</td>
<td>0.06</td>
<td>-0.03</td>
<td>0.14</td>
<td>-0.09</td>
<td>0.04</td>
<td>-0.09</td>
<td>20.68</td>
<td></td>
</tr>
<tr>
<td>DEPEND</td>
<td>3.07</td>
<td>0.41</td>
<td>0.29</td>
<td>-0.60</td>
<td>-0.40</td>
<td>-0.35</td>
<td>-0.34</td>
<td>0.01</td>
<td>20.68</td>
</tr>
</tbody>
</table>

### Display 3.2 Selected SAS Output for the Random Coefficients Model

**Fit Statistics**

-2 Res Log Likelihood 7876.0
AIC (smaller is better) 7968.0
AICC (smaller is better) 7971.5
BIC (smaller is better) 8131.4

**Solution for Fixed Effects**

| Effect     | Estimate | Error | DF  | t Value | Pr > |t| |
|------------|----------|-------|-----|---------|------|---|
| Intercept  | -9.5456  | 2.1475| 253 | -4.44   | <.0001|
| MS         | -0.3183  | 1.0664| 41  | -0.30   | 0.7668|
| HH         | -1.0514  | 1.4418| 16  | -0.73   | 0.4764|
| AGE        | -0.4027  | 1.1533| 20  | -0.35   | 0.7306|
| EMP        | -0.1498  | 0.9019| 31  | -0.17   | 0.8691|
| PREP       | -0.2156  | 0.6118| 67  | -0.35   | 0.7257|
| LNTPI      | 1.6118   | 0.3712| 257 | 4.34    | <.0001|
| DEPEND     | -0.2814  | 0.4822| 70  | -0.58   | 0.5613|
| MR         | 0.09303  | 0.2853| 250 | 0.33    | 0.7446|

### Variations of the random coefficients model

Certain variations of the two-stage interpretation of the random coefficients models lead to other forms of the random effects model in equation (3.6). To illustrate, in equation (3.6), we may take the columns of $Z_i$ to be a strict subset of the columns of $X_i$. This is equivalent to assuming that certain components of $\beta_i$ associated with $Z_i$ are stochastic whereas other components that are associated with $X_i$ (but not $Z_i$) are nonstochastic.

Note that the convention in equation (3.6) is to assume that the mean of the random effects $\alpha_i$ are known and equal to zero. Alternatively, we could assume that they are unknown with mean, say, $\alpha$, that is, $E \alpha_i = \alpha$. However, this is equivalent to specifying additional fixed effects terms $Z_i \alpha$ in equation (3.6). By convention, we absorb these additional terms into the “$X_i \beta$” potion of the model. Thus, it is customary to include those explanatory variables in the $Z_i$ design matrix as part of the $X_i$ design matrix.

Another variation of the two-stage interpretation uses known variables $B$, such that $E \beta_i = B, \beta$. Then, we have, $E y_i = X_i B, \beta$ and $\text{Var} y_i = R_i + X_i' D X_i$. This is equivalent to our equation (3.6) model replacing $X_i$ with $X_i B, \beta$ and $Z_i$ with $X_i$. Hsiao (1986E, Section 6.5) refers to this as a
variable-coefficients model with coefficients that are functions of other exogenous variables. Chapter 5 describes this approach in greater detail.

---

**Example – Lottery sales**

Section 4.5 will describe a case in which we wish to predict lottery sales. The response variable $y_{it}$ is logarithmic lottery sales in week $t$ for a geographic unit $i$. No time-varying variables are available for these data, so a basic explanation of lottery sales is through the one-way random effects ANOVA model of the form $y_{it} = \alpha_i + \epsilon_{it}$. We can interpret $\alpha_i$ to be the (conditional) mean lottery sales for the $i$th geographic unit. In addition, we will have available several time-constant variables that describe the geographic unit, including population, median household income, median home value and so forth. Denote this set of variables that describe the $i$th geographic unit as $B_i$. With the two-stage interpretation, we could use these variables to explain the mean lottery sales with the representation $\alpha_i = B_i \beta + \alpha_i$. Note that the variable $\alpha_i$ is unobservable, so this model is not estimable by itself. However, when combined with the ANOVA model, we have

$$y_{it} = \alpha_i + B_i \beta + \epsilon_{it},$$

our error components model. This combined model is estimable.

---

**Group effects**

In many applications of longitudinal data analysis, it is of interest to assess differences of responses from different groups. In this context, the term “group” refers to a category of the population. For example, in the Section 3.2 Taxpayer example, we may be interested in studying the differences in tax liability due to gender (male/female) or due to political party affiliation (democrat/republican/libertarian) and so on.

A typical model that includes group effects can be expressed as a special case of the linear mixed effects model, using $q = 1$, $z_{it} = 1$ and the expression

$$E(y_{git} | a_{gi}) = \alpha_{gi} + \delta_{g} + x_{git}' \beta.$$  

Here, the subscripts range over $g = 1, \ldots, G$ groups, $i = 1, \ldots, n_g$ subjects in each group and $t = 1, \ldots, T_{gi}$ observations of each subject. The terms $\{a_{gi}\}$ represent random, subject-specific effects and $\{\delta_{g}\}$ represent fixed differences among groups. An interesting aspect of random effects portion is that subjects need not change groups over time for the model to be estimable. To illustrate, if we were interested in gender differences in tax liability, we would not expect individuals to change gender over such a small sample. This is in contrast to the fixed effects model, where group effects are not estimable due to their collinearity with subject-specific effects.

---

**Time-constant variables**

The study of time-constant variables provides strong motivation for designing a panel, or longitudinal, study that can be analyzed as a linear mixed effects model. Within a linear mixed effects model, both the heterogeneity terms $\{\alpha_i\}$ and parameters associated with time-constant variables can be analyzed simultaneously. This was not the case for the fixed effects models, where the heterogeneity terms and time-constant variables are perfectly collinear. The group effect discussed above is a specific type of time-constant variable. Of course, it is also possible to analyze group effects where individuals switch groups over time, such as with political party affiliation. This type of problem can be handled directly using binary variables to indicate the presence or absence of a group type, and represents no particular difficulties.
We may split the explanatory variables associated with the population parameters into those that vary by time and those that do not (time-constant). Thus, we can write our linear mixed effects conditional regression function as
\[
E(y_{it} | \alpha_i) = \alpha_i' z_{it} + x_{1i}' \beta_1 + x_{2it}' \beta_2.
\]

This model is a generalization of the group effects model.

### 3.3.2 Mixed linear models

In the Section 3.3.1 linear mixed effects models, we assumed independence among subjects (Assumption RO4). This assumption is not tenable for all models of repeated observations on a subject over time, so it is of interest to introduce a generalization known as the **mixed linear model**. This model equation is given by
\[
y = Z \alpha + X \beta + \varepsilon. \tag{3.9}
\]

Here, \(y\) is a \(N \times 1\) vector of responses, \(\varepsilon\) is a \(N \times 1\) vector of errors, \(Z\) and \(X\) are \(N \times q\) and \(N \times K\) matrices of explanatory variables, respectively, and \(\alpha\) and \(\beta\) are \(q \times 1\) and \(K \times 1\) vectors of parameters.

For the mean structure, we assume \(E(y | \alpha) = Z \alpha + X \beta\) and \(E \alpha = 0\), so that \(E y = X \beta\).

For the covariance structure, we assume \(\text{Var}(y | \alpha) = R\), \(\text{Var} \alpha = D\) and \(\text{Cov}(\alpha, \varepsilon') = 0\). This yields \(\text{Var} y = Z D Z' + R = V\).

Unlike the linear mixed effects model in Section 3.3.1, the mixed linear model does not require independence between subjects. Further, the model is sufficiently flexible so that several complex hierarchical structures can be expressed as special cases of it. To see how the linear mixed effects model is a special case of the mixed linear model, take \(y = (y_1', y_2', \ldots, y_n')'\), \(\varepsilon = (\varepsilon_1', \varepsilon_2', \ldots, \varepsilon_n')'\), \(\alpha = (\alpha_1', \alpha_2', \ldots, \alpha_n')'\),

\[
X = \begin{pmatrix}
X_1 \\
X_2 \\
X_3 \\
\vdots \\
X_n
\end{pmatrix}
\quad \text{and} \quad
Z = \begin{pmatrix}
Z_1 & 0 & 0 & \cdots & 0 \\
0 & Z_2 & 0 & \cdots & 0 \\
0 & 0 & Z_3 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & Z_n
\end{pmatrix}.
\]

With these choices, the mixed linear model reduces to the linear mixed effects model.

The **two-way error components** model is an important panel data model that is not a specific type of linear mixed effects model, although it is a special case of the mixed linear model. This model can be expressed as
\[
y_{it} = \alpha_i + \lambda_i + x_{it}' \beta + \varepsilon_{it}. \tag{3.10}
\]

This is similar to the error components model but we have added a random time component, \(\lambda_i\).

We assume that \(\{\lambda_i\}, \{\alpha_i\}\) and \(\{\varepsilon_{it}\}\) are mutually independent. See Chapter 8 for additional details regarding this model.

To summarize, the mixed linear model generalizes the linear mixed effects model and includes other models that are of interest in longitudinal data analysis. Much of the estimation can be accomplished directly in terms of the mixed linear model. To illustrate, in this book many of the examples are analyzed using PROC MIXED, a procedure within the statistical package SAS specifically designed to analyze mixed linear models. The primary advantage of the linear mixed effects model is that it provides a more intuitive platform for examining longitudinal data.
Example – Income inequality


Specifically, the model was

\[
\ln y_{c,i,t} = (1-z_{1,t}) \alpha_{c,1} + z_{1,t} \alpha_{c,2} + \lambda_t + x_{c,i,t}' \beta + z_{1,t} x_{c,i,t}' \beta_2 + \epsilon_{c,i,t}.
\]

Here, \( y_{c,i,t} \) represents income for the \( i \)th subject in the \( c \)th city at time \( t \). The vector \( x_{c,i,t} \) represents several control variables that include gender, age, age squared, education, occupation and work organization (government, other public, and private firms). The variable \( z_{1,t} \) is a binary variable defined to be one if \( t \geq 1985 \) and zero otherwise. Thus, the vector \( \beta \) represents parameter estimates for the explanatory variables before 1985 and \( \beta_2 \) represents the differences after urban reform. The primary interest is in the change of the explanatory variable effects, \( \beta_2 \).

For the other variables, the random effect \( \lambda_t \) is meant to control for undetected time effects. There are two city effects: \((1-z_{1,t}) \alpha_{c,1}\) is for cities before 1985 and \( z_{1,t} \alpha_{c,2} \) is for after 1984. Note that these random effects are at the city level and not at the subject (\( i \)) level. Zhou used a combination of error components and autoregressive structure to model the serial relationships of the disturbance terms. Including these random effects accounted for clustering of responses within both cities and time periods, thus providing more accurate assessment of the regression coefficients \( \beta \) and \( \beta_2 \).

Zhou found significant returns to education and these returns increased in the post-reform era. Little change was found among organization effects, with the exception of significantly increased effects for private firms.

3.4 Inference for regression coefficients

Estimation of the linear mixed effects model proceeds in two stages. In the first stage, we estimate the regression coefficients \( \beta \), assuming knowledge of the variance components \( \tau \). Then, in the second stage, the variance components \( \tau \) are estimated. Section 3.5 discusses variance component estimation whereas this section discusses regression coefficient inference, assuming that the variance components are known.

GLS estimation

From Section 3.3, we have that the vector \( y_i \) has mean \( X_i \beta \) and variance \( Z_i D Z_i' + R_i = V_i(\tau) = V_i \). Thus, direct calculations show that the generalized least squares (GLS) estimator of \( \beta \) is

\[
b_{GLS} = \left( \sum_{i=1}^{n} X_i' V_i^{-1} X_i \right)^{-1} \sum_{i=1}^{n} X_i' V_i^{-1} y_i. \tag{3.11}
\]

The GLS estimator of \( \beta \) takes the same form as in the error components model estimator in equation (3.4) yet with a more general variance covariance matrix \( V_i \). Furthermore, direct calculation show that the variance is

\[
\text{Var} \, b_{GLS} = \left( \sum_{i=1}^{n} X_i' V_i^{-1} X_i \right)^{-1}. \tag{3.12}
\]
As with fixed effects estimators, it is possible to express $b_{GLS}$ as a weighted average of subject-specific estimators. To this end, for the $i$th subject, define the GLS estimator $b_{i,GLS} = (X_i'V_i^{-1}X_i)^{-1}X_i'V_i^{-1}y_i$ and the weight $W_{i,GLS} = X_i'V_i^{-1}X_i$. Then, we can write

$$b_{GLS} = \left(\sum_{i=1}^{n} W_{i,GLS}\right)^{-1} \sum_{i=1}^{n} W_{i,GLS} b_{i,GLS}.$$  

**Matrix inversion formula**

To simplify the calculations and to provide better intuition for our expressions, we cite a formula for inverting $V_i$. Note that the matrix $V_i$ has dimension $T_i \times T_i$. From Appendix A.5, we have

$$V_i^{-1} = (R_i + Z_i D Z_i')^{-1} = R_i^{-1} - R_i^{-1} Z_i \left(\frac{1}{\sigma^2} + \frac{Z_i' R_i^{-1} Z_i}{\sigma^2}\right)^{-1} Z_i' R_i^{-1}.$$  

(3.13)

The expression on the right-hand side of equation (3.13) is easier to compute than the left-hand side when the temporal covariance matrix $R_i$ has an easily computable inverse and the dimension $q$ is smaller than $T_i$. Moreover, because the matrix $D^{-1} + Z_i' R_i^{-1} Z_i$ is only a $q \times q$ matrix, it is easier to invert than $V_i$, a $T_i \times T_i$ matrix.

Some special cases are of interest. First, note that in the case of no serial correlation, we have $Q_i \equiv 0$ and equation (3.13) reduces to

$$V_i^{-1} = \left(\sigma^2 I + \sigma^2 Z_i Z_i'\right)^{-1} = \frac{1}{\sigma^2} \left(I_i - Z_i \left(\sigma^2 I + \sigma^2 Z_i Z_i'\right)^{-1} Z_i'\right).$$  

(3.14)

Further, in the error components model considered in Section 3.1, we have $q = 1$, $D = \sigma^2 I_i$, $Z_i = 1_i$, so that equation (3.13) reduces to

$$V_i^{-1} = \left(\sigma^2 I_i + \sigma^2 Z_i Z_i'\right)^{-1} = \frac{1}{\sigma^2} \left(I_i - \sigma^2 I_i + \sigma^2 Z_i Z_i'\right) = \frac{1}{\sigma^2} \left(I_i - \frac{T_i \sigma^2}{\sigma^2 + \sigma^2} J_i\right).$$  

(3.15)

where $\zeta_i = \frac{T_i \sigma^2}{\sigma^2 + \sigma^2}$, as in Section 3.1. This demonstrates that equation (3.4) is a special case of equation (3.11).

For another special case, consider the random coefficients model ($z_{it} = x_{it}$) with no serial correlation so that $Q_i \equiv 0$. Here, the weight $W_{i,GLS}$ takes on a simple form: $W_{i,GLS} = \left(D + \sigma^2 \left(X_i'X_i\right)\right)^{-1}$ (see Exercise 3.8). From this form, we see that subjects with “large” values of $X_i'X_i$ have a greater effect on $b_{GLS}$ than subjects with smaller values.

**Maximum likelihood estimation**

With assumption R05, the log-likelihood of a single subject is

$$l_i(\beta, \tau) = -\frac{1}{2} \left(T_i \ln(2\pi) + \ln \det V_i(\tau) + (y_i - X_i \beta)' V_i(\tau)^{-1} (y_i - X_i \beta)\right).$$  

(3.16)

With equation (3.16), the log-likelihood for the entire data set is

$$L(\beta, \tau) = \sum_{i=1}^{n} l_i(\beta, \tau).$$

The values of $\beta$ and $\tau$ that maximize $L(\beta, \tau)$ are the maximum likelihood estimators (MLEs) which we denote as $b_{MLE}$ and $\tau_{MLE}$.

**Note to Reader:** We now begin to use likelihood inference extensively. You may wish to review Appendix B for additional background on joint normality and the related likelihood function. Appendix C reviews likelihood estimation in a general context.
The score vector is the vector of derivatives of the log-likelihood taken with respect to the parameters. We denote the vector of parameters by \( \theta = (\beta', \tau')' \). With this notation, the score vector is \( \partial L(\theta)/\partial \theta \). Typically, if this score has a root, then the root is a maximum likelihood estimator. To compute the score vector, first take derivatives with respect to \( \beta \) and find the root. That is,

\[
\sum_{i=1}^{n} X_i' V_{i}(\tau)^{-1} (y_i - X_i \beta) = \sum_{i=1}^{n} X_i' V_{i}(\tau)^{-1} (y_i - X_i \beta).
\]

Setting the score vector equal to zero yields

\[
b_{MLE} = \left( \sum_{i=1}^{n} X_i' V_{i}(\tau)^{-1} X_i \right)^{-1} \sum_{i=1}^{n} X_i' V_{i}(\tau)^{-1} y_i = b_{GLS}. \tag{3.17}
\]

That is, for fixed covariance parameters \( \tau \), the maximum likelihood estimator and the general least squares estimator are the same.

**Robust estimation of standard errors**

Even without the assumption of normality, the maximum likelihood estimator \( b_{MLE} \) has desirable properties. It is unbiased, efficient and asymptotically normal with covariance matrix given in equation (3.12). However, the estimator does depend on knowledge of variance components. As an alternative, it can be useful to consider an alternative, weighted least squares estimator

\[
b_{W} = \left( \sum_{i=1}^{n} X_i' W_{i,RE} X_i \right)^{-1} \sum_{i=1}^{n} X_i' W_{i,RE} V_{i} X_i, \tag{3.18}
\]

where the weighting matrix \( W_{i,RE} \) depends on the application at hand. To illustrate, one could use the identity matrix so that \( b_{W} \) reduces to the ordinary least squares estimator. Another choice is \( Q \), from Section 2.5.3 that yields fixed effects estimators of \( \beta \). We explore this choice further in Section 7.2. The weighted least squares estimator is an unbiased estimator of \( \beta \) and is asymptotically normal, although not efficient unless \( W_{i,RE} = V_{i}^{-1} \). Basic calculations show that it has variance

\[
\text{Var } b_{W} = \left( \sum_{i=1}^{n} X_i' W_{i,RE} X_i \right)^{-1} \sum_{i=1}^{n} X_i' W_{i,RE} V_{i} W_{i,RE} X_i \left( \sum_{i=1}^{n} X_i' W_{i,RE} X_i \right)^{-1}.
\]

As in Section 2.5.3, we may consider estimators that are robust to unsuspected serial correlation and heteroscedasticity. Specifically, following a suggestion made independently by Huber (1967G), White (1980E) and Liang and Zeger (1986B), we can replace \( V_{i} \) by \( e_{i} e'_{i} \), where \( e_{i} = y_{i} - X_{i} \), \( b_{W} \) is the vector of residuals. Thus, a robust standard error of \( b_{W,j} \), the \( j \)th element of \( b_{W} \), is

\[
se(b_{W,j}) = \sqrt{j^{th \text{ diagonal element of } } \left( \sum_{i=1}^{n} X_i' W_{i,RE} X_i \right)^{-1} \sum_{i=1}^{n} X_i' W_{i,RE} e_{i} e'_{i} W_{i,RE} X_i \left( \sum_{i=1}^{n} X_i' W_{i,RE} X_i \right)^{-1}}.
\]

**Testing hypotheses**

For many statistical analyses, testing the null hypothesis that a regression coefficient equals a specified value may be the main goal. That is, the interest may be in testing \( H_0: \beta_j = \beta_{j,0} \),
where the specified value $\beta_{j,0}$ is often (although not always) equal to 0. The customary procedure is to compute the relevant t-statistic

$$t_{\text{-statistic}} = \frac{b_{j,\text{GLS}} - \beta_{j,0}}{\text{se}(b_{j,\text{GLS}})}.$$ 

Here, $b_{j,\text{GLS}}$ is the $j$th component from equation (3.17) and $\text{se}(b_{j,\text{GLS}})$ is the square root of the $j$th diagonal element of $\left(\sum_{i=1}^{n} X_i'V_i(\hat{\tau})^{-1}X_i\right)^{-1}$ where $\hat{\tau}$ is the estimator of the variance component that will be described in Section 3.5. Then, one assesses $H_0$ by comparing the $t$-statistic to a standard normal distribution.

There are two widely used variants of this standard procedure. First, one can replace $\text{se}(b_{j,\text{GLS}})$ by $\text{se}(b_{j,W})$ to get so-called “robust $t$-statistics.” Second, one can replace the standard normal distribution with a $t$-distribution with the “appropriate” number of degrees of freedom. There are several methods for calculating the degrees of freedom that depend on the data and the purpose of the analysis. To illustrate, in Display 3.2 you will see that the approximate degrees of freedom under the “DF” column is different for each variable. This is produced by the SAS default “containment method.” For the applications in this text, we typically will have large number of observations and will be more concerned with potential heteroscedasticity and serial correlation; thus, we will use robust $t$-statistics. For readers with smaller data sets interested in the second alternative, Littell et al. (1996S) describes the $t$-distribution approximation in detail.

For testing hypotheses concerning several regression coefficients simultaneously, the customary procedure is the likelihood ratio test. One may express the null hypothesis as $H_0$: $C\beta = d$, where $C$ is a $p \times K$ matrix with rank $p$, $d$ is a $p \times 1$ vector (typically $0$) and recall that $\beta$ is the $K \times 1$ vector of regression coefficients. Both $C$ and $d$ are user specified and depend on the application at hand. This null hypothesis is tested against the alternative $H_0$: $C\beta \neq d$.

<table>
<thead>
<tr>
<th>Likelihood ratio test procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Using the unconstrained model, calculate maximum likelihood estimates and the corresponding likelihood, denoted as $L_{\text{MLE}}$.</td>
</tr>
<tr>
<td>2. For the model constrained using $H_0$: $C\beta = d$, calculate maximum likelihood estimates and the corresponding likelihood, denoted as $L_{\text{Reduced}}$.</td>
</tr>
<tr>
<td>3. Compute the likelihood ratio test statistic, $L_{\text{RT}} = 2(L_{\text{MLE}} - L_{\text{Reduced}})$.</td>
</tr>
<tr>
<td>4. Reject $H_0$ if $L_{\text{RT}}$ exceeds a percentile from a $\chi^2$ (chi-square) distribution with $p$ degrees of freedom. The percentile is one minus the significance level of the test.</td>
</tr>
</tbody>
</table>

Of course, one may also use $p$-values to calibrate the significance of the test. See Appendix C.7 for more details on the likelihood ratio test.

The likelihood ratio test is the industry standard for assessing hypotheses concerning several regression coefficients. However, we note that better procedures may exist, particularly for small data sets. To illustrate, Pinheiro and Bates (2000S) recommend the use of “conditional $F$-tests” when $p$ is large relative to the sample size. As with testing individual regression coefficients, we shall be more concerned with potential heteroscedasticity for large data sets. In this case, a modification of the Wald test procedure is available.

For the case of no heteroscedasticity and/or serial correlation, the Wald procedure for testing $H_0$: $C\beta = d$ is to compute the test statistic
\[
(C_{MLE} - d)'
\left( C \left[ \sum_{i=1}^{n} X'_i (\tau_{MLE})^{-1} X_i \right] C' \right)^{-1}
(C_{MLE} - d)
\]

and compare this statistic to a chi-square distribution with \( p \) degrees of freedom. Compared to the likelihood ratio test, the advantage of the Wald procedure is that the statistic can be computed with just one evaluation of the likelihood, not two. However, the disadvantage is that for general constraints such as \( C \beta = d \), specialized software is required.

An advantage of the Wald procedure is that it is straightforward to compute robust alternatives. For a robust alternative, we use the regression coefficient estimator defined in equation (3.18) and compute

\[
(C_w - d)'
\left( C \left[ \sum_{i=1}^{n} X'_i W_{i,RE} X_i \right] ^{-1} \sum_{i=1}^{n} X'_i W_{i,RE} e_i' W_{i,RE} X_i \left( \sum_{i=1}^{n} X'_i W_{i,RE} X_i \right) ^{-1} C' \right)^{-1}
(C_w - d).
\]

We compare this statistic to a chi-square distribution with \( p \) degrees of freedom.

### 3.5 Variance components estimation

In this section, we describe several methods for estimating the variance components. The two primary methods entail maximizing a likelihood function, in contrast to moment estimators. In statistical estimation theory (Lehmann, 1991G), there are well-known trade-offs when considering moment compared to likelihood estimation. Typically, likelihood functions are maximized by using iterative procedures that require starting values. At the end of this section, we describe how to obtain reasonable starting values for the iteration using moment estimators.

#### 3.5.1 Maximum likelihood estimation

The log-likelihood was presented in Section 3.4. Substituting the expression for the generalized least squares estimator in equation (3.11) into the log-likelihood in equation (3.16) yields the concentrated or profile log-likelihood

\[
L(b_{GLS}, \tau) = -\frac{1}{2} \sum_{i=1}^{n} \left( T_i \ln(2\pi) + \ln \det V_i(\tau) + (Error SS)_i(\tau) \right),
\]

a function of \( \tau \). Here, the error sum of squares for the \( i \)th subject is

\[
(Error SS)_i(\tau) = (y_i - X_i b_{GLS})' V_i^{-1} (\tau) (y_i - X_i b_{GLS}).
\]

Thus, we now maximize the log-likelihood as a function of \( \tau \) only. In only a few special cases can one obtain closed form expressions for the maximizing variance components. Exercise 3.10 illustrates one such special case.

**Special case – Error components model**

For this special case, the variance components are \( \tau = (\sigma^2, \sigma^2_\alpha)' \). Using equation (A.5) in Appendix A.5, we have that \( \ln \det V_i = \ln \det (\sigma^2 J_i + \sigma^2 I_i) = T_i \ln \sigma^2 + \ln (1 + T_i \sigma^2 / \sigma^2) \).

From this and equation (3.9), we have that the concentrated likelihood is
Iterative estimation

In general, the variance components are estimated recursively. This can be done using either the Newton-Raphson or Fisher scoring method, see for example, Harville (1977S) and Wolfinger et al. (1994S).

Newton-Raphson. Let \( L = L(b_{GLS}(\tau), \tau) \), and use the iterative method:

\[
\tau_{NEW} = \tau_{OLD} - \left( \frac{\partial^2 L}{\partial \tau \partial \tau'} \right)^{-1} \frac{\partial L}{\partial \tau} \Bigg|_{\tau = \tau_{OLD}}.
\]

Here, the matrix \(-\frac{\partial^2 L}{\partial \tau \partial \tau'}\) is called the sample information matrix.

Fisher scoring. Define the expected information matrix \( I(\tau) = -E \left( \frac{\partial^2 L}{\partial \tau \partial \tau'} \right) \) and use

\[
\tau_{NEW} = \tau_{OLD} + I(\tau_{OLD})^{-1} \left\{ \frac{\partial L}{\partial \tau} \right\} \Bigg|_{\tau = \tau_{OLD}}.
\]

3.5.2 Restricted maximum likelihood

As the name suggests, restricted maximum likelihood (REML) is a likelihood-based estimation procedure. Thus, it shares many of the desirable properties of maximum likelihood estimators (MLEs). Because it is based on likelihoods, it is not specific to a particular design matrix, as are analysis of variance estimators (Harville, 1977S). Thus, it can be readily applied to a wide variety of models. Like MLEs, REML estimators are translation invariant.

Maximum likelihood often produces biased estimators of the variance components, \( \tau \). In contrast, estimation based on REML results in unbiased estimators of \( \tau \), at least for many balanced designs. Because maximum likelihood estimators are negatively biased, they often turn out to be negative, an intuitively undesirable situation for many users. Because of the unbiasedness of many REML estimators, there is less of a tendency to produce negative estimators (Corbeil and Searle, 1976a,S). As with MLEs, REML estimators can be defined to be parameter values for which the (restricted) likelihood achieves a maximum value over a constrained parameter space. Thus, as with maximum likelihood, it is straightforward to modify the method to produce nonnegative variance estimators.

The idea behind REML estimation is to consider the likelihood of linear combinations of the responses that do not depend on the mean parameters. To illustrate, consider the mixed linear model. We assume that the responses, denoted by the vector \( y \), are normally distributed, have mean \( E\, y = X\beta \) and variance-covariance matrix \( Var\, y = V = V(\tau) \). The dimension of \( y \) is \( N \times 1 \), and the dimension of \( X \) is \( N \times p \). With this notation, define the projection matrix \( Q = I - X (X'X)^{-1} X' \) and consider the linear combination of responses \( Qy \). Straightforward calculations show that
$Qy$ has mean $0$ and variance-covariance matrix $\text{Var}(Qy) = QVQ$. Because (i) $Qy$ has a multivariate normal distribution and (ii) the mean and variance-covariance matrix do not depend on $\beta$, the distribution of $Qy$ does not depend on $\beta$. Further, Appendix 3A.1 shows that $Qy$ is independent of the generalized least squares estimator $b_{GLS} = (X'V^{-1}X)^{-1}X'V^{-1}y$.

The vector $Qy$ is the residual vector from an ordinary least squares fit of the data. Hence, REML is also referred to as residual maximum likelihood estimation. Because the rank of $Q$ is $N - p$, we lose some information by considering this transformation of the data; this motivates the use of the descriptor restricted maximum likelihood. (There is some information about $\tau$ in the vector $b_{GLS}$ that we are not using for estimation.) Further, note that we could also use any linear transform of $Q$, such as $AQ$, in that $AQY$ also has a multivariate normal distribution with a mean and variance-covariance matrix that do not depend on $\beta$. Patterson and Thompson (1971S) and Harville (1974S, 1977S) showed that the likelihood does not depend on the choice of $A$. They introduced the “restricted” log-likelihood:

$$L_{REML}(b_{GLS}(\tau), \tau) = -\frac{1}{2} \left[ \ln \text{det}(V(\tau)) + \ln \text{det}(X'V(\tau)^{-1}X) + (\text{Error SS})(\tau) \right], \quad (3.21)$$

up to an additive constant. See Appendix 3A.2 for a derivation of this likelihood. REML estimators $\tau_{REML}$ are defined to be maximizers of the function $L_{REML}(b_{GLS}(\tau), \tau)$. Here, the error sum of squares is

$$(\text{Error SS})(\tau) = (y - Xb_{GLS}(\tau))'V(\tau)^{-1}(y - Xb_{GLS}(\tau)). \quad (3.22)$$

Analogous to equation (3.19), the usual log-likelihood is

$$L(b_{GLS}(\tau), \tau) = -\frac{1}{2} \left[ \ln \text{det}(V(\tau)) + (\text{Error SS})(\tau) \right],$$

up to an additive constant. The only difference between the two likelihoods is the term $\ln \text{det}(X'V(\tau)^{-1}X)$. Thus, iterative methods of maximization are the same—that is, using either Newton-Raphson or Fisher scoring. For linear mixed effects models, this additional term is $\ln \text{det}\left(\sum_{i=1}^{n} X_i'V_i(\tau)^{-1}X_i\right)$.

For balanced analysis of variance data ($T_i = T$), Corbeil and Searle (1976a,S) established that the REML estimation reduces to standard analysis of variance estimators. Thus, REML estimators are unbiased for these designs. However, REML estimators and analysis of variance estimators differ for unbalanced data. REML estimators achieve their unbiasedness by accounting for the degrees of freedom lost in estimating the fixed effects $\beta$; MLEs do not account for this loss of degrees of freedom. When $p$ is large, the difference between REML estimators and MLEs is significant. Corbeil and Searle (1976b,S) showed that, in terms of mean square errors, MLEs outperform REML estimators for small $p$ ($< 5$), although the situation is reversed for large $p$ with a sufficiently large sample.

Harville (1974S) gave a Bayesian interpretation of REML estimators. He pointed out that using only $Qy$ to make inferences about $\tau$ is equivalent to ignoring prior information about $\beta$ and using all the data.

Some statistical packages present maximized values of restricted likelihoods, suggesting to users that these values can be used for inferential techniques, such as likelihood ratio tests. For likelihood ratio tests, one should use “ordinary” likelihoods, even when evaluated at REML estimators, not the “restricted” likelihoods that are used to determine REML estimators. Appendix 3A.3 illustrates the potentially disastrous consequences of using REML likelihoods for likelihood ratio tests.
Starting values
Both the Newton-Raphson and Fisher scoring algorithms and the ML and REML estimation methods involve recursive calculations that require starting values. We now describe two non-recursive methods due to Swamy (1970E) and Rao (1970S), respectively. One can use the results of these non-recursive methods as starting values in the Newton-Raphson and Fisher scoring algorithms.

Swamy’s moment-based procedure appeared in the econometrics panel data literature. We consider a random coefficients model; that is, equation (3.8) with $x_{it} = z_{it}$ and $R_i = \sigma_i^2 I_i$.

<table>
<thead>
<tr>
<th>Procedure for computing moment-based variance component estimators</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Compute an ordinary least squares estimator of $\sigma_i^2$,</td>
</tr>
<tr>
<td>$s_i^2 = \frac{1}{T_i - K} y_i' \left( I_i - X_i (X_i' X_i)^{-1} X_i' \right) y_i$.</td>
</tr>
<tr>
<td>This is an ordinary least squares procedure in that it ignores $D$.</td>
</tr>
<tr>
<td>2. Next, calculate $b_{i,OLS} = (X_i' X_i)^{-1} X_i' y_i$, a predictor of $\beta + \alpha_i$.</td>
</tr>
<tr>
<td>3. Finally, estimate $D$ using</td>
</tr>
<tr>
<td>$D_{SWAMY} = \frac{1}{n-1} \sum_{i=1}^n (b_{i,OLS} - \bar{b})(b_{i,OLS} - \bar{b})' - \frac{1}{n} \sum_{i=1}^n s_i^2 (X_i' X_i)^{-1}$,</td>
</tr>
<tr>
<td>where $\bar{b} = \frac{1}{n} \sum_{i=1}^n b_{i,OLS}$.</td>
</tr>
</tbody>
</table>

The estimator of $D$ can be motivated by examining the variance of $b_{i,OLS}$, |
$\text{Var}(b_{i,OLS}) = \text{Var}\left((X_i' X_i)^{-1} X_i' (\beta + \alpha_i + \varepsilon_i)\right)$ |
$= \text{Var}(\beta + \alpha_i + (X_i' X_i)^{-1} X_i' \varepsilon_i) = D + \sigma_i^2 (X_i' X_i)^{-1}$ |
Using $\frac{1}{n-1} \sum_{i=1}^n (b_{i,OLS} - \bar{b})(b_{i,OLS} - \bar{b})'$ and $s_i^2$ as estimators of $\text{Var}(b_{i,OLS})$ and $\sigma_i^2$ respectively, yields $D_{SWAMY}$ as an estimator of $D$.

Various modifications of this estimator are possible. One can iterate the procedure by using $D_{SWAMY}$ to improve the estimators $s_i^2$, and so on. Homoscedasticity of the $\varepsilon_i$ terms could also be assumed. Hsiao (1986E) recommends dropping the second term, $n^{-1} \sum_{i=1}^n s_i^2 (X_i' X_i)^{-1}$, to ensure that $D_{SWAMY}$ is non-negative definite.

3.5.3 MIVQUE estimators
Another non-recursive method is Rao’s (1970S) minimum variance quadratic unbiased estimator (MIVQUE). To describe this method, we return to the mixed linear model $y = X \beta + \varepsilon$, in which $\text{Var} y = V = V(\tau)$. We wish to estimate the linear combination of variance components, $\sum_{k=1}^r c_k \tau_k$, where the $c_k$ are specified constants and $\tau = (\tau_1, \ldots, \tau_r)'$. We assume that $V$ is linear in the sense that |
$V = \sum_{k=1}^r \tau_k \frac{\partial^2}{\partial \tau_k} V$. |
Thus, with this assumption, we have that the matrix of second derivatives (the Hessian) of $V$ is zero (Graybill, 1969G). Although this assumption is generally viable, it is not satisfied by, for
example, autoregressive models. It is not restrictive to assume that \( \frac{\partial}{\partial \tau_k} V \) is known even though the variance component \( \tau_k \) is unknown. To illustrate, consider an error components structure so that \( V = \sigma_a^2 J + \sigma^2 I \). Then, \( \frac{\partial}{\partial \sigma_a^2} V = J \) and \( \frac{\partial}{\partial \sigma^2} V = I \) are both known.

Quadratic estimators of \( \tau_k \) are based on \( y' Ay \), where \( A \) is a symmetric matrix to be specified. The variance of \( y' Ay \), assuming normality, can easily be shown to be \( 2 \text{trace}(VAVA) \). We would like the estimator to be invariant to translation of \( \beta \). That is, we require

\[
y' A y = (y - X b_0)' A (y - X b_0) \quad \text{for each} \quad b_0.
\]

Thus, we restrict our choice of \( A \) to those that satisfy \( A X = 0 \).

For unbiasedness, we would like \( \sum_{k=1}^r c_k \tau_k = \text{E}(y' A y) \). Using \( A X = 0 \), we have

\[
\text{E}(y' A y) = \text{E}(\varepsilon' A \varepsilon) = \text{trace}(E(\varepsilon \varepsilon')A) = \text{trace}(VA)
\]

Because this equality should be valid for all variance components \( \tau_k \), we require that \( A \) satisfy

\[
c_k = \text{trace}
\left( \frac{\partial}{\partial \tau_k} V \right) A, \quad \text{for} \quad k = 1, \ldots, r. \tag{3.23}
\]

Rao showed that the minimum value of \( \text{trace}(VAVA) \) satisfying \( A X = 0 \) and the constraints in equation (3.23) is attained at

\[
A^*(V) = \sum_{k=1}^r \lambda_k V^{-1} Q \left( \frac{\partial}{\partial \tau_k} V \right) V^{-1} Q,
\]

where \( Q = Q(V) = I - X (X' V^{-1} X)^{-1} X' V^{-1} \) and \( (\lambda_1, \ldots, \lambda_r) \) is the solution of

\[
S (\lambda_1, \ldots, \lambda_r)' = (c_1, \ldots, c_r)'.
\]

Here, the \((i,j)\)th element of \( S \) is given by

\[
\text{trace}
\left( V^{-1} Q \left( \frac{\partial}{\partial \tau_i} V \right) V^{-1} Q \left( \frac{\partial}{\partial \tau_j} V \right) \right).
\]

Thus, the MIVQUE estimator of \( \tau \) is the solution of

\[
S \tau_{\text{MIVQUE}} = G, \tag{3.24}
\]

where the \(k\)th element of \( G \) is given by \( y' V^{-1} Q \left( \frac{\partial}{\partial \tau_k} V \right) V^{-1} Qy \).

When comparing Rao’s to Swamy’s method, we note that the MIVQUE estimators are available for a larger class of models. To illustrate, in the longitudinal data context, it is possible to handle serial correlation with the MIVQUE estimators. A drawback of the MIVQUE estimator is that normality is assumed; this can be weakened to zero kurtosis for certain forms of \( V \) (Swallow and Searle, 1978). Further, MIVQUE estimators require a pre-specified estimate of \( V \). A widely used specification is to use the identity matrix for \( V \) in equation (3.24). This specification produces so-called MIVQUE(0) estimators, an option in widely available statistical packages. It is the default option in PROC MIXED of the statistical package SAS.
Further reading

When compared to regression and linear models, there are fewer textbook introductions to mixed linear models, although more are becoming available. Searle, Casella and McCulloch (1992S) give an early technical treatment. A slightly less technical is Longford (1993EP). McCulloch and Searle (2001S) give an excellent recent technical treatment. Other recent contributions integrate statistical software into their exposition. Little et al. (1996D) and Verbeke and Molenberghs (2000D) introduce mixed linear models using the SAS statistical package. Pinheiro and Bates (2000D) provide an introduction using the S and S-Plus statistical packages.

Random effects in ANOVA and regression models have been part of the standard statistical literature for quite some time; see, for example, Scheffé (1959G), Searle (1971G) or Neter and Wasserman (1974G). Balestra and Nerlove (1966E) introduced the error components model to the econometric literature. The random coefficients model was described early on by Hildreth and Houck (1968S).

As described in Section 3.5, most of the development of variance component estimators occurred in the 1970’s. More recently, Baltagi and Chang (1994E) compared the relative performance of several variance components estimators for the error components model.
Appendix 3A. REML Calculations

Appendix 3A.1 Independence of Residuals and Least Squares Estimators

Assume that $y$ has a multivariate normal distribution with mean $X\beta$ and variance-covariance matrix $V$, where $X$ has dimension $N \times p$ with rank $p$. Recall that the matrix $V$ depends on the parameters $\tau$.

We use the matrix $Q = I - X(X'X)^{-1}X'$. Because $Q$ is idempotent and has rank $N - p$, we can find an $N \times (N - p)$ matrix $A$ such that $AA' = Q$ and $A'A = I_N$.

We also need $G = V^{-1}X(X'V^{-1}X)^{-1}$, an $N \times p$ matrix. Note that $G' y = b_{GLS}$, the generalized least squares estimator of $\beta$.

With these two matrices, define the transformation matrix $H = (A G)$, an $N \times N$ matrix. Consider the transformed variables

$$H' y = \begin{bmatrix} A'y \\ G'y \end{bmatrix} = \begin{bmatrix} A'y \\ b_{GLS} \end{bmatrix}.$$ 

Basic calculations show that $A'y \sim N(0, A'VA0yA')$ and $G'y = b_{GLS} \sim N(\beta, (X'V^{-1}X)^{-1})$,

in which $z \sim N(\mu, V)$ denotes that a random vector $z$ has a multivariate normal distribution with mean $\mu$ and variance $V$. Further, we have that $A'y$ and $b_{GLS}$ are independent. This is due to normality and zero covariance matrix:

$$\text{Cov}(A'y, b_{GLS}) = E(A'y y'G) = A'VG = A'X(X'V^{-1}X)^{-1} = 0.$$ 

We have $A' X = 0$ because $A' X = (A' A)A' X = A' Q X$ and $Q X = 0$. Zero covariance, together with normality, imply independence.

Appendix 3A.2 Restricted Likelihoods

To develop the restricted likelihood, we first check the rank of the transformation matrix $H$. Thus, with $H$ as in Appendix 3A.1 and equation (A.2) of Appendix A.5, we have

$$\det(H'^2) = \det(H'H) = \det \begin{bmatrix} A' & A' \\ G' & G' \end{bmatrix} = \det \begin{bmatrix} A' & A'G' \\ G'A & G'G \end{bmatrix}$$

$$= \det(A'A) \det(G'G - G'A(A'A)^{-1}A'G)$$

$$= \det(G'G - G'QG) = \det(G'X(X'X)^{-1}X'G) = \det((X'X)^{-1}),$$

using $G'X = I$. Thus, the transformation $H'$ is non-singular if and only if $X'X$ is non-singular. In this case, no information is lost by considering the transformation $H'y$.

We now develop the restricted likelihood based on the probability density function of $A'y$. We first note a relationship used by Harville (1974S), concerning the probability density function of $G'y$. We write $f_{G'y}(z, \beta)$ to denote the probability density function of the random vector $G'y$, evaluated at the (vector) point $z$ with mean (vector) parameter $\beta$. Because probability density functions integrate to 1, we have the relation
\[ 1 = \int f_{G'Y}(\mathbf{z}, \beta) d\mathbf{z} = \int \frac{1}{(2\pi)^{p/2} \det(X'X)^{-1/2}} \exp\left(-\frac{1}{2} (\mathbf{z} - \beta)'X'X(z - \beta)\right) d\mathbf{z} \]
\[ = \int \frac{1}{(2\pi)^{p/2} \det(X'X)^{-1/2}} \exp\left(-\frac{1}{2} (\mathbf{z} - \beta)'X'X(z - \beta)\right) d\beta \]
\[ = \int f_{G'Y}(\mathbf{z}, \beta) d\beta, \text{ for each } \mathbf{z}, \]

with a change of variables.

Because of the independence of \(A'\mathbf{y}\) and \(G'\mathbf{y} = \mathbf{b}_{GLS}\), we have \(f_{w'Y} = f_{A'Y}f_{G'Y}\). Here, \(f_{w'Y}, f_{A'Y}\) and \(f_{G'Y}\) are the density functions of the random vectors \(H'\mathbf{y}, A'\mathbf{y}\) and \(G'\mathbf{y}\), respectively. For notation, let \(\mathbf{y}^*\) be a potential realization of the random vector \(\mathbf{y}\). Thus, the probability density function of \(A'\mathbf{y}\) is
\[ f_{A'Y}(A'\mathbf{y}^*) = \int f_{A'Y}(A'\mathbf{y}^*)f_{G'Y}(G'\mathbf{y}^*, \beta) d\beta \]
\[ = \int f_{w'Y}(H'\mathbf{y}^*, \beta) d\beta = \int \det(H)^{-1} f_{Y}(\mathbf{y}^*, \beta) d\beta, \]

using a change of variables. Now, let \(\mathbf{b}_{GLS}^*\) be the realization of \(\mathbf{b}_{GLS}\) using \(\mathbf{y}^*\). Then, from a standard equality from analysis of variance,
\[ (\mathbf{y}^* - X\beta)'V^{-1}(\mathbf{y}^* - X\beta) = (\mathbf{y}^* - X\mathbf{b}_{GLS}^*)'V^{-1}(\mathbf{y}^* - X\mathbf{b}_{GLS}^*) + (\mathbf{b}_{GLS}^* - \beta)'X'V^{-1}X(\mathbf{b}_{GLS}^* - \beta). \]

With this equality, the probability density function \(f_{Y}\) can be expressed as
\[ f_{Y}(\mathbf{y}^*, \beta) = \frac{1}{(2\pi)^{N/2} \det(V)^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{y}^* - X\beta)'V^{-1}(\mathbf{y}^* - X\beta)\right) \]
\[ = \frac{1}{(2\pi)^{N/2} \det(V)^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{y}^* - X\mathbf{b}_{GLS}^*)'V^{-1}(\mathbf{y}^* - X\mathbf{b}_{GLS}^*)\right) \exp\left(-\frac{1}{2} (\mathbf{b}_{GLS}^* - \beta)'X'V^{-1}X(\mathbf{b}_{GLS}^* - \beta)\right) \]
\[ = \frac{(2\pi)^{p/2} \det(X'X)^{-1/2}}{(2\pi)^{N/2} \det(V)^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{y}^* - X\mathbf{b}_{GLS}^*)'V^{-1}(\mathbf{y}^* - X\mathbf{b}_{GLS}^*)\right) f_{G'Y}(\mathbf{b}_{GLS}^*, \beta). \]

Thus,
\[ f_{A'Y}(A'\mathbf{y}^*) = \frac{(2\pi)^{p/2} \det(X'X)^{-1/2}}{(2\pi)^{N/2} \det(V)^{1/2}} \det(H)^{-1} \exp\left(-\frac{1}{2} (\mathbf{y}^* - X\mathbf{b}_{GLS}^*)'V^{-1}(\mathbf{y}^* - X\mathbf{b}_{GLS}^*)\right) \int f_{G'Y}(\mathbf{b}_{GLS}^*, \beta) d\beta. \]
\[ = (2\pi)^{-(N-p)/2} \det(V)^{-1/2} \det(X'X)^{1/2} \det(X'V^{-1}X)^{-1/2} \exp\left(-\frac{1}{2} (\mathbf{y}^* - X\mathbf{b}_{GLS}^*)'V^{-1}(\mathbf{y}^* - X\mathbf{b}_{GLS}^*)\right) \]

This yields the REML likelihood in Section 3.5, after taking logarithms and dropping constants that do not involve \(\tau\).
Appendix 3A.3 Likelihood Ratio Tests and REML

Recall the likelihood ratio statistic, \( LRT = 2 \left( L(\theta_{\text{MLE}}) - L(\theta_{\text{Reduced}}) \right) \). This is evaluated using the so-called “concentrated” or “profile” log-likelihood given in equations (3.19) and (3.20). For comparison, from equation (3.21), the “restricted” log-likelihood is

\[
L_{\text{REML}}(b_{\text{GLS}}, \tau) = -\frac{1}{2} \sum_{i=1}^{n} \left( T_i \ln(2\pi) + \ln \det V_i(\tau) + (\text{Error SS})(\tau) \right) - \frac{1}{2} \ln \det \left( \sum_{i=1}^{n} X_i' V_i(\tau)^{-1} X_i \right).
\]

(3A.1)

To see why a REML likelihood does not work for likelihood ratio tests, consider the following example.

**Special Case. Testing the Importance of a Subset of Regression Coefficients.**

For simplicity, we assume that \( V_i = \sigma^2 I_i \), so that there is no serial correlation. For this special case, we have the finite, and asymptotic, distribution of the partial \( F \)-test (Chow). Because the asymptotic distribution is well known, we can easily judge whether or not REML likelihoods are appropriate.

Write \( \beta = (\beta_1', \beta_2')' \) and suppose that we wish to use the null hypothesis \( H_0: \beta_2 = 0 \).

Assuming no serial correlation, the generalized least square estimator of \( \beta \) reduces to the ordinary least squares estimator, that is, \( b_{\text{GLS}} = b_{\text{OLS}} = \left( \sum_{i=1}^{n} X_i'X_i \right)^{-1} \sum_{i=1}^{n} X_i'y_i \). Thus, from equation (3.19), the concentrated likelihood is:

\[
L(b_{\text{OLS}}, \sigma^2) = -\frac{1}{2} \sum_{i=1}^{n} \left( T_i \ln(2\pi) + T_i \ln \sigma^2 + \frac{1}{\sigma^2} (y_i - X_i b_{\text{OLS}})' (y_i - X_i b_{\text{OLS}}) \right)
\]

\[
= \frac{1}{2} \left( N \ln(2\pi) + \frac{1}{\sigma^2} \text{(Error SS)}_{\text{Full}} \right),
\]

where \( (\text{Error SS})_{\text{Full}} = \sum_{i=1}^{n} (y_i - X_i b_{\text{OLS}})' (y_i - X_i b_{\text{OLS}}) \). The maximum likelihood estimator of \( \sigma^2 \) is \( \sigma^2_{\text{MLE}} = (\text{Error SS})_{\text{Full}} / N \) so the maximum likelihood is

\[
L(b_{\text{OLS}}, \sigma^2_{\text{MLE}}) = -\frac{1}{2} \left( N \ln(2\pi) + N \ln(\text{Error SS})_{\text{Full}} - N \ln N + N \right).
\]

Now, write \( X_i = (X_{i1}, X_{i2}) \) where \( X_{i1} \) has dimension \( T_i \times (K-r) \) and \( X_{i2} \) has dimension \( T_i \times r \). Under \( H_0 \), the estimator of \( \beta_1 \) is \( b_{\text{OLS,Reduced}} = \left( \sum_{i=1}^{n} X_{i1}'X_{i1} \right)^{-1} \sum_{i=1}^{n} X_{i1}'y_i \). Thus, under \( H_0 \), the log-likelihood is:

\[
L(b_{\text{OLS,Reduced}}, \sigma^2_{\text{MLE,Reduced}}) = -\frac{1}{2} \left( N \ln(2\pi) + \frac{1}{\sigma^2} \text{(Error SS)}_{\text{Reduced}} - N \ln N + N \right),
\]

where \( (\text{Error SS})_{\text{Reduced}} = \sum_{i=1}^{n} (y_i - X_{i1} b_{\text{OLS,Reduced}})' (y_i - X_{i1} b_{\text{OLS,Reduced}}) \). Thus, the likelihood ratio test statistic is:

\[
LRT_{\text{MLE}} = 2 \left( L(b_{\text{OLS}}, \sigma^2_{\text{MLE}}) - L(b_{\text{OLS,Reduced}}, \sigma^2_{\text{MLE,Reduced}}) \right)
\]

\[
= N \ln \left( \frac{(\text{Error SS})_{\text{Reduced}} - (\text{Error SS})_{\text{Full}}}{(\text{Error SS})_{\text{Full}}} \right).
\]

From a Taylor series approximation, we have \( \ln y = \ln x + \frac{(y-x)}{x} - \frac{1}{2} \frac{(y-x)^2}{x^2} + \ldots \). Thus, we have

\[
LRT_{\text{MLE}} = N \left( \frac{(\text{Error SS})_{\text{Reduced}} - (\text{Error SS})_{\text{Full}}}{(\text{Error SS})_{\text{Full}}} \right) + \ldots
\]
which has an approximate chi-square distribution with \( r \) degrees of freedom. For comparison, from equation (3A.1), the “restricted” log-likelihood is

\[
L_{\text{REML}}(\mathbf{b}_{\text{OLS}}, \sigma^2) = -\frac{1}{2} \left( N \ln(2\pi) + (N - K) \ln \sigma^2 + \frac{1}{\sigma^2} (\text{Error SS}_{\text{Full}}) \right) - \frac{1}{2} \ln \det \left( \sum_{i=1}^{n} \mathbf{X}_i' \mathbf{X}_i \right).
\]

The restricted maximum likelihood estimator of \( \sigma^2 \) is

\[
\sigma^2_{\text{REML}} = (\text{Error SS}_{\text{Full}}) / (N - K).
\]

Thus, the restricted maximum likelihood is

\[
L_{\text{REML}}(\mathbf{b}_{\text{OLS}}, \sigma^2_{\text{REML}}) = -\frac{1}{2} \left( N \ln(2\pi) + (N - K) \ln(\text{Error SS}_{\text{Full}}) \right) - \frac{1}{2} \ln \det \left( \sum_{i=1}^{n} \mathbf{X}_i' \mathbf{X}_i \right)
+ \frac{1}{2} \left( (N - K) \ln(N - K) - (N - K) \right).
\]

Under \( H_0 \), the restricted log-likelihood is:

\[
L_{\text{REML}}(\mathbf{b}_{\text{OLS, Reduced}}, \sigma^2_{\text{REML, Reduced}})
= -\frac{1}{2} \left( N \ln(2\pi) + (N - (K - q)) \ln(\text{Error SS}_{\text{Reduced}}) \right) - \frac{1}{2} \ln \det \left( \sum_{i=1}^{n} \mathbf{X}_{i_{1}}' \mathbf{X}_{i_{1}} \right)
+ \frac{1}{2} \left( (N - (K - q)) \ln(N - (K - q)) - (N - (K - q)) \right).
\]

Thus, the likelihood ratio test statistic using a restricted likelihood is:

\[
\text{LRT}_{\text{REML}} = 2 \left( L_{\text{REML}}(\mathbf{b}_{\text{OLS, Reduced}}, \sigma^2_{\text{REML, Reduced}}) - L_{\text{REML}}(\mathbf{b}_{\text{OLS, Full}}, \sigma^2_{\text{REML, Full}}) \right)
= (N - K) \left( \ln(\text{Error SS}_{\text{Reduced}}) - \ln(\text{Error SS}_{\text{Full}}) \right) + q \ln(\text{Error SS}_{\text{Reduced}})
+ \sum_{i=1}^{n} \left( \ln \det(\mathbf{X}_{i_{1}}' \mathbf{X}_{i_{1}}) - \ln \det(\mathbf{X}_{i}' \mathbf{X}_{i}) \right) + (N - K) \ln \left( \frac{N - K}{N - (K - q)} \right) - q \ln \left( \frac{N - (K - q)}{N - (K - q)} \right)
= \frac{(N - K)}{N} \text{LRT}_{\text{MLE}}
+ \ln \left( \frac{\text{det}(\mathbf{X}_{i_{1}}' \mathbf{X}_{i_{1}})}{\text{det}(\mathbf{X}_{i}' \mathbf{X}_{i})} \right) - \ln \left( \frac{\text{det}(\mathbf{X}_{i_{1}}' \mathbf{X}_{i_{1}})}{\text{det}(\mathbf{X}_{i}' \mathbf{X}_{i})} \right)
+ q \ln \left( \frac{1}{\frac{N - (K - q)}{N - (K - q)}} \right) + (N - K) \ln \left( 1 - \frac{q}{N - (K - q)} \right).
\]

The first term is asymptotically equivalent to the likelihood ratio test, using “ordinary” maximized likelihoods. The third and fourth terms tend to constants. The second term, \( \ln \left( \frac{\text{det}(\mathbf{X}_{i_{1}}' \mathbf{X}_{i_{1}})}{\text{det}(\mathbf{X}_{i}' \mathbf{X}_{i})} \right) \), may tend to plus or minus infinity, depending on the values of the explanatory variables. For example, in the special case that \( \mathbf{X}_{i_{1}}' \mathbf{X}_{i_{2}} = 0 \), we have

\[
\ln \left( \frac{\text{det}(\mathbf{X}_{i_{1}}' \mathbf{X}_{i_{1}})}{\text{det}(\mathbf{X}_{i}' \mathbf{X}_{i})} \right) = (−1)^{q} \ln \left( \frac{\text{det}(\mathbf{X}_{2}' \mathbf{X}_{2})}{\text{det}(\mathbf{X}_{i}' \mathbf{X}_{i})} \right).
\]

Thus, this term will tend to plus or minus infinity for most explanatory variable designs.
3. Exercises and Extensions

Section 3.1

3.1. Generalized least squares (GLS) estimators

For the error components model, the variance of the vector of responses is given as

\[ V_i = \sigma^2_{\alpha} J_i + \sigma^2_{\epsilon} I_i. \]

a. By multiplying \( V_i \) by \( V_i^{-1} \), check that

\[ \begin{pmatrix} \sum_{i=1}^{n} X'_i V_i^{-1} X_i \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{n} X'_i V_i^{-1} y_i \end{pmatrix}, \]

b. Use this form of \( V_i^{-1} \) and the expression for a GLS estimator,

\[ b_{EC} = \left( \begin{pmatrix} \sum_{i=1}^{n} x'_i (1 - T_i^{-1} J_i) x_i \end{pmatrix} \right)^{-1} \begin{pmatrix} \sum_{i=1}^{n} x'_i (1 - T_i^{-1} J_i) y_i \end{pmatrix} \]

to establish the formula in equation (3.3).

c. Use equation (3.4) to show that the basic random effects estimator can be expressed as:

\[ b_{EC} = \frac{1}{\sum_{i=1}^{n} \sigma^2_{\alpha}} \begin{pmatrix} \sum_{i=1}^{n} x'_i (1 - T_i^{-1} J_i) x_i \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{n} x'_i (1 - T_i^{-1} J_i) y_i \end{pmatrix}. \]

d. Show that

\[ b = \frac{1}{\sum_{i=1}^{n} \sigma^2_{\alpha}} \begin{pmatrix} \sum_{i=1}^{n} x'_i (1 - T_i^{-1} J_i) x_i \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{n} x'_i (1 - T_i^{-1} J_i) y_i \end{pmatrix} \]

is an alternative expression for the basic fixed effects estimator given in equation (2.6).

e. Suppose that \( \sigma^2_{\alpha} \) is large relative to \( \sigma^2 \) so that we assume that \( \sigma^2_{\alpha} / \sigma^2 \rightarrow \infty \). Give an expression and interpretation for \( b_{EC} \).

f. Suppose that \( \sigma^2_{\alpha} \) is small relative to \( \sigma^2 \) so that we assume that \( \sigma^2_{\alpha} / \sigma^2 \rightarrow 0 \). Give an expression and interpretation for \( b_{EC} \).

3.2. GLS estimator as a weighted average

Consider the basic random effects model and suppose that \( K = 1 \) and that \( x_{it} = 1 \). Show that

\[ b_{EC} = \frac{\sum_{i=1}^{n} \zeta_i \bar{y}_i}{\sum_{i=1}^{n} \zeta_i}. \]

3.3. Error components with one explanatory variable

Consider the error components model, \( y_{it} = \alpha_t + \beta_0 + \beta_1 x_{it} + \epsilon_{it} \). That is, consider the model in equation (3.2) with \( K = 2 \) and \( x_{it} = (1, x_{it})' \).

a. Show that \( \zeta_i = \frac{\sigma^2_{\alpha}}{\sigma^2_{\alpha}} T_i (1 - \zeta_i) \).

b. Show that we may write the generalized least squares estimators of \( \beta_0 \) and \( \beta_1 \) as

\[ b_{1,EC} = \frac{\sum_{i,t} x_{it} y_{it} - \sum_{i,T_i} \zeta_i \bar{y}_i - \sum_{i,T_i} \zeta_i (1 - \zeta_i) T_i \bar{x}_{ti} \bar{y}_w}{\sum_{i,t} x_{it}^2 - \sum_{i,T_i} \zeta_i \bar{x}_i^2 - \sum_{i,T_i} \zeta_i (1 - \zeta_i) T_i \bar{x}_i^2} \]

and

\[ b_{0,EC} = \bar{y}_w - \bar{x}_w b_{1,EC} \]
where
\[ \bar{x}_w = \frac{\sum_i \zeta_i \bar{x}_i}{\sum_i \zeta_i} \quad \text{and} \quad \bar{y}_w = \frac{\sum_i \zeta_i \bar{y}_i}{\sum_i \zeta_i}. \]
(Hint: use the expression of $b_{1,EC}$ in Exercise 3.1c.)

c. Suppose that $\sigma_\alpha^2$ is large relative to $\sigma^2$ so that we assume that \( \frac{\sigma_\alpha^2}{\sigma^2} \to \infty \). Give an expression and interpretation for $b_{1,EC}$.

d. Suppose that $\sigma_\alpha^2$ is small relative to $\sigma^2$ so that we assume that \( \frac{\sigma_\alpha^2}{\sigma^2} \to 0 \). Give an expression and interpretation for $b_{1,EC}$.

### 3.4. Two-population slope interpretation
Consider the basic random effects model and suppose that $K = 1$ and that $x$ is binary variable. Suppose further that $x$ takes on the value of 1 for those from population 1 and –1 for those from population 2. Analogous to Exercise 2.4, let $n_{1,i}$ and $n_{2,i}$ be the number of ones and minus ones for the $i$th subject, respectively. Further, let $\bar{y}_{1,i}$ and $\bar{y}_{2,i}$ be the average response when $x$ is one and minus one, for the $i$th subject, respectively. Show that we may write the error components estimator as

\[ b_{1,EC} = \frac{\sum_{i=1}^n (w_{1,i} \bar{y}_{1,i} - w_{2,i} \bar{y}_{2,i})}{\sum_{i=1}^n (w_{1,i} + w_{2,i})}, \]

with weights $w_{1,i} = n_{1,i} \left( 1 + \zeta_i - 2 \zeta_i n_{1,i} / T_i \right)$ and $w_{2,i} = n_{2,i} \left( 1 + \zeta_i - 2 \zeta_i n_{2,i} / T_i \right)$.
(Hint: use the expression of $b_{EC}$ in Exercise 3.1c.)

### 3.5. Unbiased variance estimators
Perform the following steps to check that the variance estimators given by Baltagi and Chang (1994E) are unbiased variance estimators for the unbalanced error components model introduced in Section 3.1. For notational simplicity, assume the model follows the form $y_{it} = \mu_a + \alpha_i + x_{it}' \beta + \epsilon_{it}$, where $\mu_a$ is a fixed parameter, represent the model intercept. As described in Section 3.1, we will use the residuals $e_{it} = y_{it} - (a_i + x_{it}' b)$ using $a_i$ and $b$ according to the Chapter 2 fixed effects estimators in equations (2.6) and (2.7).

a. Show that response deviations can be expressed as $y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)' b + e_{it} - \bar{e}_i$.

b. Show that the fixed effects slope estimator can be expressed as

\[ b = \beta + \left( \sum_{i=1}^n \sum_{t=1}^{T_i} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' \right)^{-1} \sum_{i=1}^n \sum_{t=1}^{T_i} (x_{it} - \bar{x}_i) e_{it}. \]

c. Show that the residual can be expressed as $e_{it} = (x_{it} - \bar{x}_i)'(\beta - b) + e_{it} - \bar{e}_i$.

d. Show that the mean square error defined in equation (2.11) is an unbiased estimator for this model. That is, show that

\[ E(s^2) = \frac{1}{N - (n + K)} \sum_{i=1}^n \sum_{t=1}^{T_i} e_{it}^2 = \sigma^2. \]

e. Show that $a_i - \bar{a}_w = \alpha_i - \bar{\alpha}_w + (\bar{x}_i - \bar{x})'(\beta - b) + \bar{e}_i - \bar{e}$, where $\bar{a}_w = N^{-1} \sum_{i=1}^n T_i a_i$.

f. Show that $E((a_i - \bar{a}_w)^2) = \sigma^2 \left( 1 + N^{-2} \sum_i T_i^2 - 2T_i / N \right)$.

g. Show that $E(s_{\alpha}^2) = \sigma^2_{\alpha}$. 
3.6. Ordinary least squares estimator

Perform the following steps to check that the ordinary least square estimator of the slope coefficient still performs well when the error components model is true. To this end:

a. Show that the ordinary least squares estimator for the model \( y_{it} = x_{it}' \beta + \epsilon_{it} \) can be expressed as

\[
\hat{b}_{\text{OLS}} = \left( \sum_{i=1}^{n} \sum_{t=1}^{T_i} x_{it} x_{it}' \right)^{-1} \left( \sum_{i=1}^{n} \sum_{t=1}^{T_i} x_{it}' y_{it} \right).
\]

b. Assuming the error components model, \( y_{it} = \alpha_i + x_{it}' \beta + \epsilon_{it} \), show that the difference between the part (a) estimator and the vector of parameters is

\[
\hat{b}_{\text{OLS}} - \beta = \left( \sum_{i=1}^{n} \sum_{t=1}^{T_i} x_{it} x_{it}' \right)^{-1} \left( \sum_{i=1}^{n} \sum_{t=1}^{T_i} x_{it}' \left( \alpha_i + \epsilon_{it} \right) \right).
\]

c. Use part (b) to argue that the estimator given in part (a) is unbiased.

d. Calculate the variance of \( \hat{b}_{\text{OLS}} \).

e. For \( K=1 \), show that the variance calculated in part (d) is larger than the variance of the random effects estimator, \( \text{Var} \hat{b}_{\text{EC}} \), given in Section 3.1.

3.7. Pooling test

Perform the following steps to check that the test statistic for the pooling test given in Section 3.1 has an approximate chi-square distribution under the null hypothesis of a homogeneous model of the form \( y_{it} = x_{it}' \beta + \epsilon_{it} \).

a. Check that the residuals can be expressed as \( e_{it} = \epsilon_{it} + x_{it}' \left( \beta - \hat{b}_{\text{OLS}} \right) \), where \( \hat{b}_{\text{OLS}} \) is the ordinary least squares estimator of \( \beta \) in Exercise 3.5a.

b. Check that

\[
-\frac{1}{N-K} \sum_{i=1}^{n} \sum_{t=1}^{T_i} e_{it}^2 = \sigma^2.
\]

c. Check that

\[
T_i (T_i - 1) s_i = T_i^2 \bar{e}_{it}^2 - \sum_{i=1}^{n} e_{it}^2 = \sum_{r \neq s} e_{ir} e_{is},
\]

where the latter sum is over \( \{(r, s) \text{ such that } s \neq r \text{ and } r=1, \ldots, T_i, s=1, \ldots, T_i\} \).

d. Check, for \( s \neq r \), that

\[
E e_{ir} e_{is} = -h_{ir,as} \sigma^2,
\]

where \( h_{ir,as} = x_{ir}' \left( \sum_{i=1}^{n} \sum_{t=1}^{T_i} x_{it} x_{it}' \right)^{-1} x_{is} \) is an element of the hat matrix.

e. Establish conditions so that the bias is negligible. That is, check that

\[
n^{-1/2} \sum_{i=1}^{n} \sum_{r \neq s} h_{ir,as} / \sqrt{T_i (T_i - 1)} \to 0.
\]

f. Determine the approximate variance of \( s_i \) by showing that

\[
E \left( \frac{1}{T_i (T_i - 1)} \sum_{r \neq s} e_{ir} e_{is} \right)^2 = \frac{2 \sigma^4}{T_i (T_i - 1)}.
\]

g. Outline an argument to show that

\[
\frac{1}{\sigma^2 \sqrt{2n}} \sum_{i=1}^{n} s_i \sqrt{T_i (T_i - 1)} \text{ is approximately standard normal},
\]

thus, completing the argument for the behavior of the pooling test statistic under the null hypothesis.
Section 3.3
3.8. Nested models
Let \( y_{i,j,t} \) be the output of the \( j \)th firm in the \( i \)th industry for the \( t \)th time period. Assume that the error structure is given by
\[
y_{i,j,t} = E_{y_{i,j,t}} + \delta_{i,j,t},
\]
where \( \delta_{i,j,t} = \alpha_i + \nu_{i,j} + \epsilon_{i,j,t} \). Here, assume that each of \( \{ \alpha_i \} \), \( \{ \nu_{i,j} \} \) and \( \{ \epsilon_{i,j,t} \} \) are independently and identically distributed and independent of one another.

a. Let \( y_i \) be the vector of responses for the \( i \)th industry. Write \( y_i \) as a function of \( y_{i,j,t} \).

b. Use \( 2\alpha\sigma, 2\nu\sigma \) and \( 2\epsilon\sigma \) to denote the variance of each error component, respectively. Give an expression for \( \text{Var} y_i \) in terms of these variance components.

c. Consider the linear mixed effects model,
\[
y_i = Z_i \alpha_i + X_i \beta + \epsilon_i.
\]
Show how to write the quantity \( Z_i \alpha_i \) in terms of the error components \( \alpha_i \) and \( \nu_{i,j} \) and the appropriate explanatory variables.

Section 3.4
3.9. GLS estimator as a weighted average of subject-specific GLS estimators
Consider the random coefficients models and consider the weighted average expression for the GLS estimator
\[
b_{GLS} = \left( \sum_{i=1}^{n} W_{i,\text{GLS}} \right)^{-1} \sum_{i=1}^{n} W_{i,\text{GLS}} b_{i,\text{GLS}}.
\]

a. Show that the weights can be expressed as \( W_{i,\text{GLS}} = \left( D + \sigma^2 (X_i'X_i)^{-1} \right)^{-1} \).

b. Show that \( \text{Var} b_{i,\text{GLS}} = D + \sigma^2 (X_i'X_i)^{-1} \).

3.10. Matched pairs design
Consider a pair of observations that have been “matched” in some way. The pair may consist of siblings, firms with similar characteristics but from different industries, or the same entity observed before and after some event of interest. Assume that there is reason to believe that the pair of observations are dependent in some fashion. Let \((y_{i1}, y_{i2})\) be the set of responses and \((x_{i1}, x_{i2})\) be the corresponding set of covariates. Because of the sampling design, the assumption of independence between \( y_{i1} \) and \( y_{i2} \) is not acceptable.

a. One alternative is to analyze the difference between the responses. Thus, let \( y_i = y_{i1} - y_{i2} \). Assuming perfect matching, we might assume that \( x_{i1} = x_{i2} = x_i \), say, and use the model \( y_i = x_i' \gamma + \epsilon_i \). Without perfect matching, one could use the model \( y_i = x_{i1}' \beta_1 - x_{i2}' \beta_2 + \eta_i \). Calculate the ordinary least squares estimator \( \hat{\beta} = (\beta_1', \beta_2')' \), and call this estimator \( \hat{b}_{PD} \) because the responses are paired differences.

b. As another alternative, form the vectors \( y_i = (y_{i1}, y_{i2})' \) and \( \epsilon_i = (\epsilon_{i1}, \epsilon_{i2})' \), as well as the matrix
\[
X_i = \begin{pmatrix}
x_{i1} & 0 \\
0 & x_{i2}
\end{pmatrix}.
\]
Now, consider the linear mixed effects model \( y_i = X_i \beta + \epsilon_i \), where the dependence between responses is induced by the variance \( R = \text{Var} \epsilon_i \).

i Under this model specification, show that \( \hat{b}_{PD} \) is unbiased.

ii Compute the variance of \( \hat{b}_{PD} \).

iii Calculate the generalized least squares estimator of \( \beta \), say \( \hat{b}_{GLS} \).

c. For yet another alternative, assume that the dependence is induced by a common latent random variable \( \alpha_i \). Specifically, consider the error components model \( y_{ij} = \alpha_i + x_{ij}' \beta_j + \epsilon_{ij} \).

i Under this model specification, show that \( \hat{b}_{PD} \) is unbiased.
ii Calculate the variance of $b_{PD}$.

iii Let $b_{EC}$ be the generalized least squares estimator under this model. Calculate the variance.

iv Show that if $\sigma^2_a \to \infty$, then the variance of $b_{EC}$ tends to the variance of $b_{PD}$. Thus, for highly correlated data, the two estimators have the same efficiency.

d. Continue with the model in part (c). We know that $\text{Var}(b_{EC})$ is “smaller” than $\text{Var}(b_{PD})$, because $b_{EC}$ is the generalized least squares estimator of $\beta$. To quantify this in a special case, assume asymptotically equivalent matching so that $n^{-1} \sum_{i=1}^{n} x_i' x_i \to \Sigma$, $j = 1, 2$. Moreover, let $n^{-1} \sum_{i=1}^{n} x_i' x_i \to \Sigma_12$ and assume that $\Sigma_12$ is symmetric. Suppose that we are interested in differences of the respondents, so that the (vector of) parameters of interest are $\beta_1 - \beta_2$. Let $b_{EC} = (b_{1,EC}', b_{2,EC}')'$ and $b_{PD} = (b_{1,PD}', b_{2,PD}')'$.

i Show that
\[
\frac{n}{\sigma^2} \text{Var}(b_{1,EC} - b_{2,EC}) = \frac{2}{(1 - \zeta/2)} \left( \Sigma_s + z \Sigma_12 \right)^{-1},
\]
where $z = \frac{\zeta/2}{(1 - \zeta/2)}$ and $\zeta = \frac{2\sigma^2_a}{2\sigma^2_a + \sigma^2_a}$.

ii Show that
\[
\frac{n}{\sigma^2} \text{Var}(b_{1,PD} - b_{2,PD}) = 4 \left( \Sigma_s + z \Sigma_12 \right)^{-1}
\]

iii Use parts d(i) and d(ii) to quantify the relative variances. For example, if $\Sigma_12 = 0$, then the relative variances (efficiency) is $1/(2 - \zeta)$ which is between 0.5 and 1.0.

3.11. Robust standard errors

To estimate the linear mixed effects model, consider the weighted least squares estimator, given in equation (3.18). The variance of this estimator, $\text{Var}(b_W)$, is also given in Section 3.4, along with the corresponding robust standard error of the $j$th component of $b_W$, denoted as $\text{se}(b_{W,j})$. Let us re-write this as:

\[
\text{se}(b_{j,W}) = \left( j^{th} \text{ diagonal element of } \sum_{i=1}^{n} x_i' W_{i,RE} X_i \right)^{-1} \sum_{i=1}^{n} x_i' W_{i,RE} \hat{y}_i W_{i,RE} X_i \left( \sum_{i=1}^{n} x_i' W_{i,RE} X_i \right)^{-1},
\]

where $\hat{y}_i = e_i e_i'$ is an estimator of $v_i$. In particular, explain:

a. How one goes from a variance-covariance matrix to a standard error?

b. What about this standard error makes it “robust?”

c. Let’s derive a new robust standard error. For simplicity, drop the $i$ subscripts and define the “hat matrix”
\[
H_W = W_{RE}^{1/2} X (X' W_{RE}^{-1} X)^{-1} X' W_{RE}^{1/2}.
\]

i Show that the weighted residuals can be expressed as a linear combination of weighted errors. Specifically, show
\[
W_{RE}^{1/2} e = (I - H_W) W_{RE}^{1/2} e.
\]

ii Show that
\[
E \left( W_{RE}^{1/2} e e' W_{RE}^{1/2} \right) = (I - H_W) W_{RE}^{1/2} V W_{RE}^{1/2} (I - H_W).
\]

iii Show that $e e'$ is an unbiased estimator of a linear transform of $V$. Specifically, show that
Chapter 3. Models with Random Effects / 3-37

\[ E(e'e') = (I - H^*_W)V(I - H^*_W) \]

where \( H^*_W = W_{RE}^{-1/2} H_W W_{RE}^{1/2} \).

d. Explain how the result in c(iii) suggests defining an alternative estimator of \( V_{i,i} \),

\[ \hat{V}_{i,i} = (I - H^*_W) e_i' (I - H^*_W)^{-1}. \]

Use this alternative estimator to suggest a new robust estimator of the standard error of \( b_{W,j} \).
(See Frees and Jin, 2004, if you would like more details about the properties of this estimator).

Section 3.5

3.12. Bias of MLE and REML variance component estimators

Consider the basic random effects model and suppose that \( T_i = T, K = 1 \) and that \( x_{it} = 1 \).
Further, do not impose boundary conditions so that the estimators may be negative.

a. Show that the maximum likelihood estimator of \( \sigma^2 \) may be expressed as:

\[ \hat{\sigma}^2_{ML} = \frac{1}{n(T-1)} \sum_{i=1}^{n} \sum_{t=1}^{T} (y_{it} - \bar{y}_i)^2. \]

b. Show that \( \hat{\sigma}^2_{ML} \) is an unbiased estimator of \( \sigma^2 \).

c. Show that the maximum likelihood estimator of \( \sigma^2_{\alpha} \) may be expressed as:

\[ \hat{\sigma}^2_{a,ML} = \frac{1}{n} \sum_{i=1}^{n} (\bar{y}_i - \bar{y})^2 - \frac{1}{T} \hat{\sigma}^2_{ML}. \]

d. Show that \( \hat{\sigma}^2_{a,ML} \) is a biased estimator of \( \sigma^2_{\alpha} \) and determine the bias.

e. Show that the restricted maximum likelihood estimator of \( \sigma^2 \) equals the corresponding maximum likelihood estimator, that is, show \( \hat{\sigma}^2_{REML} = \hat{\sigma}^2_{ML} \).

f. Show that the restricted maximum likelihood estimator of \( \sigma^2_{\alpha} \) may be expressed as:

\[ \hat{\sigma}^2_{a,REML} = \frac{1}{n-1} \sum_{i=1}^{n} (\bar{y}_i - \bar{y})^2 - \frac{1}{T} \hat{\sigma}^2_{ML}. \]

g. Show that \( \hat{\sigma}^2_{a,REML} \) is an unbiased estimator of \( \sigma^2_{\alpha} \).

Empirical Exercises


a. Error components model

Run an error components model of CHARITY on INCOME, PRICE, DEPS, AGE and MS.
State which variables are statistically significant and justify your conclusions.

b. Re-run the step in part (a) by including the supply-side measures as additional explanatory variables.
State whether or not these variables should be included in the model. Explain your reasoning.

c. Incorporating temporal effects. Is there an important time pattern? For the model in part a(i):

i. re-run it excluding YEAR as an explanatory variable yet including an AR(1) serial component for the error.

ii. re-run it including YEAR as an explanatory variable and including AR(1) serial component for the error.

iii. re-run it including YEAR as an explanatory variable and including an unstructured serial component for the error. (This step may be difficult to achieve convergence of the algorithm!)
iv Which model do you prefer, (i), (ii), or (iii)? Justify your choice. In your justification, discuss the nonstationarity of errors.
d Variable slope models
i Re-run the model in part (a) including a variable slope for INCOME. State which of the two models is preferred and state your reason.
ii Re-run the model in part (a) including a variable slope for PRICE. State which of the two models is preferred and state your reason.

Final Part. Which model do you think is best? Do not confine yourself to the options that you tested in the preceding parts. Justify your choice.

a Run an error components model using state as the subject identifier and VEHCMILE, GSTATEP, POPDENSY, WCMPMAX, URBAN, UNEMPLOY and JSLIAB as explanatory variables.
b Re-run the error components model in part (a) and include the additional explanatory variables COLLRULE, CAPS and PUNITIVE. Test whether these additional variables are statistically significant using the likelihood ratio test. State your null and alternative hypotheses, your test statistic and decision-making rule.
c Notwithstanding your answer in part (b), re-run the model in part (a) but also include variable random coefficients associated with WCMPMAX. Which model do you prefer, the model in part (a) or this one?
d Just for fun, re-run the model in part (b) and including variable random coefficients associated with WCMPMAX
e Re-run the error components model in part (a) but include an autoregressive error of order (1). Test for the significance of this term.
f Run the model in part (a) but with fixed effects. Compare this model to the random effects version.

3.15. Housing Prices – refer to Exercise 2.21 for the problem description.
a Basic summary statistics
i Produce a multiple time series plot of NARSP.
ii Produce a multiple time series plot of YPC.
iii Produce a scatter plot of NARSP versus YPC.
iv Produce an added variable plot of NARSP versus YPC, controlling for the effects of MSA.
v Produce a scatter plot of NARSP versus YPC.
b Error components model
i Run a one-way error components model of NARSP on YPC and YEAR. State which variables are statistically significant.
ii Re-run the step in b(i) by including the supply-side measures as additional explanatory variables. State whether or not these variables should be included in the model. Explain your reasoning.
c Incorporating temporal effects. Is there an important time pattern?
i Run a one-way error components model of NARSP on YPC. Calculate residuals from this model. Produce a multiple time series plot of residuals.
ii Re-run the model in part c(i) and include an AR(1) serial component for the error. Discuss the stationarity of errors based on the output of this model fit and your analysis in part c(i).
d Variable slope models
   i Re-run the model in part c(i), including a variable slope for YPC. Assume that the two random effects (intercepts and slopes) are independent.
   ii Re-run the model in part d(i) but allow for dependence between the two random effects. State which of the two models you prefer and why.
   iii Re-run the model in part d(ii) but incorporate the time-constant supply-side variables. Estimate the standard errors using robust standard errors. State which variables are statistically significant; justify your statement.
   iv Given the discussion of non-stationarity in part (c), describe why robust variance estimators are preferred when compared to the model-based standard errors.

3.16 Capital Structure – Unstructured problem

During the 1980s, Japan’s real economy was exhibiting a healthy rate of growth. The onset of the crash in the stock and real estate markets began at the end of December 1989, and the financial crisis soon spread to Japan’s banking system. After more than ten years, the banking system remained weak and the economy struggles.

These data provide information on 361 industrial Japanese firms before and after the crash. 355 of the 361 firms in the sample are from the First Section of the Tokyo Stock Exchange; the remaining six are from the Second Section. Together, they constitute about 33 percent of the market capitalization of the Tokyo Stock Exchange.

The sample firms are classified as keiretsu or non-keiretsu. That is, the main bank system is often part of a broader, cross share-holding structure that includes corporate groups called “keiretsu.” A useful function of such corporate groupings and the main bank system is to mitigate some of the informational and incentive problems in Japan’s financial markets. An important feature of the study is to identify changes in financial structure before and after the crash, and to see how these changes are affected by whether or not a firm is classified as “keiretsu.”

For this exercise, develop a model with random effects. Assess whether or not the keiretsu structure is an important determinant of leverage. For one approach, see Paker (2000O).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TANG</td>
<td>Tangibility. Net total fixed assets as a proportion of the book value of total assets.</td>
</tr>
<tr>
<td>MTB</td>
<td>Market-to-Book. Ratio of total assets at market value to total assets at book value.</td>
</tr>
<tr>
<td>LS</td>
<td>Logarithmic Sales. The natural logarithm of the amount of sales of goods and services to third parties, relating to the normal activities of the company.</td>
</tr>
<tr>
<td>PROF</td>
<td>Profitability. Earnings before interest and taxes plus depreciation, all divided by the book value of total assets.</td>
</tr>
<tr>
<td>STD</td>
<td>Volatility. The standard deviation of weekly, unlevered stock returns during the year. This variable proxies for the business risk facing the firm.</td>
</tr>
<tr>
<td>LVB</td>
<td>Total Leverage (Book). Total debt as a proportion of total debt plus book equity. Total debt is the sum of short-term and long-term debt. This is the dependent variable.</td>
</tr>
</tbody>
</table>

Abstract. Chapters 2 and 3 introduced models with fixed and random effects, focusing on estimation and hypothesis testing. This chapter discusses the third basic type of statistical inference, prediction. Prediction is particularly important in mixed effects models where one often needs to summarize a random effects component. In many cases the predictor can be interpreted as a shrinkage estimator, that is, a linear combination of local and global effects. In addition to predicting the random effects component, this chapter also shows how to predict disturbance terms, important for diagnostic work, as well as future responses, that are the main focus in forecasting applications.

The predictors are optimal in the sense that they are derived as minimum, mean square (best) linear unbiased predictors, known as BLUPs. As an alternative to this classic frequentist setting, Bayesian predictors are also defined. Moreover, Bayesian predictors with a diffuse prior are equivalent to the minimum, mean square linear unbiased predictors, thus providing additional motivation for these predictors.

Estimators versus predictors

In the linear mixed effects model, \( y_{it} = z_{it}' \alpha_i + x_{it}' \beta + \varepsilon_{it} \), the random variables \( \{ \alpha_i \} \) describe effects that are specific to a subject. Given the data \( \{ y_{it}, z_{it}, x_{it} \} \), in some problems it is of interest to “summarize” subject effects. Chapters 2 and 3 discussed how to estimate fixed, unknown parameters. This chapter discusses applications where it is of interest to summarize random subject-specific effects. In general, we use the term predictor for an “estimator” of a random variable. Like estimators, a predictor is said to be linear if it is formed from a linear combination of the responses.

When would an analyst be interested in “estimating” a random variable? In some applications, the interest is in predicting the future values of a response, known as forecasting. To illustrate, Section 4.4 demonstrates forecasting in the context of Wisconsin lottery sales. This chapter also introduces tools and techniques for prediction in contexts other than forecasting. For example, in animal and plant breeding, one wishes to predict the production of milk for cows based on their lineage (random) and herds (fixed). In insurance, one wishes to predict expected claims for a policyholder given exposure to several risk factors (known as credibility theory). In sample surveys, one wishes to predict the size of a specific age-sex-race cohort within a small geographical area (known as small area estimation). In a survey article, Robinson (1991S) also cites (1) ore reserve estimation in geological surveys, (2) measuring quality of a production plan and (3) ranking baseball players’ abilities.
4.1 Predictions for one-way ANOVA models

To begin, recall a special case of linear mixed effects models, the traditional one-way random effects ANOVA (analysis of variance) model,

\[ y_{it} = \mu + \alpha_i + \epsilon_{it}. \]  

(4.1)

As described in Section 3.1, we assume that both \( \alpha_i \) and \( \epsilon_{it} \) are mean zero, independent quantities. Suppose now that we are interested in summarizing the (conditional) mean effect of the \( i \)th subject, \( \mu + \alpha_i \).

For contrast, recall the corresponding fixed effects model. In this case, we did not explicitly express the overall mean but used the notation \( y_{it} = \alpha_i + \epsilon_{it} \). With this notation, \( \alpha_i \) represents the mean of the \( i \)th subject. We saw that \( \bar{y}_i \) is the “best” (Gauss-Markov) estimator of \( \alpha_i \). This estimator is unbiased, that is, \( E(\bar{y}_i) = \alpha_i \). Further, it is minimum variance (“best”) among all linear unbiased estimators (known by the acronym “BLUE”).

**Shrinkage estimator**

For the model in equation (4.1), it seems intuitively plausible that \( \bar{y} \) is a desirable estimator of \( \mu \) and that \( \bar{y}_i - \bar{y} \) is a desirable “estimator” of \( \alpha_i \). Thus, \( \bar{y}_i \) is a desirable predictor of \( \mu + \alpha_i \). More generally, consider predictors of \( \mu + \alpha_i \) that are linear combinations of \( \bar{y}_i \) and \( \bar{y} \), that is, \( c_1\bar{y}_i + c_2\bar{y} \), for constants \( c_1 \) and \( c_2 \). To retain the unbiasedness, we use \( c_2 = 1 - c_1 \). Some basic calculations (see Exercise 4.1) show that the best value of \( c_1 \) that minimizes \( E(c_1\bar{y}_i + (1-c_1)\bar{y} - (\mu + \alpha_i))^2 \) is

\[ c_1 = \frac{T_i^* \sigma^2}{\sigma^2 + T_i^* \sigma^2}, \]

where

\[ T_i^* = \frac{1 - \frac{2T_i}{N} + \frac{1}{N^2} \sum_{j=1}^n T_j^2}{T_i^{-1} - \frac{1}{N^2}}. \]

Here, we use the notation \( \sigma^2 \) and \( \sigma^2 \) for the variance of \( \alpha \) and \( \epsilon \), respectively. For interpretation, it is helpful to consider the case where the number of subjects, \( n \), tends to infinity. This yields the shrinkage estimator, or predictor, of \( \mu + \alpha_i \), defined as

\[ \bar{y}_{i,s} = \zeta_i \bar{y}_i + (1 - \zeta_i)\bar{y}, \]

where

\[ \zeta_i = \frac{T_i \sigma^2}{T_i \sigma^2 + \sigma^2} = \frac{T_i}{T_i + \sigma^2 / \sigma^2} \text{ is the } i\text{th credibility factor.} \]

**Example**

Consider the following illustrative data: \( y_1 = (14, 12, 10, 12)' \), \( y_2 = (9, 16, 15, 12)' \), and \( y_3 = (8, 10, 7, 7)' \). That is, we have \( n=3 \) subjects, each of which has \( T=4 \) observations. The sample mean is \( \bar{y} = 11 \); the subject-specific sample means are \( \bar{y}_1 = 12, \bar{y}_2 = 13 \) and \( \bar{y}_3 = 8 \). We now fit the one-way random effects ANOVA model in equation (4.1). From the variance estimation procedures described in Section 3.5, we have that the REML estimates of \( \sigma^2 \) and \( \sigma^2 \) are 4.889 and 5.778, respectively. It follows that the estimated \( \zeta_i \) weight is 0.825, and the corresponding predictions for the subjects are 11.825, 12.650, and 8.525, respectively.
Figure 4.1 compares subject-specific means to the corresponding predictions. Here, we see less spread in the predictions compared to the subject-specific means; each subject’s estimate is “shrunk” to the overall mean, $\overline{y}$. These are the best predictors assuming $\alpha_i$ are random. In contrast, the subject-specific means are the best predictors assuming $\alpha_i$ are deterministic. Thus, this “shrinkage effect” is a consequence of the random effects specification.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Specific Mean</th>
<th>Predicted Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>8</td>
<td>$\overline{y}$</td>
</tr>
<tr>
<td>1</td>
<td>11</td>
<td>$\overline{y}_1$</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>$\overline{y}_2$</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>$\overline{y}_3$</td>
</tr>
<tr>
<td></td>
<td>8.525</td>
<td></td>
</tr>
<tr>
<td></td>
<td>11.825</td>
<td></td>
</tr>
<tr>
<td></td>
<td>12.650</td>
<td></td>
</tr>
<tr>
<td>$\overline{y}_{3,s}$</td>
<td></td>
<td>$\overline{y}_{1,s}$</td>
</tr>
<tr>
<td></td>
<td>8.525</td>
<td></td>
</tr>
<tr>
<td></td>
<td>11.825</td>
<td></td>
</tr>
<tr>
<td></td>
<td>12.650</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 4.1** Comparison of Subject-Specific Means to Shrinkage Estimates.  
For an illustrative data set, subject-specific and overall means are graphed on the upper scale. The corresponding shrinkage estimates are graphed on the lower scale. This figure shows the shrinkage aspect of models with random effects.

Under the random effects ANOVA model, we have that $\overline{y}_i$ is an unbiased predictor of $\mu + \alpha_i$ in the sense that $\text{E} \left[ \overline{y}_i - (\mu + \alpha_i) \right] = 0$. However, $\overline{y}_i$ is inefficient in the sense that the shrinkage estimator, $\overline{y}_{i,s}$, has a smaller mean square error than $\overline{y}_i$. Intuitively, because $\overline{y}_{i,s}$ is a linear combination of $\overline{y}_i$ and $\overline{y}$, we say that $\overline{y}_i$ has been “shrunk” towards the estimator $\overline{y}$. Further, because of the additional information in $\overline{y}$, it is customary to interpret a shrinkage estimator as “borrowing strength” from the estimator of the overall mean.

Note that the shrinkage estimator reduces to the fixed effects estimator $\overline{y}_i$ when the credibility factor, $\zeta_i$, becomes 1. It is easy to see that $\zeta_i \rightarrow 1$ as either (i) $T_i \rightarrow \infty$ or (ii) $\sigma_{\alpha}^2/\sigma^2 \rightarrow \infty$. That is, the best predictor approaches the subject mean as either (i) the number of observations per subject becomes large or (ii) the variability among subjects becomes large relative to the response variability. In actuarial language, either case supports the idea that the information from the $i$th subject is becoming more “credible.”

**Best predictors**

When the number of observations per subject varies, the shrinkage estimator defined in equation (4.2) can be improved. This is due to the fact that $\overline{y}$ is not the optimal estimator of $\mu$. Using techniques described in Section 3.1, it is easy to check that (see Exercise 3.2) the generalized least squares (GLS) estimator of $\mu$ is

$$m_{a,\text{GLS}} = \frac{\sum_{i=1}^{n} \zeta_i \overline{y}_i}{\sum_{i=1}^{n} \zeta_i}.$$  \hspace{1cm} (4.3)

In Section 4.2, we will see that the linear predictor of $\mu + \alpha_i$ that has minimum variance is
The acronym BLUP stands for best linear unbiased predictor.

**Types of predictors**

This chapter focuses on predictors for three types of random variables.

1. **Linear combinations of regression parameters and subject-specific effects.** The statistic $\bar{y}_{i,\text{BLUP}}$ provides an optimal predictor of $\mu + \alpha_i$. Thus, $\bar{y}_{i,\text{BLUP}}$ is an example of a predictor of a linear combination of a global parameter and a subject-specific effect.

2. **Residuals.** Here, we wish to “predict” $\varepsilon_{it}$. The BLUP turns out to be:

$$e_{i,\text{BLUP}} = y_{it} - \bar{y}_{i,\text{BLUP}}. \quad (4.5)$$

These quantities, called BLUP residuals, are useful diagnostic statistics for developing a model. Further, unusual residuals help us understand if an observation is “in line” with others in the data set. To illustrate, if the response is a salary or a stock return, unusual residuals may help us detect unusual salaries or stock returns.

3. **Forecasts.** Here, we wish to predict, for “$L$” lead time units into the future,

$$y_{i,t+L} = \mu + \alpha_i + \varepsilon_{i,t+L}. \quad (4.6)$$

Forecasting is similar to predicting a linear combination of global parameters and subject-specific effects, with an additional future error term. In the absence of serial correlation, we will see that the predictor is the same as the predictor of $\mu + \alpha_i$, although the mean square error turns out to be larger. Serial correlation will lead to a different forecasting formula.

In this section, we have motivated BLUP’s using minimum variance unbiased prediction. One can also motivate BLUP’s using normal distribution theory. That is, consider the case where $\alpha_i$ and $\{y_{i,1}, \ldots, y_{i,T}\}$ have a joint multivariate normal distribution. Then, it can be shown that

$$E(\mu + \alpha_i \mid y_{i,1}, \ldots, y_{i,T}) = \zeta y_i + (1 - \zeta) \mu.$$  

This calculation is of interest because, if one were interested in “estimating” the unobservable $\alpha_i$ based on the observed responses $\{y_{i,1}, \ldots, y_{i,T}\}$, then normal theory suggests that the expectation is an optimal “estimator.” That is, consider asking the question: what realization of $\mu + \alpha_i$ could be associated with $\{y_{i,1}, \ldots, y_{i,T}\}$? The expectation! The BLUP is the best linear unbiased estimator (BLUE) of $E(\mu + \alpha_i \mid y_{i,1}, \ldots, y_{i,T})$, specifically, we need only replace $\mu$ by $m_{\alpha,\text{GLS}}$. Section 4.5 will discuss these ideas more formally in a Bayesian context.

### 4.2 Best linear unbiased predictors

This section develops best linear unbiased predictors in the context of mixed linear models. Section 4.3 then specializes the consideration to linear mixed effects models. Section 8.3 will consider another specialization, to time-varying coefficient models. As described in Section 4.1, we develop BLUP’s by examining the minimum mean square error predictor of a random variable, $w$. This development is due to Harville (1976S), which also appears in his discussion of Robinson (1991S). However, the argument is originally due to Goldberger (1962E), who coined
Chapter 4. Prediction and Bayesian Inference / 4-5

the phrase best linear unbiased predictor. The acronym BLUP was first used by Henderson (1973B).

Recall the mixed linear model presented in Section 3.3.2. That is, suppose that we observe an \( N \times 1 \) random vector \( y \) with mean \( E y = X \beta \) and variance \( \text{Var} y = V \). The generic goal is to predict a random variable \( w \), such that

\[
E w = \lambda' \beta \quad \text{and} \quad \text{Var} w = \sigma_w^2.
\]

Denote the covariance between \( w \) and \( y \) as the \( 1 \times N \) vector \( \text{Cov}(w, y) = E \{ (w - E w)(y - E y)' \} \).

The choice of \( \lambda \) and \( \sigma_w^2 \), will depend on the application at hand; several examples will be given in Section 4.3.

Begin by assuming that the global regression parameters \( \beta \) are known. Then, Appendix 4A.1 shows that the best linear (in \( y \)) predictor of \( w \) is

\[
w^* = E w + \text{Cov}(w, y)V^{-1}(y - E y) = \lambda' \beta + \text{Cov}(w, y)V^{-1}(y - X \beta).
\]

As we will see in the Bayesian context in Section 4.4, if \( w, y \) have a multivariate joint normal distribution, then \( w^* \) equals \( E (w | y) \) so that \( w^* \) is a minimum mean square predictor of \( w \).

Appendix 4A.2 shows that the predictor \( w^* \) is also a minimum mean square predictor of \( w \) without the assumption of normality.

**BLUP's as predictors**

Next, we assume that the global regression parameters \( \beta \) are not known. As in Section 3.5.2, we use \( b_{\text{GLS}} = (X'V^{-1}X)^{-1}X'V^{-1}y \) to be the generalized least squares (GLS) estimator of \( \beta \). This is the best linear unbiased estimator (BLUE) of \( \beta \). Replacing \( \beta \) by \( b_{\text{GLS}} \) in the definition of \( w^* \), we arrive at an expression for the BLUP of \( w \),

\[
w_{\text{BLUP}} = \lambda' b_{\text{GLS}} + \text{Cov}(w, y)V^{-1}(y - X b_{\text{GLS}}) = (\lambda' - \text{Cov}(w, y)V^{-1}X) b_{\text{GLS}} + \text{Cov}(w, y)V^{-1}y. \quad (4.7)
\]

Appendix 4A.2 establishes that \( w_{\text{BLUP}} \) is the best linear unbiased predictor of \( w \) in the sense that it is the best linear combination of responses that is unbiased and has the smallest mean square error over all linear, unbiased predictors. From Appendix 4A.3, we also have the form for the mean square error and variance:

\[
\text{Var}(w_{\text{BLUP}} - w) = (\lambda' - \text{Cov}(w, y)V^{-1}X)(X'V^{-1}X)^{-1}(\lambda' - \text{Cov}(w, y)V^{-1}X) + \text{Cov}(w, y)' + \sigma_w^2 \quad (4.8)
\]

and

\[
\text{Var} w_{\text{BLUP}} = \text{Cov}(w, y)V^{-1}\text{Cov}(w, y)' - (\lambda' - \text{Cov}(w, y)V^{-1}X)(X'V^{-1}X)^{-1}(\lambda' - \text{Cov}(w, y)V^{-1}X)'. \quad (4.9)
\]

From equations (4.8) and (4.9), we see that

\[
\sigma_w^2 = \text{Var} w = \text{Var}(w - w_{\text{BLUP}}) + \text{Var} w_{\text{BLUP}}.
\]

Hence, the prediction error, \( w_{\text{BLUP}} - w \), and the predictor, \( w_{\text{BLUP}} \), are uncorrelated. This fact will simplify calculations in subsequent examples.

The BLUP predictors are optimal, assuming the variance components implicit in \( V \) and \( \text{Cov}(w, y) \) are known. Applications of BLUP typically require that the variance components be estimated, as described in Section 3.5. BLUP with estimated variance components are known as empirical BLUPs, or EBLUPs. The formulas in equations (4.8) and (4.9) do not account for the uncertainty in variance component estimation. Inflation factors that account for this additional uncertainty have been proposed (Kackar and Harville, 1984S, but they tend to be small, at least
for data sets commonly encountered in practice. McCulloch and Searle (2001G) and Kenward and Roger (1997B) provide further discussions.

**Special case - One-way random effects ANOVA model**

We now establish the one-way random effects model BLUPs that were described in equations (4.4)-(4.6) of Section 4.1. To do this, we first write the one-way random effects ANOVA model as a special case of the mixed linear model. We then establish the predictions as special cases of $w_{BLUP}$ given in equation (4.7).

To express equation (4.1) in terms of a mixed linear model, recall the error components formulation in Section 3.1. Thus, we write (4.1) in vector form as

$$y_i = \mu + \alpha_i + \epsilon_i,$$

where $\text{Var}(y_i) = V_i = \sigma^2 \mathbf{J}_i + \sigma^2 \mathbf{1}_i$ and $V_i^{-1} = \frac{1}{\sigma^2} \left( \mathbf{1}_i - \frac{\zeta_i}{T_i} \mathbf{J}_i \right)$. Thus, we have $y = (y', \ldots, y_n')'$, $y_i = (y_1, \ldots, y_n)'$, $X = (1, \ldots, 1_n)'$ and $V = \text{block diagonal}(V_1, \ldots, V_n)$.

To develop expressions for the best linear unbiased predictors, we begin with GLS estimator, given in equation (4.3). Thus, from equation (4.7) and the block diagonal nature of $V$, we have

$$w_{BLUP} = \lambda m_{a,\text{GLS}} + \sum_{i=1}^n \text{Cov}(w, y_i)V_i^{-1}(y_i - \mathbf{1}_i m_{a,\text{GLS}}).$$

Now, we have the relation

$$V_i^{-3}(y_i - \mathbf{1}_i m_{a,\text{GLS}}) = \frac{1}{\sigma^2} \left( \mathbf{1}_i - \frac{\zeta_i}{T_i} \mathbf{J}_i \right)(y_i - \mathbf{1}_i m_{a,\text{GLS}}) = \frac{1}{\sigma^2} \left( (y_i - \mathbf{1}_i m_{a,\text{GLS}}) - \zeta_i \mathbf{1}_i(\bar{y}_i - m_{a,\text{GLS}}) \right).$$

This yields

$$w_{BLUP} = \lambda m_{a,\text{GLS}} + \frac{1}{\sigma^2} \sum_{i=1}^n \text{Cov}(w, y_i) \left( (y_i - \mathbf{1}_i m_{a,\text{GLS}}) - \zeta_i \mathbf{1}_i(\bar{y}_i - m_{a,\text{GLS}}) \right).$$

Now, suppose that the interest is in predicting $w = \mu + \alpha_i$. Then, we have $\lambda = 1$ and $\text{Cov}(w, y_i) = 1_i' \sigma^2 \mathbf{1}_i$ for the $i$th subject, and $\text{Cov}(w, y_i) = 0$ for all other subjects. Thus, we have

$$w_{BLUP} = m_{a,\text{GLS}} + \frac{\sigma^2}{\sigma^2} 1_i' \left( (y_i - \mathbf{1}_i m_{a,\text{GLS}}) - \zeta_i \mathbf{1}_i(\bar{y}_i - m_{a,\text{GLS}}) \right)$$

$$= m_{a,\text{GLS}} + \frac{\sigma^2}{\sigma^2} T_i \left( (\bar{y}_i - m_{a,\text{GLS}}) - \zeta_i(\bar{y}_i - m_{a,\text{GLS}}) \right)$$

$$= m_{a,\text{GLS}} + \frac{\sigma^2}{\sigma^2} T_i \left( 1 - \zeta_i \right)(\bar{y}_i - m_{a,\text{GLS}}) = m_{a,\text{GLS}} + \zeta_i(\bar{y}_i - m_{a,\text{GLS}}),$$

which confirms equation (4.4).

For predicting residuals, we assume that $w = \epsilon_i$. Thus, we have $\lambda = 0$ and $\text{Cov}(\epsilon_i, y_i) = \sigma^2 \mathbf{1}_i$ for the $i$th subject, and $\text{Cov}(w, y_i) = 0$ for all other subjects. Here, $\mathbf{1}_i$ denotes a $T_i \times 1$ vector with a one in the $i$th row and zeroes otherwise. Thus, we have

$$w_{BLUP} = \sigma^2 1_i' V_i^{-1} (y_i - X_i b_{\text{GLS}}) = y_i - \bar{y}_i, w_{BLUP},$$

which confirms equation (4.5).
For forecasting, we use equation (4.6) and choose \( w = \mu + \alpha_i + \epsilon_i \). Thus, we have \( \lambda = 1 \) and \( \text{Cov}(w, y_i) = \mathbf{1}, \sigma_{\alpha}^2 \) for the \( i \)th subject, and \( \text{Cov}(w, y_i) = \mathbf{0} \) for all other subjects. With this, our expression for \( w_{BLUP} \) is the same as the case in predicting \( w = \mu + \alpha_i \).

### 4.3 Mixed model predictors

Best linear unbiased predictors for mixed linear models were presented in equation (4.7) with corresponding mean square errors and variances in equations (4.8) and (4.9), respectively. This section uses these results by presenting three broad classes of predictors that are useful for linear mixed effects models, together with a host of special cases that provide additional interpretation. In some of the special cases, we point out that these results also pertain to:

- cross-sectional models, by choosing \( D \) to be a zero matrix, and
- fixed effects models, by choosing \( D \) to be a zero matrix and incorporating \( Z_i \alpha_i \) as fixed effects into the expected value of \( y_i \).

The three broad classes of predictors are (1) linear combinations of global parameters \( \beta \) and subject-specific effects \( \alpha_i \), (2) residuals and (3) forecasts.

#### 4.3.1 Linear mixed effects model

To see how the general Section 4.2 results apply to the linear mixed effects model, recall from equation (3.5) our matrix notation of the model: \( y_i = Z_i \alpha_i + X_i \beta + \epsilon_i \). As described in equation (3.8), we stack these equations to get \( y = Z \alpha + X \beta + e \). Thus, \( E y = X \beta \) with \( X = (X_1', X_2', ..., X_n') \). Further, we have \( \text{Var} y = V \) where the variance-covariance matrix is block diagonal of the form \( V = \text{block diagonal} (V_1, V_2, ..., V_n) \), where \( V_i = Z_i D Z_i' + R_i \). Thus, from equation (4.7), we have

\[
w_{BLUP} = \lambda' b_{GLS} + \text{Cov}(w, y)V^{-1}(y - X b_{GLS}) = \lambda' b_{GLS} + \sum_{i=1}^n \text{Cov}(w, y_i)V_i^{-1}(y_i - X_i b_{GLS}) .
\]

Exercise 4.9 provides expressions for the BLUP mean square error and variance.

#### 4.3.2 Linear combinations of global parameters and subject-specific effects

Consider predicting linear combinations of the form \( w = c_1' \alpha_i + c_2' \beta \). Here, \( c_1 \) and \( c_2 \) are known vectors of constants that are user-specified. Then, with this choice of \( w \), straight-forward calculations show that \( E w = c_2' \beta \) so that \( \lambda = c_2 \). Further, we have

\[
\text{Cov}(w, y_j) = \begin{cases} c_i' D Z_j' & \text{for } j = i \\ 0 & \text{for } j \neq i \end{cases}.
\]

Putting this in equation (4.10) yields

\[
w_{BLUP} = c_1' D Z_i' V_i^{-1}(y_i - X_i b_{GLS}) + c_2' b_{GLS} .
\]

To simplify this expression, we take \( c_2 = 0 \) and use Wald’s device. This yields the BLUP of \( \alpha_i \),

\[
a_{i,BLUP} = D Z_i' V_i^{-1}(y_i - X_i b_{GLS}) .
\]

With this notation, our BLUP predictor of \( w = c_1' \alpha_i + c_2' \beta \) is

\[
w_{BLUP} = c_1' a_{i,BLUP} + c_2' b_{GLS} .
\]

Some additional special cases are of interest. For the random coefficients model introduced in Section 3.3.1, with equation (4.12) it is easy to check that the BLUP of \( \beta + \alpha_i \) is
\[ w_{\text{BLUP}} = \zeta_i b_i + (1 - \zeta_i)b_{\text{GLS}}. \]

Here, \( b_i = (X_i' V_i^{-1} X_i)^{-1} X_i' V_i^{-1} y_i \) is the subject-specific GLS estimator and \( \zeta_i = D X_i' V_i^{-1} X_i \) is a weight matrix. This result generalizes the one-way random effects predictors presented in Section 4.1.

In the case of the error components model described in Section 3.1, we have \( q = 1 \) and \( z_{it} = 1 \). Using equation (4.11), the BLUP of \( \alpha_i \) reduces to

\[ a_{i,\text{BLUP}} = \zeta_i (\bar{y}_i - \bar{x}_i' b_{\text{GLS}}). \]

For comparison, recall from Chapter 2 that the fixed effects parameter estimate is

\[ a_i = \bar{y}_i - \bar{x}_i' b. \]

The other portion, \( 1 - \zeta_i \), is “borrowing strength” from zero, the mean of \( \alpha_i \).

Section 4.6 describes further examples from insurance credibility.

**Example – Trade localization, Continued**

Feinberg, Keane and Bognano (1998E) used firm-level data to investigate U.S. based multinational corporations employment and capital allocation decisions. From Chapter 3, their model can be written as

\[
\ln y_{it} = \beta_1 i \text{CT}_{it} + \beta_2 i \text{UT}_{it} + \beta_3 i \text{Trend}_t + x_{it}' \beta + \epsilon_{it}
\]

\[
= (\beta_1 + \alpha_1 i) \text{CT}_{it} + (\beta_2 + \alpha_2 i) \text{UT}_{it} + (\beta_3 + \alpha_3 i) \text{Trend}_t + x_{it}' \beta + \epsilon_{it}
\]

\[
= \alpha_1 i \text{CT}_{it} + \alpha_2 i \text{UT}_{it} + \alpha_3 i \text{Trend}_t + x_{it}' \beta + \epsilon_{it}.
\]

where \( \text{CT}_{it} \) (UT\( _{it} \)) is a measure of Canadian (U.S.) tariffs for firm \( i \), and the response \( y \) is either employment or durable assets for the Canadian affiliate. Feinberg, Keane and Bognano presented predictors of \( \beta_1 i \) and \( \beta_2 i \) using Bayesian methods (see the Section 4.5 discussion). In our notation, they predicted the linear combinations \( \beta_1 + \alpha_1 i \) and \( \beta_2 + \alpha_2 i \). The trend term was not of primary scientific interest and was included as a control variable. One major finding was that predictors for \( \beta_1 i \) were negative for each firm, indicating that employment and assets in Canadian affiliates increased as Canadian tariffs decreased.

### 4.3.3 BLUP residuals

For the second broad class, consider predicting a linear combination of residuals, \( w = c_i' \epsilon_i \), where \( c_i \) is a vector of constants. With this choice, we have \( E w = 0 \); it follows that \( \lambda = 0 \).

Straightforward calculations show that

\[
\text{Cov}(w_j, y_i) = \begin{cases} 
 c_j' R_i & \text{for } j = i \\
 0 & \text{for } j \neq i 
\end{cases}
\]

Thus, from equation (4.10) and Wald’s device, we have the vector of BLUP residuals

\[ e_{i,\text{BLUP}} = R_i V_i^{-1} (y_i - X_i b_{\text{GLS}}), \] (4.13a)

that can also be expressed as

\[ e_{i,\text{BLUP}} = y_i - (Z_i a_{i,\text{BLUP}} + X_i b_{\text{GLS}}). \] (4.13b)

Equation (4.13a) is appealing because it allows for direct computation of BLUP residuals; equation (4.13b) is appealing because it is in the traditional “observed minus expected” form for
residuals. We remark that the BLUP residual equals the GLS residual in the case that \( D = 0 \); in this case, \( e_{i,\text{BLUP}} = y_i - X_i b_{\text{GLS}} = e_{i,\text{GLS}} \). Further recall the symbol \( 1_t \) that denotes a \( T_t \times 1 \) vector that has a “one” in the \( t \)th position and is zero otherwise. Thus, we may define the BLUP residual as

\[
e_{it,\text{BLUP}} = 1_t' e_{it,\text{BLUP}} \equiv 1_t' R_i V_i^{-1} (y_i - X_i b_{\text{GLS}}).
\]

Equations (4.13a) and (4.13b) provide a generalization of the BLUP residual for the one-way random effects model described in equation (4.5). Further, using equation (4.9), one can show that the BLUP residual has variance

\[
\text{Var} e_{it,\text{BLUP}} = 1_t' R_i V_i^{-1} \left( \sum_{i=1}^{n} X_i' V_i^{-1} X_i \right)^{-1} V_i^{-1} R_i 1_t.
\]

Taking the square root of \( \text{Var} e_{it,\text{BLUP}} \) with an estimated variance yields a standard error; this in conjunction with the BLUP residual is useful for diagnostic checking of the fitted model.

### 4.3.4 Predicting future observations

For the third broad class, suppose that the \( t \)th subject is included in the data set and we wish to predict

\[
w = y_{i,T_t + L} = z_{i,T_t + L}' a_i + x_{i,T_t + L}' \beta + \varepsilon_{i,T_t + L},
\]

for \( L \) lead time units in the future. Assume that \( z_{i,T_t + L} \) and \( x_{i,T_t + L} \) are known. With this choice of \( w \), it follows that \( \lambda = x_{i,T_t + 1} \). Further, we have

\[
\text{Cov}(w, y_j) = \begin{cases} 
z_{i,T_t + L}' D Z_i' + \text{Cov}(\varepsilon_{i,T_t + L}, \varepsilon_j) & \text{for } j = i \\
0 & \text{for } j \neq i
\end{cases}
\]

Thus, using equations (4.10), (4.12) and (4.13), we have

\[
\hat{y}_{i,T_t + L} = w_{\text{BLUP}} = \left( z_{i,T_t + L}' D Z_i' + \text{Cov}(\varepsilon_{i,T_t + L}, \varepsilon_i) \right) V_i^{-1} \left( y_i - X_i b_{\text{GLS}} \right) + x_{i,T_t + L}' b_{\text{GLS}} + z_{i,T_t + L}' a_{i,\text{BLUP}} + \text{Cov}(\varepsilon_{i,T_t + L}, \varepsilon_i) R_i^{-1} e_{i,\text{BLUP}}.
\]

Thus, the forecast is the estimate of the conditional mean plus the serial correlation correction factor \( \text{Cov}(\varepsilon_{i,T_t + L}, \varepsilon_i) R_i^{-1} e_{i,\text{BLUP}} \).

Using equation (4.8), one can show that the variance of the forecast error as

\[
\text{Var}(\hat{y}_{i,T_t + L} - y_{i,T_t + L}) = \left( x_{i,T_t + L}' D Z_i' + \text{Cov}(\varepsilon_{i,T_t + L}, \varepsilon_i) \right) V_i^{-1} \left( y_i - X_i b_{\text{GLS}} \right) + x_{i,T_t + L}' b_{\text{GLS}} + z_{i,T_t + L}' a_{i,\text{BLUP}} + \text{Cov}(\varepsilon_{i,T_t + L}, \varepsilon_i) R_i^{-1} e_{i,\text{BLUP}}.
\]

#### Special case – Autoregressive serial correlation

To illustrate, consider the special case where we have autoregressive of order 1 (AR(1)), serially correlated errors. For stationary AR(1) errors, the lag \( j \) autocorrelation coefficient \( \rho_j \) can be expressed as \( \rho_j \). Thus, with an AR(1) specification, we have
\[
R = \sigma^2 \begin{pmatrix}
1 & \rho & \rho^2 & \ldots & \rho^{T-1} \\
\rho & 1 & \rho & \ldots & \rho^{T-2} \\
\rho^2 & \rho & 1 & \ldots & \rho^{T-3} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \ldots & 1
\end{pmatrix}
\]

where we have omitted the \(i\) subscript. Straightforward matrix algebra results show that

\[
R^{-1} = \frac{1}{\sigma^2(1-\rho^2)} \begin{pmatrix}
1 & -\rho & 0 & \ldots & 0 & 0 \\
-\rho & 1+\rho^2 & -\rho & \ldots & 0 & 0 \\
0 & -\rho & 1+\rho^2 & \ldots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \ldots & 1+\rho^2 & -\rho \\
0 & 0 & 0 & \ldots & -\rho & 1
\end{pmatrix}.
\]

Further, omitting the \(i\) subscript on \(T\), we have

\[
\text{Cov}(\varepsilon_{i,T+L}, \varepsilon_i) = \sigma^2 \begin{pmatrix}
\rho^{T+L-1} & \rho^{T+L-2} & \rho^{T+L-3} & \ldots & \rho^{T+L+1} & \rho^L
\end{pmatrix}.
\]

Thus,

\[
\text{Cov}(\varepsilon_{i,T+L}, \varepsilon_i)R_i^{-1} =
\frac{1}{(1-\rho^2)} \begin{pmatrix}
1 & -\rho & 0 & \ldots & 0 & 0 \\
-\rho & 1+\rho^2 & -\rho & \ldots & 0 & 0 \\
0 & -\rho & 1+\rho^2 & \ldots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \ldots & 1+\rho^2 & -\rho \\
0 & 0 & 0 & \ldots & -\rho & 1
\end{pmatrix}^{-1} = \begin{pmatrix}
0 & 0 & 0 & \ldots & 0 & \rho^L
\end{pmatrix}.
\]

To summarize, the \(L\) step forecast is

\[
\hat{y}_{i,T+L} = x'_{i,T+L} \mathbf{b}_{GLS} + z'_{i,T+L} \mathbf{a}_{i,\text{BLUP}} + \rho^L e_{i,T,\text{BLUP}}.
\]

That is, the \(L\) step forecast equals the estimate of the conditional mean, plus a correction factor of \(\rho^L\) times the most recent BLUP residual, \(e_{i,T,\text{BLUP}}\). This result was originally given by Goldberger (1962E) in the context of ordinary regression without random effects (that is, assuming \(D = 0\)).

### 4.4 Example: Forecasting Wisconsin lottery sales

In this section, we forecast the sale of state lottery tickets from 50 postal (ZIP) codes in Wisconsin. Lottery sales are an important component of state revenues. Accurate forecasting helps in the budget planning process. Further, a model is useful in assessing the important determinants of lottery sales. Understanding the determinants of lottery sales is useful for improving the design of the lottery sales system. Additional details of this study are in Frees and Miller (2003O).
4.4.1 Sources and characteristics of data

State of Wisconsin lottery administrators provided weekly lottery sales data. We consider online lottery tickets that are sold by selected retail establishments in Wisconsin. These tickets are generally priced at $1.00, so the number of tickets sold equals the lottery revenue. We analyze lottery sales (OLSALES) over a forty-week period, April, 1998 through January, 1999, from fifty randomly selected ZIP codes within the state of Wisconsin. We also consider the number of retailers within a ZIP code for each time (NRETAIL).

A budding literature, such as Ashley, Liu and Chang (1999), suggest variables that influence lottery sales. Table 4.1 lists economic and demographic characteristics that we consider in this analysis. Much of the empirical literature on lotteries is based on annual data that examines the state as the unit of analysis. In contrast, we examine much finer economic units, the ZIP code level, and examine weekly lottery sales. The economic and demographic characteristics were abstracted from the United States census. These variables summarize characteristics of individuals within ZIP codes at a single point in time and thus are not time varying.

<p>| Table 4.1. Lottery, economic and demographic characteristics of fifty Wisconsin ZIP codes |</p>
<table>
<thead>
<tr>
<th>Lottery characteristics</th>
<th>Economic and demographic characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLSALES</td>
<td>PERPERHH</td>
</tr>
<tr>
<td>NRETAIL</td>
<td>MEDSCHYR</td>
</tr>
<tr>
<td></td>
<td>MEDHVL</td>
</tr>
<tr>
<td></td>
<td>PRCRENT</td>
</tr>
<tr>
<td></td>
<td>PRC55P</td>
</tr>
<tr>
<td></td>
<td>HHMEDAGE</td>
</tr>
<tr>
<td></td>
<td>MEDINC</td>
</tr>
<tr>
<td></td>
<td>POPULATN</td>
</tr>
<tr>
<td>Online lottery sales to individual consumers</td>
<td>Persons per household</td>
</tr>
<tr>
<td>Number of listed retailers</td>
<td>Median years of schooling</td>
</tr>
<tr>
<td></td>
<td>Median home value in $1000s for owner-occupied homes</td>
</tr>
<tr>
<td></td>
<td>Percent of housing that is renter occupied</td>
</tr>
<tr>
<td></td>
<td>Percent of population that is 55 or older</td>
</tr>
<tr>
<td></td>
<td>Household median age</td>
</tr>
<tr>
<td></td>
<td>Estimated median household income, in $1000s</td>
</tr>
<tr>
<td></td>
<td>Population, in thousands</td>
</tr>
</tbody>
</table>


Table 4.2 summarizes the economic and demographic characteristics of fifty Wisconsin ZIP codes. To illustrate, for the population variable (POPULATN), we see that the smallest ZIP code contained 280 people whereas the largest contained 39,098. The average, over fifty ZIP codes, was 9,311.04. Table 4.2 also summarizes average online sales and average number of retailers. Here, these are averages over forty weeks. To illustrate, we see that the forty-week average of online sales was as low as $189 and as high as $33,181.

| Table 4.2. Summary statistics of lottery, economic and demographic characteristics of fifty Wisconsin ZIP codes |
| Variable | Mean | Median | Standard Deviation | Minimum | Maximum |
| Average OLSALES | 6,494.83 | 2,426.41 | 8,103.01 | 189 | 33,181 |
| Average NRETAIL | 11.94 | 6.36 | 13.29 | 1 | 68.625 |
| PERPERHH | 2.71 | 2.7 | 0.21 | 2.2 | 3.2 |
| MEDSCHYR | 12.70 | 12.6 | 0.55 | 12.2 | 15.9 |
| MEDHVL | 57.09 | 53.90 | 18.37 | 34.50 | 120 |
| PRCRENT | 24.68 | 24 | 9.34 | 6 | 62 |
| PRC55P | 39.70 | 40 | 7.51 | 25 | 56 |
| HHMEDAGE | 48.76 | 48 | 4.14 | 41 | 59 |
| MEDINC | 45.12 | 43.10 | 9.78 | 27.90 | 70.70 |
| POPULATN | 9.311 | 4.405 | 11.098 | 0.280 | 39.098 |
It is possible to examine cross-sectional relationships between sales and economic demographic characteristics. For example, Figure 4.2 shows a positive relationship between average online sales and population. Further, the ZIP code corresponding to city of Kenosha, Wisconsin has unusually large average sales for its population size.

![Figure 4.2. Scatter plot of average lottery sales versus population size.](image)

Sales for Kenosha are unusually large for its population size.

However, cross-sectional relationships, such as correlations and plots similar to Figure 4.2, hide dynamic patterns of sales. Figure 4.3 is a multiple time series plot of (weekly) sales over time. Here, each line traces the sales patterns for a ZIP code. This figure shows the dramatic increase in sales for most ZIP codes, at approximately weeks eight and eighteen. For both time points, the jackpot prize of one online game, PowerBall, grew to an amount in excess of $100 million. Interest in lotteries, and sales, increases dramatically when jackpot prizes reach large amounts.

![Figure 4.3. Multiple time series plot of lottery sales.](image)

Sales at and around weeks 8 and 18 are unusually large due to large PowerBall jackpots.
Figure 4.4 shows the same information as in Figure 4.3 but on a common (base ten) logarithmic scale. Here, we still see the effects of the PowerBall jackpots on sales. However, Figure 4.4 suggests a dynamic pattern that is common to all ZIP codes. Specifically, logarithmic sales for each ZIP code are relatively stable with the same approximate level of variability. Further, logarithmic sales for each ZIP code peaks at the same time, corresponding to large PowerBall jackpots.

Another form of the response variable to consider is the proportional, or percentage, change. Specifically, define the percentage change to be

\[
pchange_t = 100 \left( \frac{sales_t}{sales_{t-1}} - 1 \right).
\]  

(4.19)

A multiple times series plot of the percentage changes, not displayed here, shows autocorrelated serial patterns. We consider models of this transformed series in the following subsection on model selection.

### 4.4.2 In-sample model specification

This subsection considers the specification of a model, a necessary component prior to forecasting. We decompose model specification criteria into two components, in-sample and out-of-sample criteria. To this end, we partition our data into two subsamples; we use the first 35 weeks to develop alternative fitted models and use the last five weeks to “predict” our held-out sample. The choice of five weeks for the out-of-sample validation is somewhat arbitrary; it was made with the rationale that lottery officials consider it reasonable to try to predict five weeks of sales based on thirty-five weeks of historical sales data.

Our first forecasting model is the pooled cross-sectional model. The model fits the data well; the coefficient of determination turns out to be \( R^2 = 69.6\% \). The estimated regression
coefficients appear in Table 4.3. From the corresponding \( t \)-statistics, we see that each variable is statistically significant.

Our second forecasting model is an error components model. Table 4.3 provides parameter estimates and the corresponding \( t \)-statistics, as well as estimates of the variance components, \( \sigma_\alpha^2 \) and \( \sigma^2 \). As we have seen in other examples, allowing intercepts to vary by subject can result in regression coefficients for other variables becoming statistically insignificant.

When comparing this model to the pooled cross-sectional model, we may use the Lagrange multiplier test described in Section 3.1. The test statistic turns out to be \( TS = 11,395.5 \), indicating that error components model is strongly preferred to the pooled cross-sectional model. Another piece of evidence is Akaike’s Information Criterion (AIC). This criterion is defined as

\[
AIC = -2 \times \ln(\text{maximized likelihood}) + 2 \times (\text{number of model parameters}).
\]

The smaller this criterion, the more preferred is the model. Appendix C.9 describes this criterion in further detail. Table 4.3 shows again that the error components model is preferred compared to the pooled cross-sectional model based on the smaller value of the AIC statistic.

### Table 4.3 Lottery model coefficient estimates

Based on in-sample data of \( n = 50 \) ZIP codes and \( T = 35 \) weeks.

The response is (natural) logarithmic sales.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Pooled cross-sectional model</th>
<th>Error components model</th>
<th>Error components model with ( AR(1) ) term</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Parameter estimate</td>
<td>t-statistic</td>
<td>Parameter estimate</td>
</tr>
<tr>
<td>Intercept</td>
<td>13.821</td>
<td>10.32</td>
<td>18.096</td>
</tr>
<tr>
<td>PERPERHH</td>
<td>-1.085</td>
<td>-6.77</td>
<td>-1.287</td>
</tr>
<tr>
<td>MEDSCHYR</td>
<td>-0.821</td>
<td>-11.90</td>
<td>-1.078</td>
</tr>
<tr>
<td>MEDHVL</td>
<td>0.014</td>
<td>5.19</td>
<td>0.007</td>
</tr>
<tr>
<td>PRCRENT</td>
<td>0.032</td>
<td>8.51</td>
<td>0.026</td>
</tr>
<tr>
<td>PRC55P</td>
<td>-0.070</td>
<td>-5.19</td>
<td>-0.073</td>
</tr>
<tr>
<td>HHMEDAGE</td>
<td>0.118</td>
<td>5.64</td>
<td>0.119</td>
</tr>
<tr>
<td>MEDINC</td>
<td>0.043</td>
<td>8.18</td>
<td>0.046</td>
</tr>
<tr>
<td>POPULATN</td>
<td>0.057</td>
<td>9.41</td>
<td>0.121</td>
</tr>
<tr>
<td>NRETAIL</td>
<td>0.021</td>
<td>5.22</td>
<td>-0.027</td>
</tr>
<tr>
<td>Var ( \alpha (\sigma_\alpha^2) )</td>
<td>0.607</td>
<td>0.528</td>
<td>0.528</td>
</tr>
<tr>
<td>Var ( \varepsilon (\sigma^2) )</td>
<td>0.700</td>
<td>0.263</td>
<td>0.263</td>
</tr>
<tr>
<td>( AR(1) ) corr (( \rho ))</td>
<td>0.555</td>
<td>25.88</td>
<td>0.555</td>
</tr>
<tr>
<td>AIC</td>
<td>4353.25</td>
<td>2862.74</td>
<td>2270.97</td>
</tr>
</tbody>
</table>

To assess further the adequacy of the error components model, residuals from the fitted model were calculated. Several diagnostic tests and graphs were made using these residuals to improve the model fit. Figure 4.5 represents one such diagnostic graph, a plot of residuals versus lagged residuals. This figure shows a strong relationship between residuals and lagged residuals which we can represent using an autocorrelation structure for the error terms. To accommodate this pattern, we also consider an error component model with an \( AR(1) \) term; the fitted model appears in Table 4.3.

Figure 4.5 also shows a strong pattern of clustering corresponding to weeks with large PowerBall jackpots. A variable that captures information about the size of PowerBall jackpots would help in developing a model of lottery sales. However, for forecasting purposes, we require one or more variables that\ anticipates \ large PowerBall jackpots. That is, because the size of
PowerBall jackpots is not known in advance, variables that proxy the event of large jackpots are not suitable for forecasting models. These variables could be developed through a separate forecasting model of PowerBall jackpots.

Other types of random effects models for forecasting lottery sales could also be considered. To illustrate, we also fit a more parsimonious version of the AR(1) version of the error components model; specifically, we re-fit this model, deleting those variables with insignificant $t$-statistics. It turned out that this fitted model did not perform substantially better in terms of overall model fit statistics such as $AIC$. We explore alternative transforms of the response when examining a held-out sample in the following subsection.

4.4.3 Out-of-sample model specification

This subsection compares the ability of several competing models to forecast values outside of the sample used for model parameter estimation. As in Section 4.4.2, we use the first 35 weeks of data to estimate model parameters. The remaining five weeks are used to assess the validity of model forecasts. For each model, we compute forecasts of lottery sales for weeks 36 through 40, by ZIP code level, based on the first 35 weeks. Denote these forecast values as $\hat{ZOLSALES}_{i,35+L}$, for $L = 1$ to 5. We summarize the accuracy of the forecasts through two statistics, the mean absolute error

$$MAE = \frac{1}{5n} \sum_{i=1}^{n} \sum_{L=1}^{5} |ZOLSALES_{i,35+L} - \hat{ZOLSALES}_{i,35+L}|$$

(4.20)

and the mean absolute percentage error
\[
\text{MAPE} = 100 \frac{\sum_{i=1}^{n} \sum_{L=1}^{5} |\text{ZOLSALES}_{i,35+L} - \hat{\text{ZOLSALES}}_{i,35+L}|}{\sum_{i=1}^{n} \sum_{L=1}^{5} \text{ZOLSALES}_{i,35+L}}.
\]

The several competing models include the three models of logarithmic sales summarized in Table 4.3. Because the autocorrelation term appears to be highly statistically significant in Table 4.3, we also fit a pooled cross-sectional model with an AR(1) term. Further, we fit two modifications of the error components model with the AR(1) term. In the first case we use lottery sales as the response (not the logarithmic version) and in the second case we use percentage change of lottery sales, defined in equation (4.19), as the response. Finally, the seventh model that we consider is a basic fixed effects model,

\[
y_{it} = \alpha_i + \varepsilon_{it},
\]

with an AR(1) error structure. Recall that for the fixed effects models, the term \(\alpha_i\) is treated as a fixed, not random, parameter. Because this parameter is time-invariant, it is not possible to include our time-invariant demographic and economic characteristics as part of the fixed effects model.

Table 4.4 presents the model forecast criteria in equations (4.20) and (4.21) for each of these seven models. We first note that Table 4.4 re-confirms the point that the AR(1) term improves each model. Specifically, for both the pooled cross-sectional and the error components model, the version with an AR(1) term outperforms the analogous model without this term. Table 4.4 also shows that the error components model dominates the pooled cross-sectional model. This was also anticipated by our pooling test, an in-sample test procedure.

Table 4.4 confirms that the error components model with an AR(1) term with logarithmic sales as the response is the preferred model, based on either the MAE or MAPE criterion. The next best model was the corresponding fixed effects model. It is interesting to note that the models with sales as the response outperformed the model with percentage change as the response based on the MAE criterion, although the reverse is true based on the MAPE criterion.

<table>
<thead>
<tr>
<th>Model forecast criteria</th>
<th>Model response</th>
<th>MAE</th>
<th>MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pooled cross-sectional model</td>
<td>logarithmic sales</td>
<td>3,012.68</td>
<td>83.41</td>
</tr>
<tr>
<td>Pooled cross-sectional model with AR(1) term</td>
<td>logarithmic sales</td>
<td>680.64</td>
<td>21.19</td>
</tr>
<tr>
<td>Error components model</td>
<td>logarithmic sales</td>
<td>1,318.05</td>
<td>33.85</td>
</tr>
<tr>
<td>Error components model with AR(1) term</td>
<td>logarithmic sales</td>
<td>571.14</td>
<td>18.79</td>
</tr>
<tr>
<td>Error components model with AR(1) term</td>
<td>sales</td>
<td>1,409.61</td>
<td>140.25</td>
</tr>
<tr>
<td>Error components model with AR(1) term</td>
<td>percentage change</td>
<td>1,557.82</td>
<td>48.70</td>
</tr>
<tr>
<td>Fixed effects model with AR(1) term</td>
<td>logarithmic sales</td>
<td>584.55</td>
<td>19.07</td>
</tr>
</tbody>
</table>

### 4.4.4 Forecasts

We now forecast using the model that provides the best fit to the data, the error components model with an AR(1) term. The forecasts and variance of forecast errors for this model are special cases of the results for the linear mixed effects model, given in equations (4.14) and (4.15), respectively. Forecast intervals are calculated, using a normal curve approximation, as the point forecast plus or minus 1.96 times the square root of the estimated variance of the forecast error.
Figure 4.6 displays the forecasts and forecast intervals. Here, we use $T = 40$ weeks of data to estimate parameters and provide forecasts for $L = 5$ weeks. Calculation of the parameter estimates, point forecasts and forecast intervals were done using logarithmic sales as the response. Then, point forecasts and forecast intervals were converted to dollars to display the ultimate impact of the model forecasting strategy.

Figure 4.6 shows the forecasts and forecast intervals for two selected postal codes. The lower forecast represents a postal code from Dane County whereas the upper represents a postal code from Milwaukee. For each postal code, the middle line represents the point forecast and the upper and lower lines represent the bounds on a 95% forecast interval. Compared to the Dane County code, the Milwaukee postal code has higher forecast sales. Thus, although standard errors on a logarithmic scale are about the same as Dane County, this higher point forecast leads to a larger interval when rescaled to dollars.

![Figure 4.6 Forecast Intervals for Two Selected Postal Codes. For each postal code, the middle line corresponds to point forecasts for five weeks. The upper and lower lines correspond to endpoints of 95% prediction intervals.](image)

### 4.5 Bayesian inference

With Bayesian statistical models, one views both the model parameters and the data as random variables. In this section, we use a specific type of Bayesian model, the normal linear hierarchical model discussed by, for example, Gelman et al. (2004S). As with the two-stage
sampling scheme described in Section 3.3.1, the hierarchical linear model is one that is specified in stages. Specifically, we consider the following two-level hierarchy:

1. Given the parameters $\beta$ and $\alpha$, the response model is $y = Z\alpha + X\beta + \epsilon$. This level is an ordinary (fixed) linear model that was introduced in Section 3.3.2. Specifically, we assume that the responses $y$ conditional on $\alpha$ and $\beta$ are normally distributed and that $E(y | \alpha, \beta) = Z\alpha + X\beta$ and $\text{Var}(y | \alpha, \beta) = R$.

2. Assume that $\alpha$ is distributed normally with mean $\mu_\alpha$ and variance $D$ and that $\beta$ is distributed normally with mean $\mu_\beta$ and variance $\Sigma_\beta$, each independent of the other.

The technical differences between the mixed linear model and the normal hierarchical linear model are:

- in the mixed linear model, $\beta$ is an unknown, fixed parameter whereas in the normal hierarchical linear model, $\beta$ is a random vector, and
- the mixed linear model is distribution-free, whereas distributional assumptions are made in each stage of the normal hierarchical linear model.

Moreover, there are important differences in interpretation. To illustrate, suppose that $\beta = 0$ with probability one. In the classic non-Bayesian, also known as the frequentist, interpretation, we think of the distribution of $\{\alpha\}$ as representing the likelihood of drawing a realization of $\alpha_i$. The likelihood interpretation is most suitable when we have a population of firms or people and each realization is a draw from that population. In contrast, in the Bayesian case, one interprets the distribution of $\{\alpha\}$ as representing the knowledge that one has of this parameter. This distribution may be subjective and allows the analyst a form mechanism to inject his or her assessments into the model. In this sense the frequentist interpretation may be regarded as a special case of the Bayesian framework.

The joint distribution of $(\alpha', \beta', y')'$ is known as the prior distribution. To summarize, the joint distribution of $(\alpha', \beta', y')'$ is

$$
\begin{pmatrix}
\alpha \\
\beta \\
y
\end{pmatrix}
\sim N
\begin{pmatrix}
\mu_\alpha \\
\mu_\beta \\
Z\mu_\alpha + X\mu_\beta
\end{pmatrix},
\begin{pmatrix}
D & 0 & DZ' \\
0 & \Sigma_\beta & \Sigma_\beta X' \\
ZD & X\Sigma_\beta & V + X\Sigma_\beta X'
\end{pmatrix},
$$

(4.16)

where $V = R + ZDZ'$.

The distribution of parameters given the data is known as the posterior distribution (see Appendix 9A). To calculate this conditional distribution, we use standard results from multivariate analysis (see Appendix B). Specifically, the posterior distribution of $(\alpha', \beta')'$ given $y$ is normal. The conditional moments are

$$
E\left(\begin{pmatrix}
\alpha \\
\beta
\end{pmatrix} | y\right) = \begin{pmatrix}
\mu_\alpha + DZ'(V + X\Sigma_\beta X')^{-1}(y - Z\mu_\alpha - X\mu_\beta) \\
\mu_\beta + \Sigma_\beta X'(V + X\Sigma_\beta X')^{-1}(y - Z\mu_\alpha - X\mu_\beta)
\end{pmatrix}
$$

(4.17)

and

$$
\text{Var}\left(\begin{pmatrix}
\alpha \\
\beta
\end{pmatrix} | y\right) = \begin{pmatrix}
D & 0 \\
0 & \Sigma_\beta
\end{pmatrix} - \begin{pmatrix}
DZ' \\
\Sigma_\beta X'
\end{pmatrix}(V + X\Sigma_\beta X')^{-1}\begin{pmatrix}
DZ' \\
\Sigma_\beta X'
\end{pmatrix}.
$$

(4.18)

Up to this point, the treatment of parameters $\alpha$ and $\beta$ has been symmetric. In longitudinal data applications, one typically has more information about the global parameters $\beta$ than subject-specific parameters $\alpha$. To see how the posterior distribution changes depending on the amount of information available, we consider two extreme cases. First, consider the case $\Sigma_\beta = 0$, so that $\beta =$
\(\mu_\beta\) with probability one. Intuitively, this means that \(\beta\) is precisely known, generally from collateral information. Then, from equations (4.17) and (4.18), we have

\[
E(\alpha | y) = \mu_\alpha + D Z' V^{-1} (y - Z \mu_\alpha - X \beta)
\]

and

\[
\text{Var}(\alpha | y) = D - D Z' V^{-1} Z D.
\]

Assuming that \(\mu_\alpha = 0\), the best linear unbiased estimator (BLUE) of \(E(\alpha | y)\) is

\[
a_{\text{BLUP}} = D Z' V^{-1} (y - X b_{\text{GLS}})
\]

Recall from equation (4.11) that \(a_{\text{BLUP}}\) is also the best linear unbiased predictor (BLUP) in the frequentist (non-Bayesian) model framework.

Second, consider the case where \(\Sigma_\beta^{-1} = 0\). In this case, prior information about the parameter \(\beta\) is vague; this is known as using a diffuse prior. To analyze the impact of this assumption, use equation (A.4) of Appendix A.5 to get

\[
(V + X \Sigma_\beta X')^{-1} = V^{-1} - V^{-1} X (X' V^{-1} X + \Sigma_\beta^{-1})^{-1} X' V^{-1}\]

\[
\to V^{-1} - V^{-1} X (X' V^{-1} X)^{-1} X' V^{-1} = Q_v,
\]

as \(\Sigma_\beta^{-1} \to 0\). Note that \(Q_v X = 0\). Thus, with \(\Sigma_\beta^{-1} = 0\) and \(\mu_\alpha = 0\) we have \(\alpha | y \sim N\) with mean

\[
E(\alpha | y) = D Z' Q_v y
\]

and variance

\[
\text{Var}(\alpha | y) = D - D Z' Q_v Z D.
\]

This summarizes the posterior distribution of \(\alpha\) given \(y\). Interestingly, from the expression for \(Q_v\), we have

\[
E(\beta | y) = D Z' (V^{-1} - V^{-1} X (X' V^{-1} X)^{-1} X' V^{-1}) y
\]

\[
= D Z' V^{-1} y - D Z' V^{-1} X b_{\text{GLS}}
\]

\[
= a_{\text{BLUP}}.
\]

Similarly, one can check that \(E(\beta | y) \to b_{\text{GLS}}\) as \(\Sigma_\beta^{-1} \to 0\).

Thus, it is interesting that in both extreme cases, we arrive at the statistic \(a_{\text{BLUP}}\) as a predictor of \(\alpha\). This analysis assumes \(D\) and \(R\) are matrices of fixed parameters. It is also possible to assume distributions for these parameters; typically, independent Wishart distributions are used for \(D'\) and \(R'\) as these are conjugate priors. (Appendix 9A introduces conjugate priors.) Alternatively, one can estimate \(D\) and \(R\) using methods described in Section 3.5. The general strategy of substituting point estimates for certain parameters in a posterior distribution is called empirical Bayes estimation.

To examine intermediate cases, we look to the following special case. Generalizations may be found in Luo, Young and Frees (2001O).

**Special case – One-way random effects ANOVA model**

We return to the model considered in Section 4.1 and, for simplicity, assume balanced data so that \(T_i = T\). The goal is to determine the posterior distributions of the parameters. For illustrative purposes, we derive the posterior means and leave the derivation of posterior variances as an exercise for the reader. Thus, with equation (4.1), the model is

\[
y_{it} = \beta + \alpha_i + \epsilon_{it},
\]

where we use the random \(\beta \sim N(\mu_\beta, \sigma_\beta^2)\) in lieu of the fixed mean \(\mu\). The prior distribution of \(\alpha_i\) is independent with \(\alpha_i \sim N(0, \sigma_\alpha^2)\).

Using equation (4.17), the posterior mean of \(\beta\) is
\[ \hat{\beta} = E(\beta | y) = \mu_\beta + \Sigma_\beta X'(V + X\Sigma_\beta X')^{-1}(y - X\mu_\beta) = \left( \frac{1}{\sigma_\beta^2} + \frac{nT}{\sigma_\beta^2 + T\sigma_\alpha^2} \right)^{-1} \left( \frac{nT}{\sigma_\beta^2 + T\sigma_\alpha^2} \bar{y} + \mu_\beta \right) \]

after some algebra. Thus, \( \hat{\beta} \) is a weighted average of the sample mean, \( \bar{y} \), and the prior mean, \( \mu_\beta \). It is easy to see that \( \hat{\beta} \) approaches the sample mean \( \bar{y} \) as \( \sigma_\beta^2 \to \infty \), that is, as prior information about \( \beta \) becomes “vague.” Conversely, \( \hat{\beta} \) approaches the prior mean \( \mu_\beta \) as \( \sigma_\beta^2 \to 0 \), that is, as information about \( \beta \) becomes “precise.”

Similarly, using equation (4.17), the posterior mean of \( \alpha \) is

\[ \hat{\alpha}_i = E(\alpha_i | y) = \zeta_i \left( (\bar{y}_i - \mu_\beta) - \zeta_\beta (\bar{y} - \mu_\beta) \right) \]

where we recall that \( \zeta = \frac{T\sigma_\alpha^2}{\sigma_\epsilon^2 + T\sigma_\alpha^2} \) and define \( \zeta_\beta = \frac{nT\sigma_\beta^2}{\sigma_\epsilon^2 + T\sigma_\alpha^2 + nT\sigma_\beta^2} \). Note that \( \zeta_\beta \) measures the precision of knowledge about \( \beta \). Specifically, we see that \( \zeta_\beta \) approaches one as \( \sigma_\beta^2 \to \infty \), and approaches zero as \( \sigma_\beta^2 \to 0 \).

Combining these two results, we have that

\[ \hat{\alpha}_i + \hat{\beta} = (1 - \zeta_\beta)(1 - \zeta)\mu_\beta + \zeta_\beta \bar{y}_i + \zeta \bar{y} \]

Thus, if our knowledge of the distribution of \( \hat{\beta} \) is vague, then \( \zeta_\beta = 1 \) and the predictor reduces to the expression in equation (4.4) (for balanced data). Conversely, if our knowledge of the distribution of \( \hat{\beta} \) is precise, then \( \zeta_\beta = 0 \) and the predictor reduces to the expression given at the end of Section 4.1. With the Bayesian formulation, we may entertain situations where knowledge is available although imprecise.

To summarize, there are several advantages of the Bayesian approach. First, one can describe the entire distribution of parameters conditional on the data, such as through equations (4.17) and (4.18). This allows one, for example, to provide probability statements regarding the likelihood of parameters. Second, this approach allows analysts to blend information known from other sources with the data in a coherent manner. In our development, we assumed that information may be known through the vector of \( \beta \) parameters, with their reliability controlled through the dispersion matrix \( \Sigma_\beta \). Values of \( \Sigma_\beta = 0 \) indicate complete faith in values of \( \mu_\beta \), whereas values of \( \Sigma_\beta^{-1} = 0 \) indicate complete reliance on the data in lieu of prior knowledge.

Third, the Bayesian approach provides for a unified approach for estimating \((\alpha,\beta)\). Chapter 3 on non-Bayesian methods required a separate section on variance components estimation. In contrast, in Bayesian methods, all parameters can be treated in a similar fashion. This is convenient for explaining results to consumers of the data analysis. Fourth Bayesian analysis is particularly useful for forecasting future responses; we develop this aspect in Chapter 10.

**Section 4.6 Credibility theory**

Credibility is a technique for pricing insurance coverages that is widely used by health, group term life and property and casualty actuaries. In the United States, the practical standards of application are described under the Actuarial Standard of Practice Number 25 published by the Actuarial Standards Board of the American Academy of Actuaries (web site: http://www.actuary.org/). Further, several insurance laws and regulation require the use of credibility.
The theory of credibility has been called a “cornerstone” of the field of actuarial science (Hickman and Heacox, 1999O). The basic idea is to use claims experience and additional information to develop a pricing formula through the relation

\[
\text{New Premium} = \zeta \times \text{Claims Experience} + (1 - \zeta) \times \text{Old Premium}.
\] (4.22)

Here, \(\zeta\) is known as the “credibility factor;” values generally lie between zero and one. The case \(\zeta = 1\) is known as “full credibility,” where claims experience is used solely to determine the premium. The case \(\zeta = 0\) can be thought of as “no credibility,” where claims experience is ignored and external information is used as the sole basis for pricing.

Credibility has long found use in practice, with applications dating back to Mowbray (1914). See Hickman and Heacox (1999O) and Venter (1996O) for historical accounts. The modern theory of credibility began with the work of Bühlmann (1967O), who showed how to express equation (4.22) in what we now call a random effects framework, thus removing the seemingly ad hoc nature of this procedure. Bühlmann expressed traditional credibility insurance prices as conditional expectations, where the conditioning is based on an unobserved risk type that he called a “structure variable.”

Applications of credibility theory are considerably enhanced by accounting for known risk factors such as trends through (continuous) explanatory variables, different risk classes through categorical explanatory variables and dynamic behavior through evolving distributions. These types of applications can be handled under the framework of mixed linear models; see Norberg (1986O) and Frees, Young and Luo (1999O). This section shows that this class of models contains the standard credibility models as a special case.

By demonstrating that many important credibility models can be viewed in a longitudinal data framework, we restrict our consideration to certain types of credibility models. Specifically, the longitudinal data models accommodate only unobserved risks that are additive. Thus, we do not address models of nonlinear random effects that have been investigated in the actuarial literature; see, for example, Taylor (1977O) and Norberg (1980). Taylor (1977O) allowed insurance claims to be possibly infinite dimensional using Hilbert space theory and established credibility formulas in this general context. Norberg (1980O) considered the more concrete context, yet still general, of multivariate claims and established the relationship between credibility and statistical empirical Bayes estimation.

By expressing credibility ratemaking applications in the framework of longitudinal data models, actuaries can realize several benefits:

- Longitudinal data models provide a wide variety of models from which to choose.
- Standard statistical software makes analyzing data relatively easy.
- Actuaries have another method for explaining the ratemaking process.
- Actuaries can use graphical and diagnostic tools to select a model and assess its usefulness.

### 4.6.1 Credibility theory models

In this subsection, we demonstrate that commonly used credibility techniques are special cases of best linear unbiased prediction applied to the longitudinal/panel data model. For additional examples, see Frees, Young and Luo (2001O).

**Special Case - Basic credibility model of Bühlmann (1967O)**

Bühlmann (1967O) considered the one-way random effects ANOVA model that we now write as \(y_{it} = \beta + \alpha_i + \epsilon_{it}\), see Section 4.1. If \(y_{it}\) represents the claims of the \(i\)th subject in period \(t\), then \(\beta\) is the grand mean of the claims over the collection of subjects (policyholders, geographical regions, occupational classes, etc.), and \(\alpha_i\) is the deviation of the \(i\)th subject’s hypothetical mean from the overall mean \(\beta\). Here, the hypothetical mean is the conditional expected value \(E(y_{it} | \alpha_i)\)
= \beta + \alpha_i. The disturbance term \( \epsilon_{it} \) is the deviation of \( y_{it} \) from its hypothetical mean. One calls \( \sigma^2_{\alpha} \) the variance of the hypothetical means and \( \sigma^2 \) the process variance.

**Special case - Heteroscedastic model of Bühlmann-Straub (1970O)**

Continue with the basic Bühlmann model and change only the variance-covariance matrix of the errors to

\[
\begin{pmatrix}
1 \\
\vdots \\
1
\end{pmatrix}
\]

By this change, we allow each observation to have a different exposure weight (Bühlmann and Straub, 1970O). For example, if a subject is a policyholder, then \( w_{it} \) measures the size of the \( i \)th policyholder’s exposure during the \( t \)th period, possibly via payroll as for workers compensation insurance.

**Special case - Regression model of Hachemeister (1975O)**

Now assume a random coefficients model so that \( x_{it} = z_{it} \). Then, with \( R \) as in the Bühlmann-Straub model, we have the regression model of Hachemeister (1975O). Hachemeister focused on the linear trend model for which \( K = q = 2, x_{it} = z_{it} = (1, t) \).

**Special case - Nested classification model of Jewell (1975O)**

Suppose \( y_{ijt} = \beta + \mu_i + \gamma_{ij} + \epsilon_{ijt} \), a sum of uncorrelated components, in which \( \beta \) is the overall expected claims, \( \mu_i \) is the deviation of the conditional expected claims of the \( i \)th sector from \( \beta, i = 1, 2, \ldots, n \), \( \gamma_{ij} \) is the deviation of the conditional expected claims of the \( j \)th subject in the \( i \)th sector from the sector expectation of \( \beta + \mu_i \), \( j = 1, 2, \ldots, n_i \), and \( \epsilon_{ijt} \) is the deviation of the observation \( y_{ijt} \) from \( \beta + \mu_i + \gamma_{ij} \), \( t = 1, 2, \ldots, T_{ij} \). If one were to apply this model to private passenger automobile insurance, for example, then the sector might be age of the insured(s) while the subject is geographical region of the insured(s). Note that one assumes with this model that the claims in region \( j \) for different ages are uncorrelated. If one believes this to be an unreasonable assumption, then a cross classification model might be appropriate; see Dannenburg, Kaas and Goovaerts (1996O). As an example for which this nested model might be more appropriate, one could let the sector be geographical region, while the subject is the policyholder.

### 4.6.2 Credibility ratemaking

In credibility ratemaking, one is interested in predicting the expected claims, conditional on the random risk parameters of the \( i \)th subject, for time period \( T_i + 1 \). In our notation, the credibility ratemaking problem is to “estimate” \( E(y_{i,T+1} | \mathbf{a}_i) = \mathbf{x}_{i,T+1}' \beta + \mathbf{z}_{i,T+1}' \mathbf{a}_i \). From Section 4.3.2, the BLUP predictor of claims is \( \mathbf{x}_{i,T+1}' \mathbf{b}_{GLS} + \mathbf{z}_{i,T+1}' \mathbf{a}_{i,BLUP} \).

In Section 4.1, we saw how to predict claims for the Bühlmann model. We now illustrate this prediction for other familiar credibility models. A summary is in Table 4.5.

**Special case - Heteroscedastic model of Bühlmann-Straub (1970O) - Continued**

In the Bühlmann-Straub case, we have \( x_{it} = z_{it} = 1 \), so that the predictor of claims is \( \mathbf{b}_{GLS} + \mathbf{a}_{i,BLUP} \). Straightforward calculations, similar to the Bühlmann case, show that the predictor of claims is

\[
(1 - \zeta) \mathbf{m}_{\alpha, GLS} + \zeta \overline{y}_{i,w}.
\]
where now the credibility factor is \( \zeta_i = \frac{\sum_{t=1}^{T_i} w_{it}}{\sum_{t=1}^{T_i} w_{it} + \sigma^2 / \sigma_a^2} \) (see Table 4.5).

**Special case - Regression model of Hachemeister (1975O) - Continued**

In the Hachemeister case, we have \( x_i = z_i \). Define \( b_i = (X'_i V_i^{-1} X_i)^{-1} X'_i V_i^{-1} y_i \) to be the GLS estimator of \( \alpha_i + \beta \) based only on the \( i \)th subject. In Exercise 4.3, we ask the reader to show that the BLUP (credibility estimator) of \( \alpha_i + \beta \) is

\[
(1 - \zeta_i) b_{GLS} + \zeta_i b_i,
\]

in which \( \zeta_i = D X'_i V_i^{-1} X_i \) is the credibility factor. As in the Bühlmann case, again we see that the credibility estimator is a weighted average of a subject-specific statistic and a statistic that summarizes information from all subjects. This example is prominent in credibility theory because one can further express the GLS estimator of \( \beta \) as a weighted average of the \( b_i \) using the credibility factors as weights, \( b_{GLS} = \left( \sum_{i=1}^{n} \zeta_i \right)^{-1} \sum_{i=1}^{n} \zeta_i b_i \).

In Table 4.5, we show how to predict expected claims in the other examples that we considered in Section 4.6.1. In each case, the predicted claims for the \( i \)th subject is a weighted average of that subject’s experience with the \( b_{GLS} \), using the \( i \)th credibility factor as a weight.
### Table 4.5 Credibility Factors and Prediction of Claims

<table>
<thead>
<tr>
<th>Notation</th>
<th>Credibility Factors</th>
<th>Prediction of Claims</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bühlmann</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \bar{y}_i )</td>
<td>( \eta_i = \frac{1}{t_i} \sum_{t=1}^{T_i} y_{it} )</td>
<td>( \xi_i = \frac{T_i}{T_i + \sigma^2/\sigma_a^2} ) For subject ( i ), ( (1-\xi_i)m_{\alpha, GLS} + \xi_i \bar{y}_i )</td>
</tr>
<tr>
<td>( m_{\alpha, GLS} )</td>
<td>( \sum_{i=1}^{n} \zeta_i \bar{y}_i )</td>
<td></td>
</tr>
</tbody>
</table>

| **Bühlmann-Straub**    |                    |                      |
| \( \bar{y}_{i,w} \)   | \( \eta_{i,w} = \frac{1}{t_{i,w}} \sum_{t=1}^{T_{i,w}} y_{it} \) | \( \xi_{i,w} = \frac{\sum_{t=1}^{T_{i,w}} w_{it} y_{it}}{\sum_{t=1}^{T_{i,w}} w_{it}} \) For subject \( i \), \( (1-\xi_{i,w})m_{\alpha, GLS} + \xi_{i,w} \bar{y}_{i,w} \) |
| \( m_{\alpha, GLS} \) | \( \sum_{i=1}^{n} \zeta_{i,w} \bar{y}_{i,w} \) |                      |

<table>
<thead>
<tr>
<th><strong>Hachemeister (Linear trend)</strong></th>
<th>( \eta )</th>
<th>( \xi )</th>
<th>( \eta_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( W_i = \left[ \sum_{j=1}^{q} w_{it} \sum_{t=1}^{T_i} t^{w_{it}} \right] )</td>
<td>( \frac{\text{det}(W_i^j)I + \sigma^2 W_i^j}{\text{det}(W_i^j) + \sigma^2 \text{trace}(W_i^j) + \sigma^4} )</td>
<td>( \xi_i )</td>
<td>( \frac{\sigma_{\mu}^2 A_i}{\sigma_{\gamma}^2 A_i + \sigma_{\mu}^2} ) and ( (1-\xi_i)(1-\xi_{i,w})b_{GLS} + \xi_{i,w} \bar{y}_{i,w} )</td>
</tr>
<tr>
<td>( \bar{y}<em>{j,w} = \frac{\sum</em>{t=1}^{T_{j,w}} w_{it} y_{it}}{\sum_{t=1}^{T_{j,w}} w_{it}} )</td>
<td>( \frac{\sigma_{\gamma}^2 \sum_{t=1}^{T_{j,w}} w_{it} y_{it}}{\sigma_{\gamma}^2 \sum_{t=1}^{T_{j,w}} w_{it} + \sigma^2} )</td>
<td>( \xi_j )</td>
<td>For sector ( i ), ( (1-\xi_j)m_{\alpha, GLS} + \xi_j \bar{y}_{j,w} )</td>
</tr>
<tr>
<td>( m_{\alpha, GLS} )</td>
<td>( \sum_{i=1}^{n} \zeta_i \left( \frac{\sum_{j=1}^{m} \zeta_j \bar{y}<em>{j,w}}{\sum</em>{j=1}^{m} \zeta_j} \right) )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Jewell</strong></th>
<th>( A_i )</th>
<th>( \eta )</th>
<th>( \xi )</th>
<th>( \eta_j )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \bar{y}<em>{j,w} = \frac{\sum</em>{t=1}^{T_{j,w}} w_{it} y_{it}}{\sum_{t=1}^{T_{j,w}} w_{it}} )</td>
<td>( \frac{\sigma_{\mu}^2 A_i}{\sigma_{\mu}^2 A_i + \sigma_{\gamma}^2} ) and ( (1-\xi_j)(1-\xi_{j,w})m_{\alpha, GLS} + \xi_{j,w} \bar{y}_{j,w} )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( m_{\alpha, GLS} )</td>
<td>( \sum_{i=1}^{n} \zeta_i \left( \frac{\sum_{j=1}^{m} \zeta_j \bar{y}<em>{j,w}}{\sum</em>{j=1}^{m} \zeta_j} \right) )</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Further Reading

For readers who would like more background in small area estimation, please refer to Ghosh and Rao (1994S).

For readers who would like more background in credibility theory, please refer to Dannenburg, Kaas and Goovaerts (1996O), Klugman, Panjer and Willmot (1998O) and Venter (1996O). The Section 4.6 introduction to credibility theory does not include the important connection to Bayesian inference that was first pointed out by Bailey (1950O). See, for example, Klugman (1992O) and Pinquet (1997O). For connections with credibility and the Chapter 8 Kalman filter, see Klugman (1992O) and Ledolter, Klugman and Lee (1991O). Bayesian inference is further described in Chapter 10.
Appendix 4A. Linear Unbiased Prediction

Appendix 4A.1 Minimum Mean Square Predictor

Let $c_1$ be an arbitrary constant and $c_2$ be a vector of constants. For this choice of $c_1$ and $c_2$, the mean square error in using $c_1 + c_2' y$ to predict $w$ is

$$MSE(c_1, c_2) = E((c_1 + c_2' y - w)^2) = Var(c_1 + c_2' y - w) + E((c_1 + c_2' E y - E w)^2).$$

Using $E y = X \beta$ and $E w = \lambda'$, we have

$$MSE(c_1, c_2) = \frac{\partial}{\partial c_1} MSE(c_1, c_2) = 2(c_1 + (c_2' X - \lambda') \beta).$$

Equating this to zero yields $c_1^* = c_1 - (\lambda' - c_2' X) \beta$. For this choice of $c_1$, we have

$$MSE(c_1^*, c_2) = E((c_2' y - (w - E w))^2) = Var(c_2' y - w) = c_2' V c_2 + \sigma^2_w - 2 \text{Cov}(w, y)c_2.$$

To find the best choice of $c_2$, we have

$$\frac{\partial}{\partial c_2} MSE(c_1^*, c_2) = 2Vc_2 - 2\text{Cov}(w, y)' \lambda.$$

Setting this equal to zero yields $c_2^* = V^{-1} \text{Cov}(w, y)'$. Thus, the minimum mean square predictor is

$$c_1^* + c_2^* y = (\lambda' - \text{Cov}(w, y) V^{-1} X) \beta + \text{Cov}(w, y) V^{-1} y$$

as required.

Appendix 4A.2 Best Linear Unbiased Predictor

To check that $w_{BLUP}$ in equation (4.7) is the best linear unbiased predictor, consider all other unbiased linear estimators of the form $w_{BLUP} + c' y$, where $c$ is a vector of constants. By the unbiasedness, we have that

$$E c' y = E w - E w_{BLUP} = 0.$$

Thus, $c' y$ is an unbiased estimator of 0. Following Harville (1976S), we require this of all possible distributions so that a necessary and sufficient condition for $E c' y = 0$ is $c' X = 0$.

We wish to minimize the mean square prediction error over all choices of $c$, so consider

$$E((w_{BLUP} + c' y - w)^2).$$

Next,

$$\text{Cov}(w_{BLUP} - w, c' y) = \text{Cov}(w_{BLUP}, y)c - \text{Cov}(w, y)c$$

$$= \text{Cov}(w, y)V^{-1} \text{Cov}(y, y)c + \left(\lambda' - \text{Cov}(w, y) V^{-1} X\right) \text{Cov}(b_{GLS}, y)c - \text{Cov}(w, y)c$$

$$= \text{Cov}(w, y)c + \left(\lambda' - \text{Cov}(w, y) V^{-1} X (X' V^{-1} X)^{-1} X' V^{-1} \text{Cov}(y, y)c - \text{Cov}(w, y)c$$

$$= \left(\lambda' - \text{Cov}(w, y) V^{-1} X\right) (X' V^{-1} X)^{-1} X' c = 0.$$ (4A.1)

The last equality follows from $c' X = 0$. Thus, we have

$$E((w_{BLUP} + c' y - w)^2) = \text{Var}(w_{BLUP} - w) + \text{Var}(c' y),$$

that can be minimized by choosing $c = 0$. 
Appendix 4A.3 BLUP Variance

First note that \( \text{Cov}(y, \text{Cov}(w, y)V^{-1}y - w) = \text{Cov}(y, y)V^{-1}\text{Cov}(w, y)' - \text{Cov}(y, w) = 0. \)

Then, we have
\[
\text{Var}(w_{\text{BLUP}} - w) = \text{Var}(\lambda' - \text{Cov}(w, y)V^{-1}X)b_{\text{GLS}} + \text{Cov}(w, y)V^{-1}y - w) \\
= \text{Var}(\lambda' - \text{Cov}(w, y)V^{-1}X)b_{\text{GLS}} + \text{Var}(\text{Cov}(w, y)V^{-1}y - w).
\]

Also, we have
\[
\text{Var}(\text{Cov}(w, y)V^{-1}y - w) = \text{Var}(\text{Cov}(w, y)V^{-1}y) + \text{Var}(w) - 2\text{Cov}(\text{Cov}(w, y)V^{-1}y, w) \\
= \text{Cov}(w, y)V^{-1}\text{Cov}(w, y)' + \text{Var}(w) - 2\text{Cov}(w, y)V^{-1}\text{Cov}(y, w) \\
= \sigma_w^2 - \text{Cov}(w, y)V^{-1}\text{Cov}(w, y)'.
\]

Thus,
\[
\text{Var}(w_{\text{BLUP}} - w) = (\lambda' - \text{Cov}(w, y)V^{-1}X)(X'V^{-1}X)^{-1}(\lambda' - \text{Cov}(w, y)V^{-1}X)' \\
- \text{Cov}(w, y)V^{-1}\text{Cov}(w, y) + \sigma_w^2,
\]

as in equation (4.8).

From equation (4A.1), we have \( \text{Cov}(w_{\text{BLUP}} - w, w_{\text{BLUP}}) = 0. \) Thus,
\[
\sigma_w^2 = \text{Var}(w - w_{\text{BLUP}}) + \text{Var}(w_{\text{BLUP}}).
\]

With equation (4A.2), we have
\[
\text{Var} w_{\text{BLUP}} = \sigma_w^2 - \text{Var}(w - w_{\text{BLUP}}) \\
= \text{Cov}(w, y)V^{-1}\text{Cov}(w, y)' - (\lambda' - \text{Cov}(w, y)V^{-1}X)(X'V^{-1}X)^{-1}(\lambda' - \text{Cov}(w, y)V^{-1}X)',
\]

which is equation (4.9).
4. Exercises and Extensions

Section 4.1

4.1. Shrinkage estimator

Consider the Section 4.1 one-way random effects model with \( K=1 \) so that \( y_{it} = \mu + \alpha_i + \epsilon_{it} \).

a. Show that \( \text{E}(c_2 (\overline{y} - \overline{y}_i) + \overline{y}_i - \alpha_i)^2 \) is minimized over choices of \( c_2 \) at

\[
c_2 = -\frac{\text{Cov}(\overline{y} - \overline{y}_i, \overline{y}_i - \alpha_i)}{\text{Var}(\overline{y} - \overline{y}_i)}.
\]

b. Show that \( \text{Var}(\overline{y}) = \sigma^2 + \frac{T_i \sigma^2}{N} \), \( \text{Cov}(\overline{y}_i, \alpha_i) = \sigma^2 \), \( \text{Cov}(\overline{y}_i, \alpha_i) = \frac{T_i}{N} \sigma^2 \),

\[
\text{Cov}(\overline{y}, \overline{y}_i) = \frac{\sigma^2 + T_i \sigma^2}{N} \quad \text{and} \quad \text{Var}(\overline{y}) = \frac{\sigma^2}{N} + \frac{\sigma^2}{N^2} \sum_{j=1}^{n} T_j^2.
\]

c. Use part (b) to show that \( \text{Cov}(\overline{y} - \overline{y}_i, \overline{y}_i - \alpha_i) = \sigma^2 \left( \frac{1}{N} - \frac{1}{T_i} \right) \).

d. Use part (b) to show that \( \text{Var}(\overline{y} - \overline{y}_i) = \sigma^2 \left( \frac{1}{T_i} - \frac{1}{N} \right) \sigma^2 \left( 1 - \frac{2T_i}{N} + \frac{1}{N^2} \sum_{j=1}^{n} T_j^2 \right) \).

e. Use parts (a), (c) and (d) to show that the optimal choice of \( c_2 \) is \( c_2 = \frac{\sigma^2}{\sigma^2 + T_i \sigma^2} \) and

\[
c_1 = 1 - c_2 = \frac{T_i \sigma^2}{\sigma^2 + T_i \sigma^2}, \quad \text{where} \quad T_i^* = \frac{1 - \frac{2T_i}{N} + \frac{1}{N^2} \sum_{j=1}^{n} T_j^2}{\frac{1}{T_i} - \frac{1}{N}}.
\]

f. Use part (e) to show that, for balanced data with \( T_i = T \), we have \( T_i^* = T \).

g. Use part (e) to show that, as \( N \to \infty \), we have \( T_i^* = T_i \).

Section 4.3

4.2. BLUP predictor of random effects – error components model

Consider the Section 4.1 one-way random ANOVA model. Use equation (4.11) to show that the BLUP predictor of \( \alpha_i \) is

\[
a_{i,\text{BLUP}} = \zeta_i (\overline{y}_i - X_i b_{\text{GLS}}).
\]

4.3. BLUP prediction – random coefficients model

Consider the random coefficients model with \( K=q \) and \( z_{it} = x_{it} \). Use equation (4.12) to show that the BLUP predictor of \( \beta + \alpha_i \) is \( w_{\text{BLUP}} = \zeta_i b_i + (1 - \zeta_i) b_{\text{GLS}} \), where \( b_i = (X_i' V_i^{-1} X_i)^{-1} X_i' V_i^{-1} y_i \), \( X_i' V_i^{-1} y_i \) is the subject-specific GLS estimator and \( \zeta_i = D X_i' V_i^{-1} X_i b_{\text{GLS}} \).

4.4. BLUP residuals

Use equations (4.11) and (4.13) to show \( a_{i,\text{BLUP}} = D Z_i' R_i^{-1} e_{i,\text{BLUP}} \).
4.5. **BLUP subject-specific effects**

Use equations (4.11) and (A.4) of Appendix A.5 to show

\[
a_{i,\text{BLUP}} = \left( D^{-1} + Z_i^T R_i^{-1} Z_i \right)^{-1} Z_i^T R_i^{-1} (y_i - X_i b_{\text{GLS}}). \]

(For this alternative expression, one needs to only invert \( R_i \) and \( q \times q \) matrices, not a \( T_i \times T_i \) matrix.)

4.6. **BLUP residuals**

Use equation (4.11) to show that the BLUP residual in equation (4.13a) can be expressed as equation (4.13b), that is, as

\[
e_{i,\text{BLUP}} = y_i - (Z_i a_{i,\text{BLUP}} + X_i b_{\text{GLS}}). \]

4.7. **Covariance**

Use equation (4.11) to show that

\[
\text{Cov}(a_{i,\text{BLUP}}, b_{\text{GLS}}) = 0. \]

4.8. **BLUP forecasts – random walk**

Assume a random walk serial error correlation structure (Section 8.2.2 will provide additional motivation). Specifically, suppose that the subject-level dynamics are specified through \( \varepsilon_{it} = \varepsilon_{i,t-1} + \eta_{it} \). Here, \( \{ \eta_{it} \} \) is an i.i.d. sequence with \( \text{Var} \eta_{it} = \sigma^2 \eta \) and assume that \( \{ \varepsilon_{0i} \} \) are unobserved constants. Then, we have \( \text{Var} \varepsilon_{it} = r \sigma^2 \eta \) and \( \text{Cov}(\varepsilon_{ir}, \varepsilon_{is}) = \text{Var} \varepsilon_{ir} = r \sigma^2 \eta \), for \( r < s \). This yields \( R = \sigma^2 \eta R_{RW} \), where

\[
R_{RW} = \begin{pmatrix}
1 & 1 & 1 & \cdots & 1 \\
1 & 2 & 2 & \cdots & 2 \\
1 & 2 & 3 & \cdots & 3 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 2 & 3 & \cdots & T
\end{pmatrix}.
\]

a. Show that

\[
R_{RW}^{-1} = \begin{pmatrix}
2 & -1 & 0 & \cdots & 0 & 0 \\
-1 & 2 & -1 & \cdots & 0 & 0 \\
0 & -1 & 2 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 2 & -1 \\
0 & 0 & 0 & \cdots & -1 & 1
\end{pmatrix}.
\]

b. Show that

\[
\text{Cov}(\varepsilon_{i,T_i+L}, \varepsilon_i)' = \sigma^2 \eta \begin{pmatrix} 1 & 2 & \cdots & T_i \end{pmatrix}.
\]

c. Determine the 1 step forecast, that is, determine the BLUP predictor of \( y_{i,T_i+1} \).

d. Determine the \( L \) step forecast, that is, determine the BLUP predictor of \( y_{i,T_i+L} \).

4.9. **BLUP mean square errors – linear mixed effects model**

Consider the linear mixed effects model introduced in Section 4.3.

a. Use the general expression for the BLUP mean square error to show that the mean square error for the linear mixed effects model can be expressed as:
\[
\text{Var}(w_{\text{BLUP}} - w) = \left( \lambda' - \sum_{i=1}^{n} \text{Cov}(w_i, y_i) V_i^{-1} X_i \right)^{-1} \left( \lambda' - \sum_{i=1}^{n} \text{Cov}(w_i, y_i) V_i^{-1} X_i \right)'
\]
\[\quad - \sum_{i=1}^{n} \text{Cov}(w_i, y_i) V_i^{-1} \text{Cov}(w, y_i)' + \sigma_w^2.
\]

b. Use the general expression for the BLUP variance to show that the variance for the linear mixed effects model can be expressed as:
\[
\text{Var} w_{\text{BLUP}} = \sum_{i=1}^{n} \text{Cov}(w_i, y_i) V_i^{-1} \text{Cov}(w, y_i)'.
\]

\[\text{Var}(w_{\text{BLUP}} - w) = \left( \lambda' - \sum_{i=1}^{n} \text{Cov}(w_i, y_i) V_i^{-1} X_i \right)^{-1} \left( \lambda' - \sum_{i=1}^{n} \text{Cov}(w_i, y_i) V_i^{-1} X_i \right)'
\]
\[\quad - \sum_{i=1}^{n} \text{Cov}(w_i, y_i) V_i^{-1} \text{Cov}(w, y_i)' + \sigma_w^2.
\]

c. Now suppose that the BLUP of interest is a linear combination of global parameters and subject-specific effects of the form \( w = c_1' \alpha_i + c_2' \beta \). Use part (a) to show that the mean square error is
\[
\text{Var}(w_{\text{BLUP}} - w) = \left( c_2' - c_i' DZ_i X_i' V_i^{-1} X_i \right)^{-1} \left( c_2' - c_i' DZ_i X_i' V_i^{-1} X_i \right)'
\]
\[\quad - c_i' DZ_i X_i' V_i^{-1} Z_i Dc_i + c_i' Dc_i.
\]

d. Use direct calculations to show that the variance of the BLUP of \( \alpha_i \) is
\[
\text{Var} \alpha_{i,\text{BLUP}} = DZ_i X_i' V_i^{-1} \left( I_j - X_i' X_i^{-1} X_i' \right)^{-1} X_i' V_i^{-1} Z_i D.
\]

e. Use part (b) to establish the form of the variance of the BLUP residual in Section 4.3.3.

f. Use part (a) to establish the variance of the forecast error in equation (4.15).

4.10. Henderson’s mixed linear model justification of BLUPs

Consider the model in equation (3.8), Section 3.3.2,
\[
y = Z \alpha + X \beta + \epsilon.
\]

In addition, assume that \( \alpha, \epsilon \) are jointly multivariate normally distributed such that
\[
y | \alpha \sim N(Z\alpha + X\beta, R)
\]
and
\[
\alpha \sim N(0, D).
\]

a. Show that the joint logarithmic probability density function of \( y, \alpha \) is
\[
\ln(l(y, \alpha)) = -\frac{1}{2} \left( N \ln(2\pi) + \ln \det R + (y - (Z\alpha + X\beta))' R^{-1} (y - (Z\alpha + X\beta)) \right)
\]
\[-\frac{1}{2} \left( q \ln(2\pi) + \ln \det D + \alpha' D^{-1} \alpha \right).
\]

b. Treat this as a function of \( \alpha \) and \( \beta \). Take partial derivatives with respect to \( \alpha, \beta \) to yield Henderson’s (1984B) “mixed model equations”
\[
X' R^{-1} X\beta + X' R^{-1} Z\alpha = X' R^{-1} y
\]
\[
Z' R^{-1} X\beta + (Z' R^{-1} Z + D^{-1}) \alpha = Z' R^{-1} y.
\]

c. Show that solving Henderson’s “mixed model equations” for unknowns \( \alpha, \beta \) yields
\[ b_{GLS} = \left( X'V^{-1}X \right)^{-1} X'V^{-1} y \]
\[ a_{BLUP} = DZV^{-1} \left( y - Xb_{GLS} \right) \]

(Hint: Use equation (A.4) of Appendix A.5.)

**Empirical Exercises**

4.11. Housing Prices – refer to Exercise 2.19 for the problem description.

Here, we will calculate 95% prediction intervals for Chicago, the 11th metropolitan area. Below are the 9 annual values of NARSP, PERPYC and PERPOP for Chicago.

**a** Assume that you have fit a one-way fixed effects model:

\[
NARSP_t = \alpha_i + \beta_1 \text{PERPYC}_t + \beta_2 \text{PERPOP}_t + \beta_3 \text{YEAR}_t + \epsilon_t
\]

You have fit the model using least squares and arrived at the estimates \( b_1 = -0.008565, b_2 = -0.004347, b_3 = 0.036750, \alpha_{11} = 0.285 \) and \( \sigma = 0.0738 \). Assume that next year’s (1995) values for the explanatory variables are \( \text{PERPYC}_{11,10} = 3.0 \) and \( \text{PERPOP}_{11,10} = 0.20 \). Calculate a 95% prediction interval for Chicago’s 1995 average sale price. When expressing your final answer, convert it to dollars in lieu of logarithmic dollars.

**b** Assume that you have fit an error components model that you have estimated using generalized least squares. You have fit the model using generalized least squares and arrived at the estimates \( b_1 = -0.01, b_2 = -0.004, b_3 = 0.0367, \sigma_a = 0.10 \) and \( \sigma_e = 0.005 \). Assume that next year’s (1995) values for the explanatory variables are \( \text{PERPYC}_{11,10} = 3.0 \) and \( \text{PERPOP}_{11,10} = 0.20 \). Calculate a 95% prediction interval for Chicago’s 1995 average sale price. When expressing your final answer, convert it to dollars in lieu of logarithmic dollars.

**c** Assume that you have fit an error components model with an AR(1) autocorrelation structure that you have estimated using generalized least squares. You have fit the model using generalized least squares and arrived at the estimates \( b_1 = -0.01, b_2 = -0.004, b_3 = 0.0367, \rho = 0.1, \sigma_a = 0.10 \) and \( \sigma_e = 0.005 \). Assume that next year’s (1995) values for the explanatory variables \( \text{PERPYC}_{11,10} = 3.0 \) and \( \text{PERPOP}_{11,10} = 0.20 \). Calculate a 95% prediction interval for Chicago’s 1995 average sale price. When expressing your final answer, convert it to dollars in lieu of logarithmic dollars.

<table>
<thead>
<tr>
<th>YEAR</th>
<th>NARSP</th>
<th>PERPYC</th>
<th>PERPOP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.45551</td>
<td>5.83817</td>
<td>0.19823</td>
</tr>
<tr>
<td>2</td>
<td>4.50866</td>
<td>5.95691</td>
<td>0.32472</td>
</tr>
<tr>
<td>3</td>
<td>4.48864</td>
<td>7.80832</td>
<td>0.13056</td>
</tr>
<tr>
<td>4</td>
<td>4.67283</td>
<td>7.17689</td>
<td>0.33683</td>
</tr>
<tr>
<td>5</td>
<td>4.76046</td>
<td>5.90655</td>
<td>0.47377</td>
</tr>
<tr>
<td>6</td>
<td>4.87596</td>
<td>2.02724</td>
<td>0.99697</td>
</tr>
<tr>
<td>7</td>
<td>4.91852</td>
<td>-0.27135</td>
<td>-0.77503</td>
</tr>
<tr>
<td>8</td>
<td>4.95583</td>
<td>3.80041</td>
<td>0.19762</td>
</tr>
<tr>
<td>9</td>
<td>4.96564</td>
<td>3.66127</td>
<td>0.19723</td>
</tr>
</tbody>
</table>

Consider an example from workers’ compensation insurance, examining losses due to permanent, partial disability claims. The data are from Klugman (1992), who considers Bayesian model representations, and are originally from the National Council on Compensation Insurance. We consider $n = 121$ occupation, or risk, classes, over $T = 7$ years. To protect the data sources, further information on the occupation classes and years are not available.

The response variable of interest is the pure premium (PP), defined to be losses due to permanent, partial disability per dollar of PAYROLL. The variable PP is of interest to actuaries because worker compensation rates are determined and quoted per unit of payroll. The exposure measure, PAYROLL, is one of the potential explanatory variables. Other explanatory variables are YEAR ($= 1, ..., 7$) and occupation class.

For this exercise, develop a random effects model. Use this model to provide forecasts of the conditional mean of pure premium (known as credibility estimates in the actuarial literature). For one approach, see Frees et al. (2001).


We now consider claims data provided by a Wisconsin-based credit insurer. The data contains claims and exposure information for 88 Florida credit unions. These are “life savings” claims from a contract between the credit union and their members that provides a death benefit based on the member’s savings deposited in the credit union. The dependent variable is $\ln(LSTC) = \ln(1 + LSTC/1,000)$, where LSTC is the annual total claims from the life savings contract. The exposure measure is $\ln(LSCV) = \ln(1 + LSCV/1,000,000)$, where LSCV is the annual coverage for the life savings contract. Also available is the contract YEAR.

For this exercise, develop a random effects model. Use this model to provide forecasts of the conditional mean of pure premium (known as credibility estimates in the actuarial literature). For one approach, see Frees et al. (2001).
Chapter 5. Multilevel Models

Abstract. This chapter describes a conditional modeling framework that takes into account hierarchical and clustered data structures. The data and models, known as multilevel, are used extensively in educational science and related disciplines in the social and behavioral sciences. We show that a multilevel model can be viewed as a linear mixed effects model and hence, the statistical inference techniques introduced in Chapter 3 are readily applicable. By considering multilevel data and models as a separate unit, we expand the breadth of applications that linear mixed effects models enjoy.

5.1 Cross-sectional multilevel models

Educational systems are often described by structures in which the units of observation at one level are grouped within units at a higher level of structure. To illustrate, suppose that we are interested in assessing student performance based on an achievement test. Students are grouped in classes, classes are grouped in schools and schools are grouped into districts. At each level, there are variables that may affect responses from a student. For example, at the class level, education of the teacher may be important, at the school level, the school size may be important, and at the district level, funding may be important. Further, each level of grouping may be of scientific interest. Finally, there may be not only relationships among variables within each group but also across groups that should be considered.

The term multilevel is used for this nested data structure. In the above situation, we consider students to be the basic unit of observation; they are known as the “level-1” units of observation. The next level up is called “level-2” (classes in this example), and so forth.

We can imagine multilevel data being collected by a cluster sampling scheme. A random sample of districts is identified. For each district selected, a random sample of schools is chosen. From each school, a random sample of classes is taken and from each class selected, a random sample of students. Mechanisms other than random sampling may be used, and this will influence the model selected to represent the data. Multilevel models are specified through conditional relationships, where the relationships described at one level are conditional on (generally unobserved) random coefficients of upper levels. Because of this conditional modeling framework, multilevel data and models are also known as hierarchical.

5.1.1 Two-level models

To illustrate the important features of the model, initially consider only two levels. Suppose that we have a sample of \( n \) schools and, for the \( i \)th school, we randomly select \( n_i \) students (omitting class for the moment). For the \( j \)th student in the \( i \)th school, we assess the student’s performance on an achievement test, \( y_{ij} \), and information on the student’s socio-economic status, \( z_{ij} \), for example, the total family income. To assess achievement in terms of socio-economic status, we could begin with a simple model of the form

\[
y_{ij} = \beta_{0i} + \beta_{1i} z_{ij} + \epsilon_{ij}.
\]

Equation (5.1) describes a linear relation between socio-economic status and expected performance, although we allow the linear relationship to vary by school through the notation \( \beta_{0i} \).
and \( \beta_{1i} \) for school-specific intercepts and slopes. Equation (5.1) summarizes the “level-1” model that concerns student performance as the unit of observation.

If we have identified a set of schools that are of interest, then we may simply think of the quantities \( \{ \beta_{0i}, \beta_{1i} \} \) as fixed parameters of interest. However, in educational research, it is customary to consider these schools to be a sample from a larger population; the interest is in making statements about this larger population. Thinking of the schools as a random sample, we model \( \{ \beta_{0i}, \beta_{1i} \} \) as random quantities. A simple representation for these quantities is:

\[
\beta_{0i} = \beta_0 + a_{0i} \quad \text{and} \quad \beta_{1i} = \beta_1 + a_{1i} ,
\]

(5.2)

where \( a_{0i}, a_{1i} \) are mean zero random variables. Display (5.2) represents a relationship about the schools and summarizes the “level-2” model.

Displays (5.1) and (5.2) describe models at two levels. For estimation, we combine (5.1) and (5.2) to yield

\[
y_{ij} = (\beta_0 + a_{0i}) + (\beta_1 + a_{1i}) z_{ij} + \epsilon_{ij} \\
= a_{0i} + a_{1i} z_{ij} + \beta_0 + \beta_1 z_{ij} + \epsilon_{ij} .
\]

(5.3)

Equation (5.3) shows that the two-level model may be written as a single linear mixed effects model. Specifically, we define \( \mathbf{a}_i = (a_{0i}, a_{1i})' \), \( \mathbf{z}_{ij} = (1, z_{ij})' \), \( \mathbf{b} = (\beta_0, \beta_1)' \) and \( \mathbf{x}_{ij} = z_{ij} \), to write

\[
y_{ij} = \mathbf{z}_{ij}' \mathbf{a}_i + \mathbf{x}_{ij}' \mathbf{b} + \epsilon_{ij} ,
\]

similar to equation (3.5). Because we can write the combined multilevel model as a linear mixed effects model, we can use the Chapter 3 techniques to estimate the model parameters. Note that we are now using the subscript “\( \ell \)” to denote replications within a stratum such as a school. This is because we interpret the replication to have no time ordering; generally we will assume no correlation among replications (conditional on the subject). Section 5.2 will re-introduce the “\( t \)” subscript when we consider time-ordered repeated measurements.

One desirable aspect of the multilevel model formulation is that we may modify conditional relationships at each level of the model, depending on the research interests of the study. To illustrate, we may wish to understand how characteristics of the school affect student performance. For example, Raudenbush and Bryk (2002EP) discussed an example where \( x_i \) indicates whether the school was a Catholic based or a public school. A simple way to introduce this information is to modify the level-2 model in display (5.2) to

\[
\beta_{0i} = \beta_0 + \beta_{01} x_i + a_{0i} \quad \text{and} \quad \beta_{1i} = \beta_1 + \beta_{11} x_i + a_{1i} .
\]

(5.2a)

There are two level-2 regression models in display (5.2a); analysts find it intuitively appealing to specify regression relationships that capture additional model variability. Note, however, that for each model, the left-hand side quantities are not observed. To emphasize this, Raudenbush and Bryk (2002EP) call these models “intercepts-as-outcomes” and “slopes-as-outcomes.” In Section 5.3, we will learn how to predict these quantities.

Combining display (5.2a) with the level-1 model in equation (5.1), we have

\[
y_{ij} = (\beta_0 + \beta_{01} x_i + a_{0i}) + (\beta_1 + \beta_{11} x_i + a_{1i}) z_{ij} + \epsilon_{ij} \\
= a_{0i} + a_{1i} z_{ij} + \beta_0 + \beta_{01} x_i + \beta_1 z_{ij} + \beta_{11} x_i z_{ij} + \epsilon_{ij} .
\]

(5.4)

By defining \( \mathbf{a}_i = (a_{0i}, a_{1i})' \), \( \mathbf{z}_{ij} = (1, z_{ij})' \), \( \mathbf{b} = (\beta_0, \beta_{01}, \beta_1, \beta_{11})' \) and \( \mathbf{x}_{ij} = (1, x_i, z_{ij}, x_i z_{ij})' \), we may again express this multilevel model as a single linear mixed effects model.

The term \( \beta_{11} x_i z_{ij} \), interacting between the level-1 variable \( z_{ij} \) and the level-2 variable \( x_i \), is known as a cross-level interaction. For this example, suppose that we use \( x = 1 \) for Catholic schools and \( x = 0 \) for public schools. Then, \( \beta_{11} \) represents the difference between the marginal change in achievement scores, per unit of family income, between Catholic and public schools.
Many researchers (see, for example, Raudenbush and Bryk, 2002EP) argue that understanding cross-level interactions is a major motivation for analyzing multilevel data.

**Centering of variables**

It is customary in educational science to “center” explanatory variables in order to enhance the interpretability of model coefficients. To illustrate, consider the hierarchical models in (5.1), (5.2a) and (5.4). Using the “natural” metric for $z_{ij}$, we interpret $\beta_{0i}$ to be the mean (conditional on the $i$th subject) response when $z = 0$. In many applications such as where $z$ represents total income or test scores, a value of zero falls outside a meaningful range of values.

One possibility is to center level-1 explanatory variables about their overall mean and use $z_{ij} - \bar{z}$ as an explanatory variable in equation (5.1). In this case, we may interpret the intercept $\beta_{0i}$ to be the expected response for an individual with a score equal to the grand mean. This can be interpreted as an adjusted mean for the $i$th group.

Another possibility is to center each level-1 explanatory variable about its level-2 mean and use $z_{ij} - \bar{z}_i$ as an explanatory variable in equation (5.1). In this case, we may interpret the intercept $\beta_{0i}$ to be the expected response for an individual with a score equal to the mean of the $i$th group.

For longitudinal applications, you may wish to center the level-1 explanatory variables so that the intercept equals the expected random coefficient at a specific point in time, for example, at the start of a training program (see, for example, Kreft and deLeeuw, 1998).

**Extended two-level models**

To consider many explanatory variables, we extend equations (5.1) and (5.2). Consider a level-1 model of the form

$$y_{ij} = z_{1,ij}' \beta_i + x_{1,ij}' \beta_1 + \epsilon_{ij}. \quad (5.5)$$

Here, $z_{1,ij}$ and $x_{1,ij}$ represent the set of level-1 variables associated with varying (over level-1) and fixed coefficients, respectively. The level-2 model is of the form:

$$\beta_i = X_{2,i} \beta_2 + \alpha_i, \quad (5.6)$$

where $E \alpha_i = 0$. With this notation, the term $X_{2,i} \beta_2$ forms another set of effects with parameters to be estimated. Alternatively, we could write equation (5.5) without explicitly recognizing the fixed coefficients $\beta_1$ by including them in the random coefficients equation (5.6) but with zero variance. However, we prefer to recognize their presence explicitly because this helps in translating equations (5.5) and (5.6) into computer statistical routines for implementation. Combining equations (5.5) and (5.6) yields

$$y_{ij} = z_{1,ij}' (X_{2,i} \beta_2 + \alpha_i) + x_{1,ij}' \beta_1 + \epsilon_{ij} = z_{ij}' \alpha_i + x_{ij}' \beta + \epsilon_{ij}, \quad (5.7)$$

with the notation $x_{ij}' = (x_{1,ij}' z_{1,ij}' X_{2,i})$, $z_{ij} = z_{1,ij}$ and $\beta = (\beta_1' \beta_2')'$. Again, equation (5.7) expresses this multilevel model in our usual linear mixed effects model form.

It will be helpful to consider a number of special cases of equations (5.5)-(5.7). To begin, suppose that $\beta_i$ is a scalar and that $z_{1,ij} = 1$. Then, the model in equation (5.7) reduces to the error components model introduced in Section 3.1. Raudenbush and Bryk (2002EP) further discuss the special case, where equation (5.5) does not contain the fixed effects $x_{1,ij}' \beta_1$ portion. In this case, equation (5.7) reduces to

$$y_{ij} = \alpha_i + X_{2,i} \beta_2 + \epsilon_{ij},$$
that Raudenbush and Bryk refer to as the “means-as-outcomes” model. This model, with only level-2 explanatory variables available, can be used to predict the means, or expected values, of each group \(i\). We will study this prediction problem formally in Section 5.3.

Another special case of equations (5.5)-(5.7) is the random coefficients model. Here, we omit the level-1 fixed effects portion \(x_{i,j}' \beta_1\) and use the identity matrix for \(X_{2,i}\). Then, equation (5.7) reduces to

\[
y_{ij} = z_{ij}' (\beta_2 + \alpha_i) + \epsilon_{ij}.
\]

**Example**

As reported in Lee (2000EP), Lee and Smith (1997EP) studied 9,812 Grade 12 students in 1992 who attended 789 public, Catholic, and elite private high schools, drawn from a nationally representative sample from the National Education Longitudinal Study. The responses were achievement gains in reading and mathematics over four years of high school. The main variable of interest was a school level variable, size of the high school. Educational research had emphasized that larger schools enjoy economies of scale and are able to offer a broader curriculum whereas smaller schools offer more positive social environments, as well as a more homogenous curriculum. Lee and Smith sought to investigate the optimal school size. To control for additional student level effects, level-1 explanatory variables included gender, minority status, ability and socio-economic status. To control for additional school level characteristics, level-2 explanatory variables included school average minority concentration, school average socio-economic status and type of school (Catholic, public and elite private). Lee and Smith found that a middle school size, of approximately 600-900 students, produced the best achievement results.

**Motivation for multilevel models**

As we have seen, multilevel models allow analysts to assess the importance of cross-level effects. Specifically, the multilevel approach allows and/or forces researchers to hypothesize relationships at each level of analysis. Many different “units of analysis” within the same problem are possible, thus permitting modeling of complex systems. The ability to estimate cross-level effects is one advantage of multilevel modeling when compared to an alternate research strategy calling for the analysis of each level in isolation of the others.

As described in the introductory Chapter 1, multilevel models allow analysts to address problems of heterogeneity with samples of repeated measurements. Within the educational research literature, not accounting for heterogeneity from individuals is known as aggregation bias; see for example, Raudenbush and Bryk (2002EP). Even if the interest is in understanding level-2 relationships, we will get a better picture by incorporating a level-1 model of individual effects. Moreover, multilevel modeling allows us to predict quantities at both level-1 and level-2; Section 5.3 describes this prediction problem.

Second and higher levels of multilevel models also provide us with an opportunity to estimate the variance structure using a parsimonious, parametric structure. Improved estimation of the variance structure provides a better understanding of the entire model and will often result in improved precision of our usual regression coefficient estimators. Moreover, as discussed above, often these relationships at the second and higher levels are of theoretical interest and may represent the main focus of the study. However, technical difficulties arise when testing certain hypotheses about variance components. These difficulties, and solutions, are presented in Section 5.4.
5.1.2 Multiple level models

Extensions to more than two levels follow the same pattern as two level models. To be explicit, we give a three-level model based on an example from Raudenbush and Bryk (2002EP). Consider modeling a student’s achievement as the response \( y \). The level-1 model is

\[
y_{i,j,k} = z_{1,i,j,k}' \beta_{i,j} + x_{1,i,j,k}' \beta_1 + \epsilon_{1,i,j,k},
\]  

(5.8)

where there are \( i = 1, \ldots, n \) schools, \( j = 1, \ldots, J_i \) classrooms in the \( i \)th school and \( k = 1, \ldots, K_{ij} \) students in the \( j \)th classroom (within the \( i \)th school). The explanatory variables \( z_{1,i,j,k} \) and \( x_{1,i,j,k} \) may depend on the student (gender, family income and so on), classroom (teacher characteristics, classroom facilities and so on) or school (organization, structure, location and so on). The parameters that depend on either school \( i \) or classroom \( j \) appear as part of the \( \beta_{i,j} \) vector whereas parameters that are constant appear in the \( \beta_1 \) vector. This dependence is made explicit in the higher-level model formulation. Conditional on the classroom and school, the disturbance term \( \epsilon_{1,i,j,k} \) is mean zero and has a variance that is constant over all students, classrooms and schools.

The level-2 model describes the variability at the classroom level. Again, the parameters \( \gamma_{i,j} \) may be (i) varying but nonstochastic or (ii) stochastic. With this notation, we use a zero variance to model parameters that are varying but nonstochastic.

The level-3 model describes the variability at the school level. Again, the level-2 parameters \( \gamma_{i,j} \) may be varying but nonstochastic or stochastic. The level-3 model is of the form

\[
\gamma_i = X_{3,i} \beta_3 + \epsilon_{3,i}.
\]  

(5.10)

Again, the explanatory variables \( X_{3,i} \) may depend on the school. Conditional on the school, the disturbance term \( \epsilon_{3,i} \) is mean zero and has a variance that is constant over schools.

Putting equations (5.8)-(5.10) together, we have

\[
y_{i,j,k} = z_{1,i,j,k}' \beta_{i,j} + z_{1,i,j,k}' x_{2,i,j} \beta_2 + z_{1,i,j,k}' x_{1,i,j,k}' \beta_1 + z_{1,i,j,k}' x_{1,i,j,k}' \beta_1 + \epsilon_{1,i,j,k}
\]

\[
= x_{1,i,j,k}' \beta_1 + z_{1,i,j,k}' x_{2,i,j} \beta_2 + z_{1,i,j,k}' z_{2,i,j} x_{3,i} \beta_3 + z_{1,i,j,k}' z_{2,i,j} x_{3,i} \beta_3 + z_{1,i,j,k}' z_{2,i,j} \gamma_i + z_{1,i,j,k}' \epsilon_{2,i,j} + \epsilon_{1,i,j,k}
\]  

(5.11)

where \( x_{1,i,j,k}' = (x_{1,i,j,k}' z_{1,i,j,k}' z_{2,i,j}' x_{3,i}) \) and \( \beta = (\beta_1 \beta_2 \beta_3)' \). We have already specified the usual assumption of homoscedasticity for each random quantity \( \epsilon_{1,i,j,k}, \epsilon_{2,i,j} \) and \( \epsilon_{3,i} \). Moreover, it is customary to assume that these quantities are uncorrelated with one another. Our main point is that, as with the two-level model, equation (5.11) expresses the three-level model as a linear mixed effects model. (Converting the model in equation (5.11) into the linear mixed effects model in equation (3.5) is a matter of defining vector expressions carefully. Section 5.3 provides further details.) Thus, parameter estimation is a direct consequence of our Chapter 3 results. Many variations of the basic assumptions that we have described are possible. In Section 5.2 on longitudinal multilevel models, we will give a more detailed description of an example of a three-level model. Appendix 5A extends the discussion to higher order multilevel models.
For applications, several statistical software packages exist (such as HLM, MLwiN, and MIXREG) that allow analysts to fit multilevel models without combining the several equations into a single expression such as equation (5.11). However, these specialized packages may not have all of the features that the analyst wishes to display in his or her analysis. As pointed out by Singer (1998EP), an alternative, or supplementary, approach is to use a general purpose mixed linear effects package (such as SAS PROC MIXED) and rely directly on the fundamental mixed linear model theory.

5.1.3 Multilevel modeling in other fields

The field of educational research has been an area of active development of cross-sectional multilevel modeling although it by no means has a corner on the market. This subsection describes examples where these models have been used in other fields of study.

One type of study that is popular in economics is data based on a matched pairs sample. For example, we might select a set of families for level-2 sample and, for each family, observe the behavior of siblings (or twins). The idea underlying this design is that by observing more than one family member we will be able to control for unobserved family characteristics. See Wooldridge (2002E) and Exercise 3.10 for further discussion of this design.

In insurance and actuarial science, it is possible to model claims distributions using a hierarchical framework. Typically, the level-2 unit of analysis is based on an insurance customer, and explanatory variables may include characteristics of the customer. The level-1 model uses claims amounts as the response (typically over time) and typical time-varying explanatory variables include time trends. For example, Klugman (1992O) gives a Bayesian perspective of this problem. For a frequentist perspective, see Frees, Young and Luo (1999O).

5.2 Longitudinal multilevel models

This section shows how to use the conditional modeling framework to represent longitudinal (time-ordered) data. The key change in the modeling set-up is that we now will typically consider the individual as the level-2 unit of analysis and observations at different time points as the level-1 units. The goal is now also substantially different; typically, in longitudinal studies the assessment of change is the key research interest. As with Section 5.1, we begin with the two-level model and then discuss general multilevel extensions.

5.2.1 Two-level models

Following the notation established in Section 5.1, we consider level-1 models of the form

\[ y_{it} = z_{1, i} \beta_i + x_{1, i} \beta_1 + \epsilon_{it}. \]  

(5.12)

This is a model of \( t = 1, \ldots, T_i \) responses over time for the \( i \)th individual. The unit of analysis for the level-1 model is an observation at a point in time, not the individual as in Section 5.1. Thus, we use the subscript \( \cdot t \) as an index for time. Most other aspects of the model are as in Section 5.1.1; \( z_{1, i} \) and \( x_{1, i} \) represent sets of level-1 explanatory variables. The associated parameters that may depend on the \( i \)th individual appear as part of the \( \beta_i \) vector whereas parameters that are constant appear in the \( \beta_1 \) vector. Conditional on the subject, the disturbance term \( \epsilon_{it} \) is mean zero random variable that is uncorrelated with \( \beta_i \).

An important feature of the longitudinal multilevel model that distinguishes it from its cross-sectional counterpart is that time generally enters the level-1 specification. There are a number of ways that this can happen. One way is to let one or more of the explanatory variables be a function of time. This is the approach historically taken in growth curve modeling, described below. Another approach is to let one of the explanatory variables be a lagged response variable.
Chapter 5. Multilevel Models / 5-7

This approach is particularly prevalent in economics and will be further explored in Chapter 6. Yet another approach is to model the serial correlation through the variance covariance-matrix of the vector of disturbance \( \epsilon_i = (\epsilon_{i1} \ldots \epsilon_{iT})' \). Specifically, in Sections 2.5.1 and 3.3.1 we developed the notation \( \text{Var} \{\epsilon_i \} = R_i \) to represent the serial covariance structure. This approach is widely adopted in biostatistics and educational research and will be further developed here.

Like the cross-sectional model, the level-2 model can be represented as \( \beta_i = X_{2i} \beta_2 + \alpha_i \); see equation (5.6). Now, however, we interpret the unobserved \( \beta_i \) to be the random coefficients associated with the \( i \)th individual. Thus, although the mathematical representation is similar to the cross-sectional setting, our interpretations of individual components of the model are quite different. Yet, as with equation (5.7), we may still combine level-1 and level-2 models to get

\[
y_{it} = z_{1it}' (X_{2it} \beta_2 + \alpha_i) + x_{1it}' \beta_1 + \epsilon_{it}
\]

using the notation \( x_{it}' = (x_{1it}' \ z_{1it}' X_{2it}) \), \( z_{it} = z_{1it} \) and \( \beta = (\beta_1' \beta_2')' \). This is the linear mixed effects model introduced in Section 3.3.1.

**Growth curve models**

To develop intuition, we now consider *growth curve models*, models that have a long history of applications. The idea behind growth curve models is that we seek to monitor the natural development or aging of an individual. This development is typically monitored without intervention and the goal is to assess differences among groups. In growth curve modeling, one uses a polynomial function of age or time to track growth. Because growth curve data may reflect observations from a development process, it is intuitively appealing to think of the expected response as a function of time. Parameters of the function vary by individual, so that one can summarize an individual’s growth through the parameters. To illustrate, we now consider a classic example.

**Example - Dental Data**

This example is originally due to Potthoff and Roy (1964B); see also Rao (1987B). Here, \( y \) is the distance, measured in millimeters, from the center of the pituitary to the pteryomaxillary fissure. Measurements were taken on 11 girls and 16 boys at ages 8, 10, 12, and 14. The interest is in the relation between the distance and age, specifically, in how the distance grows with age and whether there is a difference between males and females.

Table 5.1 shows the data and Figure 5.1 gives a graphical impression of the growth over time. From Figure 5.1, we can see that the measurement length grows as each child ages, although it is difficult to detect differences between boys and girls. In Figure 5.1, we use open circular plotting symbols for girls and opaque plotting symbols for boys. Figure 5.1 does show that the ninth boy has an unusual growth pattern; this pattern can also be seen in Table 5.1.
Table 5.1 Dental measurements of 11 girls and 16 boys.
Measurements are in millimeters.

<table>
<thead>
<tr>
<th>Number</th>
<th>Girls</th>
<th>Age in years</th>
<th>Boys</th>
<th>Age in years</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>8 10 12 14</td>
<td></td>
<td>8 10 12 14</td>
</tr>
<tr>
<td>1</td>
<td>21</td>
<td>20 21.5 22.5</td>
<td>26</td>
<td>25 29 31</td>
</tr>
<tr>
<td>2</td>
<td>21</td>
<td>21.5 24 25</td>
<td>21.5</td>
<td>22.5 23 26.5</td>
</tr>
<tr>
<td>3</td>
<td>22.5</td>
<td>24.5 25.5 26</td>
<td>26.5</td>
<td>22.5 24 27.5</td>
</tr>
<tr>
<td>4</td>
<td>23.5</td>
<td>24.5 25 26.5</td>
<td>25.5</td>
<td>27.5 26.5 27</td>
</tr>
<tr>
<td>5</td>
<td>21.5</td>
<td>23 22.5 23.5</td>
<td>20</td>
<td>23.5 22.5 26</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>21 21 22.5</td>
<td>24.5</td>
<td>25.5 27 28.5</td>
</tr>
<tr>
<td>7</td>
<td>21.5</td>
<td>22.5 23 25</td>
<td>22</td>
<td>22 24.5 26.5</td>
</tr>
<tr>
<td>8</td>
<td>23</td>
<td>23 23.5 24</td>
<td>24</td>
<td>21.5 24.5 25.5</td>
</tr>
<tr>
<td>9</td>
<td>20</td>
<td>21 22 21.5</td>
<td>23</td>
<td>20.5 31 26</td>
</tr>
<tr>
<td>10</td>
<td>16.5</td>
<td>19 19 19.5</td>
<td>27.5</td>
<td>28 31 31.5</td>
</tr>
<tr>
<td>11</td>
<td>24.5</td>
<td>25 28 28</td>
<td>23</td>
<td>23 23.5 25</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td>21.5 23.5 24</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td>17 24.5 26</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td>22.5 25.5 26</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td>23 24.5 26</td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td>22 21.5 23.5</td>
</tr>
</tbody>
</table>


Figure 5.1 Multiple Time Series Plot of Dental Measurements.
Open circles represent girls; opaque circles represent boys.
A level-1 model is

\[ y_{it} = \beta_{0i} + \beta_{1i} z_{1it} + \epsilon_{it}, \]

where \( z_{1it} \) is the age of the child \( i \) on occasion \( t \). This model relates the dental measurement to the age of the child, with parameters that are specific to the child. Thus, we may interpret the quantity \( \beta_{1i} \) to be the growth rate for the \( i \)th child. A level-2 model is

\[ \beta_{0i} = \beta_{00} + \beta_{01} \text{GENDER}_i + \alpha_{0i} \quad \text{and} \quad \beta_{1i} = \beta_{10} + \beta_{11} \text{GENDER}_i + \alpha_{1i}. \]

Here, \( \beta_{00}, \beta_{01}, \beta_{10} \) and \( \beta_{11} \) are fixed parameters to be estimated. Suppose that we use a binary variable for gender, say, coding the GENDER variable 1 for females and 0 for males. Then, we may interpret \( \beta_{10} \) to be the expected male growth rate and \( \beta_{11} \) to be the difference in growth rates between females and males.

Table 5.2 shows the parameter estimates for this model. Here, we see that the coefficient associated with linear growth is statistically significant, over all models. Moreover, the rate of increase for girls is lower than boys. The estimated covariance between \( \alpha_{0i} \) and \( \alpha_{1i} \) (which is also the estimated covariance between \( \beta_{0i} \) and \( \beta_{1i} \)) turns out to be negative. One interpretation of the negative covariance between initial status and growth rate is that subjects who start at a low level tend to grow more quickly than those who start at higher levels, and vice versa.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Error Components Model</th>
<th>Growth Curve Model</th>
<th>Growth Curve Model deleting the 9th boy</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_{00} )</td>
<td>16.341</td>
<td>16.341</td>
<td>16.470</td>
</tr>
<tr>
<td>Age (( \beta_{01} ))</td>
<td>0.784</td>
<td>0.784</td>
<td>0.772</td>
</tr>
<tr>
<td>GENDER (( \beta_{01} ))</td>
<td>1.032</td>
<td>1.032</td>
<td>0.903</td>
</tr>
<tr>
<td>AGE*GENDER (( \beta_{11} ))</td>
<td>-0.305</td>
<td>-0.305</td>
<td>-0.292</td>
</tr>
<tr>
<td>Var ( \epsilon_{it} )</td>
<td>1.922</td>
<td>1.716</td>
<td>0.971</td>
</tr>
<tr>
<td>Var ( \alpha_{0i} )</td>
<td>3.299</td>
<td>5.786</td>
<td>11.005</td>
</tr>
<tr>
<td>Var ( \alpha_{1i} )</td>
<td>0.033</td>
<td>0.073</td>
<td></td>
</tr>
<tr>
<td>Cov (( \alpha_{0i}, \alpha_{1i} ))</td>
<td>-0.290</td>
<td>-0.734</td>
<td></td>
</tr>
<tr>
<td>-2 Log Likelihood</td>
<td>433.8</td>
<td>432.6</td>
<td>388.5</td>
</tr>
<tr>
<td>AIC</td>
<td>445.8</td>
<td>448.6</td>
<td>404.5</td>
</tr>
</tbody>
</table>

For comparison purposes, Table 5.2 shows the parameter estimates with the 9th boy deleted. The effects of this subject deletion on the parameter estimates are small. Table 5.2 also shows parameter estimates of the error components model. This model employs the same level-1 model but with level-2 models

\[ \beta_{0i} = \beta_{00} + \beta_{01} \text{GENDER}_i + \alpha_{0i} \quad \text{and} \quad \beta_{1i} = \beta_{10} + \beta_{11} \text{GENDER}_i. \]

With parameter estimates calculated using the full data set, there again is little change in the parameter estimates. Because the results appear to be robust to both unusual subjects and model selection, we have greater confidence in our interpretations.
5.2.2 Multiple level models

Longitudinal versions of multiple level models follow the same notation as the cross-sectional models in Section 5.1.2 except that the level-1 replications are over time. To illustrate, we consider a 3-level model in the context of a social work application by Guo and Hussey (1999EP).

Guo and Hussey examined subjective assessments of children’s behavior made by multiple raters at two or more time points. That is, the level-1 repeated measurements are over time $t$, where the assessment was made by rater $j$ on child $i$. Raters assessed $n = 144$ seriously emotionally disturbed children receiving services through a large child mental health treatment agency located in Cleveland, Ohio. For this study, the assessment is the response of interest $y$; this response is the Deveroux Scale of Mental Disorders, a score made up of 111 items. Ratings were taken over a two-year period by parents and teachers; at each time point, assessments may be made either by the parent, teacher or both. The time of the assessment was recorded as $\text{TIME}_{i,j,t}$, measured in days since the inception of the study. The variable $\text{PROGRAM}_{i,j,t}$ was recorded as a 1 if the child was in program residence at the time of the assessment and 0 if the child was in day treatment or day treatment combined with treatment foster care. The variable $\text{RATER}_{i,j}$ was recorded as a 1 if rater was a teacher and 0 if the rater was a caretaker.

Analogous to equation (5.8), the level-1 model is

$$y_{i,j,t} = \mathbf{z}_{1,i,j,t}^\top \beta_{1,i,j} + \mathbf{x}_{1,i,j,t}^\top \beta_{1} + \epsilon_{1,i,j,t}, \quad (5.14)$$

where there are $i = 1, \ldots, n$ children, $j = 1, \ldots, J_{i}$ raters and $t = 1, \ldots, T_{i,j}$ evaluations. Specifically, Guo and Hussey (1999EP) used $\mathbf{x}_{1,i,j,t} = \text{PROGRAM}_{i,j,t}$ and $\mathbf{z}_{1,i,j,t} = (1 \times \text{TIME}_{i,j,t})^\top$. Thus, their level-1 model can be written as

$$y_{i,j,t} = \beta_{0,i,j} + \beta_{1,i,j} \times \text{TIME}_{i,j,t} + \beta_{1} \times \text{PROGRAM}_{i,j,t} + \epsilon_{1,i,j,t}.$$ 

The variables associated with the intercept and the coefficient for time may vary over child and rater whereas the program coefficient is constant over all observations.

The level-2 model is the same as equation (5.9)

$$\beta_{i,j} = \mathbf{Z}_{2,i,j}^\top \gamma_{i} + \mathbf{X}_{2,i,j}^\top \beta_{2} + \epsilon_{2,i,j},$$

where there are $i = 1, \ldots, n$ children and $j = 1, \ldots, J_{i}$ raters. The level-2 model of Guo and Hussey can be written as

$$\beta_{0,i,j} = \beta_{0,i,0} + \beta_{0,1,i} \times \text{RATER}_{i,j} + \epsilon_{2,i,j}$$

and

$$\beta_{1,i,j} = \beta_{2,0} + \beta_{2,1} \times \text{RATER}_{i,j}.$$ 

Again, we leave it as an exercise for the reader to show how this formulation is a special case of equation (5.9).

The level-3 model is the same as equation (5.10)

$$\gamma_{i} = \mathbf{X}_{3,i}^\top \beta_{3} + \epsilon_{3,i}.$$ 

To illustrate, the level-3 model of Guo and Hussey can be written as

$$\beta_{0,i,0} = \beta_{0,0,0} + \beta_{0,1,0} \times \text{GENDER}_{i} + \epsilon_{3,i}.$$ 

where GENDER$_{i}$ is a binary variable indicating the gender of the child.

As with the cross-sectional models in Section 5.1.2, one combines the three levels to form a single equation representation, as in equation (5.14). The hierarchical framework allows analysts to develop hypotheses that are interesting to test. The combined model allows for simultaneous, over all levels, estimation of parameters that is more efficient than estimating each level in isolation of the others.
5.3 Prediction

In Chapter 4, we distinguished between the concepts of estimating model parameters as compared to predicting random variables. In multilevel models, the dependent variables at second and higher levels are unobserved random coefficients. Because it is often desirable to understand their behavior, we wish to predict these random coefficients. To illustrate, if the unit of analysis at the second level is a school, we may wish to use predictions of second level coefficients to rank schools. It may also be of interest to use predictions of second (or higher) level coefficients for prediction in a first level model. To illustrate, if we are studying a child’s development over time, we may wish to make predictions about the future status of a child’s development.

This subsection shows how to use the best linear unbiased predictors (BLUPs) developed in Chapter 4 for these prediction problems. Best linear unbiased predictors, by definition, have the smallest variance among all unbiased predictors. In Chapter 4, we showed that these predictors can also be interpreted as empirical Bayes estimators. Moreover, they often have desirable interpretations as shrinkage “estimators.” Because we have expressed multilevel models in terms of linear mixed effects models, we will not need to develop new theory but will be able to rely directly on the Chapter 4 results.

Two-level models

We begin our prediction discussion with the two-level model, introduced in equations (5.5)-(5.7). To make the multilevel model notation consistent with Chapters 3 and 4, use \( D = \text{Var} \alpha_i \) and \( R_i = \text{Var} \epsilon_i \), where \( \epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{iT_i})' \). Suppose that we wish to predict \( \beta_i \). Using the results in Section 4.3.2, it is easy to check that the best linear unbiased predictor (BLUP) of \( \beta_i \) is

\[
\hat{\beta}_i,\text{BLUP} = \hat{\alpha}_i,\text{BLUP} + X_{2i} \hat{\beta}_{2,\text{GLS}},
\]

where \( \hat{\beta}_{2,\text{GLS}} \) is the generalized least squares estimator of \( \beta_2 \) and, from equation (4.11),

\[
\hat{\alpha}_i,\text{BLUP} = D Z_i' V_i^{-1} (y_i - X_i \hat{\beta}_{2,\text{GLS}}).
\]

Recall that \( z_{it} = z_{i,\alpha} \), so that \( Z_i = (z_{i1}, z_{i2}, \ldots, z_{i,T_i})' \). Further, \( \hat{\beta}_{GLS} = (\hat{\beta}_{1,\text{GLS}}, \hat{\beta}_{2,\text{GLS}})' \), \( V_i = R_i + Z_i D Z_i' \) and \( X_i = (x_{i1}, x_{i2}, \ldots, x_{iT_i})' \), where \( x_{it}' = (x_{it}, z_{it}' X_{2i}) \). Thus, it is easy to compute these predictors.

Chapter 4 discussed interpretation in some special cases, the error components and the random coefficients models. Suppose that we have the error components model, so that \( z_{ij} = z_{1,ij} = 1 \) and \( R_i \) is a scalar times the identity matrix. Further suppose that there are no level-1 explanatory variables. Then, one can check that the BLUP of the conditional mean of the level-1 response, \( E (y_{it} | \alpha_i) = \alpha_i + X_{2i} \hat{\beta}_2 \), is

\[
\hat{\alpha}_i,\text{BLUP} + X_{2j} \hat{\beta}_{2,\text{GLS}} = \zeta_i (\bar{y}_{ij} - X_{2j} \hat{\beta}_{2,\text{GLS}}) + X_{2j} \hat{\beta}_{2,\text{GLS}} = \zeta_i \bar{y}_{ij} + (1- \zeta_i) X_{2j} \hat{\beta}_{2,\text{GLS}},
\]

where \( \zeta_i = \frac{T_i}{T_i + (\text{Var} \epsilon_i)/(\text{Var} \alpha)} \). Thus, the predictor is a weighted average of the level-2 ith unit’s average, \( \bar{y}_{ij} \), and the regression estimator which is an estimator derived from all level-2 units. As noted in Section 5.1.1, Raudenbush and Bryk (2002EP) refer to this as the “means-as-outcomes” model.

As described in Section 4.3.4, one can also use the BLUP technology to predict the future development of a level-1 response. From equation (4.14), we have that the forecast \( L \) lead times in the future of \( y_{iT_j} \) is

\[
\hat{y}_{i,T_j+L} = z_{i,j,T_j+L}' \hat{\beta}_i,\text{BLUP} + x_{i,j,T_j+L}' \hat{\beta}_{1,\text{GLS}} + \text{Cov}(\epsilon_{i,j,T_j+L}, \epsilon_i) R_i^{-1} \epsilon_i,\text{BLUP},
\]
where $e_{i,\text{BLUP}}$ is the vector of BLUP residuals, given in equation (4.13a). As we saw in Section 4.3.4, in the case where the disturbances follow an autoregressive model of order 1 (AR(1)) with parameter $\rho$, we have

$$\hat{y}_{i,T_i+L} = z'_{i,T_i+L} b_{i,\text{BLUP}} + x'_{i,T_i+L} b_{i,\text{GLS}} + \rho^L e_{i,T_i,\text{BLUP}}.$$  

To illustrate, consider the Section 5.1.2 Dental example. Here, there is no serial correlation (so that $R$ is a scalar times the identity matrix), no level-1 fixed parameters and $T_i = 4$ observations for all children. Thus, the $L$ step forecast for the $i$th child is

$$\hat{y}_{i,4+L} = b_{0,i,\text{BLUP}} + b_{1,i,\text{BLUP}} z_{i,4+L},$$

where $z_{i,4+L}$ is the age of the child at time $4+L$.

**Multiple level models**

For three and higher level models, the approach is the same as with two-level models although it becomes more difficult to interpret the results. Nonetheless, for applied work, the idea is straightforward.

**Procedure for forecasting future level-1 responses**

1. Hypothesize a model at each level.
2. Combine all level models into a single model.
3. Estimate the parameters of the single model, using generalized least squares and variance components estimators, as described in Sections 3.4 and 3.5, respectively.
4. Determine best linear unbiased predictors of each unobserved random coefficient for levels two and higher, as described in Section 4.3.
5. Use the parameter estimators and random coefficient predictors to form forecasts of future level-1 responses.

To illustrate, let’s see how this procedure works for the three-level longitudinal data model.

Step 1. We will use the level-1 model described in equation (5.14), together with the level-2 and level-3 models in equations (5.9) and (5.10), respectively. For the level-1 model, let $R_{ij} = \text{Var} e_{i,j}$, where $e_{i,j} = (e_{i,j} \ldots e_{i,j,T})'$.

Step 2. The combined model is equation (5.11), except that now we use a “t” subscript for time in lieu of the “k” subscript. Assuming the level-1, 2 and 3 random quantities are uncorrelated with one another, we define

$$\text{Var} a_{i,j} = \text{Var} \left( e_{2,i,j} \right) = \begin{pmatrix} \text{Var} e_{2,j} & 0 \\ 0 & \text{Var} e_{3,j} \end{pmatrix} = \begin{pmatrix} D_2 & 0 \\ 0 & D_3 \end{pmatrix} = D_v$$

and

$$\text{Cov} (a_{i,j}, a_{i,k}) = \begin{pmatrix} \text{Cov} (e_{2,i,j}, e_{2,i,k}) & \text{Cov} (e_{2,i,j}, e_{3,i,k}) \\ \text{Cov} (e_{3,i,j}, e_{2,i,k}) & \text{Var} e_{3,i,k} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & D_3 \end{pmatrix} = D_c.$$
Stacking vectors, we write $\mathbf{y}_{i,j} = (y_{i,j,1} \ldots y_{i,j,T_{ij}})'$, $\mathbf{y}_i = (y_{i,1} \ldots y_{i,T_{ij}})'$, $\mathbf{e}_i = (e_{i,1} \ldots e_{i,T_{ij}})'$ and $\mathbf{a}_i = (a_{i,1} \ldots a_{i,T_{ij}})'$. Stacking matrices, we have $\mathbf{X}_{i,j} = (x_{i,j,1} \ldots x_{i,j,T_{ij}})'$, $\mathbf{Z}_{i,j} = (z_{i,j,1} \ldots z_{i,j,T_{ij}})'$, $\mathbf{X}_i = (x_{i,1} \ldots x_{i,T_{ij}})'$ and $\mathbf{Z}_i = (z_{i,1} \ldots z_{i,T_{ij}})'$.

With this notation, we may write equation (5.11) in a linear mixed effects model form as $\mathbf{y}_i = \mathbf{Z}_i \mathbf{a}_i + \mathbf{X}_i \mathbf{b} + \mathbf{e}_i$. Note the form of $\mathbf{R}_i = \text{Var} \mathbf{e}_i = \text{blockdiagonal}(\mathbf{R}_{i,1}, \ldots, \mathbf{R}_{i,J_i})$ and

$$
\mathbf{D} = \text{Var} \mathbf{a}_i = \begin{pmatrix}
\mathbf{D}_V & \mathbf{D}_C & \cdots & \mathbf{D}_C \\
\mathbf{D}_C & \mathbf{D}_V & \cdots & \mathbf{D}_C \\
\vdots & \vdots & \ddots & \vdots \\
\mathbf{D}_C & \mathbf{D}_C & \cdots & \mathbf{D}_V 
\end{pmatrix}.
$$

Step 3. Having coded the explanatory variables and the form of the variance matrices $\mathbf{D}$ and $\mathbf{R}_i$, parameter estimates follow directly from the Sections 3.4 and 3.5 results.

Step 4. The BLUP predictors are formed beginning with predictors for $\mathbf{a}_i$ of the form $\mathbf{a}_{i,\text{BLUP}} = \mathbf{D}^{-1} \mathbf{Z}_i' \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i \mathbf{b}_{\text{GLS}})$. This yields the BLUPs for $\mathbf{a}_{i,j} = (e_{2,i,j} \mathbf{e}_{3,i,j})'$, say $\mathbf{a}_{i,j,\text{BLUP}} = (e_{2,i,j,\text{BLUP}}' \mathbf{e}_{3,i,j,\text{BLUP}}')'$. These BLUPs allow us to predict the second and higher level random coefficients through the relations

$$
\mathbf{g}_{i,\text{BLUP}} = \mathbf{X}_{3,i} \mathbf{b}_{3,\text{GLS}} + \mathbf{e}_{3,i,\text{BLUP}}
$$

and

$$
\mathbf{b}_{i,j,\text{BLUP}} = \mathbf{Z}_{2,i,j} \mathbf{g}_{i,\text{BLUP}} + \mathbf{X}_{2,i,j} \mathbf{b}_{2,\text{GLS}} + \mathbf{e}_{2,i,j,\text{BLUP}},
$$

corresponding to equations (5.10) and (5.9), respectively.

Step 5. If desired, we may forecast future level-1 responses. From equation (4.14), for an $L$-step forecast, we have

$$
\hat{y}_{i,j,T_{ij}+L} = \mathbf{Z}'_{i,j,T_{ij}+L} \mathbf{b}_{i,j,\text{BLUP}} + \mathbf{x}'_{i,j,T_{ij}+L} \mathbf{b}_{i,1,\text{GLS}} + \text{Cov}(e_{i,j,T_{ij}+L}, e_{i,j,L}) \mathbf{R}_y^{-1} e_{i,j,\text{BLUP}},
$$

For AR(1) level-1 disturbances, this simplifies to

$$
\hat{y}_{i,j,T_{ij}+L} = \mathbf{z}'_{i,j,T_{ij}+L} \mathbf{b}_{i,j,\text{BLUP}} + \mathbf{x}'_{i,j,T_{ij}+L} \mathbf{b}_{i,1,\text{GLS}} + \rho L e_{i,j,T_{ij},\text{BLUP}}.
$$

## 5.4 Testing variance components

Multilevel models implicitly provide a representation for the variance as a function of explanatory variables. To illustrate, consider the cross-sectional two-level model summarized in equations (5.5)-(5.7). With equation (5.7), we have

$$
\text{Var} y_{ij} = \mathbf{z}_{ij}' \text{Var} \mathbf{a}_i \mathbf{z}_{ij} + \text{Var} \mathbf{e}_{ij},
$$

and

$$
\text{Cov} (y_{ij}, y_{ik}) = \mathbf{z}_{ij}' \text{Var} \mathbf{a}_i \mathbf{z}_{ik}.
$$

Thus, even if the random quantities $\mathbf{a}_i$ and $\mathbf{e}_{ij}$ are homoscedastic, the variance is a function of the explanatory variables $\mathbf{z}_{ij}$. Particularly in education and psychology, researchers wish to test theories by examining hypotheses concerning these variance functions.
Unfortunately, the usual likelihood ratio testing procedure is not valid for testing many variance components of interest. In particular, the concern is for testing parameters where the null hypothesis is on the boundary of possible values. As a general rule, the standard hypothesis testing procedures favors the simpler null hypothesis more often than it should.

To illustrate the difficulties with boundary problems, let’s consider the classic example of i.i.d. random variables $y_1, \ldots, y_n$ where each random variable is distributed normally with known mean zero and variance $\sigma^2$. Suppose that we wish to test the null hypothesis $H_0: \sigma^2 = \sigma_0^2$, where $\sigma_0^2$ is a known positive constant. It is easy to check that the maximum likelihood estimator of $\sigma^2$ is $n^{-1} \sum_{i=1}^{n} y_i^2$. As we have seen, a standard method of testing hypotheses is the likelihood ratio test procedure (described in more detail in Appendix A.7). Here, one computes the likelihood ratio test statistic, which is twice the difference between the unconstrained maximum log-likelihood and the maximum log-likelihood under the null hypothesis, and compares this statistic to a chi-square distribution with one degree of freedom. Unfortunately, this procedure is not available when $\sigma_0^2 = 0$ because the log-likelihoods are not well defined. Because $\sigma_0^2 = 0$ is on the boundary of the parameter space $[0, \infty)$, the regularity conditions of our usual test procedures are not valid.

However, $H_0: \sigma^2 = 0$ is still a testable hypothesis; a simple test is to reject $H_0$ if the maximum likelihood estimator exceeds zero. This procedure will always reject the null hypothesis when $\sigma^2 > 0$ and accept when $\sigma^2 = 0$. Thus, this test procedure has power 1 versus all alternatives and a significance level of zero, a very good test!

For an example closer to longitudinal data models, consider the Section 3.1 error components model with variance parameters $\sigma^2$ and $\sigma_\alpha^2$. In the Exercise 5.4, we outline the proof to establish that the likelihood ratio test statistic for assessing $H_0: \sigma_\alpha^2 = 0$ is $\frac{1}{2} \chi^2_1$, where $\chi^2_1$ is a chi-square random variable with 1 degree of freedom. In the usual likelihood ratio procedure for testing one variable, the likelihood ratio test statistic has a $\chi^2_1$ distribution under the null hypothesis. This means that using nominal values, we will accept the null hypothesis more often than we should; thus, we will sometimes use a simpler model than suggested by the data.

The critical point of this exercise is that we define maximum likelihood estimators to be non-negative, arguing that a negative estimator of variance components is not valid. Thus, the difficulty is that the usual regularity conditions (see, for example, Serfling, 1980G) require that the hypotheses that we test lie on the interior of a parameter space. For most variances, the parameter space is $[0, \infty)$. By testing that the variance equals zero, we are on the boundary and the usual asymptotic results are not valid. This does not mean that tests for all variance components are not valid. For example, for testing most correlations and autocorrelations, the parameter space is $[-1,1]$. Thus, for testing correlations (and covariances) equal to zero, we are in the interior of the parameter space and so the usual test procedures are valid.

In contrast, in Exercise 5.3, we allow negative variance estimators. In this case, by following the outline of the proof, you will see that the usual likelihood ratio test statistic for assessing $H_0: \sigma_\alpha^2 = 0$ is $\chi^2_1$, the customary distribution. Thus, it is important to know the constraints underlying the software package that you are using.

A complete theory for testing variance components has yet to be developed. When only one variance parameter needs to be assessed for equality to zero, results similar to the error components model discussed above have been worked out. For example, Balagi and Li (1990E) developed a test for a second (independent) error component representing time; this model will be described in Chapter 8. More generally, checking for the presence of an additional random effect in the model implicitly means checking that not only the variance, but also the covariances, are
equal to zero. For example, for the linear mixed effects model with a \( q \times 1 \) vector of variance components \( \alpha_i \), we might wish to assess the null hypothesis

\[
H_0 : \mathbf{D} = \text{Var} \alpha_i = \begin{pmatrix} \text{Var}(\alpha_{i,1}, \ldots, \alpha_{i,q-1}) & 0 \\ 0' & 0 \end{pmatrix}.
\]

In this case, based on the work of Self and Liang (1987S), Stram and Lee (1994S) showed that the usual likelihood ratio test statistic has asymptotic distribution \( \frac{1}{2} \chi^2_{q-1} + \frac{1}{2} \chi^2_q \), where \( \chi^2_{q-1} \) and \( \chi^2_q \) are independent chi-square random variables with \( q-1 \) and \( q \) degrees of freedom, respectively. The usual procedure for testing means comparing the likelihood ratio test statistic to \( \chi^2_q \) because we are testing a variance parameter and \( q-1 \) covariance parameters. Thus, if one rejects using the usual procedure, one will reject using the mixture distribution corresponding to \( \frac{1}{2} \chi^2_{q-1} + \frac{1}{2} \chi^2_q \). Put another way, the actual \( p \)-value (computed using the mixture distribution) is less than the nominal \( p \)-value (computed using the standard distribution). Based on this, we see that the standard hypothesis testing procedures favors the simpler null hypothesis more often than it should.

No general rules for checking for the presence of several additional random effects are available although simulation methods are always possible. The important point is that analysts should not quickly quote \( p \)-values associated with testing variance components without carefully considering the model and estimator.

**Further reading**

Appendix 5A – High Order Multilevel Models

Despite their widespread application, standard treatments that introduce the multilevel model use at most only three levels, anticipating that users will be able to intuit patterns (and hopefully equation structures) to higher levels. In contrast, this appendix describes a high order multilevel model using “k” levels.

To motivate the extensions, begin with the three-level model in Section 5.1.2. Extending equation (5.8), the level-1 model is expressed as

\[ y_{i_1,i_2,...,i_k} = Z_{i_1,...,i_k}^{(l)} \beta_{i_1,...,i_k}^{(l)} + X_{i_1,...,i_k}^{(l)} \beta_1 + \varepsilon_{i_1,...,i_k}^{(l)}. \]

Here, we might use \( i_k \) as a time index, \( i_{k-1} \) is a student index, \( i_{k-2} \) is classroom index, and so on. We denote the observation set by \( \{i(k) = (i_1, i_2, ..., i_k) : y_{i_1,i_2,...,i_k} \text{ is observed}\} \). More generally, define

\[ i(k-s) = (i_1, i_2, ..., i_{k-s}) : y_{i_1,i_2,...,i_{k-s}} \text{ is observed for some } j_{k+1}, ..., j_k \}, \]

for \( s = 0, 1, ..., k-1 \). We will let \( i(k) = \{i_1, i_2, ..., i_k\} \) be a typical element of \( i(k) \) and use \( i(k-s) = \{i_1, ..., i_{k-s}\} \) for the corresponding element of \( i(k-s) \).

With this additional notation, we are now in a position to provide a recursive specification of high order multilevel models.

Recursive specification of high order multilevel models

1. The level-1 model is
   \[ y_{i(k)} = Z_{i(k)}^{(l)} \beta_{i(k)}^{(l)} + X_{i(k)}^{(l)} \beta_1 + \varepsilon_{i(k)}^{(l)}, \quad i(k) \in i(k). \] (5A.1)

   The level-1 fixed parameter vector \( \beta_1 \) has dimension \( K_1 \times 1 \) and the level-1 vector of parameters that may vary over higher levels, \( \beta_{i(k-1)}^{(l)} \), has dimension \( q_1 \times 1 \).

2. For \( g = 2, ..., k-1 \), the level-g model is
   \[ \beta_{i(k+1-g)}^{(g)} = Z_{i(k+1-g)}^{(g)} \beta_{i(k-g)}^{(g)} + X_{i(k+1-g)}^{(g)} \beta_{g} + \varepsilon_{i(k+1-g)}^{(g)}, \]
   for \( g = 2, ..., k-1 \). (5A.2)

   Here, the level-2 fixed parameter vector \( \beta_{g} \) has dimension \( K_g \times 1 \) and the level-g varying parameter vector \( \beta_{i(k+1-g)}^{(g)} \) has dimension \( q_g \times 1 \). Thus, for the covariates, \( Z_{i(k+1-g)}^{(g)} \) has dimension \( q_{g+1} \times q_g \) and \( X_{i(k+1-g)}^{(g)} \) has dimension \( q_{g+1} \times K_g \).

3. The level-1 model is
   \[ \beta_{i(k)}^{(k-1)} = X_{i(k)}^{(k-1)} \beta_k + e_{i(k)}^{(k-1)}. \] (5A.3)

We assume that all disturbance terms \( \varepsilon \) are mean zero and are uncorrelated with one another.

Further, define \( D_g = \text{Var}(\varepsilon_{i(k+1-g)}^{(g)}) = \sigma_g^2 I_{q_{g+1}}, \) for \( g \geq 2 \).

We now show how to write the multilevel model as a linear mixed effects model. We do this by recursively inserting the higher level models from equation (5A.2) into the level-1 equation (5A.1). This yields

\[ y_{i(k)} = \varepsilon_{i(k)}^{(l)} + X_{i(k)}^{(l)} \beta_1 + Z_{i(k)}^{(l)} \left( Z_{i(k-1)}^{(2)} \beta_{i(k-2)}^{(2)} + X_{i(k-1)}^{(2)} \beta_2 + e_{i(k-1)}^{(2)} \right). \]
\[ \begin{align*}
&= \varepsilon_{i(k)}^{(1)} + X_{i(k)}^{(1)}\beta_1 + Z_{i(k)}^{(1)}\left(\varepsilon_{i(k-1)}^{(2)} + X_{i(k-1)}^{(2)}\beta_2\right) \\
&\quad + Z_{i(k)}^{(2)}\left(Z_{i(k-2)}^{(3)}\beta_3 + \varepsilon_{i(k-2)}^{(3)}\right) \\
&\quad + \cdots + \varepsilon_{i(k)}^{(1)} + X_{i(k)}^{(1)}\beta_1 + \sum_{s=1}^{k-1} \left(\sum_{j=1}^{r} Z_{i(k-1-j)}^{(j)}\varepsilon_{i(k-s)}^{(s+1)} + X_{i(k-s)}^{(s+1)}\beta_{s+1}\right).
\end{align*} \]

To simplify notation, define the \(1 \times q_s\) vector
\[ Z_{s,i(k)} = \prod_{j=1}^{s} Z_{i(k-1-j)}^{(j)} . \quad (5A.4) \]

Further, define \(K = K_1 + \ldots + K_k\), the \(K \times 1\) vector \(\beta = (\beta_1', \beta_2', \ldots, \beta_k')\) and the \(1 \times K\) vector
\[ X_{i(k)} = \left(X_{i(k)}^{(1)} Z_{i(k)}^{(1)} X_{i(k)}^{(2)} \cdots Z_{k-1,i(k)} X_{i(k)}^{(k)}\right). \]

This yields
\[ X_{i(k)}\beta = X_{i(k)}^{(1)}\beta_1 + \sum_{s=1}^{k-1} Z_{s,i(k)} X_{i(k-s)}^{(s+1)}\beta_{s+1} . \]

Thus, we may express the multilevel model as
\[ y_{i(k)}' = X_{i(k)}\beta + \varepsilon_{i(k)}^{(1)} + \sum_{s=1}^{k-1} Z_{s,i(k)}\varepsilon_{i(k-s)}^{(s+1)} . \quad (5A.5) \]

To write equation (5A.5) as a mixed linear model, we require some additional notation. For a fixed set \(\{i_1, \ldots, i_{k-1}\} = i(k-1)\), let \(n(i(k-1))\) denote the number of observed responses of the form \(y_{i,1,\ldots,1,j}'\), for some \(j\). Denote the set of observed responses as
\[ y_{i(k-1)} = \begin{pmatrix} y_{i,1,\ldots,1,1} \\ \vdots \\ y_{i,1,\ldots,1,n(i(k-1))} \end{pmatrix} = \begin{pmatrix} y_{i(k-1),1} \\ \vdots \\ y_{i(k-1),n(i(k-1))} \end{pmatrix} . \]

For each \(s=1, \ldots, k-1\), consider a set \(\{i_1, \ldots, i_{k-s}\} = i(k-s)\) and let \(n(i(k-s))\) denote the number of observed responses of the form \(y_{i(k-s),j}'\), for some \(j\). Thus, we define
\[ y_{i(k-s)} = \begin{pmatrix} y_{i(k-s),1} \\ \vdots \\ y_{i(k-s),n(i(k-s))} \end{pmatrix} . \]

Finally, let \(y = (y_1', \ldots, y_{n(i(k))})'\). Use a similar stacking scheme for \(X\) and \(\varepsilon^{(s)}\), for \(s = 1, \ldots, k\). We may also use this notation when stacking over the first level of \(Z\). Thus, define
\[ Z_{s,i(k-1)} = \begin{pmatrix} Z_{s,i(k-1),1} \\ \vdots \\ Z_{s,i(k-1),n(i(k-1))} \end{pmatrix} , \text{ for } s = 1, \ldots, k-1. \]

With this notation, when stacking over the first level, we may express equation (5A.5) as
\[ y_{i(k-1)} = X_{i(k-1)}\beta + \varepsilon_{i(k-1)}^{(1)} + \sum_{s=1}^{k-1} Z_{s,i(k-s)}\varepsilon_{i(k-s)}^{(s+1)} . \]
For the next level, define
\[
Z_{s.i(k-2)} = \begin{pmatrix}
Z_{s,i(k-2),1} \\
\vdots \\
Z_{s,i(k-2),n(i(k-2))}
\end{pmatrix}, \text{for } s = 2, \ldots, k-1.
\]
and
\[
Z_{i,(k-2)} = \text{blkdiag}(Z_{1.i(k-2),1}, \ldots, Z_{n(i(k-2))})
\]

With this notation, we have
\[
Z_{i,(k-2)}e_{i(k-2)} = \begin{pmatrix}
Z_{1.i(k-2),1} & 0 & \cdots & 0 \\
0 & Z_{1.i(k-2),2} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & Z_{n(i(k-2)),n(i(k-2))}
\end{pmatrix}
\begin{pmatrix}
e^{(2)}_{i(k-2),1} \\
e^{(2)}_{i(k-2),2} \\
\vdots \\
e^{(2)}_{i(k-2),n(i(k-2))}
\end{pmatrix} = \begin{pmatrix}
Z_{1.i(k-2),1}e^{(2)}_{i(k-2),1} \\
Z_{1.i(k-2),2}e^{(2)}_{i(k-2),2} \\
\vdots \\
Z_{n(i(k-2)),n(i(k-2))}e^{(2)}_{i(k-2),n(i(k-2))}
\end{pmatrix}
\]

Thus, we have
\[
y_{i(k-2)} = X_{i(k-2)}\beta + e^{(1)}_{i(k-2)} + Z_{i,(k-2)}e^{(2)}_{i(k-2)} + \sum_{s=2}^{k-1} Z_{s,i(k-2)}e^{(s+1)}_{i(k-s)}.
\]

Continuing, at the gth stage, we have
\[
Z_{s,i(k-g)} = \begin{pmatrix}
Z_{s,i(k-g),1} \\
\vdots \\
Z_{s,i(k-g),n(i(k-g))}
\end{pmatrix}, \text{for } s \geq g
\]
and
\[
Z_{s,i(k-g)} = \text{blkdiag}(Z_{s,i(k-g),1}, \ldots, Z_{s,i(k-g),n(i(k-g))}), \text{for } s < g.
\]

This yields
\[
y_{i(k-g)} = X_{i(k-g)}\beta + e^{(1)}_{i(k-g)} + \sum_{s=1}^{g} Z_{s,i(k-g)}e^{(s+1)}_{i(k-g)} + \sum_{s=g+1}^{k-1} Z_{s,i(k-g)}e^{(s+1)}_{i(k-s)}.
\]

Taking g = k – 1, we have
\[
y_{i(1)} = X_{i(1)}\beta + e^{(1)}_{i(1)} + \sum_{s=1}^{k-1} Z_{s,i(1)}e^{(s+1)}_{i(1)}, \quad (5A.6)
\]

an expression for the usual linear mixed effects model.

The system of notation takes us directly from the multilevel model in equations (5A.1)-(5A.3) to the linear mixed effects model in equation (5A.6). Properties of parameter estimates for linear mixed effects model are well established. Thus, parameter estimators of the multilevel model also enjoy these properties. Moreover, by showing how to write multilevel models as linear mixed effects model, no special statistical software is required. One may simply use software written for linear mixed effects models for multilevel modeling.
5. Exercises and Extensions

Section 5.3

5.1. Two-level model

Consider the two-level model described in Section 5.1.1 and suppose that we have the error components model, so that

\[ z_{ij} = \zeta_i \xi_j + (1 - \zeta_i) \mathbf{b}_{2GLS} \]

where \( \zeta_i = \frac{T_i}{T_i + (\text{Var} \varepsilon)/(\text{Var} \alpha)} \).

5.2. Random intercepts three-level model

Assume that we observe \( i = 1, \ldots, n \) school districts. Within each school district, we observe \( j = 1, \ldots, J \) students. For each student, we have \( t = 1, \ldots, T_{ij} \) observations. Assume that the model is given by

\[ y_{ij,t} = \mathbf{x}_{ijt}^\prime \beta + \alpha_i + \nu_{ij} + \varepsilon_{ij,t} \]

Here, assume that each of \( \{\alpha_i\}, \{\nu_{i1}\}, \ldots, \{\nu_{iJ}\} \) and \( \{\varepsilon_{ij,t}\} \) are independently and identically distributed as well as independent of one another. Also assume that \( \{\alpha_i\}, \{\nu_{i1}\}, \ldots, \{\nu_{iJ}\} \) and \( \{\varepsilon_{ij,t}\} \) are mean zero with variances \( \sigma_{\alpha_i^2}, \sigma_{\nu_{i1}^2}, \ldots, \sigma_{\nu_{iJ}^2} \) and \( \sigma_{\varepsilon_{ij,t}^2} \), respectively.

Define \( \mathbf{z}_{ij} \) to be a \( T_{ij} \times (J+1) \) matrix with ones in the first and \( j+1 \)st columns and zeroes elsewhere. For example, we have

\[ \begin{pmatrix} 1 & 0 & 1 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 1 & \cdots & 0 \end{pmatrix} \]

Further define

\[ \mathbf{Z}_i = \left( \mathbf{z}_{i1}' \quad \mathbf{z}_{i2}' \quad \cdots \quad \mathbf{z}_{ij}' \right)' \]

\[ \mathbf{a}_i = \left( \alpha_i \quad \nu_{i1} \quad \cdots \quad \nu_{iJ} \right)' \]

\[ \mathbf{D}_\nu = \text{diag}(\sigma_{\nu_{i1}^2}, \sigma_{\nu_{i2}^2}, \ldots, \sigma_{\nu_{iJ}^2}) \]

a. Define \( \mathbf{y}_i, \mathbf{X}_i \) and \( \varepsilon_i \) in terms of \( \{y_{ij,t}\}, \{x_{ijt}\} \) and \( \{\varepsilon_{ij,t}\} \), so that we may write

\[ \mathbf{y}_i = \mathbf{Z}_i \mathbf{a}_i + \mathbf{X}_i \beta + \varepsilon_i, \]

using the usual notation.

b. For the appropriate choice of \( \mathbf{R}_i \), show that

\[ \mathbf{Z}_i^\prime \mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i \mathbf{b}_{GLS}) = \frac{1}{\sigma_{\xi_i}^2} \begin{pmatrix} T_i \bar{\varepsilon}_i^t & T_i \bar{\varepsilon}_i^t & \cdots & T_i \bar{\varepsilon}_i^t \end{pmatrix} \]

where \( \bar{\varepsilon}_{ij} = y_{ij,t} - x_{ij,t}^\prime \mathbf{b}_{GLS}, \bar{\varepsilon}_i = T_i^{-1} \sum_{t=1}^{T_i} \bar{\varepsilon}_{ij} \) and \( T_i \bar{\varepsilon}_i = \sum_{j=1}^{J} \sum_{t=1}^{T_{ij}} \varepsilon_{ij,t} \).

c. Show that

\[ \left( \mathbf{D}^{-1} + \mathbf{Z}_i^\prime \mathbf{R}_i^{-1} \mathbf{Z}_i \right)^{-1} = \begin{pmatrix} \mathbf{C}^{-1}_{11} & -\mathbf{C}^{-1}_{11} \xi'_0 \\ -\mathbf{C}^{-1}_{11} \xi_0 & \mathbf{C}^{-1}_{22} \end{pmatrix} \]
where \( \zeta_i = \frac{\sigma_\epsilon^2 T_i}{\sigma_\epsilon^2 + \sigma_\alpha^2} \), \( \zeta_{ij} = \frac{\sigma_\epsilon^2 T_{ij}}{\sigma_\epsilon^2 + \sigma_\gamma^2} \), \( \zeta_i = (\zeta_{i1}, \zeta_{i2}, \ldots, \zeta_{ik})' \),

\[
T_{i,2} = \text{diag}(T_{i1}, T_{i2}, \ldots, T_{ik}), \quad C_{11}^{-1} = \sum_{j=1}^J T_{ij} (1 - \zeta_i \zeta_{ij})
\]

\[
C_{22}^{-1} = (D_{1o}^{-1} + \sigma_\epsilon^2 T_{i,2})^{-1} + C_{11}^{-1} \sigma_\epsilon \zeta_i'
\]

d. With the notation \( a_{i,\text{BLUP}} = (a_{i,\text{BLUP}1}, a_{i,\text{BLUP2}}, \ldots, a_{i,\text{BLUPJ}})' \), show that

\[
a_{i,\text{BLUP}} = \zeta_i \left( \sum_{j=1}^J T_{ij} (1 - \zeta_i \zeta_{ij}) \right) \\
v_{ij,\text{BLUP}} = \zeta_{ij} \left( \zeta_i - a_{i,\text{BLUP}} \right)
\]

Section 5.4

5.3. MLE variance estimators without boundary conditions

Consider the basic random effects model and suppose that \( T_i = T, K = 1 \) and that \( x_i = 1 \). Parts (a) and (b) are the same as Exercise 3.10 (a) and (b). As there, we now ignore boundary conditions so that the estimator may become negative with positive probability.

a. Show that the maximum likelihood estimator of \( \sigma_\epsilon^2 \) may be expressed as:

\[
\hat{\sigma}_{\epsilon,\text{ML}}^2 = \frac{1}{n(T-1)} \sum_{i=1}^n \sum_{t=1}^T (y_{it} - \bar{y}_i)^2
\]

b. Show that the maximum likelihood estimator of \( \sigma_\alpha^2 \) may be expressed as:

\[
\hat{\sigma}_{\alpha,\text{ML}}^2 = \frac{1}{n} \sum_{i=1}^n (\bar{y}_i - \bar{y})^2 - \frac{1}{T} \hat{\sigma}_{\epsilon,\text{ML}}^2
\]

c. Show that the maximum likelihood may be expressed as:

\[
L(\hat{\sigma}_{\alpha,\text{ML}}^2, \hat{\sigma}_{\epsilon,\text{ML}}^2) = -\frac{n}{2} \left[ T \ln(2\pi) + T + (T-1) \ln \hat{\sigma}_{\epsilon,\text{ML}}^2 + \ln(T \hat{\sigma}_{\alpha,\text{ML}}^2 + \hat{\sigma}_{\epsilon,\text{ML}}^2) \right]
\]

d. Consider the null hypothesis \( H_0: \sigma_\alpha^2 = 0 \). Under this null hypothesis, show that the maximum likelihood estimator of \( \sigma_\epsilon^2 \) may be expressed as:

\[
\hat{\sigma}_{\epsilon,\text{Reduced}}^2 = \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T (y_{it} - \bar{y})^2
\]

e. Under the null hypothesis \( H_0: \sigma_\alpha^2 = 0 \), show that the maximum likelihood may be expressed as:

\[
L(0, \hat{\sigma}_{\epsilon,\text{Reduced}}^2) = -\frac{n}{2} \left[ T \ln(2\pi) + T + \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T (y_{it} - \bar{y})^2 \right]
\]

f. Use a second order approximation of the logarithm function to show that twice the difference of log-likelihoods may be expressed as:

\[
2 \left( L(\hat{\sigma}_{\alpha,\text{ML}}^2, \hat{\sigma}_{\epsilon,\text{ML}}^2) - L(0, \hat{\sigma}_{\epsilon,\text{Reduced}}^2) \right) = \frac{1}{2nT(T-1)} \left( SSW - (T-1)SSB \right)^2
\]
where \( SSW = \sum_{i=1}^{n} \sum_{t=1}^{T} (y_{it} - \bar{y}_i)^2 \) and \( SSB = T \sum_{i} (\bar{y}_i - \bar{y})^2 \).

f. Assuming normality of the responses and the null hypothesis \( H_0: \sigma^2 = 0 \), show that
\[
2\left(L(\hat{\sigma}^2_{\alpha,ML}, \hat{\sigma}^2_{e,ML}) - L(0, \hat{\sigma}^2_{e,Reduced})\right) \xrightarrow{D} \chi^2(1) ,
\]
as \( n \to \infty \).

5.4. MLE variance estimators with boundary conditions

Consider the basic random effects model and suppose that \( T_i = T, K=1 \) and that \( x_{it} = 1 \). Unlike problem 5.3, we now impose boundary conditions so that variance estimators must be nonnegative.

a. Using the notation of Exercise 5.3, show that the maximum likelihood estimators of \( \sigma^2_e \) and \( \sigma^2_\alpha \) may be expressed as:
\[
\hat{\sigma}^2_{\alpha,CML} = \begin{cases} 
\hat{\sigma}^2_{\alpha,ML} & \text{if } \hat{\sigma}^2_{\alpha,ML} > 0 \\
0 & \text{if } \hat{\sigma}^2_{\alpha,ML} \leq 0
\end{cases}
\]
and
\[
\hat{\sigma}^2_{e,CML} = \begin{cases} 
\hat{\sigma}^2_{e,ML} & \text{if } \hat{\sigma}^2_{e,ML} > 0 \\
\hat{\sigma}^2_{e,Reduced} & \text{if } \hat{\sigma}^2_{e,ML} \leq 0
\end{cases}
\]
An early reference for this result is Herbach (1959G).

b. Show that the maximum likelihood may be expressed as:
\[
L(\hat{\sigma}^2_{\alpha,CML}, \hat{\sigma}^2_{e,CML}) = \begin{cases} 
L(\hat{\sigma}^2_{\alpha,ML}, \hat{\sigma}^2_{e,ML}) & \text{if } \hat{\sigma}^2_{\alpha,ML} > 0 \\
L(0, \hat{\sigma}^2_{e,Reduced}) & \text{if } \hat{\sigma}^2_{\alpha,ML} \leq 0
\end{cases}
\]
c. Define the cut-off \( c_n = (T-1) \frac{SSB}{SSW} - 1 \). Check that \( c_n > 0 \) if and only if \( \hat{\sigma}^2_{\alpha,ML} > 0 \). Confirm that we may express the likelihood ratio statistic as
\[
2\left(L(\hat{\sigma}^2_{\alpha,CML}, \hat{\sigma}^2_{e,CML}) - L(0, \hat{\sigma}^2_{e,Reduced})\right) = \begin{cases} 
\left[n \left(T \ln \left(1 + \frac{c_n}{T}\right) - \ln(1 + c_n)\right)\right] & \text{if } c_n > 0 \\
0 & \text{if } c_n \leq 0
\end{cases}
\]
d. Assuming normality of the responses and the null hypothesis \( H_0: \sigma^2_\alpha = 0 \), show that the cut-off \( c_n \to_p 0 \) as \( n \to \infty \).

e. Assuming normality of the responses and the null hypothesis \( H_0: \sigma^2_\alpha = 0 \), show that
\[
\sqrt{nc_n} \to_D N\left(0, \frac{2T}{T-1}\right) \text{ as } n \to \infty ,
\]
where \( \Phi \) is the standard normal distribution function.

f. Assume normality of the responses and the null hypothesis \( H_0: \sigma^2_\alpha = 0 \). Show, for \( a > 0 \), that
\[
\Pr\left[2\left(L(\hat{\sigma}^2_{\alpha,CML}, \hat{\sigma}^2_{e,CML}) - L(0, \hat{\sigma}^2_{e,Reduced})\right) > a\right] \to_D 1 - \Phi\left(\sqrt{a}\right) \text{ as } n \to \infty .
\]
g. Assume normality of the responses and the null hypothesis \( H_0: \sigma^2_\alpha = 0 \). Summarize the results above to establish that the likelihood ratio test statistic asymptotically has a distribution that is 50% equal to 0 and 50% a chi-square distribution with one degree of freedom.
Empirical Exercise

5.5. Student Achievement

These data were gathered to assess the relationship between student achievement and education initiatives. Moreover, they can also be used to address related interesting questions, such as how one can rank the performance of schools or how one can forecast a child’s future performance on achievement tests based on their early test scores.

Webb et al. (2002EP) investigated relationships between student achievement and Texas school district participation in the National Science Foundation Statewide Systemic Initiatives program between 1994 and 2000. They focused on the effects of systemic reform on performance on a state mathematics test. We consider here a subset of these data to model trajectories of students’ mathematics achievement over time. This subset consists of a random sample of 20 elementary schools in Dallas, with 20 students randomly selected from each school. All available records for these 400 students during elementary school are included. In Dallas, Grades 3 through 6 correspond to elementary school.

Although there exists a natural hierarchy at each time point (students are nested within schools), this hierarchy was not maintained completely over time. Several students switched schools (see variable SWITCH_SCHOOLS) and many students were not promoted (see variable RETAINED). To maintain the hierarchy of students within schools, a student was associated with a school at the time of selection. To maintain a hierarchy over time, a cohort variable was defined as 1, 2, 3, 4 for those in grades 6, 5, 4 and 3, respectively, in 1994, and a 5 for those in grade 3 in 1995, and so on up to a 10 for those in grade 3 in 2000. The variable FIRST_COHORT attaches a student to a cohort during the first year of observation whereas the variable LAST_COHORT attaches a student to a cohort during the last year of observation.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRADE</td>
<td>Grade when assessment was made (3-6)</td>
</tr>
<tr>
<td>YEAR</td>
<td>Year of assessment (1994-2000)</td>
</tr>
<tr>
<td>TIME</td>
<td>Observed repeated occasions for each student</td>
</tr>
<tr>
<td>RETAINED</td>
<td>Retained in grade for a particular year (1=yes, 0=no)</td>
</tr>
<tr>
<td>SWITCH_SCHOOLS</td>
<td>Switched schools in a particular year (1=yes, 0=no)</td>
</tr>
<tr>
<td>DISADVANTAGED</td>
<td>Economically disadvantaged (1= free/reduced lunch, 0=no)</td>
</tr>
<tr>
<td>TLI_MATH</td>
<td>Texas Learning Index on mathematics – assessment measure</td>
</tr>
<tr>
<td>CHILDID</td>
<td>Student identification number</td>
</tr>
<tr>
<td>MALE</td>
<td>Gender of students (1=male, 0=female)</td>
</tr>
<tr>
<td>ETHNICITY</td>
<td>White, black, hispanic, other (“other” includes asian as well as mixed races)</td>
</tr>
<tr>
<td>FIRST_COHORT</td>
<td>First observed cohort membership</td>
</tr>
<tr>
<td>LAST_COHORT</td>
<td>Last observed cohort membership</td>
</tr>
<tr>
<td>SCHOOLID</td>
<td>School identification number</td>
</tr>
<tr>
<td>MATH_SESSIONS</td>
<td>Number of teachers attended mathematics sessions</td>
</tr>
<tr>
<td>N_TEACHERS</td>
<td>Total number of teachers in the school</td>
</tr>
</tbody>
</table>

a  Basic summary statistics
   i  Summarize the school level variables. Produce a table to summarize the frequency of the USI variable. For MATH_SESSIONS and N_TEACHERS, provide the mean, median, standard deviation, minimum and the maximum.
   ii Summarize the child level variables. Produce tables to summarize the frequency of gender, ethnicity and the cohort variables.
   iii Provide basic relationships among level 2 and 3 variables. Summarize the number of teachers by gender. Examine ethnicity by gender.
   iv Summarize the level 1 variables. Produce means for the binary variables RETAINED, SWITCH_SCHOOLS and DISAVANTAGED. For TLI_MATH, provide the mean, median, standard deviation, minimum and the maximum.
   v Summarize numerically some basic relationships between TLI_MATH and the explanatory variables. Produce tables of means of TLI_MATH by GRADE, YEAR, RETAINED, SWITCH_SCHOOLS and DISAVANTAGED.
   vi Summarize graphically some basic relationships between TLI_MATH and the explanatory variables. Produce boxplots of TLI_MATH by GRADE, YEAR, RETAINED, SWITCH_SCHOOLS and DISAVANTAGED. Comment on the trend over time and grade.
   vii Produce a multiple time series plot of TLI_MATH. Comment on the dramatic declines of some students in year-to-year test scores.

b  Two-level - Error components model
   i  Ignoring the school level information, run an error components model using child as the second level unit of analysis. Use the level 1 categorical variables GRADE and YEAR and binary variables RETAINED and SWITCH_SCHOOLS. Use the level 2 categorical variables ETHNICITY and the binary variable MALE.
   ii Repeat your analysis in part b(i) but include the variable DISAVANTAGED. Describe the advantages and disadvantages of including this variable in the model specification.
   iii Repeat your analysis in part b(i) but include an AR(1) specification of the error. Does this improve the model specification?
   iv Repeat your analysis in part b(iii) but include a (fixed) school level categorical variable. Does this improve the model specification?

c  Three-level model
   i  Now incorporate school level information into your model in b(i). At the first level, the random intercept varies by child and school. We also include GRADE, YEAR, RETAINED and SWITCH_SCHOOLS as level 1 explanatory variables. For the second level model, the random intercept varies by school and includes ETHNICITY and MALE as level 2 explanatory variables. At the third level, we include USI, MATH_SESSIONS and N_TEACHERS as level 3 explanatory variables. Comment on the appropriate of this fit.
   ii Is the USI categorical variable statistically significant? Re-run the part c(i) model without USI and use a likelihood ratio test statistic to respond to this question.
   iii Repeat your analysis in part c(i) but include an AR(1) specification of the error. Does this improve the model specification?
Appendix 5A

5.6. BLUP predictors for a general multilevel model

Consider the general multilevel model developed in Appendix 5A and the mixed linear model representation in equation (5A.6). Let $V_{i(1)} = \text{Var}(y_{i(1)})$.

a. Using best linear unbiased prediction introduced in Section 4.2, show that we can express the BLUP predictors of the residuals as

\[ \begin{align*}
\epsilon_{i(k+1-g),BLUP}^{(g)} &= \text{Cov}(\epsilon_{i(1)}, \epsilon_{i(k+1-g)}) Z_{g-1,i(1)} V_{i(1)}^{-1}(y_{i(1)} - X_{i(1)} b_{GLS}), \\
\end{align*} \]

for $g = 2, \ldots, k$, and, for $g=1$,

\[ \begin{align*}
\epsilon_{i(k),BLUP}^{(1)} &= \text{Cov}(\epsilon_{i(1)}, \epsilon_{i(k)}) V_{i(1)}^{-1}(y_{i(1)} - X_{i(1)} b_{GLS}).
\end{align*} \]

b. Show that the BLUP predictor of $b_{i(k+1-g)}^{(g-1)}$ is

\[ \begin{align*}
b_{i(k+1-g),BLUP}^{(g-1)} &= X_{i(k+1-g)}^{(g)} b_{g,GLS} + Z_{i(k+1-g)}^{(g)} b_{i(k+1-g),BLUP} + \epsilon_{i(k+1-g),BLUP}^{(g-1)}.
\end{align*} \]

c. Show that the BLUP forecast of $V_{h_{l_1,l_2,\ldots,l_{k+L}}}$ is

\[ \begin{align*}
\hat{y}_{h_{l_1,l_2,\ldots,l_{k+L}}} &= Z_{h_{l_1,l_2,\ldots,l_{k+L}},BLUP}^{(1)} b_{l_1,GLS} + X_{h_{l_1,l_2,\ldots,l_{k+L}},BLUP}^{(1)} b_{GLS} \\
&\quad + \text{Cov}(\epsilon_{h_{l_1,l_2,\ldots,l_{k+L}},1}, \epsilon_{i(1)}) V_{i(1)}^{-1}(y_{i(1)} - X_{i(1)} b_{GLS}).
\end{align*} \]
Chapter 6. Stochastic Regressors

Abstract. In many applications of interest, explanatory variables, or regressors, cannot be thought of as fixed quantities but rather are modeled stochastically. In some applications, it can be difficult to determine what variables are being predicted and what variables are doing the prediction! This chapter summarizes several models that incorporate stochastic regressors. The first consideration is to identify under what circumstances we can safely condition on stochastic regressors and use the results from prior chapters. We then discuss exogeneity, formalizing the idea that a regressor influences the response variable and not the other way around. Finally, this chapter introduces situations where more than one response is of interest, thus permitting us to investigate complex relationships among responses.

Section 6.1. Stochastic regressors in non-longitudinal settings

Up to this point, we have assumed that the explanatory variables, $X_i$ and $Z_i$, are non-stochastic. This convention follows a long-standing tradition in the statistics literature. Pedagogically, this tradition allows for simpler verification of properties of estimators than the stochastic convention. Moreover, in classical experimental or laboratory settings, treating explanatory variables as non-stochastic allows for intuitive interpretations, such as when $X$ is under the control of the analyst.

However, for other applications, such as the analysis of survey data that are drawn as a probability sample from a population, the assumption of non-stochastic variables is more difficult to interpret. For example, when drawing a sample of individuals to understand each individual’s health care decisions, we may wish to explain their health care services utilization in terms of their age, gender, race and so on. These are plausible explanatory variables and it seems sensible to model them as stochastic in that the sample values are determined by a random draw from a population.

In some ways, the study of stochastic regressors subsumes that of non-stochastic regressors. First, with stochastic regressors, we can always adopt the convention that a stochastic quantity with zero variance is simply a deterministic, or non-stochastic, quantity. Second, we may make inferences about population relationships conditional on values of stochastic regressors, essentially treating them as fixed. However, the choice of variables on which we condition depends on the scientific interest of the problem, making the difference between fixed and stochastic regressors dramatically different in some cases.

Understanding the best ways to use stochastic regressors in longitudinal settings is still a developing research area. Thus, before presenting techniques useful for longitudinal data, this section reviews known and proven methods that are useful in non-longitudinal settings, either for the cross-section or the time dimension. Subsequent sections in this chapter focus on longitudinal settings that incorporate both the cross-section and time dimensions.

Section 6.1.1 Endogenous stochastic regressors

An exogenous variable is one that can be taken as “given” for the purposes at hand. As we will see, exogeneity requirements vary depending on the context. An endogenous variable is one that fails the exogeneity requirement. In contrast, it is customary in economics to use the term
“endogenous” to mean a variable that is determined within an economic system whereas an exogenous variable is determined outside the system. Thus, the accepted econometric/statistical usage differs from the general economic meaning.

To develop exogeneity and endogeneity concepts, we begin by thinking of \( \{(x_i, y_i)\} \) as a set of observations from the same distribution. For example, this assumption is appropriate when the data arise from a survey where information is collected using a simple random sampling mechanism. We suppress the \( t \) subscript because we are considering only one dimension in this section. Thus, \( i \) may represent either the cross-sectional identifier or time period. For non-longitudinal data, we do not consider the \( z_i \) variables.

For independent observations, we can write the assumptions of the linear model as in Chapter 2, adding only that we are conditioning on the stochastic explanatory variables when writing down the moments of the response \( y_i \). To begin, for the regression function, we assume that \( \mathbb{E}(y_i | x_i) = x_i' \beta \) and the conditional variance is \( \text{Var}(y_i | x_i) = \sigma^2 \).

To handle additional sampling mechanisms, we now introduce a more general setting. Specifically, we condition on all of the explanatory variables in the sample, not just the ones associated with the \( i \)th draw. Define \( X = (x_1, \ldots, x_n) \) and work with the following assumptions.

### Assumptions of the Linear Regression Model with Strictly Exogenous Regressors

1. \( \mathbb{E}(y_i | X) = x_i' \beta \).
2. \( \{x_1, \ldots, x_n\} \) are stochastic variables.
3. \( \text{Var}(y_i | X) = \sigma^2 \).
4. \( \{y_i | X\} \) are independent random variables.
5. \( \{y_i\} \) is normally distributed, conditional on \( \{X\} \).

Assuming for the moment that \( \{(x_i, y_i)\} \) are mutually independent, then \( \mathbb{E}(y_i | X) = \mathbb{E}(y_i | x_i) \), \( \text{Var}(y_i | X) = \text{Var}(y_i | x_i) \) and \( (y_i | x_i) \) are independent. Thus, the assumptions SE1-4 are certainly useful in the random sampling context.

Moreover, the assumptions SE1-4 are the appropriate stochastic regressor generalization of the fixed regressors model assumptions that ensure that we retain most of the desirable properties of the ordinary least squares estimators of \( \beta \). For example, the unbiasedness and the Gauss-Markov properties of ordinary least squares estimators of \( \beta \) hold under SE1-SE4. Moreover, assuming conditional normality of the responses, then the usual \( t \) and \( F \) statistics have their customary distributions, regardless as to whether or not \( X \) is stochastic; see, for example Greene (2002E) or Goldberger (1991E).

It turns out that the usual ordinary least squares estimators also have desirable asymptotic properties under assumptions SE1-4. For many social science applications, data sets are large and researchers are primarily interested in asymptotic properties of estimators. If achieving desirable asymptotic properties is the goal, then the stochastic regressor model assumptions SE1-4 can be relaxed, thus permitting a wider scope of applications.

For discussion purposes, we now focus on the first assumption. Using the linear model framework, we define the disturbance term to be \( \varepsilon_i = y_i - x_i' \beta \) and write SE1 as \( \mathbb{E}(\varepsilon_i | X) = 0 \). This assumption on the regressors is known as strict exogeneity in the econometrics literature (see, for example, Hayashi, 2000E). If the index \( i \) represents independent cross-sectional draws, such as with simple random sampling, then strict exogeneity is an appropriate assumption. However, if the index \( i \) as represents time, then the strict exogeneity is not useful for many applications; it assumes that the time \( i \) disturbance term is orthogonal to all regressors in the past, contemporaneous and the future.

An alternative assumption is

\[
\text{SE1p. } \mathbb{E}(\varepsilon_i | x_i) = \mathbb{E}(y_i - x_i' \beta | x_i) = 0.
\]
If SE1p holds, then the regressors are said to be *predetermined*. Because SE1p implies zero covariance between the regressors and the disturbances, we say that predetermined regressors are uncorrelated with contemporaneous disturbances. Another way of expressing assumption SE1p is through the *linear projection*

\[ L(\varepsilon_i | x_i) = 0. \]

See Appendix 6A for definitions and properties of linear projections. This alternative method will be useful as we explore longitudinal extensions of the notion of endogeneity in Section 6.3.

The assumption SE1p is weaker than SE1. Only the weaker assumption SE1p (and conditions analogous to those in SE2-4) is required for the asymptotic property of consistency of the ordinary least squares estimators of \( \beta \). We will be more specific in our discussion of longitudinal data beginning in Section 6.2. For specifics regarding non-longitudinal data settings, see, for example, Hayashi (2000E).

To reiterate, the strict exogeneity assumption SE1 is sufficient for the ordinary least squares estimators of \( \beta \) to retain finite sample properties such as unbiasedness whereas only the weaker predetermined assumption SE1p is required for consistency. For asymptotic normality, we require an assumption that is somewhat stronger than SE1p. A sufficient condition is:

SE1m. \( E(\varepsilon_i | \varepsilon_{i-1}, \ldots, \varepsilon_1, x_i, \ldots, x_1) = 0 \) for all \( i \).

When SE1m holds, then \( \{\varepsilon_i\} \) satisfies the requirements for a *martingale difference* sequence. We note, using the law of iterated expectations, that SE1m implies SE1p. For time series data where the index \( i \) represents time, we see that both Assumptions SE1p and SE1m do not rule out the possibility that the current error term \( \varepsilon_i \) will be related to future regressors, as does the strict exogeneity assumption SE1.

### Section 6.1.2 Weak and strong exogeneity

We began Section 6.1.1 by expressing two types of exogeneity, strict exogeneity and predeterminedness, in terms of conditional means. This is appropriate for linear models because it gives precisely the conditions needed for inference and is directly testable. We begin this subsection by generalizing these concepts to assumptions regarding the entire distribution, not just the mean function. Although stronger than the conditional mean versions, these assumptions are directly applicable to nonlinear models. Moreover, we use this distribution framework to introduce two new types of exogeneity, weak and strong exogeneity.

A stronger version of strict exogeneity in SE1 is

SE1’ \( \varepsilon_i \) is independent of \( X \).

Here, we are using the convention that the zero mean disturbances are defined as \( \varepsilon_i = y_i - x_i' \beta \). Note that SE1’ implies SE1; SE1’ is a requirement on the joint distribution of \( \varepsilon_i \) and \( X \), not just the conditional mean. Similarly, a stronger version of SE1p is

SE1p’. \( \varepsilon_i \) is independent of \( x_i \).

A drawback of SE1 and SE1p is that the reference to parameter estimability is only implicit. An alternative set of definitions introduced by Engle, Hendry and Richard (1983E) explicitly defines exogeneity in terms of parametric likelihood functions. Intuitively, a set of variables are said to be *weakly exogenous* if, when we condition on them, there is no loss of information for the parameters of interest. If, in addition, the variables are “not caused” by the endogenous variables, then they are said to be *strongly exogenous*. Weak exogeneity is sufficient for efficient estimation. Strong exogeneity is required for conditional predictions in forecasting of endogenous variables.
Specifically, suppose that we have random variables \((x_1, y_1), \ldots, (x_T, y_T)\) with joint probability density (or mass) function for \(f(y_1, \ldots, y_T, x_1, \ldots, x_T)\). Using \(t\) for the (time) index, we can always write this conditionally as

\[
\prod_{t=1}^{T} f(y_t, x_t \mid y_{1:t-1}, x_{1:t-1})
\]

Here, when \(t=1\) the conditional distributions are the marginal distributions of \(y_1\) and \(x_1\), as appropriate. Now, suppose that this joint distribution is characterized by vectors of parameters \(\theta\) and \(\psi\) such that

\[
\begin{pmatrix}
\prod_{t=1}^{T} f(y_t \mid y_{1:t-1}, x_{1:t-1}, \theta)
\prod_{t=1}^{T} f(x_t \mid y_{1:t-1}, x_{1:t-1}, \psi)
\end{pmatrix}
\]

In this case, we can ignore the second term for inference about \(\theta\), treating the \(x\) variables as essentially fixed. If the relationship \(\text{SE1w}\) holds, then we say that the explanatory variables are weakly exogenous.

Suppose, in addition, that

\[
f(x_t \mid y_{1:t-1}, x_{1:t-1}, \psi) = f(x_t \mid x_{1:t-1}, \psi), \quad (6.1a)
\]

that is, conditional on \(x_1, \ldots, x_{t-1}\), that the distribution of \(x_t\) does not depend on past values of \(y_1, y_{1:t-1}\). Then, we say that \(\{y_1, \ldots, y_{t-1}\}\) does not Granger-cause \(x_t\). This condition, together with \(\text{SE1w}\), suffices for strong exogeneity. We note that Engle et al. (1983E) also introduce a so-called super exogeneity assumption for policy analysis purposes; we will not consider this type of exogeneity.

Section 6.1.3 Causal effects

Issues of when variables are endogenous or exogenous are important to researchers that use statistical models as part of their arguments for assessing whether or not causal relationships hold. Researchers are interested in causal effects, often more so than measures of association among variables.

Traditionally, statistics has contributed to making causal statements primarily through randomization. This tradition goes back to the work of Fisher and Neyman in the context of agricultural experiments. In Fisher's work, treatments were randomly allocated to experimental units (plots of land). Because of this random assignment, differences in responses (crop yields from the land) could be reasonably ascribed to treatments without fear of underlying systematic influences from unknown factors. Data that arise from this random assignment mechanism are known as experimental.

In contrast, most data from the social sciences are observational, where it is not possible to use random mechanisms to randomly allocate observations according to variables of interest. However, it is possible to use random mechanisms to gather data through probability samples and thus to estimate stochastic relationships among variables of interest. The primary example of this is to use a simple random sample mechanism to collect data and estimate a conditional mean through regression methods. The important point is that this regression function measures
In the economics literature, Goldberger (1972E) defines a structural model as a stochastic model representing a causal relationship, not a relationship that simply captures statistical associations. In contrast, a sampling based model is derived from our knowledge of the mechanisms used to gather the data. The sampling based model directly generates statistics that can be used to estimate quantities of interest and thus is also known as an estimable model. To illustrate, suppose that \( \{(x_i, y_i)\} \) represents a random sample from a population. Then, we can always estimate \( E(y \mid x) \) nonparametrically. Moreover, we might assume that \( E(y \mid x) = x' \beta \), for some vector \( \beta \). This requires no appeal to the theory from an underlying functional field. We use only the assumption of the data generating mechanism and thus refer to this as a sampling based model.

As an example of a structural model, Duncan (1969EP) considers the following model equations that relate one’s self-esteem \((y_{it}, t=1, 2)\) to delinquency \((x_{it}, t=1, 2)\):

\[
\begin{align*}
y_{i2} &= \beta_0 + \beta_1 y_{i1} + \beta_2 x_{i1} + \epsilon_{i1} \\
x_{i2} &= \gamma_0 + \gamma_1 y_{i1} + \gamma_2 x_{i1} + \epsilon_{i2}.
\end{align*}
\]

In this model, current period \((t=2)\) self-esteem and delinquency are affected by the prior period’s self-esteem and delinquency. This model specification relies on theory from the functional field. This is an example of a structural equations model that Sections 6.4 and 6.5 will discuss in more detail.

Particularly for observational data, causal statements are based primarily on substantive hypotheses in which the researcher carefully develops. Causal inference is theoretically driven. Causal processes generally cannot be demonstrated directly from the data; the data can only present relevant empirical evidence serving as a link in a chain of reasoning about causal mechanisms.

Longitudinal data are much more useful in establishing causal relationships than (cross-sectional) regression data. This is because, for most disciplines, the “causal” variable must precede the “effect” variable in time. To illustrate, Lazarsfeld and Fiske (1938O) considered the effect of radio advertising on product sales. Traditionally, hearing radio advertisements was thought to increase the likelihood of purchasing a product. Lazarsfeld and Fiske considered whether those that bought the product would be more likely to hear the advertisement, thus positing a reverse in the direction of causality. They proposed repeatedly interviewing a set of people (the ‘panel’) to clarify the issue.

Notions of randomization have been extended by Rubin (1976G, 1978G, 1990G) to observational data through the concept of potential outcomes. This is an area that is rapidly developing; we refer to Angrist, Imbens and Rubin (1996G) for further discussions.

**Section 6.1.4 Instrumental variable estimation**

According to Wooldridge (2002E, page 83), instrumental variable estimation “is probably second only to ordinary least squares in terms of methods used in empirical economics research.” Instrumental variable estimation is a general technique that is widely used in economics and related fields to handle problems associated with the disconnect between the structural model and a sampling based model.

To introduce instrumental variable estimation, this subsection assumes that the index \(i\) represents cross-sectional draws and that these draws are independent. The instrumental variable technique can be used in instances where the structural model is specified by a linear equation of the form

\[
y_i = x_i' \beta + \epsilon_i,
\]

where \(y_i\) and \(x_i\) are the dependent and independent variables, respectively, and \(\epsilon_i\) is the error term.
yet not all of the regressors are predetermined, that is, \( E(\varepsilon_i x_i) \neq 0 \). The instrumental variable technique employs a set of predetermined variables, \( w_i \), that are correlated with the regressors specified in the structural model. Specifically, we assume

\[
\text{IV1. } E(\varepsilon_i w_i) = E( (y_i - x_i' \beta) w_i) = 0 
\]

and

\[
\text{IV2. } E(w_i w_i') \text{ is invertible.}
\]

With these additional variables, an instrumental variable estimator of \( \beta \) is

\[
b_{IV} = (X' P_W X)^{-1} X' P_W y,
\]

where \( P_W = W (W' W)^{-1} W' \) is a projection matrix and \( W = (w_1, \ldots, w_n)' \) is the matrix of instrumental variables. Instrumental variable estimators can be expressed as special cases of generalized method of moment estimators; see Appendix C.6 for further details.

To illustrate, we now describe three commonly encountered situations where the instrumental variable technique has proven to be useful.

The first situation concerns situations where important variables have been omitted from the sampling model. In this situation, we write the structural regression function as \( E(y_i | x_i, u_i) = x_i' \beta + \gamma' u_i \), where \( u_i \) represents important unobserved variables. However, the sampling based model uses only \( E(y_i | x_i) = x_i' \beta \), thus omitting the unobserved variables. For example, in his discussion of omitted variable bias, Wooldridge (2002E) discusses an application by Card (1995E) concerning a cross-section of men where the interest is in studying (logarithmic) wages in relation to years of education. Additional control variables include years of experience (and its square), regional indicators, racial indicators and so forth. The concern is that the structural model omits an important variable, the man’s “ability” \( (u) \), that is correlated with years of education. Card introduces a variable to indicate whether a man grew up in the vicinity of a four-year college as an instrument for years of education. The motivation behind this choice is that this variable should be correlated with education yet uncorrelated with ability. In our notation, we would define \( w_i \) to be the same set of explanatory variables used in the structural equation model but with the vicinity variable replacing the years of education variable. Assuming positive correlation between the vicinity and years of education variables, we expect assumption IV2 to hold. Moreover, assuming that vicinity to be uncorrelated with ability, we expect assumption IV1 to hold.

The second situation where the instrumental variable technique has proven useful concerns important explanatory variables that have been measured with error. Here, the structural model is given as in equation 6.1 but estimation is based on the model

\[
y_i = x_i'^* \beta + \varepsilon_i, \quad (6.2)
\]

where \( x_i'^* = x_i + \eta_i \) and \( \eta_i \) is an error term. That is, the observed explanatory variables \( x_i'^* \) are measured with error yet the underlying theory is based on the “true” explanatory variables \( x_i \). Measurement error causes difficulties because even if the structural model explanatory variables are predetermined, such that \( E( (y_i - x_i' \beta) x_i) = 0 \), this does not guarantee that the observed variables will be because \( E( (y_i - x_i'^* \beta) x_i'^*) \neq 0 \). For example, in Card’s (1995E) returns to schooling example described above, it is often maintained that years of education records are fraught with errors due to lack of recall and other reasons. One strategy is to replace years of education by a more reliable instrument such as completion of high school or not. As with omitted variables, the goal is to select instruments that are highly related to the suspect endogenous variables yet are unrelated to model deviations.

A third important application of instrumental variable techniques regards the endogeneity induced by systems of equations. We will discuss this topic further in Section 6.4.
In many situations, instrumental variable estimators can be easily computed using two-stage least squares. In the first stage, one regresses each endogenous regressor on the set of exogenous explanatory variables and calculates fitted values of the form \( \hat{X} = P_w X \). In the second stage, one regresses the dependent variable on the fitted values using ordinary least squares to get the instrumental variable estimator, that is, \( (\hat{X}'\hat{X})^{-1}\hat{X}'y = b_{IV} \). However, Wooldridge (2002E, page 98) recommends for empirical work that researchers use statistical packages that explicitly incorporate a two-stage least squares routine; some of the sums of squares produced in the second stage that would ordinarily be used for hypothesis testing are not appropriate in the two-stage setting.

The choice of instruments is the most difficult decision faced by empirical researchers using instrumental variable estimation. Theoretical results are available concerning the optimal choice of instruments (White, 1984E). For practical implementation of these results, empirical researchers should essentially try to choose instruments that are highly correlated with the endogenous explanatory variables. Higher correlation means that the bias as well as standard error of \( b_{IV} \) will be lower (Bound, Jaeger and Baker, 1995E). For additional background reading, we refer the reader to virtually any graduate econometrics text (see, for example, Greene, 2002E, Hayashi, 2000E, Wooldridge, 2002E).

**Section 6.2. Stochastic regressors in longitudinal settings**

This section describes estimation in longitudinal situations that can be readily handled using techniques already described in the text. Section 6.3 follows by considering more complex models that require specialized techniques.

**Section 6.2.1 Longitudinal data models without heterogeneity terms**

As we will see in the following subsections, the introduction of heterogeneity terms \( \alpha_i \) complicates the endogeneity questions in longitudinal and panel data models considerably. Conversely, without heterogeneity terms, longitudinal and panel data models carry few features that would not allow us to apply the Section 6.1 techniques directly. To begin, we may write a model with strictly exogenous regressors as:

| Assumptions of the Longitudinal Data Model with Strictly Exogenous Regressors |
| SE1. \( \text{E}(y_{it} | X) = x_{it}' \beta \). |
| SE2. \( \{x_{it}\} \) are stochastic variables. |
| SE3. \( \text{Var}(y_i | X) = R_t \). |
| SE4. \( \{y_{it} | X\} \) are independent random vectors. |
| SE5. \( \{y_{it}\} \) is normally distributed, conditional on \( \{X\} \). |

Recall that \( X = \{X_1, \ldots, X_n\} \) is the complete set of regressors over all subjects and time periods. Because this set of assumptions includes those in the Section 6.1.1 non-longitudinal setting, we still refer to the set as Assumptions SE1-5.

With longitudinal data, we have repeatedly noted the important fact that observations from the same subject tend to be related. Often, we have used the heterogeneity term \( \alpha_i \) to account for this relationship. However, one can also use the covariance structure of the disturbances (\( R_t \)) to account for these dependencies; see Section 7.1. Thus, SE3 allows analysts to choose a correlation structure such as arises from an autoregressive or compound symmetry structure to account for these intra-subject correlations. This formulation, employing strictly exogenous
variables, means that the usual least squares estimators have desirable finite, as well as asymptotic, properties.

As we saw in Section 6.1.1, the strict exogeneity assumption does not permit lagged dependent variables, another widely used approach for incorporating intra-subject relationships among observations. Still, without heterogeneity terms, we can weaken the assumptions on the regressors to the assumption of predetermined regressors, as in Section 6.1.1, and still achieve consistent regression estimators. With the longitudinal data notation, this assumption can be written as:

$$\text{SE1p. } E(\varepsilon_{it} x_{it}) = E((y_{it} - x_{it}' \beta) x_{it}) = 0.$$ 

Using linear projection notation (Appendix 6A), we can also express this assumption as $L(\varepsilon_{it} | x_{it}) = 0$, assuming $E x_{it}x_{it}'$ is invertible. Writing the corresponding martingale difference sequence assumption that allows for asymptotic normality is slightly more cumbersome because of the two indices for the observations in the conditioning set. We leave this as an exercise for the reader.

The important point of this subsection is to emphasize that longitudinal and panel data models have the same endogeneity concerns as the cross-sectional models. Moreover, often the analyst may use well-known techniques for handling endogeneity developed in cross-sectional analysis for longitudinal data. However, when employing these techniques, the longitudinal data models should not possess heterogeneity terms. Instead, devices such as a correlation structure for the conditional response or lagging the dependent variable can be used to account for heterogeneity in longitudinal data, thus allowing the analyst to focus on endogeneity concerns.

Section 6.2.2 Longitudinal data models with heterogeneity terms and strictly exogenous regressors

As we saw in Section 6.1 for non-longitudinal data, the remedies that account for endogenous stochastic regressors require knowledge of a functional field. The formulation of an underlying structural model is by definition field-specific and this formulation affects the determination of the best model estimators. For longitudinal data, in many disciplines it is customary to incorporate a subject-specific heterogeneity term to account for intra-subject correlations, either from knowledge of the underlying data generating process being studied or by tradition within the field of study. Thus, it is often important to understand the effects of regressors when a heterogeneity term $\alpha_i$ is present in the model.

To define “endogeneity” in the panel and longitudinal data context, we again begin with the simpler concept of strict exogeneity. Recall the linear mixed effects model

$$y_{it} = z_{it}' \alpha_i + x_{it}' \beta + \varepsilon_{it}$$

and its vector version

$$y_i = Z_i \alpha_i + X_i \beta + \varepsilon_i.$$ 

To simplify the notation, let $X^* = \{X_1, Z_1, \ldots, X_n, Z_n\}$ be the collection of all observed explanatory variables and $\alpha = (\alpha_1', \ldots, \alpha_n')'$ be the collection of all subject-specific terms. We now consider:
Assumptions of the Linear Mixed Effects Model with Strictly Exogenous Regressors

Conditional on the Unobserved Effect

SEC1. \( \mathbb{E}(y_i | \alpha, X^*) = Z_i \alpha_i + X_i \beta \).
SEC2. \{X^*\} are stochastic variables.
SEC3. \( \text{Var}(y_i | \alpha, X^*) = R_i \).
SEC4. \{y_i\} are independent random vectors, conditional on \{\alpha\} and \{X^*\}.
SEC5. \{y_i\} is normally distributed, conditional on \{\alpha\} and \{X^*\}.
SEC6. \( \mathbb{E}(\alpha_i | X^*) = 0 \) and \( \text{Var}(\alpha_i | X^*) = D \). Further, \{\alpha_1, ..., \alpha_n\} are mutually independent, conditional on \{X^*\}.
SEC7. \{\alpha_i\} is normally distributed, conditional on \{X^*\}.

These assumptions are readily supported by a random sampling scheme. For example, suppose that \((x_1, z_1, y_1), \ldots, (x_n, z_n, y_n)\) represents a random sample from a population. Each draw \((x_i, z_i, y_i)\) has associated with it an unobserved, latent vector \(\alpha_i\) that is part of the conditional regression function. Then, because \{\{(\alpha_i, x_i, z_i, y_i)\}\} are identically and independently distributed, we immediately have SEC2 and SEC4, as well as the conditional independence of \{\alpha_i\} in SEC6. Assumptions SEC1 and SEC2, and the first part of SEC6, are moment conditions and thus depend on the conditional distributions of the draws. Further, assumptions SEC5 and SEC7 are also assumptions about the conditional distribution of a draw.

Assumption SEC1 is a stronger assumption than strict exogeneity (SE1). Using the disturbance term notation, we may re-write this as \( \mathbb{E}(\varepsilon_i | \alpha, X^*) = 0 \). By the law of iterated expectations, this implies that both \( \mathbb{E}(\varepsilon_i | X^*) = 0 \) and \( \mathbb{E}(\varepsilon_i \alpha' | X^*) = 0 \) hold. That is, this condition requires both that the regressors are strictly exogenous and that the unobserved effects are uncorrelated with the disturbance terms. In the context of an error components model with random sampling (the case where \( q=1, z_{it}=1 \) and random variables from different subjects are independent), SEC1 may be expressed as:

\[
\mathbb{E}(y_{it} | \alpha_i, x_{i1}, \ldots, x_{iT}) = \alpha_i + x_{it}' \beta, \text{ for each } t.
\]

Chamberlain (1982E, 1984E) introduced conditional strict exogeneity in this context.

The first part of Assumption SEC6, \( \mathbb{E}(\alpha_i | X^*) = 0 \), also implies that the unobserved, time-constant effects and the regressors are uncorrelated. Many econometric panel data applications use an error components model such as in Section 3.1. In this case, it is customary to interpret \( \alpha_i \) to be an unobserved, time-constant, effect that influences the expected response. This is motivated by the relation \( \mathbb{E}(y_{it} | \alpha_i, X^*) = \alpha_i + x_{it}' \beta \). In this case, we interpret this part of Assumption SEC6 to mean that this unobserved effect is not correlated with the regressors. Sections 7.2 and 7.3 will discuss ways of testing, and relaxing, this assumption.

Example – Tax liability – Continued

Section 3.2 describes an example where we use a random sample of taxpayers and examine their tax liability \((y)\) in terms of demographic and economic characteristics, summarized in Table 3.1. Because the data were gathered using a random sampling mechanism, we can interpret the regressors as stochastic and assume that observable variables from different taxpayers are mutually independent. In this context, the assumption of strict exogeneity implies that we are assuming that tax liability will not affect any of the explanatory variables. For example, the demographic characteristics such as number of dependents and marital status may affect the tax liability but that the reverse implication is not true. In particular, note that the total personal income is based on positive income items from the tax return; exogeneity concerns dictated using this variable in contrast to an alternative such as net income, a variable that may be affected by prior year’s tax liability.
One potentially troubling variable is the use of the tax preparer; it may be reasonable to assume that the tax preparer variable is predetermined, although not strictly exogenous. That is, we may be willing to assume that this year’s tax liability does not affect our decision to use a tax preparer because we do not know the tax liability prior to this choice, making the variable predetermined. However, it seems plausible that the prior year’s tax liability will affect our decision to retain a tax preparer, thus failing the strict exogeneity test. In a model without heterogeneity terms, consistency may be achieved by assuming only that the regressors are predetermined.

For a model with heterogeneity terms, consider the error components model in Section 3.2. Here, we interpret the heterogeneity terms to be unobserved subject-specific (taxpayer) characteristics, such as “aggressiveness,” that would influence the expected tax liability. For strict exogeneity conditional on the unobserved effects, one needs to argue that the regressors are strictly exogenous and that the disturbances, representing “unexpected” tax liabilities, are uncorrelated with the unobserved effects. Moreover, Assumption SEC6 employs the condition that the unobserved effects are uncorrelated with the observed regressor variables. One may be concerned that individuals with high earnings potential who have historically high levels of tax liability (relative to their control variables) may be more likely to use a tax preparer, thus violating this assumption.

As in Chapter 3, the assumptions based on distributions conditional on unobserved effects lead to the following conditions that are the basis of statistical inference.

<table>
<thead>
<tr>
<th>Observables Representation of the Linear Mixed Effects Model with Strictly Exogenous Regressors Conditional on the Unobserved Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>SE1. $E(y_i</td>
</tr>
<tr>
<td>SE2. ${X^*}$ are stochastic variables.</td>
</tr>
<tr>
<td>SE3a. $\text{Var}(y_i</td>
</tr>
<tr>
<td>SE4. ${y_i}$ are independent random vectors, conditional on ${X^*}.$</td>
</tr>
<tr>
<td>SE5. ${y_i}$ is normally distributed, conditional on ${X^*}.$</td>
</tr>
</tbody>
</table>

These conditions are virtually identical to the assumptions of the longitudinal data mixed model with strictly exogenous regressors that does not contain heterogeneity terms. The difference is in the conditional variance component, SE3. In particular, the inference procedures described in Chapters 3 and 4 can be readily used in this situation.

Fixed effects estimation

As we saw in the above example that discussed exogeneity in terms of the income tax liability, there are times when the analyst is concerned with Assumption SEC6. Among other things, this assumption implies that the unobserved effects are uncorrelated with the observed regressors. Although readily accepted as the norm in the biostatistics literature, this assumption is often questioned in the economics literature. Fortunately, Assumptions SEC1-4 (and SEC5, as needed) are sufficient to allow for consistent (as well as asymptotic normality) estimation, using the fixed effects estimators described in Chapter 2. Intuitively, this is because the fixed effects estimation procedures “sweep out” the heterogeneity terms and thus do not rely on the assumption that they are uncorrelated with observed regressors. See Mundlak (1978aE) for an early contribution; Section 7.2 provides further details.

These observations suggest a strategy that is commonly used by analysts. If there is no concern that unobserved effects may be correlated with observed regressors, use the more efficient inference procedures in Chapter 3 based on mixed models and random effects. If there is a concern, use the more robust fixed effects estimators. Some analysts prefer to test the
assumption of correlation between unobserved and observed effects by examining the difference between these two estimators. This is the subject of Sections 7.2 and 7.3 where we will examine inference for the unobserved, or “omitted,” variables.

In some applications, researchers have partial information about the first part of Assumption SEC6. Specifically, we may re-arrange the observables into two pieces, \( o_i = (o_i^{(1)}, o_i^{(2)}) \) so that \( \text{Cov}(\alpha_i, o_i^{(1)}) \neq 0 \) and \( \text{Cov}(\alpha_i, o_i^{(2)}) = 0 \). That is, the first piece of \( o_i \) is correlated to the unobservables whereas the second piece is not. In this case, estimators that are neither fixed nor random effects estimators have been developed in the literature. This idea, due to Hausman and Taylor (1981E), is further pursued in Section 7.3.

**Section 6.3 Longitudinal data models with heterogeneity terms and sequentially exogenous regressors**

For some economic applications such as production function modeling (Keane and Runkle, 1992E), the assumption of strict exogeneity even when conditioning on unobserved heterogeneity terms is limiting. This is because strict exogeneity rules out current values of the response \( (y_{it}) \) feeding back and influencing future values of the explanatory variables (such as \( x_{it+1} \)). An alternative assumption introduced by Chamberlain (1992E) allows for this feedback. To introduce this assumption, we follow Chamberlain and assume random sampling so that random variables from different subjects are independent. Following this econometric literature, we assume \( q = 1 \) and \( z_{it} = 1 \) so that the heterogeneity term is an intercept. We say that the regressors are *sequentially exogenous conditional on the unobserved effects* if

\[
E (e_{it} | \alpha_i, x_{i1}, \ldots, x_{it}) = 0.
\]

(6.3)

This implies

\[
E (y_{it} | \alpha_i, x_{i1}, \ldots, x_{it}) = \alpha_i + x_{it}' \beta.
\]

After controlling for \( \alpha_i \) and \( x_{it} \), past values of regressors do not affect the expected value of \( y_{it} \).

**Lagged dependent variable model**

In addition to feedback models, this formulation allows us to consider lagged dependent variables as regressors. For example, the equation

\[
y_{it} = \alpha + \gamma y_{i,t-1} + x_{it}' \beta + e_{it},
\]

(6.4)

satisfies equation (6.3) by using the set of regressors \( o_i = (1, y_{i,t-1}, x_{it}') \) and \( E (e_{it} | \alpha_i, y_{i1}, \ldots, y_{i,t-1}, x_{i1}, \ldots, x_{it}) = 0 \). The explanatory variable \( y_{i,t-1} \) is not strictly exogenous so that the Section 6.2.2 discussion does not apply.

As will be discussed in Section 8.1, this model differs from the autoregressive error structure, the common approach in the longitudinal biomedical literature. Judged by the number of applications, this is an important dynamic panel data model in econometrics. The model is appealing because it is easy to interpret the lagged dependent variable in the context of economic modeling. For example, if we think of \( y \) as the demand of a product, it is easy to think of situations where a strong demand in the prior period \( (y_{i,t-1}) \) has a positive influence on the current demand \( (y_{it}) \), suggesting that \( \gamma \) be a positive parameter.

**Estimation difficulties**

Estimation of the model in equation (6.4) is difficult because the parameter \( \gamma \) appears in both the mean and variance structure. It appears in the variance structure because

\[
\text{Cov} (y_{it}, y_{i,t-1}) = \text{Cov} (\alpha + \gamma y_{i,t-1} + x_{it}' \beta + e_{it}, y_{i,t-1}) = \text{Cov} (\alpha, y_{i,t-1}) + \gamma \text{Var} (y_{i,t-1}).
\]
To see that it appears in the mean structure, consider equation (6.4). By recursive substitution, we have:
\[
E y_{it} = \gamma E y_{i,t+1} + x_{it}' \beta = \gamma (E y_{i,t+2} + x_{i,t+1}' \beta) + x_{it}' \beta
\]
\[
= \ldots = (x_{it}' + \gamma x_{i,t-1}' + \ldots + \gamma^{t-2} x_{i,2}' \beta) + \gamma^{t-1} E y_{i,1}.
\]
Thus, \( E y_{it} \) clearly depends on \( \gamma \).

Special estimation techniques are required for the model in equation (6.4); it is not possible to treat the lagged dependent variables as explanatory variables, either using a fixed or random effects formulation for the heterogeneity terms \( \alpha_i \). We first examine the fixed effects form, beginning with an example from Hsiao (2002E).

**Special case (Hsiao, 2002E).** Suppose that \( \alpha_i \) is treated as fixed parameter (effect) and, for simplicity, take \( K = 1 \) and \( x_{it,1} = 1 \) so that equation (6.4) reduces to
\[
y_{it} = \alpha_i^* + \gamma y_{i,t-1} + \varepsilon_{it}, \tag{6.5}
\]
where \( \alpha_i^* = \alpha_i + \beta \). The ordinary least squares estimator of \( \gamma \) turns out to be
\[
\hat{\gamma} = \frac{\sum_{t=1}^{n} \sum_{t'=2}^{T} (y_{it} - \bar{y}_{i,t})(y_{i,t-1} - \bar{y}_{i,t-1})}{\sum_{t=1}^{n} \sum_{t'=2}^{T} (y_{i,t-1} - \bar{y}_{i,t-1})^2} = \frac{\sum_{t=1}^{n} \sum_{t'=2}^{T} \varepsilon_{it} y_{i,t-1} - \bar{y}_{i,t-1}}{\sum_{t=1}^{n} \sum_{t'=2}^{T} (y_{i,t-1} - \bar{y}_{i,t-1})^2},
\]
where \( \bar{y}_{i,t-1} = \frac{\sum_{t'=1}^{T-1} y_{it}}{(T-1)} \). Now, we can argue that \( E (\varepsilon_{it} y_{i,t-1}) = 0 \) by conditioning on information available at time \( t-1 \). However, it is not true that \( E (\varepsilon_{it} \bar{y}_{i,t-1}) = 0 \), suggesting that \( \hat{\gamma} \) is biased. In fact, Hsiao demonstrates that the asymptotic bias is
\[
\lim_{n \to \infty} E \hat{\gamma} - \gamma = \frac{-1 + \gamma}{T-1} \left( 1 - \frac{1 - \gamma^T}{T} \right) \left( 1 - \frac{1 - \gamma^T}{T} \right).
\]
This bias is small for large \( T \) and tends to zero as \( T \) tends to infinity. Further, it is interesting that the bias is nonzero even when \( \gamma = 0 \).

To see the estimation difficulties in the context of the random effects model, now consider the model in equation (6.4) where \( \alpha_i \) are treated as random variables that are independent of the error terms, \( \varepsilon_{it} \). It is tempting to treat lagged response variables as explanatory variables and use the usual generalized least squares (GLS) estimators. However, this procedure also induces bias. To see this, we note that it is clear that \( y_{it} \) is a function of \( \alpha_i \) and, thus, so is \( y_{i,t+1} \). However, GLS estimation procedures implicitly assume independence of the random effects and explanatory variables. Thus, this estimation procedure is not optimal.

Although the usual generalized least squares estimators are not desirable, alternative estimators are available. To illustrate, taking first differences of the model in equation (6.5) yields
\[
y_{it} - y_{i,t-1} = \gamma (y_{i,t+1} - y_{i,t+2}) + \varepsilon_{it} - \varepsilon_{i,t-1},
\]
eliminating the heterogeneity term. Note that \( y_{i,t+1} \) and \( \varepsilon_{i,t-1} \) are clearly dependent; thus, using ordinary least squares with regressors \( \Delta y_{i,t+1} = y_{i,t+1} - y_{i,t+2} \) produces biased estimators of \( \gamma \). We can,
however, use $\Delta y_{i,t-2}$ as an instrument for $\Delta y_{i,t-1}$, because $\Delta y_{i,t-2}$ is independent of the (differenced) disturbance term $\varepsilon_{it} - \varepsilon_{i,t-1}$. This approach of differencing and using instrumental variables is due to Anderson and Hsiao (1982E). Of course, this estimator is not efficient, because the differenced error terms will usually be correlated.

Thus, first differencing proves to be a useful device for handling the heterogeneity term. To illustrate how first differencing by itself can fail, consider the following special case.

**Special case – Feedback.** Consider the error components $y_{it} = \alpha_i + x_{it}' \beta + \varepsilon_{it}$, where $\{\varepsilon_{it}\}$ are i.i.d. with variance $\sigma^2$. Suppose that the current regressors are influenced by the “feedback” from the prior period’s disturbance through the relation $x_{it} = x_{i,t-1} + \upsilon_i \varepsilon_{i,t-1}$, where $\{\upsilon_i\}$ is an i.i.d. random vector that is independent of $\{\varepsilon_{it}\}$. Taking differences of the model, we have

$$\Delta y_{it} = y_{it} - y_{i,t-1} = \Delta x_{it}' \beta + \Delta \varepsilon_{it},$$

where $\Delta \varepsilon_{it} = \varepsilon_{it} - \varepsilon_{i,t-1}$ and $\Delta x_{it} = x_{it} - x_{i,t-1} = \upsilon_i \varepsilon_{i,t-1}$. Using first differences, the ordinary least squares estimator of $\beta$ is

$$b_{FD} = \left( \frac{1}{n} \sum_{i=1}^{n} \Delta X_i \Delta X_i' \right)^{-1} \sum_{i=1}^{n} \Delta X_i \Delta y_i = \beta + \left( \frac{1}{n} \sum_{i=1}^{n} \Delta X_i \Delta X_i' \right)^{-1} \sum_{i=1}^{n} \Delta X_i \Delta \varepsilon_i,$$

where $\Delta y_i = (\Delta y_{i,2} , \ldots , \Delta y_{i,T})'$, $\Delta \varepsilon_i = (\Delta \varepsilon_{i,2} , \ldots , \Delta \varepsilon_{i,T})'$ and $\Delta X_i = (\Delta x_{i,2} , \ldots , \Delta x_{i,T}) = (\upsilon_i \varepsilon_{i,2} , \ldots , \upsilon_i \varepsilon_{i,T})$.

Straightforward calculations show that

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \Delta X_i \Delta X_i' = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \upsilon_i \upsilon_i' \sum_{t=2}^{T} \varepsilon_{it}^2 = (T-1)\sigma^2 \mathbb{E} \upsilon_i \upsilon_i',$$

and

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \Delta X_i \Delta \varepsilon_i = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \left[ \begin{array}{c} \upsilon_i \varepsilon_{i,2} \\ \vdots \\ \upsilon_i \varepsilon_{i,T-1} \\ \upsilon_i \varepsilon_{i,T} - \varepsilon_{i,T-1} \end{array} \right] = -(T-1)\sigma^2 \mathbb{E} \upsilon_i,$$

both with probability one. With probability one, this yields the asymptotic bias

$$\lim_{n \to \infty} b_{FD} - \beta = -(\mathbb{E} \upsilon_i \upsilon_i')^{-1} \mathbb{E} \upsilon_i.$$

One strategy for handling sequentially exogenous regressors with heterogeneity terms is to use a transform, such as first differencing or fixed effects, to sweep out the heterogeneity and then use instrumental variable estimation. One such treatment has been developed by Arellano and Bond (1991E). For this treatment, we assume that the responses follow the model equation

$$y_{it} = \alpha_i + x_{it}' \beta + \varepsilon_{it},$$

yet the regressors are potentially endogenous. We also assume that there exist two sets of instrumental variables. The first set, of the form $w_{1,it}$, are strictly exogenous so that

$$L( \varepsilon_{it} \mid w_{1,i1}, \ldots , w_{1,iT} ) = 0, \quad t = 1, \ldots , T_i. \quad (6.6)$$

The second set, of the form $w_{2,it}$, satisfies the following sequential exogeneity conditions

$$L( \varepsilon_{it} \mid w_{2,i1}, \ldots , w_{2,iT} ) = 0, \quad t = 1, \ldots , T_i. \quad (6.7)$$

The dimensions of $\{w_{1,it}\}$ and $\{w_{2,it}\}$ are $p_1 \times 1$ and $p_2 \times 1$, respectively. Because we will remove the heterogeneity term via transformation, we need not specify this in our linear projections. Note
that equation (6.7) implies that current disturbances are uncorrelated with current as well as past instruments.

Time-constant heterogeneity parameters are handled via sweeping out their effects, so let $K_i$ be a $(T_i - 1) \times T_i$ upper triangular matrix such that $K_i 1_i = 0_i$. For example, Arellano and Bover (1995E) recommend the matrix (suppressing the $i$ subscript)

$$K_{FOD} = \text{diag} \left( \frac{T}{\sqrt{T - 1}}, \ldots, \frac{1}{\sqrt{2}} \right) \begin{pmatrix}
1 & -\frac{1}{T-1} & -\frac{1}{T-1} & \cdots & -\frac{1}{T-1} & -\frac{1}{T-1} & -\frac{1}{T-1}
0 & 1 & -\frac{1}{T-2} & \cdots & -\frac{1}{T-2} & -\frac{1}{T-2} & -\frac{1}{T-2}
\vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots
0 & 0 & 0 & \cdots & 1 & -\frac{1}{2} & -\frac{1}{2}
0 & 0 & 0 & \cdots & 0 & 1 & -1
\end{pmatrix}.$$  

Defining $\varepsilon_{it,FOD} = K_{FOD} \varepsilon_i$, the $t$th row is

$$\varepsilon_{it,FOD} = \sqrt{\frac{T-t}{T-t+1}} \left( \varepsilon_{it} - \frac{1}{T-t} (\varepsilon_{i,t+1} + \cdots + \varepsilon_{i,T}) \right),$$

for $t = 1, \ldots, T-1$. These are known as forward orthogonal deviations. If the original disturbances are serially uncorrelated and constant variance, then so are orthogonal deviations. Preserving this structure is the advantage of forward orthogonal deviations when compared to simple differences.

To define the instrumental variable estimator, let $W_i^*$ be a block diagonal matrix with the $t$th block given by $(w_{1,t,1} \ w_{2,t,1} \ \cdots \ w_{2,t,T-1})$. That is, define

$$W_i^* = \begin{pmatrix}
w_{1,t,1} & 0 & \cdots & 0 \\
0 & (w_{1,t,1} \ w_{2,t,1} \ w_{2,t,2}) & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & (w_{1,t,1} \ w_{2,t,1} \ w_{2,t,2} \ \cdots \ w_{2,t,T-1})
\end{pmatrix}.$$  

With this notation, it can be shown that the sequentially exogeneity assumption in equations (6.6) and (6.7) implies $E W_i^* K_i \varepsilon_i = 0$. Let $W_i = (W_i^* : 0_i)$, where $W_i$ has dimensions $(T_i - 1) \times (p_1 T_i + p_2 T (T + 1)/2)$, and $T = \max (T_1, \ldots, T_n)$. This zero matrix augmentation is needed when we have unbalanced data; it is not needed if $T_i = T$.

**Special case - Feedback – Continued.** A natural set of instruments is to choose $w_{it} = x_{1,it}$. For simplicity, also use first differences in our choice of $K$. Thus,

$$K_{FD} \varepsilon_i = \begin{pmatrix}
-1 & 1 & 0 & \cdots & 0 & 0 & 0 \\
0 & -1 & 1 & \cdots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & -1 & 1 & 0 \\
0 & 0 & 0 & \cdots & 0 & -1 & 1
\end{pmatrix} \begin{pmatrix}
\varepsilon_{i0} \\
\varepsilon_{i1} \\
\vdots \\
\varepsilon_{iT} \\
\varepsilon_{i,T-1}
\end{pmatrix} = \begin{pmatrix}
\varepsilon_{i1} - \varepsilon_{i0} \\
\varepsilon_{i2} - \varepsilon_{i1} \\
\vdots \\
\varepsilon_{i,T} - \varepsilon_{i,T-1}
\end{pmatrix}.$$  

With these choices, the $t$th block of $E W_i^* K_{FD} \varepsilon_i$ is

$$E \begin{pmatrix}
x_{i0} \\
\vdots \\
x_{iT}
\end{pmatrix} (\varepsilon_{i,t+1} - \varepsilon_{i,t}) = \begin{pmatrix}
0 \\
\vdots \\
0
\end{pmatrix}.$$
so the exogeneity assumption in equation (6.6) is satisfied.

**Special case – Lagged dependent variable.** Consider the model in equation (6.5). In this case, one can choose \( w_{2,t} = y_{t-1} \), and, for simplicity, use first differences in our choice of \( K \). With these choices, the \( r \)th block of \( \mathbb{E} W_2 K_{FD} \varepsilon_t \) is

\[
\begin{bmatrix}
y_{1t} \\
y_{1,t-1} \\
\vdots \\
\vdots \\
0
\end{bmatrix} = \begin{bmatrix}
0 \\
0 \\
\vdots \\
\vdots \\
1
\end{bmatrix},
\]

so the sequentially exogeneity assumption in equation (6.7) is satisfied.

Now, define the matrices \( M_{wX} = \sum_{i=1}^{n} W_i' K_i X_i \) and \( M_{wY} = \sum_{i=1}^{n} W_i' K_i y_i \). With this notation, we define the instrumental variable estimator as

\[
b_{IV} = \left( M_{wX}' \Sigma_{IV}^{-1} M_{wX} \right)^{-1} M_{wX}' \Sigma_{IV}^{-1} M_{wY}, \tag{6.8}
\]

where \( \Sigma_{IV} = \mathbb{E} \left( W_i' K_i \varepsilon_t K_i' W_i \right) \). This estimator is consistent and asymptotically normal, with asymptotic covariance matrix

\[
\text{Var} b_{IV} = \left( M_{wX}' \Sigma_{IV}^{-1} M_{wX} \right)^{-1}.
\]

Both the estimator and the asymptotic variance rely on the unknown matrix \( \Sigma_{IV} \). To compute a first stage estimator, we may assume that the disturbances are serially uncorrelated and homoscedastic and thus use

\[
b_{IV} = \left( M_{wX}' \hat{\Sigma}_{IV,1}^{-1} M_{wX} \right)^{-1} M_{wX}' \hat{\Sigma}_{IV,1}^{-1} M_{wY},
\]

where \( \hat{\Sigma}_{IV,1} = \sum_{i=1}^{n} W_i' K_i' K_i W_i \). As is the usual case with generalized least squares estimators, this estimator is invariant to the estimator of the scale \( \text{Var} \varepsilon \). To estimate this scale parameter, use the residuals \( e_i = y_i - x_i' b_{IV} \). An estimator of \( \text{Var} b_{IV} \) that is robust to the assumption of no serial correlation and homoscedasticity of the disturbances is

\[
\hat{\Sigma}_{IV,2} = \frac{1}{n} \sum_{i=1}^{n} W_i' K_i e_i e_i' K_i' W_i,
\]

where \( e_i = (e_{i1}, \ldots, e_{iT})' \).

**Example – Tax liability – Continued**

In Section 6.2.2, we suggested that a heterogeneity term may be due to an individual’s earning potential and that this may be correlated with the variable that indicates use of a professional tax preparer. Moreover, there was concern that tax liabilities from one year may influence the choice in subsequent year’s choice of whether or not to use a professional tax preparer. Instrumental variable estimators provide protection against these endogeneity concerns.

Table 6.1 summarizes the fit of two dynamic models. Both models use heterogeneity terms and lagged dependent variables. One model assumes that all demographic and economic variables are strictly exogenous (so are used as \( w_{1,i} \)); the other models assumes that demographic variables are strictly exogenous (so are used as \( w_{1,i,n} \)) but that the economic variables are
sequentially exogenous (so are used as $w_{2,t}$). For the second model, we also present robust $t$-statistics (based on the variance-covariance matrix $\hat{\Sigma}_{IV,2}$) in addition to the usual “model-based” $t$-statistics (based on the variance-covariance matrix $\hat{\Sigma}_{IV,1}$). These models were fit using the statistical/econometric package STATA.

Table 6.1 shows that the lagged dependent variable was statistically significant for both models and methods of calculating $t$-statistics. Of the demographic variables, only the head of household variables was statistically significant, and this was not even true under the model treating economic variables as sequentially exogenous and using robust $t$-statistics. Of the economic variables, neither EMP nor PREP was statistically significant, whereas MR was statistically significant. The other measure of income, LNTPI, was statistically significant when treated as strictly exogenous but not when treated as sequentially exogenous. Because the main purpose of this study was to study the effect of PREP on LNTAX, the effects of LNTPI were not further investigated.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Parameter Estimates</th>
<th>Model-based $t$-statistic</th>
<th>Parameter Estimates</th>
<th>Model-based $t$-statistic</th>
<th>Robust $t$-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lag LNTAX</td>
<td>0.205</td>
<td>4.26</td>
<td>0.108</td>
<td>3.13</td>
<td>2.48</td>
</tr>
<tr>
<td>Demographic Variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MS</td>
<td>-0.351</td>
<td>-0.94</td>
<td>-0.149</td>
<td>-0.42</td>
<td>-0.49</td>
</tr>
<tr>
<td>HH</td>
<td>-1.236</td>
<td>-2.70</td>
<td>-1.357</td>
<td>-3.11</td>
<td>-1.71</td>
</tr>
<tr>
<td>AGE</td>
<td>-0.160</td>
<td>-0.34</td>
<td>0.010</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>DEPEND</td>
<td>0.026</td>
<td>0.21</td>
<td>0.084</td>
<td>0.73</td>
<td>0.68</td>
</tr>
<tr>
<td>Economic Variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LNTPI</td>
<td>0.547</td>
<td>4.53</td>
<td>0.340</td>
<td>1.91</td>
<td>1.07</td>
</tr>
<tr>
<td>MR</td>
<td>0.116</td>
<td>8.40</td>
<td>0.143</td>
<td>7.34</td>
<td>5.21</td>
</tr>
<tr>
<td>EMP</td>
<td>0.431</td>
<td>1.22</td>
<td>0.285</td>
<td>0.48</td>
<td>0.36</td>
</tr>
<tr>
<td>PREP</td>
<td>-0.272</td>
<td>-1.20</td>
<td>-0.287</td>
<td>-0.78</td>
<td>-0.68</td>
</tr>
<tr>
<td>INTERCEPT</td>
<td>0.178</td>
<td>4.21</td>
<td>0.215</td>
<td>4.90</td>
<td>3.41</td>
</tr>
</tbody>
</table>

As with other instrumental variable procedures, a test of the exogeneity assumption $E W_i' K_i e_i = 0$, is available. The robust version of the test statistic is

$$TS_{IV} = \left( \sum_{i=1}^{n} e_i' K_i W_i \right) \hat{\Sigma}_{IV,2}^{-1} \left( \sum_{i=1}^{n} W_i' K_i e_i \right).$$

Under the null hypothesis of $E W_i' K_i e_i = 0$, this test statistic has an asymptotic chi-square distribution with $(p_1 T + p_2 (T + 1)/2) - K$ degrees of freedom (see Arellano and Honoré, 2001E). Moreover, one can use incremental versions of this test statistic to assess the exogeneity of selected variables in the same manner as partial $F$-test. This is important because the number of moment conditions increases substantially as one considers modeling a variable as strictly exogenous (that uses $T$ moment conditions) compared to the less restrictive sequential exogeneity assumption (that uses $T (T + 1)/2$ moment conditions). For additional discussion on testing exogeneity using instrumental variable estimators, we refer to Arellano (2003E), Baltagi (2002E) and Wooldridge (2002E).
Section 6.4 Multivariate responses

As with Section 6.1, we begin by reviewing ideas from a non-longitudinal setting, specifically, cross-sectional data. Section 6.4.1 introduces multivariate responses in the context of multivariate regression. Section 6.4.2 describes the relation between multivariate regression and sets of regression equations. Section 6.4.3 introduces simultaneous equations methods. Section 6.4.4 then applies these ideas to systems of equation with error components that have been proposed in the econometric literature.

Section 6.4.1 Multivariate regression

A classic method of modeling multivariate responses is through a multivariate regression model. We start with a general expression

\[ Y = X \Gamma' + \varepsilon, \]  

(6.9)

where \( Y \) is an \( n \times G \) matrix of responses, \( X \) is an \( n \times K \) matrix of explanatory variables, \( \Gamma \) is a \( G \times K \) matrix of parameters and \( \varepsilon \) is an \( n \times G \) matrix of disturbances. To provide intuition, consider the (transpose of the) \( i \)th row of equation (6.9),

\[ y_i = \Gamma x_i + \varepsilon_i, \quad i = 1, \ldots, n. \]  

(6.10)

Here, \( G \) responses are measured for each subject, \( y_i = (y_{1i}, \ldots, y_{Gi})' \), whereas the vector of explanatory variables, \( x_i \), is \( K \times 1 \). We initially assume that the disturbance terms are identically and independently distributed with variance-covariance matrix \( \text{Var} \varepsilon_i = \Sigma \).

Example 6.1 Supply and demand

To illustrate, consider a sample of \( n \) countries; for each country, we measure \( G = 2 \) responses. Specifically, we wish to relate \( y_1 \), the price of a good provided by suppliers, and \( y_2 \), the quantity of this good demanded, to several exogenous measures (\( x \)'s) such as income, the price of substitute goods, and so on. We assume that price and quantity may be related through

\[ \text{Var} y_i = \text{Var} \begin{pmatrix} y_{1i} \\ y_{2i} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} = \Sigma. \] 

Specifically, \( \sigma_{12} \) measures the covariance between price and quantity. Using \( \Gamma = (\beta_1, \beta_2)' \), the expression for equation (6.10) is

\[ y_{1i} = \beta_1' x_i + \varepsilon_{1i} \quad \text{price} \]

\[ y_{2i} = \beta_2' x_i + \varepsilon_{2i} \quad \text{quantity}. \]

The ordinary least squares regression coefficient estimator of \( \Gamma \) is

\[ G_{OLS} = \left( \sum_{i=1}^{n} y_i x_i' \right) \left( \sum_{i=1}^{n} x_i x_i' \right)^{-1}. \]  

(6.11)

Somewhat surprisingly, it turns out that this estimator is also the generalized least squares estimator and, hence, the maximum likelihood estimator.

Now, let \( \beta_g' \) be the \( g \)th row of \( \Gamma \) so that \( \Gamma = (\beta_1, \beta_2, \ldots, \beta_g)' \). With this notation, the \( g \)th row of equation (6.10) is

\[ y_{gi} = \beta_g' x_i + \varepsilon_{gi}, \quad i = 1, \ldots, n. \]

We can calculate the ordinary least squares estimator of \( \beta_g \) as
Thus, the estimator $\mathbf{b}_{g,OLS}$ can be calculated on a row-by-row basis, that is, using standard (univariate response) multiple linear regression software. Nonetheless, the multivariate model structure has important features. To illustrate, by considering sets of responses simultaneously in equation (6.10), we can account for relationships among responses in the covariance of the regression estimators. For example, with equation (6.10), it is straightforward to show that

$$\text{Cov}(\mathbf{b}_{g,OLS}, \mathbf{b}_{k,OLS}) = \sigma_{gk} \left( \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \right)^{-1}.$$

Multivariate regression analysis can be applied directly to longitudinal data by treating each observation over time as one of the $G$ responses. This allows for a variety of serial correlation matrices through the variance matrix $\Sigma$. However, this perspective requires balanced data ($G = T$).

Section 6.4.2 Seemingly unrelated regressions

Even for cross-sectional data, a drawback of the classic multivariate regression is that the same set of explanatory variables is required for each type of response. To underscore this point, we return to our supply and demand example.

Example 6.1 Supply and demand - continued

Suppose now that the expected price of a good ($y_1$) depends linearly on $x_1$, the purchasers’ income; similarly, quantity ($y_2$) depends on $x_2$, the suppliers’ wage rate. That is, we wish to consider

$$y_{1i} = \beta_{10} + \beta_{11} x_{1i} + \epsilon_{1i}, \quad \text{(price)}$$
$$y_{2i} = \beta_{20} + \beta_{21} x_{2i} + \epsilon_{2i}, \quad \text{(quantity)}$$

so that different explanatory variables are used in different equations.

Let us reorganize our observed variables so that $\mathbf{y}_g = (y_{1g}, \ldots, y_{ng})'$ is the vector of the $g$th response and $\mathbf{X}_g$ is the $n \times K_g$ matrix of explanatory variables. This yields two sets of regression equations

$$\mathbf{y}_1 = \mathbf{X}_1 \mathbf{b}_1 + \mathbf{e}_1, \quad \text{(price)}$$
$$\mathbf{y}_2 = \mathbf{X}_2 \mathbf{b}_2 + \mathbf{e}_2, \quad \text{(quantity)}$$

representing $2n$ equations. Here, $\text{Var}\mathbf{y}_1 = \sigma_{11}^2 \mathbf{I}_n$, $\text{Var}\mathbf{y}_2 = \sigma_{22}^2 \mathbf{I}_n$, and $\text{Cov}(\mathbf{y}_1, \mathbf{y}_2) = \sigma_{12} \mathbf{I}_n$. Thus, there is zero covariance between different countries yet a common covariance within a country between price and quantity of the good.

If we run separate regressions, we get regression coefficient estimators

$$\mathbf{b}_{g,OLS} = (\mathbf{X}_g' \mathbf{X}_g)^{-1} \mathbf{X}_g' \mathbf{y}_g \quad g = 1, 2.$$

These are ordinary least squares estimators; they do not account for the information in $\sigma_{12}$.

The seemingly unrelated regressions technique, attributed to Zellner (1962E), combines different sets of regression equations and uses generalized least squares estimation. Specifically,
suppose that we start with \( G \) sets of regression equations of the form \( y_g = X_g \beta_g + \varepsilon_g \). To see how to combine these, we work with \( G = 2 \) and define

\[
\begin{bmatrix}
  y_1 \\
  y_2 
\end{bmatrix}, \quad \begin{bmatrix}
  X_1 & 0 \\
  0 & X_2 
\end{bmatrix}, \quad \begin{bmatrix}
  \beta_1 \\
  \beta_2 
\end{bmatrix} \quad \text{and} \quad \begin{bmatrix}
  \varepsilon_1 \\
  \varepsilon_2 
\end{bmatrix}.
\]

(6.12)

Thus, we have

\[
\text{Var } y = \text{Var } \varepsilon = \begin{pmatrix}
  \sigma_1^2 I_n & \sigma_{12} I_n \\
  \sigma_{12}^* I_n & \sigma_2^2 I_n
\end{pmatrix}
\]

and, with this, the generalized least squares estimator is

\[
b_{GLS} = (X' \text{Var } y)^{-1} X' \text{Var } y^{-1} y.
\]

These are known as the \( SUR \), for seemingly unrelated regression, estimators. It is easy to check that \( b_{GLS} = \begin{pmatrix} b_{1,OLS} \\ b_{2,OLS} \end{pmatrix} \) if either (i) \( \sigma_{12} = 0 \) or (ii) \( X_1 = X_2 \) holds. In either case, we have that the GLS estimator is equivalent to the OLS estimator.

On one hand, the seemingly unrelated regression set-up can be viewed as a special case of multiple linear, and hence multivariate, regression, with display (6.12). On the other hand, seemingly unrelated regressions can be viewed as a way of extending multivariate regressions to allow for explanatory variables that depend on the type of response. As we will see, another way of allowing type specific explanatory variables is to restrict the parameter matrix.

**Section 6.4.3 Simultaneous equations models**

In our supply and demand example, we assumed that price and quantity were potentially related through covariance terms. However, it is often the case that researchers wish to estimate alternative models that allow for relationships among responses directly through the regression equations.

**Example 6.1 Supply and demand - continued**

Consider a demand-supply model that yields the following two sets of equations:

\[
\begin{align*}
  y_{1i} &= \beta_1 y_{2i} + \gamma_{10} + \gamma_{11} x_{1i} + \varepsilon_{1i} & \text{(price)} \\
  y_{2i} &= \beta_2 y_{1i} + \gamma_{20} + \gamma_{21} x_{2i} + \varepsilon_{2i} & \text{(quantity)}
\end{align*}
\]

(6.13)

Here, we assume that quantity linearly affects price, and vice-versa. As before, both \( x \)'s are assumed to be exogenous for the demand and supply equations.

In Section 6.1.3, we saw that using only ordinary least squares in a single equation produced biased estimators due to the endogenous regressor variables. That is, when examining the price equation, the quantity variable \( (y_{2i}) \) is clearly endogenous because it is influenced by price, as is seen in the quantity equation. One can use similar reasoning to argue that SUR estimators also yield biased and inconsistent estimators; seemingly unrelated regression techniques improve upon the efficiency of ordinary least squares but do not change the nature of the bias in estimators.

To introduce estimators for the equations in display (6.13), we collect the dependent variables with the matrix \( B = \begin{pmatrix} 0 & \beta_1 \\ \beta_2 & 0 \end{pmatrix} \). Thus, we may express display (6.13) as \( y_i = B y_i + \Gamma x_i \).
+ \epsilon_i$, where $\epsilon_i = (\epsilon_{i1} \epsilon_{i2})'$, $\mathbf{x}_i = (1 \ x_{1i} \ x_{2i})'$ and $\Gamma = \begin{pmatrix} \gamma_{10} & \gamma_{11} & 0 \\ \gamma_{20} & 0 & \gamma_{21} \end{pmatrix}$. This expression looks very much like multivariate regression model in equation (6.10), the difference being that we now have included a set of endogenous regressors, $\mathbf{B}_i \mathbf{y}_i$. As noted in Section 6.4.2, we have incorporated different regressors in different equations by defining a “combined” set of explanatory variables and imposing the appropriate restrictions on the matrix of coefficients, $\beta$.

This subsection considers systems of regression equations where responses from one equation may serve as endogenous regressors in another equation. Specifically, we consider model equations of the form

$$y_i = \mathbf{B}_i \mathbf{y}_i + \Gamma \mathbf{x}_i + \epsilon_i. \quad (6.14)$$

Here, we assume that $\mathbf{I} - \mathbf{B}$ is a $G \times G$ non-singular matrix, that $\Gamma$ has dimension $G \times K$ and the vector of explanatory variables, $\mathbf{x}_i$, is $K \times 1$ and that $\text{Var} \ \epsilon_i = \Sigma$. With these assumptions, we may write the so-called “reduced form”

$$y_i = \Pi \mathbf{x}_i + \eta_i,$$

where $\Pi = (\mathbf{I} - \mathbf{B})^{-1} \Gamma$, $\eta_i = (\mathbf{I} - \mathbf{B})^{-1} \epsilon_i$ and $\text{Var} \ \eta_i = \text{Var} \ [(\mathbf{I} - \mathbf{B})^{-1} \epsilon_i] = (\mathbf{I} - \mathbf{B})^{-1} \Sigma (\mathbf{I} - \mathbf{B})^{-1} \epsilon_i$. For example, in our supply-demand example, we have

$$\Pi = \frac{1}{1 - \beta_1 \beta_2} \begin{pmatrix} \gamma_{10} + \beta_1 \gamma_{20} & \gamma_{11} \\ \beta_2 \gamma_{10} + \gamma_{20} & \beta_2 \gamma_{11} \end{pmatrix} \begin{pmatrix} 1 \\ \beta_2 \end{pmatrix}$$

The reduced form is simply a multivariate regression model as in equation (6.10). We will assume sufficient conditions on the observables to consistently estimate the reduced form coefficients $\Pi$ and the corresponding variance-covariance matrix $\Omega$. Thus, we will have information on the $GK$ elements in $\Pi$ and the $G(G+1)/2$ elements of $\Omega$. However, this information in and of itself will not allow us to properly identify all the elements of $\Gamma$, $\mathbf{B}$ and $\Sigma$. There are $GG$, $G^2$ and $G(G+1)/2$ elements in these matrices, respectively. Additional restrictions, generally from economic theory, are required. To illustrate, in our supply demand example, there are six structural parameters of interest in $\Gamma$ and six elements of $\Pi$. Thus, we need to check to see that this provides sufficient information to recover the relevant structural parameters. This process is known as identification. Detailed treatments of this topic are available in many sources; see, for example, Greene (2002E), Hayashi (2000E) and Wooldridge (2002E).

Estimates of $\Pi$ allow us to recover the structural parameters in $\Gamma$ and $\mathbf{B}$. This method of estimating the structural parameters is known as indirect least squares. Alternatively, it is possible to estimate equation (6.14) directly using maximum likelihood theory. However, this becomes complex because the parameters in $\mathbf{B}$ appear in both the mean and variance.

Not surprisingly, many alternative estimation procedures are available. A commonly used method is two-stage least squares, introduced in Section 6.1.3. For the first stage of this procedure, one runs all the exogenous variables to fit the responses. That is, using equation (6.11), calculate

$$\hat{\mathbf{y}}_i = \mathbf{G}_{OLS} \mathbf{x}_i = \left( \sum_{i=1}^{n} \mathbf{y}_i \mathbf{x}_i' \right) \left( \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \mathbf{x}_i. \quad (6.15)$$

For the second stage, assume that we can write the $g$th row of equation (6.14) as

$$y_{gi} = \mathbf{B}_g' \mathbf{y}_{(g)} + \beta'_{g} \mathbf{x}_i + \epsilon_{gi} \quad \quad (6.16)$$
where \( y_{ig} \) is \( y_i \) with the \( g \)th row omitted and \( B_g \) is the transpose of the \( g \)th row of \( B \), omitting the diagonal element. Then, we may calculate ordinary least squares estimators corresponding to the equation

\[
y_{gi} = B_g' \hat{y}_{i(g)} + \beta_g' x_i + \text{residual}.
\]  

(6.17)

The fitted values, \( \hat{y}_{i(g)} \), are determined from \( y_i \) in equation (6.15), after removing the \( g \)th row. Ordinary least squares in equation (6.16) is inappropriate because of the endogenous regressors in \( y_{ig} \). Because the fitted values, \( \hat{y}_{i(g)} \), are linear combinations of the exogenous variables, there is no such endogeneity problem.

Using equation (6.17), we can express the two-stage least squares estimators of the structural parameters as

\[
\begin{pmatrix} \hat{B}_g \\ \hat{\beta}_g \end{pmatrix} = \left( \sum_{i=1}^{n} \begin{pmatrix} \hat{y}_{i(g)} \\ x_i \end{pmatrix} \begin{pmatrix} \hat{y}_{i(g)} \\ x_i \end{pmatrix}' \right)^{-1} \sum_{i=1}^{n} \begin{pmatrix} \hat{y}_{i(g)} \\ x_i \end{pmatrix} y_{gi}, \quad g = 1, \ldots, G.
\]  

(6.18)

Note that for this estimation methodology to work, the number of exogenous variables excluded from the \( g \)th row must be at least as large as the number of endogenous variables that appear in \( B_g' \hat{y}_{i(g)} \).

**Example 6.1 Supply and demand - continued**

To illustrate, we return to our demand-supply example. Then, for \( g = 1 \), we have \( B_g = y_1 \), \( y_{gi} = y_{2i} \), \( B_g' = (\gamma_1 0) \), and \( x_i = (1 x_{1i} x_{2i})' \). We calculate fitted values for \( y_{1i} \) as

\[
\hat{y}_{1i} = \left( \sum_{i=1}^{n} y_{ii} x_i' \right)^{-1} \sum_{i=1}^{n} y_i x_i'.
\]

Similar calculations hold for \( g = 2 \). Then, straightforward substitution into equation (6.18) yields the two-stage least squares estimators.

**Section 6.4.4 Systems of equations with error components**

This section describes panel data extensions to systems of equations involving error components to model the heterogeneity. We first examine seemingly unrelated regression models and then simultaneous equation models. Only one-way error components are dealt with. Interested readers will find additional details and summaries of extensions to two-way error components in Baltagi (2001E) and Krishnakumar (Chapter 9 of Mátyás and Sevestre, 1996E).

**Seemingly unrelated regression models with error components**

Like seemingly unrelated regressions in the cross-sectional setting, we now wish to study several error component models simultaneously. Following the notation of Section 3.1, our interest is in studying situations that can be represented by

\[
y_{gt} = \alpha_g + x_{gt}' \beta_g + \epsilon_{gt}, \quad g = 1, \ldots, G.
\]

In our supply demand example, \( g \) represented the price and quantity equations whereas \( i \) represented the country. We now assume that we follow countries over time, so that \( t = 1, \ldots, T_i \). Assuming that the \( x \)'s are the only exogenous variables and that type and country specific random
effects, \( \alpha_{gi} \), are independent of the disturbances terms, one can always use ordinary least squares estimators of \( \beta_{gi} \); these are unbiased and consistent.

To compute the more efficient generalized least squares, we begin by stacking over the \( G \) responses,

\[
\begin{pmatrix}
  y_{1t} \\
  \vdots \\
  y_{Gt}
\end{pmatrix} = \begin{pmatrix}
  \alpha_{i1} \\
  \vdots \\
  \alpha_{Gi}
\end{pmatrix} + \begin{pmatrix}
  x_{1t}' & 0 & 0 \\
  \vdots & \ddots & \vdots \\
  0 & \cdots & 0
\end{pmatrix} \begin{pmatrix}
  \beta_1 \\
  \vdots \\
  \beta_G
\end{pmatrix} + \begin{pmatrix}
  \epsilon_{1t} \\
  \vdots \\
  \epsilon_{Gt}
\end{pmatrix},
\]

that we write as

\[
y_{it} = \alpha_i + X_{it} \beta + \epsilon_{it}. \tag{6.19}
\]

Here, we assume that \( \beta \) has dimension \( K \times 1 \) so that \( X_{it} \) has dimension \( G \times K \). Following conventional seemingly unrelated regressions, we may allow for covariances among responses through the notation \( \text{Var} \epsilon_{it} = \Sigma \). We may also allow for covariances through \( \text{Var} \alpha_i = \mathbf{D} \). Stacking over \( t \), we have

\[
\begin{pmatrix}
  y_{i1} \\
  \vdots \\
  y_{iT_i}
\end{pmatrix} = \begin{pmatrix}
  \alpha_i \\
  \vdots \\
  \alpha_{iT_i}
\end{pmatrix} + \begin{pmatrix}
  X_{i1} \\
  \vdots \\
  X_{iT_i}
\end{pmatrix} \beta + \begin{pmatrix}
  \epsilon_{i1} \\
  \vdots \\
  \epsilon_{iT_i}
\end{pmatrix},
\]

that we write as

\[
y_i = \alpha_i \otimes \mathbf{1}_{T_i} + X_i \beta + \epsilon_i. \tag{6.20}
\]

With this notation, note that \( \text{Var} \epsilon_i = \text{blkdiagonal}(\text{Var} \epsilon_{i1}, \ldots, \text{Var} \epsilon_{iT_i}) = \Sigma \otimes \mathbf{I}_{T_i} \) and that \( \text{Var} (\alpha_i \otimes \mathbf{1}_{T_i}) = \mathbf{E} (\alpha_i \otimes \mathbf{1}_{T_i}) (\alpha_i \otimes \mathbf{1}_{T_i})' = \mathbf{D} \otimes \mathbf{J}_{T_i} \). Thus, the generalized least squares estimator for \( \beta \) is

\[
b_{\text{GLS}} = \left( \sum_{i=1}^n X_i' (\mathbf{D} \otimes \mathbf{J}_{T_i} + \Sigma \otimes \mathbf{I}_{T_i})^{-1} X_i \right)^{-1} \left( \sum_{i=1}^n X_i' (\mathbf{D} \otimes \mathbf{J}_{T_i} + \Sigma \otimes \mathbf{I}_{T_i})^{-1} y_i \right). \tag{6.20}
\]

Avery (1977E) and Baltagi (1980E) considered this model.

**Simultaneous equation models with error components**

Extending the notation of equation (6.14), we now consider

\[
y_{it} = \alpha_i + \mathbf{B} y_{it} + \Gamma x_{it} + \epsilon_{it}. \tag{6.21}
\]

The subject-specific term is \( \alpha_i = (a_{i1}, a_{i2}, \ldots, a_{ic})' \) that has mean zero and variance-covariance matrix \( \text{Var} \alpha_i = \mathbf{D} \). We may re-write equation (6.21) in reduced form as

\[
y_{it} = \Pi x_{it} + \eta_{it}
\]

where \( \Pi = (\mathbf{I} - \mathbf{B})^{-1} \Gamma \) and \( \eta_{it} = (\mathbf{I} - \mathbf{B})^{-1} (\alpha_i + \epsilon_{it}) \). With this formulation, we see that the panel data mean effects are the same as the model in equation (6.14) without subject-specific effects. Specifically, as pointed out by Hausman and Taylor (1981E), without additional restrictions on the variance or covariance parameters, the identification issues are the same with and without subject-specific effects. In addition, estimation of the reduced form is similar; details are provided in the review by Krishnakumar (Chapter 9 of Mátyás and Sevestre, 1996E).

We now consider direct estimation of the structural parameters using two-stage least squares. To begin, note that the two-stage least squares estimators described in equation (6.18) still provide unbiased, consistent estimators. However, they do not account for the error variance-covariance structure, and thus, may be inefficient. Nonetheless, these estimators can be
used to calculate estimators of the variance components that will be used in the following estimation procedure.

For the first stage, we need to calculate fitted values of the responses, using only the exogenous variables as regressors. Note that

\[
(\mathbf{I} - \mathbf{B})\Gamma \mathbf{x}_{it} = \begin{pmatrix}
\beta'_{it} \mathbf{x}_{it} \\
\vdots \\
\beta'_{G} \mathbf{x}_{it}
\end{pmatrix}
= \begin{pmatrix}
\begin{pmatrix}
\mathbf{x}'_{it} & 0 & 0
\end{pmatrix} \\
\begin{pmatrix}
0 & \ddots & 0
\end{pmatrix} \\
0 & 0 & \mathbf{x}'_{it}
\end{pmatrix}
= \mathbf{X}_{it}' \beta^*.
\]

Thus, with equation (6.21), we may express the reduced form as

\[
\mathbf{y}_i = (\mathbf{I} - \mathbf{B})^\dagger \mathbf{a}_i + (\mathbf{I} - \mathbf{B})^\dagger \Gamma \mathbf{x}_{it} + (\mathbf{I} - \mathbf{B})^\dagger \mathbf{e}_i = \mathbf{a}_i^* + \mathbf{X}_{it}^* \mathbf{\beta}^* + \mathbf{e}_i^*.
\]

This has the same form as the seemingly unrelated regression with error components model in equation (6.19). Thus, one can use equation (6.20) to get fitted regression coefficients and thus fitted values. Alternatively, we have seen that ordinary least squares provides unbiased and consistent estimates for this model. Thus, this technique would also serve for computing the first stage fitted values.

For the second stage, write the model in equation (6.21) in the same fashion as equation (6.16) to get

\[
\mathbf{y}_{git} = \mathbf{y}_{i(g)}' \mathbf{B}_g + \mathbf{x}_{it}^* \mathbf{\beta}_g + \mathbf{a}_{gi} + \mathbf{e}_{git}.
\]

Recall that \( \mathbf{B}_g \) is a \((G-1) \times 1\) vector of parameters and \( \mathbf{y}_{i(g)} \) is \( \mathbf{y}_i \) with the \( g \)th row omitted. Stacking over \( t = 1, \ldots, T \) yields

\[
\mathbf{y}_{gi} = \mathbf{Y}_{i(g)} \mathbf{B}_g + \mathbf{X}_i \mathbf{\beta}_g + \mathbf{a}_{gi} \mathbf{1}_T + \mathbf{e}_{gi}.
\]

Let \( \text{Var}(\mathbf{a}_{gi} \mathbf{1}_T + \mathbf{e}_{gi}) = \sigma^2_{a_i} \mathbf{J}_T + \sigma^2_{e_i} \mathbf{I}_T \). Replacing \( \mathbf{Y}_{i(g)} \) by \( \hat{\mathbf{Y}}_{i(g)} \) yields the two-stage least squares estimators

\[
\begin{pmatrix}
\hat{\mathbf{B}}_g' \\
\hat{\mathbf{\beta}}_g
\end{pmatrix}
= \sum_{i=1}^{n}\begin{pmatrix}
\hat{\mathbf{Y}}_{i(g)} \\
\mathbf{X}_i
\end{pmatrix}' \begin{pmatrix}
\sigma^2_{a_i} \mathbf{J}_T + \sigma^2_{e_i} \mathbf{I}_T
\end{pmatrix}^{-1} \begin{pmatrix}
\hat{\mathbf{Y}}_{i(g)} \\
\mathbf{X}_i
\end{pmatrix}^{-1} \sum_{i=1}^{n}\begin{pmatrix}
\hat{\mathbf{Y}}_{i(g)} \\
\mathbf{X}_i
\end{pmatrix}.'
\]

Section 6.5 Simultaneous equation models with latent variables

Simultaneous equation models with latent variables comprise a broad framework for handling complex systems of equations, as well as longitudinal data models. This framework, that originated in the work of Jöreskog, Keesling and Wiley (see Bollen, 1989EP, page 6), is widely applied in sociology, psychology and educational sciences. Like the introductions to simultaneous equations in Sections 6.4.3 and 6.4.4, the systems of equations may be used for both different types of responses (multivariate) as well as different times of responses (longitudinal), or both. The estimation techniques preferred in psychology and education are known as covariance structure analysis.

To keep our treatment self-contained, we first outline the framework in Section 6.5.1 in the cross-sectional context. Section 6.5.2 then describes longitudinal data applications.
Section 6.5.1 Cross-sectional models

We begin by assuming that \((x_i, y_i)\) are identically and independently distributed observable draws from a population. We treat \(x\) as the exogenous vector and \(y\) is the endogenous vector.

The \textit{x-measurement equation} is

\[ x_i = \tau_x + \Lambda_x \xi_i + \delta_i. \quad (6.22) \]

Here, \(\delta\) is the disturbance term, \(\xi\) is a vector of latent (unobserved) variables and \(\Lambda_x\) is a matrix of regression coefficients that relates \(\xi\) to \(x\). In a measurement error context, we might use \(\Lambda_x = I\) and interpret \(\tau_x + \xi\) to be the “true” values of the observables that are corrupted by the “error” \(\delta\). In other applications, ideas from classic factor analysis drive the model set-up. That is, there may be many observable exogenous measurements that are driven by relatively few underlying latent variables. Thus, the dimension of \(x\) may be large relative to the dimension of \(\xi\). In this case, we can reduce the dimension of the problem by focusing on the latent variables. Moreover, the latent variables more closely correspond to social science theory than do the observables.

To complete the specification of this measurement model, we assume that \(E \xi_i = \mu\) and \(E \delta_i = 0\) so that \(E x_i = \tau_x + \Lambda_x \mu\). Further, we use \(\text{Var} \delta_i = \Theta_\delta\) and \(\text{Var} \xi_i = \Phi_i\) so that \(\text{Var} x_i = \Lambda_x \Phi \Lambda_x' + \Theta_\delta\). That is, \(\xi\) and \(\delta\) are mutually uncorrelated.

Specification of the \textit{y-measurement equation} is similar. Define

\[ y_i = \tau_y + \Lambda_y \eta_i + \epsilon_i. \quad (6.23) \]

Here, \(\epsilon\) is the disturbance term, \(\eta\) is a vector of latent variables and \(\Lambda_y\) is a matrix of regression coefficients that relates \(\eta\) to \(y\). We use \(\Theta_\epsilon = \text{Var} \epsilon_i\) and assume that \(\eta\) and \(\epsilon\) are mutually uncorrelated. Note that equation (6.23) is not a usual multiple linear regression equation because regressor \(\eta\) is unobserved.

To link the exogenous and endogenous latent variables, we have the \textit{structural equation}

\[ \eta_i = \tau_\eta + B \eta_i + \Gamma \xi_i + \zeta_i. \quad (6.24) \]

Here, \(\tau_\eta\) is a fixed intercept, \(B\) and \(\Gamma\) are matrices of regression parameters and \(\zeta\) is a mean zero disturbance term with second moment \(\text{Var} \zeta = \Psi\).

A drawback of a structural equation model with latent variables, summarized in equations (6.22)-(6.24), is that it is overparameterized and too unwieldy to use, without further restrictions. The corresponding advantage is that this model formulation captures a number of different models under a single structure, thus making it a desirable framework for analysis. Before describing important applications of the model, we first summarize the mean and variance parameters that are useful in this model formulation.

**Mean Parameters**

With \(E \xi_i = \mu\), from equation (6.24), we have

\[ E \eta_i = \tau_\eta + B E \eta_i + E (\Gamma \xi_i + \zeta_i) = \tau_\eta + B \eta_i + \Gamma \mu, \]

Thus, \(E \eta_i = (I - B)^{-1} (\tau_\eta + \Gamma \mu)\), assuming that \(I - B\) is invertible. Summarizing, we have

\[
\begin{bmatrix}
E x \\
E y
\end{bmatrix} =
\begin{bmatrix}
\tau_x + \Lambda_x \mu \\
\tau_y + \Lambda_y (I - B)^{-1} (\tau_\eta + \Gamma \mu)
\end{bmatrix}.
\]

**Covariance Parameters**

From equation (6.24), we have \(E \eta_i = (I - B)^{-1} (\tau_\eta + \Gamma \xi_i + \zeta_i)\) and

\[
\text{Var} \eta = \text{Var}(I - B)^{-1} (\Gamma \xi_i + \zeta_i) = (I - B)^{-1} \text{Var}((\Gamma \xi_i + \zeta_i))(I - B)^{-1}
\]

\[
= (I - B)^{-1} (\Gamma \Phi \Gamma' + \Psi)(I - B)^{-1}.
\]
Thus, with equation (6.23), we have
\[
\text{Var} y = \Lambda_y (\text{Var} \eta) \Lambda'_y + \Theta_\varepsilon = \Lambda_y (I - B)^{-1} (\Gamma \Phi \Gamma' + \Psi)(I - B)^{-1}' \Lambda'_y + \Theta_\varepsilon.
\]

With equation (6.22), we have
\[
\text{Cov}(y, x) = \text{Cov}(\Lambda_y \eta + \varepsilon, \Lambda_x \xi + \delta) = \Lambda_y \text{Cov}(\eta, \xi) \Lambda'_x = \Lambda_y (I - B)^{-1} \Gamma \Phi \Lambda'_x.
\]

Summarizing, we have
\[
\begin{pmatrix}
\text{Var} y & \text{Cov}(y, x) \\
\text{Cov}(y, x)' & \text{Var} x
\end{pmatrix}
= \begin{pmatrix}
\Lambda_y (I - B)^{-1} (\Gamma \Phi \Gamma' + \Psi)(I - B)^{-1}' \Lambda'_y + \Theta_\varepsilon & \Lambda_y (I - B)^{-1} \Gamma \Phi \Lambda'_x \\
\Lambda_x \Phi \Gamma'(I - B)^{-1}' \Lambda'_y & \Lambda_x \Phi \Lambda'_x + \Theta_\delta
\end{pmatrix}. \quad (6.25)
\]

**Identification issues**

With the random sampling assumption, one can consistently estimate the means and covariances of the observables, specifically, the left-hand sides of equation (6.5.4) and (6.25). The model parameters are given in terms of the right-hand sides of these equations. There are generally more parameters that can be uniquely identified by the data. Identification is demonstrated by showing that the unknown parameters are functions only of the means and covariances and that these functions lead to unique solutions. In this case, we say that the unknown parameters are identified. Otherwise, they are said to be underidentified.

There are many approaches available for this process. We will illustrate a few in conjunction with some special cases, described below. Detailed broad treatments of this topic are available in many sources; see, for example, Bollen (1989EP).

**Special cases**

As noted above, the model summarized in equations (6.22)-(6.24) is overparameterized and too unwieldy to use, although it does encompass many special cases that are directly relevant for applications. To provide focus and intuition, we now summarize a few of these special cases.

- Consider only the x-measurement equation. This is the classic factor analysis model (see, for example, Johnson and Wichern, 1999G).
- Assume that both x and y are used directly in the structural equation model without any additional latent variables. (That is, assume \( x_i = \xi_i \) and \( y_i = \eta_i \).) Then, equation (6.24) represents a structural equation model based on observables, introduced in Section 6.4.3.
- Moreover, assuming the \( B = 0 \), the structural equation model with observables reduces to the multivariate regression model.
- Assume that y is used directly in the structural equation model but that x is measured with error so that \( x_i = \tau_i + \xi_i + \delta_i \). Assuming no feedback effects (for y, so that \( B = 0 \)), then equation (6.24) represents the classic errors in variables model.

Many other special cases appear in the literature. Our focus is on longitudinal special cases in Section 6.5.2.
Path diagrams

The popularity of structural equation models with latent variables in education and psychology is due in part to path diagrams. Path diagrams, due to Sewall Wright (1918B), are pictorial representations of system of equations. These diagrams show the relations among all variables, including disturbances and errors. These graphical relations allow many users to readily understand the consequences of modeling relationships. Moreover, statistical software routines have been developed to allow analysts to specify the model graphically, without resorting to algebraic representations. Table 6.2 summarizes the primary symbols used to make path diagrams.

<table>
<thead>
<tr>
<th>Table 6.2 Primary symbols used in path diagrams</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image.png" alt="Diagram" /></td>
</tr>
<tr>
<td>$x_1$</td>
</tr>
<tr>
<td>Rectangular or square box signifies an observed variable</td>
</tr>
<tr>
<td>$\eta_1$</td>
</tr>
<tr>
<td>Circle or ellipse signifies a latent variable</td>
</tr>
<tr>
<td>$\zeta_1$</td>
</tr>
<tr>
<td>Unenclosed variable signifies a disturbance term (error in either the structural equation or measurement equation)</td>
</tr>
<tr>
<td>$\eta_1$ $y_1$</td>
</tr>
<tr>
<td>Straight arrow signifies an assumption that the variable at the base of arrow “causes” variable at head of arrow</td>
</tr>
<tr>
<td>$\varepsilon_1$</td>
</tr>
<tr>
<td>Curved two-headed arrow signifies an association between two variables</td>
</tr>
<tr>
<td>$\eta_1$ $\eta_2$</td>
</tr>
<tr>
<td>Two straight single-headed arrows connecting two variables signifies a feedback relation or reciprocal causation</td>
</tr>
</tbody>
</table>

Source: Bollen (1989EP)

Estimation techniques

Estimation is typically done using maximum likelihood assuming normality; sometimes using instrumental variable estimation for initial values. Descriptions of alternative techniques, including generalized least squares and “unweighted” least squares, can be found in Bollen (1989EP).

Interestingly, the likelihood estimation is customary done by maximizing the likelihood over all the observables. Specifically, assuming that $(x_i, y_i)$ are jointly multivariate normal with moments given in equations (6.5.4) and (6.25), one maximizes the likelihood over the entire sample.

In contrast, most of the maximum likelihood estimation presented in this text has been for the likelihood of the response (or endogenous) variables, conditional on the exogenous observables. Specifically, suppose that observables consist of exogenous variables $x$ and endogenous variables $y$. Let $\theta_1$ be a vector of parameters that indexes the conditional distribution
of the endogenous variables given the exogenous variables, say \( p_1(y \mid x, \theta_1) \). Assume that there is another set of parameters, \( \theta_2 \), that are unrelated to \( \theta_1 \) and that indexes the distribution of the exogenous variables, say \( p_2(x, \theta_2) \). With this set-up, the complete likelihood is given by \( p_1(y \mid x, \theta_1) p_2(x, \theta_2) \). If our interest is only in the parameters that influence the relationship between \( x \) and \( y \), we can be content with maximizing the likelihood with respect to \( \theta_1 \). Thus, the distribution of the exogenous variables, \( p_2(x, \theta_2) \), is not relevant to the interest at hand, and may be ignored. Because of this philosophy, in our prior examples in this text, we did not concern ourselves with the sampling distribution of the \( x \)'s. (See Engle et al. 1983.) Section 7.4.2 will discuss this further.

Although the requirement that the two sets of parameters be unrelated is a restrictive assumption (that is generally not tested), it provides the analyst some important freedoms. With this assumption, the sampling distribution of the exogenous variables does not provide information about the conditional relationship under investigation. Thus, we need not make restrictive assumptions about the shape of this sampling distribution of the exogenous variables. As a consequence, we need not model the exogenous variables as multivariate normal or even require that they be continuous. To illustrate, a major distinction between the multiple linear regression model and the general linear model is that the latter formulation easily handles categorical regressors. The general linear model is about the conditional relationship between the response and the regressors, imposing few restrictions on the behavior of the regressors.

For the structural equation model with latent variables, the parameters associated with the distribution of exogenous variables are \( \theta_2 = \{ \tau_x, \mu_\xi, \Lambda_x, \Theta_\delta, \Phi \} \). Assuming, for example, multivariate normality, one can use equations (6.5.4) and (6.25) to compute the conditional likelihood of \( y \mid x \) (Appendix B). However, it is difficult to write down a set of parameters \( \theta_1 \) that are a subset of the full model parameters that are not related to \( \theta_2 \). Thus, maximum likelihood for the structural equations model with latent variables requires maximizing the full likelihood over all observables. This is in contrast to our first look at structural equations models in Section 6.5.3 (without measurement equations 6.5.1 and 6.5.2) where we could isolate the conditional model parameters from the sampling distribution of the exogenous variables.

To summarize, maximum likelihood estimation for structural equation models with latent variables customarily employs maximum likelihood estimation, where the likelihood is with respect to all the observables. A full model of both the exogenous and endogenous variables is specified; maximum likelihood estimators are well known to enjoy many optimality properties. A consequence of this specification of the exogenous variables is that it is difficult to handle categorical variables; multivariate distributions for discrete variables are much less well understood than the continuous variable case.

**Section 6.5.2 Longitudinal data applications**

The structural equation model with latent variables provides a broad framework for modeling longitudinal data. This section provides a series of examples to demonstrate the breadth of this framework.

**Special case - Autoregressive model**

Suppose that \( y_{it} \) represents the reading ability of the \( i \)th child; assume that we observe the set of \( n \) children over \( t =1, \ldots, 4 \) time periods. We might represent a child’s reading ability as

\[
y_{it} = \beta_0 + \beta_1 y_{i,t-1} + \xi_{it}, \text{ for } t = 2, 3, 4,
\]

(6.26)

using \( t=1 \) as the baseline measurement. We summarizes these relations using
Thus, this model is a special case of equations (6.22)-(6.24) by choosing $y_i = \eta_i$ and $\Gamma = 0$. Graphically, we can express this model as:

\begin{figure}
\centering
\begin{tikzpicture}

\node (y1) at (0,0) {$y_1$};
\node (y2) at (2,0) {$y_2$};
\node (y3) at (4,0) {$y_3$};
\node (y4) at (6,0) {$y_4$};

\node (e1) at (0,-1) {$\xi_1$};
\node (e2) at (2,-1) {$\xi_2$};
\node (e3) at (4,-1) {$\xi_3$};
\node (e4) at (6,-1) {$\xi_4$};

\path[->] (y1) edge (y2);
\path[->] (y2) edge (y3);
\path[->] (y3) edge (y4);
\path[->] (e1) edge (y1);
\path[->] (e2) edge (y2);
\path[->] (e3) edge (y3);
\path[->] (e4) edge (y4);
\end{tikzpicture}
\caption{Path diagram for the model in equation (6.26)}
\end{figure}

This basic model could be extended in several ways. For example, one may wish to consider evaluation at unequally spaced time points, such as 4th, 6th, 8th and 12th grades. This suggests using slope coefficients that depend on time. In addition, one could also use more than one lag predictor variable. For some other extensions, see the following continuation of this basic example.

**Special case - Autoregressive model – continued.** *Autoregressive model with latent variables and multiple indicators*

Suppose now that “reading ability” is considered a latent variable denoted by $\eta_i$, and that we have two variables, $y_{1it}$ and $y_{2it}$, that measure this ability, known as indicators. The indicators follow a measurement error model,

$$y_{jit} = \lambda_0 + \lambda_1 \eta_{i,t} + \varepsilon_{jit}, \quad (6.27a)$$

and the latent reading variable follows an autoregressive model,

$$\eta_{i,t} = \beta_0 + \beta_1 \eta_{i,t-1} + \varsigma_{it}. \quad (6.27b)$$

With the notation $y_i = (y_{1i1}, y_{2i2}, y_{1i2}, y_{1i3}, y_{2i3}, y_{1i4}, y_{2i4})'$, $\varepsilon_i$ defined similarly, and $\eta_i = (\eta_{i1}, \eta_{i2}, \eta_{i3}, \eta_{i4})'$, we can express equation (6.27a) as in equation (6.23). Here, $\tau_i$ is $\lambda_0$ times a vector of ones and $\Lambda_i$ is $\lambda_1$ times an identity matrix. Equation (6.27b) can be expressed as the structural equation (6.24) using notation similar to equation (6.26b).

Graphically, we can express equations (6.27a) and (6.27b) as:
Special case - Autoregressive model - continued.

Autoregressive model with latent variables, multiple indicators and predictor variables.

Continuing with this example, we might also wish to incorporate exogenous predictor variables, such as \(x_{1, it}\), the number of hours reading, and \(x_{2, it}\), a rating of emotional support given in a child’s home. One way of doing this is to simply add this as a predictor in the structural equation, replacing equation (6.5.7b) by

\[
\eta_{it} = \beta_0 + \beta_1 \eta_{i,t-1} + \gamma_1 x_{1, it} + \gamma_2 x_{2, it} + \varsigma_{it}.
\]

Jöreskog and Goldberger (1975S) introduced a model with multiple indicators of a latent variable that could be explained by multiple causes; this is now widely known as a MIMIC model.

Growth curve models

Integrating structural equation modeling with latent variables and longitudinal data modeling has been the subject of extensive research in recent years; see for example Duncan et al. (1999EP). One widely adopted approach concerns growth curve modeling.

Example 5.1 – Dental data – continued.

In the Section 5.2 Dental data example, we used \(y_{it}\) to represent the dental measurement of the \(i\)th child, measures at four times corresponding to ages 8, 10, 12, and 14 years of age. Using structural equation modeling notation, we could represent the \(y\)-measurement equation as

\[
y_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \\ y_{i4} \end{pmatrix} = \begin{pmatrix} 1 & 8 \\ 1 & 10 \\ 1 & 12 \\ 1 & 14 \end{pmatrix} \begin{pmatrix} \beta_{10i} \\ \beta_{12i} \end{pmatrix} + \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \epsilon_{i3} \\ \epsilon_{i4} \end{pmatrix} = \Lambda_y \eta_i + \epsilon_i.
\]

Note that this representation assumes that all children are evaluated at the same ages and that these ages correspond to a known “parameter” matrix \(\Lambda_y\). Alternatively, one could form groups so that all children within a group are measured at the same set of ages and let \(\Lambda_y\) vary by group.

Taking \(\tau_\eta = 0, B = 0\) and using the \(x\)-measurement directly (without error), we can write the structural equation as
\[ \eta_i = \left( \begin{array}{c} \beta_{i0} \\ \beta_{i1} \end{array} \right) = \left( \begin{array}{cc} \beta_{00} & \beta_{01} \\ \beta_{10} & \beta_{11} \end{array} \right) \left( \begin{array}{c} 1 \\ \text{GENDER}_i \end{array} \right) + \left( \begin{array}{c} \xi_{i0} \\ \xi_{i1} \end{array} \right) = \Gamma x_i + \zeta_i. \]

Thus, this model serves to express intercepts and growth rates associated with each child, \( \beta_{0i} \) and \( \beta_{1i} \), as a function of gender.

Willet and Sayer (1994EP) introduced growth curve modeling in the context of structural equations with latent variables. There are several advantages and disadvantages when using structural equations to model growth curves when compared to our Chapter 5 multilevel models. The main advantage of structural equation models is that it is straightforward to incorporate multivariate responses. To illustrate, in our dental example, there may be more than one dental measurement of interest, or it may be of interest to model dental and visual acuity measurements simultaneously.

The main disadvantage of structural equation models also relates to its multivariate response nature; it is difficult to handle unbalanced structure with this approach. If children came into the clinic for measurements at different ages, this would complicate the design considerably. Moreover, if not all observations were not available, issues of missing data are more difficult to deal with in this context. Finally, we have seen that structural equations with latent variables implicitly assume continuous data that can be approximated by multivariate normality; if the predictor variables are categorical (such as gender), this poses additional problems.

**Further reading**

Other introductions to the concept of exogeneity can be found in most graduate econometrics texts; see, for example, Greene (2002E) and Hayashi (2000E). The text by Wooldridge (2002E) gives an introduction with a special emphasis on panel data. Arellano and Honoré (2001E) provide a more sophisticated overview of panel data exogeneity. The collection of chapters in Mátyás and Sevestre (1996E) provide another perspective, as well as an introduction to structural equations with error components. For other methods for handling endogenous regressors with heterogeneity terms, we refer to Arellano (2003E), Baltagi (2002E) and Wooldridge (2002E).

There are many sources that introduce structural equations with latent variables. Bollen (1989EP) is a widely cited source that has been available for many years.
Appendix 6A. Linear Projections

Suppose that \( \{x', y\} \) is a random vector with finite second moments. Then, the linear projection of \( y \) onto \( x \) is \( x' (E(x'x'))^{-1} E(x'y) \), provided that \( E(x'x') \) is invertible. To ease notation, we define \( \beta = (E(x'x'))^{-1} E(x'y) \) and denote this projection as
\[ L(y | x) = x' \beta. \]

Others, such as Goldberger (1991E), use \( E^*(y | x) \) to denote this projection. By the linearity of expectations, it is easy to check that \( L(. | x) \) is a linear operator in \( y \). As a consequence, if we define \( \epsilon \) through the equation
\[ y = x' \beta + \epsilon, \]
then \( E^*(\epsilon | x) = 0 \). Note that this result is a consequence of the definition, not a model assumption that requires checking.

Suppose that we now assume that the model is of the form \( y = x' \beta + \epsilon \) and that \( E(x'x') \) is invertible. Then, the condition \( E(\epsilon x) = 0 \) is equivalent to checking the condition that \( L(\epsilon | x) = 0 \). That is, checking for a correlation between the disturbance term and the predictors is equivalent to checking that the linear projection of the disturbance term on the predictors is zero.

Linear projections can be justified as minimum mean square predictors. That is, the choice of \( \beta \) that minimizes \( E(y - x'\beta)^2 \) is \( \beta = (E(x'x'))^{-1} E(x'y) \). As an example, suppose that we have a data set of the form \( \{x_i', y_i\}, i=1, \ldots, n \), and define the expectation operator to be that probability distribution that assigns \( 1/n \) probability to each outcome (the empirical distribution).

Then, we are attempting to minimize \( E(y - x'\beta)^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 \). The solution is
\[ (E(xx'))^{-1} E(xy) = \left( \frac{1}{n} \sum_{i=1}^{n} x_i x_i' \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} x_i y_i \right), \]
the familiar ordinary least squares estimator.
Chapter 7. Modeling Issues

Abstract. As introduced in Chapter 1, longitudinal and panel data are often heterogeneous and may suffer from problems of attrition. This chapter describes models for handling these tendencies, as well as models designed to handle omitted variable bias.

Heterogeneity may be induced by (1) fixed effects, (2) random effects or (3) within subject covariances. Distinguishing among these mechanisms can be practically difficult although, as the chapter points out, is not always necessary. Moreover, the chapter describes the well-known Hausman test for distinguishing between estimators based on fixed versus random effects. As pointed out by Mundlak (1978aE), the Hausman test provides a test of the significance of time-constant omitted variables. This chapter emphasizes that an important feature of longitudinal and panel data is the ability to handle certain types of omitted variables.

This ability is one of the important benefits of using longitudinal and panel data; in contrast, attrition is one of the main drawbacks. The chapter reviews methods for detecting biases arising due to attrition and introduces models that provide corrections for attrition difficulties.

7.1 Heterogeneity

Heterogeneity is a common feature of many longitudinal and panel data sets. When we think of longitudinal data, we think of repeated measurements on subjects. This text emphasizes repeated observations over time although other types of clustering are of interest. For example, one could model the family unit as a “subject” and have individual measurements of family members as the repeated observations. Similarly, one could have a geographic area (such as a state) as the subject and have individual measurements of towns as the repeated observations. Regardless of the nature of the repetition, the common theme is that different observations from the same subject, or observational unit, tend to be related to one another. In contrast, the word “heterogeneity” refers to things that are unlike, or dissimilar. Thus, when discussing heterogeneity in the context of longitudinal data analysis, we mean that observations from different subjects tend to be dissimilar whereas observations from the same subject tend to be similar. We refer to models without heterogeneity components as homogenous.

Two approaches to modeling heterogeneity

In multivariate analysis, there are many methods for quantifying relationships among random variables. The goal of each method is to understand the joint distribution function of random variables; distribution functions provide all the details on the possible outcomes of random variables, both in isolation of one another and as they occur jointly. There are several methods for constructing multivariate distributions, see Hougaard (1987G) and Hutchinson and Lai (1990G) for detailed reviews. For applied longitudinal data analysis, we focus on two different methods of generating jointly dependent distribution functions, through (1) common variables and (2) covariances.

With the “variables-in-common” technique for generating multivariate distributions, a common element serves to induce dependencies among several random variables. We have
already used this modeling technique extensively, beginning in Chapters 2 and 3. Here, we used
the vector of parameters $\alpha_i$ to denote time-constant characteristics of a subject. In Chapter 2, the
fixed parameters induced similarities among different observations from the same subject through
the mean function. In Chapter 3, the random vectors $\alpha_i$ induced similarities through the
covariance function. In each case, $\alpha_i$ is common to all observations within a subject and thus
induces a dependency among these observations.

Although subject-specific common variables are widely used for modeling heterogeneity,
they do not cover all the important longitudinal data applications. We have already discussed
time-specific variables in Chapter 2 (denoted as $\lambda_t$) and will extend this discussion to cross-
classified data in Chapter 8, that is, incorporating both subject-specific and time-specific common
variables. Another important area of applications involves clustered data, described in Chapter 5.

We can also account for heterogeneity by directly modeling the covariance among
observations within a subject. To illustrate, in Section 3.1 on error components, we saw that a
common random $\alpha_i$ induced a positive covariance among observations within a subject. We also
saw that we could model this feature using the compound symmetry correlation matrix. The
advantage of the compound symmetry covariance approach is that it also allows for models of
negative dependence. Thus, modeling the joint relation among observations directly using
covariances can be simpler than an approach using common variables and may also cover
additional distributions. Further, for serial correlations, modeling covariances directly is much
simpler than alternative approaches. We know that for normally distributed data, modeling the
covariance function, together with the mean, is sufficient to specify the entire joint distribution
function. Although this is not true in general, correctly specifying the first two moments suffices
for much applied work. We take up this issue further in Chapters 9 and 10 when discussing the
generalized estimating equations approach.

Practical identification of heterogeneity may be difficult

For many longitudinal data sets, an analyst could consider many alternative strategies for
modeling the heterogeneity. One could use subject-specific intercepts that may be fixed or
random. One could use subject-specific slopes that, again, may be fixed or random. Alternatively,
one could use covariance specifications to model the tendency for observations from the same
subject to be related. As the following illustration from Jones (1993S) shows, it may be difficult
to distinguish among these alternative models when only using the data to aid model
specification.

Figure 7.1 shows panels of times series plots for $n = 3$ subjects. The data are generated
with no serial correlation over time but with three different subject-specific parameters, $\alpha_1 = 0$, $\alpha_2 = 2$, and $\alpha_3 = -2$. With perfect knowledge of the subject-specific parameters, one would correctly
use a scalar times the identity matrix for the covariance structure. However, if these subject-
specific variables are ignored, a correlation analysis shows a strong positive serial correlation.
That is, from the first panel in Figure 7.1, we see that observations tend to oscillate about the
overall mean of zero in a random fashion. However, the second panel shows that all observations
are above zero and the third panel indicates that almost all observations are below zero. Thus, an
analysis without subject-specific terms would indicate strong positive autocorrelation. Although
not the “correct” formulation, a time series model such as the AR(1) model would serve to capture
the heterogeneity in the data.
Figure 7.1. Different subject-specific parameters can induce positive serial correlation.

Theoretical identification with heterogeneity may be impossible

Thus, identifying the correct model formulation to represent heterogeneity can be difficult. Moreover, in the presence of heterogeneity, identifying all the model components may be impossible. For example, our training in linear model theory has led us to believe that with \( N \) observations and \( p \) linear parameters, we require only that \( N > p \) in order to identify variance components. Although this is true in cross-sectional regression, the more complex longitudinal data setting requires additional assumptions. The following example, due to Neyman and Scott (1948E), illustrates some of these complexities.

Example – Neyman-Scott on identification of variance components

Consider the fixed effects model

\[
y_{it} = \alpha_i + \epsilon_{it}, \quad i=1, \ldots, n, \quad t=1,2,
\]

where \( \text{Var} \epsilon_{it} = \sigma^2 \) and \( \text{Cov} (\epsilon_{i1}, \epsilon_{i2}) = \sigma^2 \rho \). The ordinary least squares estimator of \( \alpha_i \) is \( \bar{y}_i = (y_{i1} + y_{i2})/2 \). Thus, the residuals are \( e_{i1} = y_{i1} - \bar{y}_i = (y_{i1} - y_{i2})/2 \) and \( e_{i2} = y_{i2} - \bar{y}_i = (y_{i2} - y_{i1})/2 = - e_{i1} \). Because of these relations, it turns out that \( \rho \) cannot be estimated, despite having \( 2n - n = n \) degrees of freedom available for estimating the variance components.

Estimation of regression coefficients without complete identification is possible

Fortunately, complete model identification is not required for all inference purposes. To illustrate, if our main goal is to estimate or test hypotheses about the regression coefficients, then we do not require knowledge of all aspects of the model. For example, consider the one-way fixed effects model

\[
y_i = \alpha_i + X_i \beta + \epsilon_i.
\]
For the balanced case, \( T_i = T \), Kiefer (1980E) showed how to consistently estimate all components of \( \text{Var} \varepsilon_i = R \) that are needed for inference about \( \beta \). That is, apply the common transformation matrix \( Q = I - T^{-1} J \) to each equation to get

\[
y_i^* = Q y_i = Q X_i \beta + Q \varepsilon_i = X_i^* \beta + \varepsilon_i^* ,
\]

because \( Q 1 = 0 \). Note that \( \text{Var} \varepsilon_i^* = Q R Q = R^* \). With this transformed equation, the population parameters \( \beta \) can be consistently (and root-\( n \)) estimated. Further, elements of \( R^* \) can be consistently estimated and used to get feasible generalized least squares estimators.

**Example - Neyman-Scott - continued**

Here, we have \( T=2, Q = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1/2 \\ 1 \end{pmatrix} = \begin{pmatrix} 1/2 \\ 1 \\ -1 \\ 1 \end{pmatrix} \) and

\[
R = \begin{pmatrix} \sigma^2 & \sigma^2 \rho \\ \sigma^2 \rho & \sigma^2 \end{pmatrix} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} .
\]

Thus,

\[
R^* = Q R Q = \sigma^2 \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} = \frac{\sigma^2 (1-\rho)}{2} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}
\]

Using moment based estimators, we can estimate the quantity \( \sigma^2 (1-\rho) \) yet cannot separate the terms \( \sigma^2 \) and \( \rho \).

This example shows that, in the balanced basic fixed effects model, feasible generalized least squares estimation is possible, even without complete identification of all variance components. More generally, consider the case where we have unbalanced data and variable slopes, represented with the model

\[
y_i = Z_i \alpha_i + X_i \beta + \varepsilon_i,
\]

where \( \text{Var} \varepsilon_i = R_i \). For this model, in Section 2.5.3 we introduced the transformation \( Q_i = I_i - Z_i (Z_i' Z_i)^{-1} Z_i' \). Applying this transform to the model yields

\[
y_i^* = Q_i y_i = Q_i X_i \beta + Q_i \varepsilon_i = X_i^* \beta + \varepsilon_i^* ,
\]

because \( Q_i Z_i = 0 \). Note that \( \text{Var} \varepsilon_i^* = Q_i R_i Q_i = R_i^* \). For this model, we see that an ordinary least squares estimator of \( \beta \) is unbiased and (root-\( n \)) consistent. Without knowledge of the variance components in \( R_i^* \), one can still use robust standard errors to get asymptotically correct confidence intervals and tests of hypotheses.

The case for feasible generalized least squares is more complex. Now, if \( R_i = \sigma^2 I_i \), then \( R_i^* \) is known up to a constant; in this case, the usual generalized least squares estimator, given in equation (2.16), is applicable. For other situations, Kung (1996O) provided sufficient conditions for the identification and estimation of a feasible generalized least squares estimator.

### 7.2 Comparing fixed and random effects estimators

In Section 3.1, we introduced the sampling and inferential bases for choosing a random effects model. However, there are many instances when these bases do not provide sufficient guidance to dictate which type of estimator, fixed or random effects, the analyst should employ. In Section 3.1, we saw that the random effects estimator is derived using generalized least squares and thus has minimum variance among all unbiased linear estimators. However, in Section 6.2 we
saw that fixed effects estimators do not rely on assumption SEC6, zero correlation between the
time-constant heterogeneity variables and the regressor variables. Oftentimes, analysts look to
features of the data to provide additional guidance. This section introduces the well-known
Hausman test for deciding whether to use a fixed or random effects estimator. The Hausman
(1978E) test is based on an interpretation due to Mundlak (1978aE) that the fixed effects
estimator is robust to certain omitted variable model specifications. Throughout this section, we
maintain assumption SEC1, the strict exogeneity of the regressor variables conditional on the
unobserved effects.

To introduce the Hausman test, we first return to a version of our Section 3.1 error
components model
\[ y_{it} = \alpha_i + x_{it}' \beta + \varepsilon_{it} + u_i. \]  
(3.1)*

Here, as in Section 3.1, \( \alpha_i \) is a random variable that is uncorrelated with the disturbance term \( \varepsilon_{it} \)
(and the explanatory variables \( x_{it} \), see Chapter 6). However, we have also added \( u_i \), a term for
unobserved omitted variables. Unlike \( \alpha_i \), the concern is that the \( u_i \) quantity may represent a fixed
effect, or a random effect that is correlated with either the disturbance terms or explanatory
variables. If \( u_i \) is present, then the heterogeneity term \( \alpha_i^* = \alpha_i + u_i \) does not satisfy the usual
assumptions required for unbiased and consistent regression coefficient estimators.

We do, however, restrict the omitted variables to be time-constant; this assumption
allows us to derive unbiased and consistent estimators, even in the presence of omitted variables.
Taking averages over time in equation (3.1)*, we have
\[ \overline{y}_i = \alpha_i + \overline{x}_i' \beta + \overline{\varepsilon}_i + u_i. \]

Subtracting this from equation (3.1)* yields
\[ y_{it} - \overline{y}_i = (x_{it} - \overline{x}_i)' \beta + \varepsilon_{it} - \overline{\varepsilon}_i. \]

Based on these deviations, we have removed the effects of the unobserved variable \( u_i \). Thus, the
equation (2.6) fixed effects estimator
\[ b_{FE} = \left( \sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \overline{x}_i)(x_{it} - \overline{x}_i)' \right)^{-1} \left( \sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \overline{x}_i)(y_{it} - \overline{y}_i) \right) \]
is not corrupted by \( u_i \); it turns out to be unbiased and consistent even in the presence of omitted
variables. For notational purposes, we have added the subscript \( FE \) to suggest the motivation of
this estimator even though there may not be any fixed effects (\( \alpha_i \)) in equation (3.1)*.

The Hausman test statistic compares the robust estimator \( b_{FE} \) to the generalized least
squares estimator \( b_{EC} \). Under the null hypothesis of no omitted variables, \( b_{EC} \) is more efficient
(because it is a generalized least squares estimator). Under the alternative hypothesis of (time-
constant) omitted variables, \( b_{FE} \) is still unbiased and consistent. Hausman (1978E) showed that the statistic
\[ \chi^2_{FE} = (b_{FE} - b_{GLS})' (\text{Var}(b_{FE}) - \text{Var}(b_{GLS}))^{-1} (b_{FE} - b_{GLS}) \]
has an asymptotic chi-square distribution with \( K \) degrees of freedom under the null hypothesis. It
is a widely used statistic for detecting omitted variables.
Example: Income tax payments

To illustrate the performance of the fixed effects estimators and omitted variable tests, this section examines data on determinants of income tax payments introduced in Section 3.2. Specifically, we begin with the error components model with $K = 8$ coefficients estimated using generalized least squares. The parameter estimates, as well as the corresponding $t$-statistics, appear in Table 7.1.

Also in Table 7.1 are the corresponding fixed effects estimators. The fixed effects estimators are robust to time-constant omitted variables. The standard errors and corresponding $t$-statistics are computed in a straightforward fashion using the procedures described in Chapter 2. Comparing the fixed and random effects estimators in Table 7.1, we see that the coefficients are qualitatively similar. For each variable, the estimators have the same sign and similar orders of magnitude. They also indicate the same order of statistical significance. To illustrate, the two measures of income, LNTPI and MR, are both strongly statistically significant under both estimation procedures. The one exception is the EMP variable; this is strongly statistically significantly negative using the random effects estimation but is not statistically significant using the fixed effects estimation.

To assess the overall differences among the coefficient estimates, we may use the omitted variable test statistic due to Hausman (1978E). As indicated in Table 7.1, this test statistic turns out to be $\chi^2_{FE} = 6.021$. Comparing this test statistic to a chi-square distribution with $K = 8$ degrees of freedom, the $p$-value associated with this test statistic is $\text{Prob}(\chi^2 > 6.021) = 0.6448$. This does not provide enough evidence to indicate a serious problem with omitted variables. Thus, the random effects estimator is preferred.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Robust Fixed Effects Estimation</th>
<th>Random Effects Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Parameter Estimates</td>
<td>$t$-statistic</td>
</tr>
<tr>
<td>LNTPI</td>
<td>0.717</td>
<td>9.30</td>
</tr>
<tr>
<td>MR</td>
<td>0.122</td>
<td>13.55</td>
</tr>
<tr>
<td>MS</td>
<td>0.072</td>
<td>0.28</td>
</tr>
<tr>
<td>HH</td>
<td>-0.707</td>
<td>-2.17</td>
</tr>
<tr>
<td>AGE</td>
<td>0.002</td>
<td>0.01</td>
</tr>
<tr>
<td>EMP</td>
<td>-0.244</td>
<td>-0.99</td>
</tr>
<tr>
<td>PREP</td>
<td>-0.030</td>
<td>-0.18</td>
</tr>
<tr>
<td>DEPEND</td>
<td>-0.069</td>
<td>-0.83</td>
</tr>
<tr>
<td>$\chi^2_{FE}$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Because of the complexities and the widespread usage of this test in the econometrics literature, we split the remainder of the discussion into two parts. The first part, Section 7.2.1, introduces some important additional ideas in the context of a special case. Here, we show a relationship between fixed and random effects estimators, introduce Mundlak’s alternative hypothesis, derive the fixed effects estimator under this hypothesis and discuss Hausman’s test of omitted variables. The second part, Section 7.2.2, extends the discussion to incorporate (1) unbalanced data, (2) many variables, (3) variable slopes as well as (4) potential serial correlation and heteroscedasticity. Section 7.3 will discuss alternative sampling bases.
7.2.1 A special case

Consider the Section 3.1 error components model with \( K = 2 \) so that

\[
y_{it} = \alpha_i + \beta_0 + \beta_1 x_{it,1} + \epsilon_{it},
\]

(7.1)

where both \( \{\alpha_i\} \) and \( \{\epsilon_{it}\} \) are i.i.d. as well as independent of one another. For simplicity, we also assume balanced data so that \( T_i = T \). Ignoring the variability in \( \{\alpha_i\} \), the usual ordinary least squares estimator of \( \beta_1 \) is

\[
b_{1,HOM} = \frac{\sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{x})(y_{it} - \bar{y})}{\sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{x})^2}.
\]

Because this estimator excludes the heterogeneity component \( \{\alpha_i\} \), we label it using the subscript “HOM” for homogeneous. In contrast, from Exercise 7.5, an expression for the generalized least squares estimator is:

\[
b_{1,EC} = \frac{\sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it}^* - \bar{x}^*)(y_{it}^* - \bar{y}^*)}{\sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it}^* - \bar{x}^*)^2},
\]

where \( x_{it}^* = x_{it} - \bar{x}, \) \( y_{it}^* = y_{it} - \bar{y} \),

\[
1 - \left( \frac{\sigma_{\alpha}^2}{T\sigma_{\alpha}^2 + \sigma^2} \right)^{1/2} \quad \text{and} \quad 1 - \left( \frac{\sigma_{\alpha}^2}{T\sigma_{\alpha}^2 + \sigma^2} \right)^{1/2},
\]

As described in Section 3.1, both estimators are unbiased, consistent and asymptotically normal. Because \( b_{1,EC} \) is a generalized least squares estimator, it has a smaller variance than \( b_{1,HOM} \). That is, \( \text{Var} \ b_{1,EC} \leq \text{Var} \ b_{1,HOM} \) where

\[
\text{Var} \ b_{1,EC} = \sigma^2 \left( \sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it}^* - \bar{x}^*)^2 \right)^{-1}.
\]

(7.2)

Also note that \( b_{1,EC} \) and \( b_{1,HOM} \) are approximately equivalent when the heterogeneity variance is small. Formally, because \( x_{it} \rightarrow x_i^{\ast} \) and \( y_{it} \rightarrow y_i^{\ast} \) as \( \sigma_{\alpha}^2 \rightarrow 0 \), we have that \( b_{1,EC} \rightarrow b_{1,HOM} \) as \( \sigma_{\alpha}^2 \rightarrow 0 \).

This section also considers an alternative estimator

\[
b_{1,FE} = \frac{\sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{x})(y_{it} - \bar{y})}{\sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{x})^2}.
\]

(7.3)

This estimator could be derived from the model in equation (7.1) by assuming that the terms \( \{\alpha_i\} \) are fixed, not random, components. In the notation of Chapter 2, we may assume that \( \alpha_i^{\ast} = \alpha_i + \beta_0 \) are the fixed components that are not centered about zero. An important point of this section is that the estimator defined in equation (7.3) is unbiased, consistent and asymptotically normal under the model that includes random effects in equation (7.1). Further, straightforward calculations show that

\[
\text{Var} \ b_{1,FE} = \sigma^2 \left( \sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{x})^2 \right)^{-1}.
\]

(7.4)

We note that \( b_{1,EC} \) and \( b_{1,FE} \) are approximately equivalent when the heterogeneity variance is large. Formally, because \( x_{it} \rightarrow x_i - \bar{x} \) and \( y_{it} \rightarrow y_i - \bar{y} \) as \( \sigma_{\alpha}^2 \rightarrow \infty \), we have that \( b_{1,EC} \rightarrow b_{1,FE} \) as \( \sigma_{\alpha}^2 \rightarrow \infty \).

To relate the random and fixed effects estimators, we define the so-called “between groups” estimator,
7. Modeling Issues

\[ b_{1,b} = \frac{\sum_{i=1}^{n} (\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})}{\sum_{i=1}^{n} (\bar{x}_i - \bar{x})^2}. \]  

(7.5)

This estimator can be motivated by averaging all observations from a subject and then computing an ordinary least squares estimator using the data \( \{\bar{x}_i, \bar{y}_i\}_{i=1}^{n} \). As with the other estimators, this estimator is unbiased, consistent and asymptotically normal under the equation (7.1) model. Further, straightforward calculations show that

\[ \text{Var} \ b_{1,b} = \left( T \sigma_\alpha^2 + \sigma_\gamma^2 \right) \left( T \sum_{i=1}^{n} (\bar{x}_i - \bar{x})^2 \right)^{-1}. \]  

(7.6)

To interpret the relations among \( b_{1,EC}, b_{1,FE}, \) and \( b_{1,b} \), we cite the following decomposition due to Maddala (1971E),

\[ b_{1,EC} = (1 - \Delta) b_{1,FE} + \Delta b_{1,b}. \]  

(7.7)

Here, the term \( \Delta = \frac{\text{Var} \ b_{1,EC}}{\text{Var} \ b_{1,b}} \) measures the relative precision of the two estimators of \( \beta \).

Because \( b_{1,EC} \) is the generalized least squares estimator, we have that \( \text{Var} \ b_{1,EC} \leq \text{Var} \ b_{1,b} \) so that \( 0 \leq \Delta \leq 1 \).

### Omitted variables – model of correlated effects

Thus, assuming the random effects model in equation (7.1) is an adequate representation, we expect each of the four estimators of \( \beta \) to be close to one another. However, in many data sets these estimators can differ dramatically. To explain these differences, Mundlak (1978aE) proposed what we will call a model of “correlated effects.” Here, we interpret \( \alpha \) to represent time-constant, or “permanent,” characteristics of \( y_i \) that are unobserved and hence “omitted.” Mundlak introduced the possibility that \( \{\alpha_i\} \) are correlated with the observed variables \( x_i \). That is, the latent variables \( \alpha \) fail the exogeneity assumption SE6 described in Section 6.2.2. To express the relationship between \( \alpha \) and \( x_i \), we consider the function \( \text{E} [\alpha_i | x_i] \). Specifically, for our special case, Mundlak assumed that \( \alpha = \eta_i + \gamma \bar{x}_{i,1} \), where \( \{\eta_i\} \) is an i.i.d. sequence that is independent of \( \{\varepsilon_{it}\} \). Thus, the model of correlated effects is

\[ y_{it} = \eta_i + \beta_0 + \beta_1 x_{it} + \gamma \bar{x}_{i,1} + \nu_{i,1} + \varepsilon_{it}. \]  

(7.8)

Under this model, one can show that the generalized least squares estimator of \( \beta_1 \) is \( b_{1,FE} \). Further, the estimator \( b_{1,FE} \) is unbiased, consistent and asymptotically normal. In contrast, the estimators \( b_{1,HOMS}, b_{1,b} \) and \( b_{1,EC} \) are biased and inconsistent.

To compare the model of correlated effects in equation (7.8) with the baseline model in equation (7.1), we need only examine the null hypothesis \( H_0: \gamma = 0 \). This is customarily done using the Hausman (1978E) test statistic

\[ \chi^2_{FE} = \frac{(b_{1,EC} - b_{1,FE})^2}{\text{Var} b_{1,FE} - \text{Var} b_{1,EC}}. \]  

(7.9)

Under the null hypothesis of the model in equation (7.1), this test statistic has an asymptotic (as \( n \to \infty \)) chi-square distribution with 1 degree of freedom. This provides the basis for comparing the two models. Moreover, we see that the test statistic will be large when there is a large difference between the fixed and random effects estimators. In addition, it is straightforward to construct the test statistic based on a fit of the random effects model in equation (7.1) (to get \( b_{1,EC} \) and \( \text{Var} b_{1,EC} \)) and a fit of the corresponding fixed effects model (to get \( b_{1,FE} \) and \( \text{Var} b_{1,FE} \)). Thus, one need not construct the “augmented” variable \( \bar{x}_{i,1} \) in equation (7.8).
7.2.2 General case

Extension to the general case follows directly. To incorporate intercepts, we use a modification of the linear mixed effects model in equation (3.5). Thus, we assume that $E\alpha_i = \alpha$ and re-write the model as

$$y_i = Z_i \alpha + X_i \beta + \varepsilon_i^*,$$

where $\varepsilon_i^* = \varepsilon_i + Z_i (\alpha_i - \alpha)$ and $\text{Var} \varepsilon_i^* = Z_i D Z_i' + R_i = V_i$. Straightforward calculations (Exercise 7.3) show that the generalized least squares estimator of $\beta$ is

$$b_{GLS} = C_{GLS}^{-1} \sum_{i=1}^{n} \left( X_i' V_i^{-1} - \left( \sum_{i=1}^{n} X_i' V_i^{-1} Z_i \right) \left( \sum_{i=1}^{n} Z_i' V_i^{-1} Z_i \right)^{-1} Z_i' V_i^{-1} \right) y_i,$$

with

$$C_{GLS} = \left( \sum_{i=1}^{n} X_i' V_i^{-1} X_i \right) - \left( \sum_{i=1}^{n} X_i' V_i^{-1} Z_i \right) \left( \sum_{i=1}^{n} Z_i' V_i^{-1} Z_i \right)^{-1} \left( \sum_{i=1}^{n} Z_i' V_i^{-1} X_i \right).$$

From equation (2.16), we have that the corresponding fixed effects estimator is

$$b_{FE} = C_{FE}^{-1} \sum_{i=1}^{n} X_i' R_i^{-1/2} Q_{Z,i} R_i^{-1/2} y_i,$$

where $C_{FE} = \sum_{i=1}^{n} X_i' R_i^{-1/2} Q_{Z,i} R_i^{-1/2} X_i$ and $Q_{Z,i} = I_i - R_i^{-1/2} Z_i (Z_i' R_i^{-1} Z_i)^{-1} Z_i' R_i^{-1/2}$. From Exercise (7.5), the between-groups estimator is

$$b_B = C_B^{-1} \sum_{i=1}^{n} \left( X_i' V_i^{-1} Z_i \left( Z_i' V_i^{-1} Z_i \right)^{-1} Z_i' V_i^{-1} \right) y_i - \left( \sum_{i=1}^{n} X_i' V_i^{-1} Z_i \right) \left( \sum_{i=1}^{n} Z_i' V_i^{-1} Z_i \right)^{-1} \left( \sum_{i=1}^{n} Z_i' V_i^{-1} X_i \right).$$

To relate these three estimators, the extension of Maddala’s (1971E) result is

$$b_{GLS} = (I - \Delta) b_{FE} + \Delta b_B,$$

where $\Delta = (\text{Var} b_{GLS}) (\text{Var} b_B)^{-1}$, $\text{Var} b_B = C_B^{-1}$ and $\text{Var} b_{GLS} = C_{GLS}^{-1}$ (see Exercise 7.5).

Again, the matrix $\Delta$ is a weight matrix that quantifies the relative precision of the two estimators, $b_{GLS}$ and $b_B$. 
Correlated effects model

For a model of correlated effects that describes the correlation between $\{a_i\}$ and $\{X_i\}$, let $x_i = \text{vec}(X_i')$, a $KT \times 1$ vector built by stacking vectors $\{x_{i1}', \ldots, x_{iT}'\}$. For notation, we denote the observed independent variables by $o_{it} = (z_{it}', x_{it}')'$, a $(q+K) \times 1$ vector of observed effects, and let $o_i$ be the associated column vector, that is, $o_i = (o_{i1}', \ldots, o_{iT}')'$. We assume that the relationship between $a_i$ and $X_i$ can be described through the conditional moments

$$E(a_i | o_i) = \Sigma_a \Sigma_x^{-1}(x_i - E(x_i))$$ and $\text{Var}[a_i | o_i] = D^*$, \hspace{1cm} (7.11)

where $\Sigma_a = \text{Cov}(a_i, x_i) = E(a_i (x_{i1}', \ldots, x_{iT}')')$ and $\Sigma_x = \text{Var} x_i$. Equation (7.11) can be motivated by joint normality of $a_i$ and $o_i$ but is also useful when some components of $o_i$ are categorical. With display (7.11), we have

$$E[y_{ij} | o_{i}] = E[E[y_{ij} | a_i, o_i] | o_i] = Z_j \Sigma_a \Sigma_x^{-1}(x_i - E(x_i)) + X_j \beta$$ \hspace{1cm} (7.12)

and

$$\text{Var}[y_{ij} | o_{i}] = \text{Var}(E[y_{ij} | a_i, o_i] | o_i) + \text{Var}(E[y_{ij} | a_i, o_i] | o_i) = R_i + Z_j D^* Z_j'.$$ \hspace{1cm} (7.13)

The correlated effects alters the form of the regression function in equation (7.12) but not the conditional variance in equation (7.13).

Now, under the model of correlated effects summarized in equations (7.12) and (7.13), it is easy to see that the random effects estimator is generally biased. In contrast, the fixed effects estimator is unbiased and has variance

$$\text{Var} b_{FE} = \left(\sum_{i=1}^n X_i' R_i^{-1/2} Q_i R_i^{-1/2} X_i\right)^{-1},$$

which is the same as under the fixed effects model formulation; see Section 2.5.3.

Again, an extension of the Hausman (1978E) test allows us to compare the baseline model and the model of correlated effects. The test statistic is

$$\chi^2_{FE} = (b_{FE} - b_{GLS})' \left(\text{Var}(b_{FE}) - \text{Var}(b_{GLS})\right)^{-1}(b_{FE} - b_{GLS}).$$ \hspace{1cm} (7.14)

Under the null hypothesis of the model in equation (7.10), this test statistic has an asymptotic (as $n \rightarrow \infty$) chi-square distribution with $K$ degrees of freedom. As in Section 7.2.1, this test statistic is intuitively pleasing in that large differences between the fixed and random effects estimators allow us to reject the null hypothesis of no correlated effects.

To summarize, the fixed effects estimator is easy to compute and is robust to omitted variable bias. The estimator has desirable properties under a variation of the random effects model that we call a model of correlated effects. Under this model of correlated effects formulation, the many subject-specific fixed parameters generally associated with fixed effects models need not be computed. In addition to the estimator itself, standard errors associated with this estimator are easy to compute. Further, the equation (7.14) test statistic provides a simple method for assessing the adequacy of the random effects model; this could lead to further follow-up investigations that may in turn lead to an improved model specification.
**Example: Income tax payments - continued**

The analysis above is based on the error components model. Many additional features could be fit to the data. After additional model exploration, we retained the variable intercepts and also used subject-specific, variable slopes for the income variables, LNTPI and MR. In addition, to accommodate serial patterns in tax liabilities, we specified an $AR(1)$ component for the errors. Part of the rationale comes from the nature of tax liabilities. That is, we hypothesize that the tax liability increases with income. Moreover, because individuals have different attitudes towards tax-sheltering programs, live in different states that have their own tax programs (that affect the amount of the federal tax), and so on, we expect coefficients associated with income to differ among individuals. A similar argument can be made for MR because this is simply a nonlinear function of income.

The random effects (estimated generalized least squares) and robust fixed effects estimators are given in Table 7.2. We now see some important differences in the two estimation methodologies. To illustrate, examining the coefficients associated with the MS and EMP variables, the random effects estimator indicates that each variable is strongly statistically negatively significant whereas the robust estimators do not signal statistical significance. This is also true, but to a lesser extent, of the coefficient associated with AGE.

To compare the vector of random versus fixed effects estimators, the omitted variable test statistic turns out to be $\chi^2_{FE} = 13.628$. Using $K = 6$ degrees of freedom, the $p$-value associated with this test statistic is $\text{Prob}(\chi^2 > 13.628) = 0.0341$. In contrast to the error components model, this provides evidence to indicate a serious problem with omitted variables. The result of the hypothesis test suggests using the robust estimators. Interestingly, both random and fixed effects estimation indicate that use of a tax preparer (PREP) significantly lowers the tax liability of a taxpayer, when controlling for income and demographic characteristics. This was not a finding in the error components model.

By introducing two variable slopes, the number of estimator comparisons dropped from eight to six. Examining equation (7.10), we see that the variables included in the random effects formulation are no longer included in the “$X\beta$” portion. Thus, in equation (7.10), the number of rows of $\beta$, $K$, refers to the number of variables not associated with the random effects portion; we can think of these variables as associated with only fixed effects. This implies, among other things, that the Section 7.2.3 omitted variable test is not available for the random coefficients model where there are no variables associated with only fixed effects. Thus, Section 7.3 introduces a test that will allow us to consider the random coefficients model.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Robust Fixed Effects Estimation</th>
<th>Random Effects Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>LNTPI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MS</td>
<td>-0.197</td>
<td>-0.603</td>
</tr>
<tr>
<td>HH</td>
<td>-1.870</td>
<td>-0.729</td>
</tr>
<tr>
<td>AGE</td>
<td>-0.464</td>
<td>-0.359</td>
</tr>
<tr>
<td>EMP</td>
<td>-0.198</td>
<td>-0.661</td>
</tr>
<tr>
<td>PREP</td>
<td>-0.474</td>
<td>-0.300</td>
</tr>
<tr>
<td>DEPEND</td>
<td>-0.304</td>
<td>-0.138</td>
</tr>
<tr>
<td>$AR(1)$</td>
<td>0.454</td>
<td>0.153</td>
</tr>
</tbody>
</table>

$\chi^2_{FE} = 13.628$
7.3 Omitted variables

Particularly in the social sciences where observational, in lieu of experimental, data are predominant, problems of omitted variables abound. The possibility of unobserved, omitted variables that affect both the response and explanatory variables encourage analysts to distinguish between “cause” and “effect.” We have already seen one approach for handling cause and effect analysis through multiple systems of equations in Sections 6.4 and 6.5. Further, the structure of longitudinal data allows us to construct estimators that are less susceptible to bias arising from omitted variables than common alternatives. For example, in Section 7.2 we saw that fixed effects estimators are robust to certain types of time-constant omitted variables. This section introduces estimators that are robust to other types of omitted variables. These omitted variable robust estimators do not provide protection from all types of omitted variables; they are sensitive to the nature of the variables being omitted. Thus, as a matter of practice, analysts should always attempt to collect as much information as possible regarding the nature of the omitted variables.

Specifically, Section 7.2 showed how a fixed effects estimator is robust to assumption SEC6, zero correlation between the time-constant heterogeneity variables and the regressor variables. Unfortunately, the fixed effects transform sweeps out time-constant variables and these variables may be the focus of a study. To remedy this, this section shows how to use partial information about the relationship between the unobserved heterogeneity variables and the regressor variables. This idea of partial information is due to Hausman and Taylor (1981E) who developed an instrumental variable estimation procedure. We consider a broader class of omitted variables models that, under certain circumstances, also allow for time-varying omitted variables. Here, we will see that the fixed effects estimator is a special case of a class that we call “augmented regression” estimators. This class not only provides extensions but also gives a basis for providing heteroscedasticity-consistent standard errors of the estimators.

To set the stage for additional analyses, we first return to a version of the Section 3.1 error components model

\[ y_{it} = \alpha_i + \mathbf{x}_{it}^{(1)\prime} \beta_1 + \mathbf{x}_{it}^{(2)\prime} \beta_2 + \epsilon_{it} + u_i \]  

We have now split up the \( \mathbf{x}_{it} \) into two portions, one for time-varying explanatory variables (\( \mathbf{x}_{it}^{(1)} \) and \( \beta_1 \)) and one for time-constant explanatory variables (\( \mathbf{x}_{it}^{(2)} \) and \( \beta_2 \)). As before, the \( u_i \) term presents unobserved, omitted variables, that may be a fixed effect, or a random effect that is correlated with either the disturbance terms or explanatory variables. As pointed out in Section 2.3, if an explanatory variable is time-constant, then the fixed effects estimator is no longer estimable. Thus, the techniques introduced in Section 7.2 no longer immediately apply. However, by examining deviations from the mean,

\[ y_{it} - \bar{y}_i = (\mathbf{x}_{it}^{(1)} - \bar{\mathbf{x}}_i^{(1)}) \beta_1 + \epsilon_{it} - \bar{\epsilon}_i, \]

we see that we can still derive unbiased and consistent estimators of \( \beta_1 \), even in the presence of omitted variables. To illustrate, one such estimator is

\[ \mathbf{b}_{1,FE} = \left( \sum_{i=1}^{n} \sum_{t=1}^{T} (\mathbf{x}_{it}^{(1)} - \bar{\mathbf{x}}_i^{(1)})(\mathbf{x}_{it}^{(1)} - \bar{\mathbf{x}}_i^{(1)})' \right)^{-1} \left( \sum_{i=1}^{n} \sum_{t=1}^{T} (\mathbf{x}_{it}^{(1)} - \bar{\mathbf{x}}_i^{(1)})(y_{it} - \bar{y}_i) \right). \]

Moreover, with the additional assumption that \( u_i \) is not correlated with \( \mathbf{x}_{it}^{(2)} \), we will be able to provide consistent estimators of \( \beta_2 \). This is a strong assumption; still, the interesting aspect is that with longitudinal/panel data, we can derive estimators with desirable properties even in the presence of omitted variables.
When hypothesizing relationships among variables, the breakdown between time-varying and time-constant variables can be artificial. Thus, in our discussion below, we refer to explanatory variables that either are or are not related to omitted variables.

7.3.1 Models of omitted variables

It is helpful to review the sampling basis for the model in order to describe how variations in sampling may induce omitted variable bias. The sampling can be described in two stages, see, for example, Section 3.1 and Ware (1985S). Specifically, we have:

Stage 1. Draw a random sample of \( n \) subjects from a population. The vector of subject-specific parameters \( \alpha_i \) is associated with the \( i \)th subject.

Stage 2. Conditional on \( \alpha_i \), draw realizations \( \{y_{it}, z_{it}, x_{it}\} \), for \( t = 1, \ldots, T_i \) for the \( i \)th subject. Summarize these draws as \( \{y_i, Z_i, X_i\} \).

We follow notation introduced in Section 7.2 and denote the observed independent variables by \( o_i = (z_{it}', x_{it}')' \), a \( (q+K) \times 1 \) vector of observed effects, and let \( o_i \), be the associated column vector, that is, \( o_i = (o_{it1}', \ldots, o_{itq}')' \).

We now use an “unobserved variable model” also considered by Palta and Yao (1991B), Palta, Yao and Velu (1994B), and others; Frees (2001S) provides a summary. (In the biological sciences, omitted variables are known as unmeasured confounders.) Here, we assume that the latent vector \( \alpha_i \) is independent of the observed variables, \( o_i \), yet we have omitted important, possibly time-varying, variables in the second sampling stage. Specifically, assume at the second stage we have

Stage 2*. Conditional on \( \alpha_i \), draw realizations \( \{y_i, Z_i, X_i, U_i\} \), for the \( i \)th subject.

Here, \( o_i = \{Z_i, X_i\} \) represents observed, independent variables whereas \( U_i \) represents the unobserved, independent variables. Moments of the dependent variables are specified as

\[
E[y_i | \alpha_i, o_i, U_i] = Z_i \alpha_i + X_i \beta + U_i \gamma
\]  

and

\[
Var[y_i | \alpha_i, o_i, U_i] = R_i.  
\]

Thus, \( \gamma \) is a \( g \times 1 \) vector of parameters that signals the presence of omitted variables \( U_i \). Using the same notation convention as \( o_i \), let \( u_i \) be the column vector associated with \( U_i \).

To estimate the model parameters, simplifying assumptions are required. One route is to make full distributional assumptions, such as multivariate normality, that permits estimation using maximum likelihood methods. In the following, we instead use assumptions on the sampling design because this permits estimation procedures that link back to Section 7.2 without the necessity of assuming that a distribution follows a particular parametric model.

To specify the model that is used for inference, further use \( \Sigma_{o_o} = \text{Cov}(u_i, o_i) \) to capture the correlation between unobserved and observed effects. For simplicity, we drop the \( i \) subscript on the \( T_i g \times T_i (q+K) \) matrix \( \Sigma_{o_o} \). Now, we only need be concerned with those observed variables that are related to the unobservables. Specifically, we may re-arrange the observables into two pieces, \( o_i = (o_i^{(1)}, o_i^{(2)}) \) so that \( \Sigma_{o_o} = (\text{Cov}(u_i, o_i^{(1)}) \text{Cov}(u_i, o_i^{(2)})) = (\Sigma_{o_o}^{(1)} 0) \). That is, the first piece of \( o_i \) is correlated to the unobservables whereas the second piece is not. To prevent an indirect relation between \( u_i \) and \( o_i^{(2)} \), we also assume that \( o_i^{(1)} \) and \( o_i^{(2)} \) have zero covariance. With these conventions, we assume that

\[
E[u_i | \alpha_i, o_i] = \Sigma_{o_o}^{(1)} (\text{Var } o_i^{(1)})^{-1} (o_i^{(1)} - \text{E} o_i^{(1)}).
\]  

(7.17)
A sufficient condition for (7.17) is joint normality of $o_i^{(1)}$ and $u_i$, conditional on $a_i$. An advantage of assuming equation (7.17) directly is that it also allows us to handle categorical variables within $o_i$. An implication of (7.17) is that $\text{Cov}(u_i, o_i^{(2)}) = 0$; however, no implicit distributional assumptions for $o_i^{(2)}$ are required.

For the sampling design of the observables, we assume that the explanatory variables are generated from an error components model. Specifically, we assume that

$$
\begin{align*}
\mathbb{E}[y_i | a_i, o_i] &= z_i' a_i + x_i' \beta + \gamma' o_i^{(2)} + \alpha_i\\
\text{Var}[y_i | o_i] &= R_i + Z_i D Z_i',
\end{align*}
$$

(7.18)

where $\beta_0 = -\gamma' o_i^{(2)}$ and $\alpha_i$ that we hypothesize will be correlated with the unobserved variables. This is a testable hypothesis. Further, equation (7.18) suggests that, by incorporating the terms $z_i^{(1)}$ and $x_i^{(1)}$ as regression terms in the analysis, we may avoid omitted variable bias. This is the topic of Section 7.3.2.

**Special case: Error components model with time-constant explanatory variables**

With the error components model, we have $q = 1$ and $z_i = 1$, so that equation (7.18) reduces to

$$
\mathbb{E}[y_i | a_i, x_i] = \beta_0^* + a_i + x_i' \beta + x_i^{(1)} \gamma_1^* + x_i^{(2)} \gamma_2^*.
$$

(7.19)

To estimate the coefficients of this conditional regression equation, we require that the explanatory variables not be linear combinations of one another. In particular, if there are any time-constant variables in $x_i$, they must not be included in $x_i^{(1)}$. In other words, we require that time-constant variables be uncorrelated with omitted variables.

7.3.2 **Augmented regression estimation**

This subsection considers an “augmented” regression model of the form

$$
\mathbb{E}[y_i | o_i] = X_i \beta + G_i \gamma
$$

(7.19)

and

$$
\text{Var}[y_i | o_i] = R_i + Z_i D Z_i'.
$$

(7.20)

Here, $\gamma$ is a $g \times 1$ vector of coefficients and $G_i$ is a known function of independent variables $\{o_i\}$, such as in equation (7.18). Further, $D$ is a positive definite variance-covariance matrix to be estimated. One may simply use weighted least squares to determine estimates of $\beta$ and $\gamma$. Explicitly, we use minimizers of the weighted sum of squares

$$
WSS(\beta, \gamma) = \sum_{i=1}^n (y_i - (X_i \beta + G_i \gamma))' W_i^{-1} (y_i - (X_i \beta + G_i \gamma))
$$

(7.21)
to define estimators of $\beta$ and $\gamma$. We denote these estimators as $b_{AR}$ and $\hat{\gamma}_{AR}$, respectively. Here, the “$AR$” subscript denotes “artificial regression” as in Davidson and MacKinnon (1990E) or “augmented regression” as in Arellano (1993E). The point is that no specialized software is required for the omitted variable estimator ($\hat{\gamma}_{AR}$) or omitted variable bias corrected regression coefficient estimator ($b_{AR}$).

Different choices of the $G_i$ permit us to accommodate different data features. To illustrate, it is easy to see, using $W_i = V_i$ and omitting $G_i$, that $b_{AR}$ reduces to $b_{GLS}$. Further, Frees (2001S) shows that $b_{AR}$ reduces to $b_{FE}$ when $W_i = V_i$. As another example, consider the random coefficients design. Here, we assume that $q = K$ and $x_{it} = z_{it}$. Thus, from equation (7.18), we have $E[y_{it} | a_i, 0] = \beta_0^* + x_{it}'(a_i + \beta) + \bar{x}_{i}(1)^*\gamma^*$. Hypothesizing that all variables are potentially related to omitted variables, we require $g = K$ and $G_i = 1_i\bar{x}'_i$. An advantage of the augmented regression formulation, compared to the Section 7.2.2 omitted variable test, is that it permits a direct assessment of the hypothesis of omitted variables $H_0: \gamma = 0$. This can be done directly using a Wald test statistic of the form $\chi^2_{AR} = \hat{\gamma}_{AR}'(\text{Var}\hat{\gamma}_{AR})^{-1}\hat{\gamma}_{AR}$. Here, $\text{Var}\hat{\gamma}_{AR}$ can be estimated based on the variance specification in equation (7.16) or using a robust alternative as introduced in Section 3.4.

Example - Income tax payments - Continued

We continue with the example from Section 7.2.3 by first considering the model with variable intercepts, two variable slopes and an $AR(1)$ serial correlation coefficient for the errors. The usual random effects (generalized least squares) estimators are presented in Table 7.3. These coefficients also appear in Tables 7.1 and 7.2, where we learned that this model suffered from omitted variable bias. Table 7.3 presents the fits from the augmented regression model, where we have augmented the regression using the averages from all the explanatory variables, $\bar{x}_i$. Table 7.3 shows that the averages of both of the income variables, LNTPI and MR, are statistically significantly different from zero. Further, from the overall test of $H_0: \gamma = 0$, the test statistic is $\chi^2_{AR} = 57.511$. Using a chi-square with 8 degrees of freedom, the $p$-value associated with this test is less than 0.0001. As in Section 7.2.3, this indicates serious potential problems from omitted variables and suggests using the robust, bias-corrected estimators in Table 7.3. Compared to the fixed effects estimators in Table 7.1, we see that the results are qualitatively similar. The variables DEPEND, PREP and HH are statistically significantly negative for both models whereas EMP and AGE are not statistically significant for either estimation procedure. The estimation procedures give different results for the marital status (MS) variable, with the augmented regression procedure. The advantage of the augmented regression procedure is that it permits estimation of a mean for the LNTPI and MR variables, something that was not possible using the fixed effects robust estimators.

The augmented regression procedures also allow us to give robust estimators for the random coefficients model. Table 7.3 summarizes estimation of the random coefficients model with an $AR(1)$ serial correlation coefficient for the errors. Table 7.3 also shows the robust estimators, calculated using augmentation with the averages from all the explanatory variables, $\bar{x}_i$. Again, the test statistic of $\chi^2_{AR} = 56.629$ indicates serious potential problems from omitted variables. The bias-corrected estimators are qualitatively similar to those of the augmented regression results for the variable intercept and two variable slope model.
Table 7.3. Comparison of Random Effects Estimators to Robust Alternatives.
Based on the Section 3.2 Example.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Random Effects Estimation</th>
<th>Augmented Regression Estimation</th>
<th>Random Effects Estimation</th>
<th>Augmented Regression Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Parameter Estimates</td>
<td>t-statistic</td>
<td>Parameter Estimates</td>
<td>t-statistic</td>
</tr>
<tr>
<td>LNTPI</td>
<td>2.270</td>
<td>13.40</td>
<td>2.351</td>
<td>11.94</td>
</tr>
<tr>
<td>MR</td>
<td>0.005</td>
<td>0.46</td>
<td>0.031</td>
<td>2.77</td>
</tr>
<tr>
<td>MS</td>
<td>-0.603</td>
<td>-3.86</td>
<td>-0.563</td>
<td>-3.11</td>
</tr>
<tr>
<td>HH</td>
<td>-0.729</td>
<td>-3.75</td>
<td>-1.056</td>
<td>-2.67</td>
</tr>
<tr>
<td>AGE</td>
<td>-0.359</td>
<td>-2.15</td>
<td>-0.386</td>
<td>-1.08</td>
</tr>
<tr>
<td>EMP</td>
<td>-0.661</td>
<td>-5.05</td>
<td>-0.352</td>
<td>-1.01</td>
</tr>
<tr>
<td>PREP</td>
<td>-0.300</td>
<td>-3.21</td>
<td>-0.296</td>
<td>-2.15</td>
</tr>
<tr>
<td>DEPEND</td>
<td>-0.138</td>
<td>-2.84</td>
<td>-0.177</td>
<td>-2.19</td>
</tr>
<tr>
<td>LNTPIAVG</td>
<td>-0.109</td>
<td>-7.08</td>
<td>-0.149</td>
<td>-7.08</td>
</tr>
<tr>
<td>MRAVG</td>
<td>-0.265</td>
<td>-1.07</td>
<td>-0.057</td>
<td>-0.21</td>
</tr>
<tr>
<td>MSAVG</td>
<td>0.398</td>
<td>0.95</td>
<td>0.057</td>
<td>0.95</td>
</tr>
<tr>
<td>HHAVG</td>
<td>-0.053</td>
<td>-0.13</td>
<td>0.565</td>
<td>0.79</td>
</tr>
<tr>
<td>AGEAVG</td>
<td>-0.489</td>
<td>-1.22</td>
<td>-0.351</td>
<td>-1.21</td>
</tr>
<tr>
<td>EMPAVG</td>
<td>0.039</td>
<td>0.22</td>
<td>0.089</td>
<td>0.53</td>
</tr>
<tr>
<td>DEPAVG</td>
<td>-0.007</td>
<td>-0.07</td>
<td>0.038</td>
<td>-0.43</td>
</tr>
<tr>
<td>PREPAVG</td>
<td>-0.153</td>
<td>3.38</td>
<td>0.118</td>
<td>2.70</td>
</tr>
</tbody>
</table>

\[ \chi^2_{AC} = 57.511 \]

\[ \chi^2_{AR} = 56.629 \]
Section 7.4 Sampling, selectivity bias and attrition

7.4.1 Incomplete and rotating panels

Historically, the primary approaches to panel/longitudinal data analysis in the econometric and biostatistics literatures assumed balanced data of the following form.

\[
\begin{array}{ccc}
\vdots & \vdots & \vdots \\
y_{11} & \cdots & y_{n1} \\
\vdots & \vdots & \vdots \\
y_{1T} & \cdots & y_{nT}
\end{array}
\]

This form suggested using techniques from multivariate analysis as described by, for example, Rao (1987B) and Chamberlain (1982E).

However, there are many ways in which a complete, balanced, set of observations may not be available due to delayed entry, early exit and intermittent nonresponse. This section begins by considering unbalanced situations where the lack of balanced is planned or designed by the analyst. Section 7.4.2 then discusses unbalanced situations that are not planned. When unplanned, we call the nonresponses missing data. We will be particularly concerned with situations in which the mechanisms for missingness are related to the response, discussed in Section 7.4.3.

To accommodate missing data, for subject \( i \) we use the notation \( T_i \) for the number of observations and \( t_{ij} \) to denote the time of the \( j \)th observation. Thus, the times for the available set of observations are \( \{t_{i1}, \ldots, t_{iT_i}\} \). For this chapter, we assume that these times are discrete and a subset of \( \{1, 2, \ldots, T\} \). Chapter 6 described the continuous time case. Define \( M_i \) to be the \( T_i \times T \) design matrix that has a “1” in the \( t_{ij} \)th column and \( j \)th row, and is zero otherwise, \( j = 1, \ldots, T_i \).

Specifically, with this design matrix, we have

\[
\begin{bmatrix}
t_{i1} \\
t_{i2} \\
\vdots \\
t_{iT_i}
\end{bmatrix}
= M_i \begin{bmatrix} 1 \\ \vdots \\ T \\ 2 \\ \vdots \\ 1 \end{bmatrix}.
\]  

(7.22)

Using this design matrix, most of the formulas in Chapters 1 through 6 carry through for planned missing data. The case of time-varying parameters turns out to be more complex; Appendix 8A.1 will illustrate this point in the context of the two-way model. For simplicity, in this text we use the notation \( \{1, 2, \ldots, T_i\} \) to denote an unbalanced observation set.

Panel surveys of people provide a good illustration of planned missing data; these studies are often designed to have incomplete observations. This is because we know that people become tired of responding to surveys on a regular basis. Thus, panel surveys are typically designed to “replenish” a certain proportion of the sample at each interview time. Specifically, for a rolling or rotating panel, a fixed proportion of individuals enter and exit the panel during each interview time.

Figure 7.4.1 illustrates three waves of a rotating panel. At time \( t_1 \), the first \( 3n \) individuals that comprise Wave 1 are interviewed. At time \( t_2 \), \( n \) individuals from Wave 1 are dropped and \( n \) new individuals that comprise Wave 2 are added. At time \( t_3 \), \( n \) individuals from Wave 1 are dropped and \( n \) new individuals that comprise Wave 3 are added. This pattern is continued, with \( n \) individuals dropped and added at each interview time. Thus, at each time, \( 3n \) individuals are interviewed. Each person stays in the survey for at most 3 time periods.
Figure 7.4.1. Three Waves of a Rotating Panel.
The solid lines indicate the individuals that are interviewed at each of five interview times.

<table>
<thead>
<tr>
<th>Wave 1</th>
<th>$i=1$</th>
<th>$i=n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wave 1</td>
<td>$i=n+1$</td>
<td>$i=2n$</td>
</tr>
<tr>
<td>Wave 1</td>
<td>$i=2n+1$</td>
<td>$i=3n$</td>
</tr>
<tr>
<td>Wave 2</td>
<td>$i=3n+1$</td>
<td>$i=4n$</td>
</tr>
<tr>
<td>Wave 3</td>
<td>$i=4n+1$</td>
<td>$i=5n$</td>
</tr>
</tbody>
</table>

Interview time: $t_1, t_2, t_3, t_4, t_5$

Number of subjects interviewed: $3n, 3n, 3n, 2n, n$

7.4.2 Unplanned nonresponse

Few difficulties arise when the data are not balanced due to a planned design. However, when the data are unbalanced due to unforeseen events, this lack of balance represents a potential source of bias. To provide some specific examples, we again return to panel surveys of people. Verbeek and Nijman (in Chapter 18 of Mátyás and P. Sevestre, 1996) provide the following list of types of unplanned nonresponse.

Types of panel survey nonresponse

- **Initial nonresponse.** A subject contacted cannot, or will not, participate. Because of limited information, this potential problem is often ignored in the analysis.
- **Unit nonresponse.** A subject contacted cannot, or will not, participate even after repeated attempts (in subsequent waves) to include the subject.
- **Wave nonresponse.** A subject does not respond for one or more time periods but does respond in the preceding and subsequent times (for example, the subject may be on vacation).
- **Attrition.** A subject leaves the panel after participating in at least one survey.

It is also of concern to deal with survey item nonresponse where the items are treated as covariates in the analysis. Here, information on one or more variables is missing. For example, individuals may not wish to report income, age, and so on. However, we restrict consideration to missing responses.

To understand the mechanisms that lead to unplanned nonresponse, we model it stochastically. Let $r_{ij}$ be an indicator variable for the $ij$th observation, with a one indicating that this response is observed and a zero indicating that the response is missing. Let $r = (r_{11}, \ldots, r_{1T}, \ldots, r_{n1}, \ldots, r_{nT})'$ summarize the data availability for all subjects. The interest is in whether or not the responses influence the missing data mechanism. For notation, we use $Y = (y_{11}, \ldots, y_{1T}, \ldots, y_{n1}, \ldots, y_{nT})'$ to be the collection of all potentially observed responses.
Missing data models

In the case where $Y$ does not affect the distribution of $r$, we follow Rubin (1976G) and call this case missing completely at random (MCAR). Specifically, the missing data are MCAR if $f(r \mid Y) = f(r)$, where $f(.)$ is a generic probability mass function. An extension of this idea is in Little (1995G), where the adjective “covariate dependent” is added when $Y$ does not affect the distribution of $r$, conditional on the covariates. If the covariates are summarized as $\{X, Z\}$, then the condition corresponds to the relation $f(r \mid Y, X, Z) = f(r \mid X, Z)$. To illustrate this point, consider an example of Little and Rubin (1987G) where $X$ corresponds to age and $Y$ corresponds to income of all potential observations. If the probability of being missing does not depend on income, then the missing data are MCAR. If the probability of being missing varies by age but does not by income over observations within an age group, then the missing data are covariate dependent MCAR. Under the latter specification, it is possible for the missing data to vary by income. For example, younger people may be less likely to respond to a survey. This shows that the “missing at random” feature depends on the purpose of the analysis. Specifically, it is possible that an analysis of the joint effects of age and income may encounter serious patterns of missing data whereas an analysis of income controlled for age suffers no serious bias patterns.

When the data are MCAR, Little and Rubin (1987G, Chapter 3) describe several approaches for handling partially missing data. One option is to treat the available data as if nonresponses were planned and use unbalanced estimation techniques. Another option is to utilize only subjects with a complete set of observations by discarding observations from subjects with missing responses. A third option is to impute values for missing responses. Little and Rubin note that each option is generally easy to carry out and may be satisfactory with small amounts of missing data. However, the second and third options may not be efficient. Further, each option implicitly relies heavily on the MCAR assumption.

Little and Rubin (1987G) advocate modeling the missing data mechanisms; they call this the model-based approach. To illustrate, consider a likelihood approach using a selection model for the missing data mechanism. Now, partition $Y$ into observed and missing components using the notation $Y = (Y_{obs}, Y_{miss})$. With the likelihood approach, we base inference on the observed random variables. Thus, we use a likelihood proportional to the joint function $f(r, Y_{obs})$. We also specify a selection model by specifying the conditional mass function $f(r \mid Y_{obs})$.

Suppose that the observed responses and the selection model distributions are characterized by a vectors of parameters $\theta$ and $\psi$, respectively. Then, with the relation $f(r, Y_{obs}, \theta, \psi) = f(Y_{obs}, \theta) \times f(r \mid Y_{obs}, \psi)$, we may express the log likelihood of the observed random variables as

$$L(\theta, \psi) = \log f(r, Y_{obs}, \theta, \psi) = \log f(Y_{obs}, \theta) + \log f(r \mid Y_{obs}, \psi).$$

In the case where the data are MCAR, then $f(r \mid Y_{obs}, \psi) = f(r \mid \psi)$ does not depend on $Y_{obs}$. Little and Rubin (1987G) also consider the case where the selection mechanism model distribution does not depend on $Y_{miss}$ but may depend on $Y_{obs}$. In this case, they call the data missing at random (MAR).

In both the MAR and MCAR cases, we see that the likelihood may be maximized over the parameters, separately for each case. In particular, if one is only interested in the maximum likelihood estimator of $\theta$, then the selection model mechanism may be “ignored.” Hence, both situations are often referred to as the ignorable case.
Example – Income tax payments

Let $y$ represent tax liability and $x$ represent income. Consider the following five selection mechanisms.

- The taxpayer is not selected (missing) with probability $\psi$ without regard to the level of tax liability. In this case, the selection mechanism is MCAR.
- The taxpayer is not selected if the tax liability is less than $100. In this case, the selection mechanism depends on the observed and missing response. The selection mechanism cannot be ignored.
- The taxpayer is not selected if the income is less than $20,000. In this case, the selection mechanism is MCAR, covariate dependent. That is, assuming that the purpose of the analysis is to understand tax liabilities conditional on knowledge of income, stratifying based on income does not seriously bias the analysis.
- The probability of a taxpayer being selected decreases with tax liability. For example, suppose the probability of being selected is $\logit(-\psi_y)$. In this case, the selection mechanism depends on the observed and missing response. The selection mechanism cannot be ignored.
- The taxpayer is followed over $T = 2$ periods. In the second period, a taxpayer is not selected if the first period tax is less than $100. In this case, the selection mechanism is MAR. That is, the selection mechanism is based on an observed response.

The second and fourth selection mechanisms represent situations where the selection mechanism must be explicitly modeled; these are known as non-ignorable cases. In these situations without explicit adjustments, procedures that ignore the selection effect may produce seriously biased results. To illustrate a correction for selection bias in a simple case, we outline an example due to Little and Rubin (1987G). Section 7.4.3 describes additional mechanisms.

Example- Historical heights

Little and Rubin (1987G) discuss data due to Wachter and Trusell (1982G) on $y$, the height of men recruited to serve in the military. The sample is subject to censoring in that minimum height standards were imposed for admission to the military. Thus, the selection mechanism is

\[ r_j = \begin{cases} 1 & \text{if } y_j > c_i \\ 0 & \text{otherwise} \end{cases}, \]

where $c_i$ is the known minimum height standard imposed at the time of recruitment. The selection mechanism is non-ignorable because it depends on the individual’s height.

For this example, additional information is available to provide reliable model inference. Specifically, based on other studies of male heights, we may assume that the population of heights is normally distributed. Thus, the likelihood of the observables can be written down and inference may proceed directly. To illustrate, suppose that $c_i = c$ is constant. Let $\mu$ and $\sigma$ denote the mean and standard deviation of $y$. Further suppose that we have a random sample of $n + m$ men in which $m$ men fall below the minimum standard height $c$ and we observe $Y_{obs} = (y_1, \ldots, y_n)'$. The joint distribution for observables is

\[
&\mathbb{P}(r, Y_{obs}, \mu, \sigma) = \mathbb{P}(Y_{obs}, \mu, \sigma) \times \mathbb{P}(r | Y_{obs}) \\
&= \prod_{i=1}^{n} \{f(y_i | y_i > c) \times \text{Prob}(y_i > c)\} \times \left(\text{Prob}(y \leq c)\right)^m.
\]

Now, let $\phi$ and $\Phi$ represent the density and distribution function for the standard normal distribution. Thus, the log-likelihood is
\[ L(\mu, \sigma) = \log f(r, Y_{obs}, \mu, \sigma) = \sum_{i=1}^{n} \log \left( \frac{1}{\sigma} \phi \left( \frac{y_i - \mu}{\sigma} \right) \right) + m \log \left( 1 - \Phi \left( \frac{c - \mu}{\sigma} \right) \right). \]

This is easy to maximize in \( \mu \) and \( \sigma \). If one ignored the censoring mechanisms, then one would derive estimates of the observed data from the “log likelihood,”

\[ \sum_{i=1}^{n} \log \left( \frac{1}{\sigma} \phi \left( \frac{y_i - \mu}{\sigma} \right) \right). \]

yielding different, and biased, results.

### 7.4.3 Non-ignorable missing data

For non-ignorable missing data, Little (1995G) recommends:

- Avoid missing responses whenever possible by using appropriate follow-up procedures.
- Collect covariates that are useful for predicting missing values.
- Collect as much information as possible regarding the nature of the missing data mechanism.

For the third point, if little is known about the missing data mechanism, then it is difficult to employ a robust statistical procedure to correct for the selection bias.

There are many models of missing data mechanisms. A general overview appears in Little and Rubin (1987G). Verbeek and Nijman (in Chapter 18 of Mátyás and P. Sevestre, 1996E) survey more recent econometric panel data literature. Little (1995G) surveys the problem of attrition. Rather than survey this developing literature, we give a few models of non-ignorable missing data.

**Heckman two-stage procedure**

Heckman (1976E) developed this procedure in the context of cross-sectional data. Because it relies on correlations of unobserved variables, it is also applicable to fixed effects panel data models. Thus, assume that the response model follows a one-way fixed effects. As introduced in Chapter 2, this model can be expressed as

\[ y_{it} = \alpha_i + x_{it}' \beta + \epsilon_{it}. \]

Further assume that the sampling response mechanism is governed by the latent (unobserved) variable \( r_{it}^{*} \) where

\[ r_{it}^{*} = w_{it}' \gamma + \eta_{it}. \]

The variables in \( w_{it} \) may or may not include the variables in \( x_{it} \). We observe \( y_{it} \) if \( r_{it}^{*} \geq 0 \), that is, if \( r_{it}^{*} \) crosses the threshold 0. Thus, we observe

\[ r_{it} = \begin{cases} 1 & \text{if } r_{it}^{*} \geq 0 \\ 0 & \text{otherwise} \end{cases}. \]

To complete the specification, we assume that \( \{(\epsilon_{it}, \eta_{it})\} \) are identically and independently distributed, and that the joint distribution of \( (\epsilon_{it}, \eta_{it}) \) is bivariate normal with means zero, variances \( \sigma^2 \) and \( \sigma^2 \eta \) and correlation \( \rho \). Note that if the correlation parameter \( \rho \) equals zero, then the response and selection models are independent. In this case, the data are MCAR and the usual estimation procedures are unbiased and asymptotically efficient.

Under these assumptions, basic multivariate normal calculations show that

\[ E(y_{it} | r_{it}^{*} \geq 0) = \alpha_i + x_{it}' \beta + \beta_{it}(w_{it}' \gamma), \]
where $\beta_\lambda = \rho \sigma$ and $\lambda(a) = \frac{\phi(a)}{\Phi(a)}$. Here, $\lambda$ is the inverse of the so-called “Mills ratio.” This calculation suggests the following two-step procedure for estimating the parameters of interest.

**Heckman’s two-stage procedure**

1. Use the data $\{ (r_{it}, w_{it}) \}$ and a probit regression model to estimate $\gamma$. Call this estimator $g_H$.
2. Use the estimator from stage 1 to create a new explanatory variable, $x_{it,K+1} = \lambda(w_{it}' g_H)$. Run a one-way fixed effects model using the $K$ explanatory variables $x_{it}$ as well as the additional explanatory variable $x_{it,K+1}$. Use $b_H$ and $b_\lambda H$ to denote the estimators of $\beta$ and $\beta_\lambda$, respectively.

Section 9.1.1 will introduce probit regressions. We also note that the two-step method does not work in absence of covariates to predict the response and, for practical purposes, requires variables in $w$ that are not in $x$ (see Little and Rubin, 1987).

To test for selection bias, we may test the null hypothesis $H_0: \beta_\lambda = 0$ in the second stage due to the relation $\beta_\lambda = \rho \sigma$. When conducting this test, one should use heteroscedasticity corrected standard errors. This is because the conditional variance $\text{Var}(y_{it} | r_{it}^* \geq 0)$ depends on the observation. Specifically, $\text{Var}(y_{it} | r_{it}^* \geq 0) = \sigma^2 (1 - \rho^2 \delta_{it})$, where $\delta_{it} = \lambda_{it} (\lambda_{it} + w_{it}' \gamma)$ and $\lambda_{it} = \phi(w_{it}' \gamma)/\Phi(w_{it}' \gamma)$.

This procedure assumes normality for the selection latent variables to form the augmented variables. Other distribution forms are available in the literature, including the logistic and uniform distributions. A deeper criticism, raised by Little (1985G), is that the procedure relies heavily on assumptions that cannot be tested using the data available. This criticism is analogous to the historical heights example where we relied heavily on the normal curve to infer the distribution of heights below the censoring point. Despite these criticisms, Heckman’s procedure is widely used in the social sciences.

**Hausman and Wise procedure**

To see how to extend the Heckman procedure to error component panel data models, we now describe a procedure originally due to Hausman and Wise (1979E); see also the development in Verbeek and Nijman (in Chapter 18 of Mátýás and P. Sevestre, 1996E). For simplicity, we work with the error components model described in Section 3.1, $y_{it} = \alpha_i + x_{it}' \beta + \epsilon_{it}$. We also assume that the sampling response mechanism is governed by the latent variable of the form $r_{it}^* = \xi_i + w_{it}' \gamma + \eta_{it}$.

This is also an error components model. The variables in $w_{it}$ may or may not include the variables in $x_{it}$. As before, $r_{it}$ indicates whether $y_{it}$ is observed and $r_{it} = \begin{cases} 1 & \text{if } r_{it}^* \geq 0 \\ 0 & \text{otherwise} \end{cases}$. The random variables $\alpha_i$, $\epsilon_{it}$, $\xi_i$ and $\eta_{it}$ are assumed to be jointly normal, each with mean zero and variance

$$
\text{Var} \left( \begin{array}{c} \alpha_i \\ \epsilon_{it} \\ \xi_i \\ \eta_{it} \end{array} \right) = \begin{pmatrix}
\sigma_\alpha^2 & 0 & \sigma_{\alpha \xi} & 0 \\
0 & \sigma_\epsilon^2 & 0 & \sigma_{\epsilon \eta} \\
\sigma_{\alpha \xi} & 0 & \sigma_\xi^2 & 0 \\
0 & \sigma_{\epsilon \eta} & 0 & \sigma_\eta^2
\end{pmatrix}
$$

If $\sigma_{\alpha \xi} = \sigma_{\epsilon \eta} = 0$, then the selection process is independent of the observation process.
It is easy to check that $b_{EC}$ is unbiased and consistent if $E(y_{it} \mid r_i) = x_i' \beta$. Under conditional normality, one can check that

$$E(y_{it} \mid r_i) = x_i' \beta + \frac{\sigma_\zeta^2}{T \sigma_\zeta^2 + \sigma_\eta^2} g_{it} + \frac{\sigma_\eta^2}{\sigma_\eta^2} \left( g_{it} - \frac{\sigma_\zeta^2}{T \sigma_\zeta^2 + \sigma_\eta^2} \sum_{i=1}^T g_{it} \right),$$

where $g_{it} = E(\xi_i + \eta_i \mid r_i)$. The calculation of $g_{it}$ involves the multivariate normal distribution that requires numerical integration. This calculation is straightforward although computationally intensive. Following this calculation, testing for ignorability and producing bias-corrected estimators proceeds as in the Heckman case. Chapter 9 will discuss the fitting of binary dependent responses to error component models. For other additional details, we refer the reader to Verbeek and Nijman (in Chapter 18 of Mátyás and P. Sevestre, 1996).

**EM algorithm**

Section 7.4 has focused on introducing specific models of non-ignorable nonresponse. General robust models of nonresponse are not available. Rather, a more appropriate strategy is to focus on a specific situation, collect as much information as possible regarding the nature of the selection problem and then develop a model for this specific selection problem.

The EM algorithm is a computational device for computing model parameters. Although specific to each model, it has found applications in a wide variety of models involving missing data. Computationally, the algorithm iterates between the “E,” for conditional expectation, and “M,” for maximization, steps. The E step finds the conditional expectation of the missing data given the observed data and current values of the estimated parameters. This is analogous to the time-honored tradition of imputing missing data. A key innovation of the EM algorithm is that one imputes sufficient statistics for missing values, not the individual data points. For the M step, one updates parameter estimates by maximizing an observed log likelihood. Both the sufficient statistics and the log likelihood depend on the model specification.

Many introductions of the EM algorithm are available in the literature. Little and Rubin (1987G) provide a detailed treatment.
7. Exercises and Extensions

Section 7.2

7.1 Fuller-Battese Transform

Consider the Section 3.1 error components model with $y_i = \alpha_i + X_i \beta + \epsilon_i$ and $\text{Var}(y_i) = \sigma^2 I_i + \sigma_\alpha^2 J_i$. Recall from Section 2.5.3 that $Q_i = I_i - Z_i (Z_i' Z_i)^{-1} Z_i'$.

- Define $\psi_i = \left( \frac{\sigma^2}{T_i \sigma_\alpha^2 + \sigma^2} \right)^{1/2}$ and $P_i = I_i - Q_i$. Show that $V_i^{-1/2} = \sigma^{-1}(P_i + \psi_i Q_i)$. 

- Transform the error components model using $y_i^* = (P_i + \psi_i Q_i) y_i$ and $X_i^* = (P_i + \psi_i Q_i) X_i$. Show that the generalized least squares estimator of $\beta$ is $b_{EC} = \left( \sum_{i=1}^n X_i' X_i^* \right)^{-1} \sum_{i=1}^n X_i^* y_i^*$. 

- Now consider the special case of the error components model in equation (7.1). Show that the generalized least squares estimator of $\beta_1$ is

$$b_{1,EC} = \frac{\sum_{i=1}^n \sum_{t=1}^T x_{it}^* y_{it}^* \left( \sum_{i=1}^n \psi_i^2 T_i \right) - \left( \sum_{i=1}^n \psi_i T_i \bar{y}_i \right) \left( \sum_{i=1}^n \psi_i T_i \bar{x}_i \right) \left( \sum_{i=1}^n \psi_i T_i \bar{y}_i \right) \left( \sum_{i=1}^n \psi_i T_i \bar{x}_i \right)}{\left( \sum_{i=1}^n \psi_i^2 T_i \right)^2 - \left( \sum_{i=1}^n \psi_i T_i \bar{x}_i \right)^2},$$

where $x_{it}^* = x_{it} - \bar{x}_i$ and $y_{it}^* = y_{it} - \bar{y}_i$. 

- Now consider the balanced data case so that $T_i = T$ for each $i$. Show that $b_{1,EC} = \frac{\sum_{i=1}^n \sum_{t=1}^T (x_{it}^* - \bar{x}_i) (y_{it}^* - \bar{y}_i) \left( \sum_{i=1}^n (x_{it}^* - \bar{x}_i) \right)^2}{\sum_{i=1}^n \sum_{t=1}^T (x_{it}^* - \bar{x}_i)^2}$. 

- For the model considered in part (d), show that variance of $b_{1,EC}$ is as given in equation (7.2).

7.2 Maddala’s decomposition

Consider the special case of the error components model in equation (7.1). 

- Show equation (7.6). That is, show

$$\text{Var}(b_{1,B}) = \left( \frac{T \sigma_\alpha^2 + \sigma_\epsilon^2}{\sum_{i=1}^n (\bar{x}_i - \bar{x})^2} \right) \left( T \sum_{i=1}^n (\bar{x}_i - \bar{x})^2 \right)^{-1}$$

where $b_{1,B} = \frac{\sum_{i=1}^n (\bar{x}_i - \bar{x}) (\bar{y}_i - \bar{y})}{\sum_{i=1}^n (\bar{x}_i - \bar{x})^2}$, as in equation (7.5). 

- Show equation (7.4). That is, show

$$\text{Var}(b_{1,FE}) = \sigma_\epsilon^2 \left( \frac{T}{\sum_{i=1}^n \sum_{t=1}^T (x_{it} - \bar{x}_i)^2} \right)^{-1}$$

where $b_{1,FE} = \frac{\sum_{i=1}^n \sum_{t=1}^T (x_{it} - \bar{x}_i) (y_{it} - \bar{y}_i)}{\sum_{i=1}^n \sum_{t=1}^T (x_{it} - \bar{x}_i)^2}$, as in equation (7.3). 

- Use parts (a) and (b) and the expressions for $b_{1,EC}$ and $\text{Var}(b_{1,EC})$ in Section 7.2.1 to show

$$\frac{1}{\text{Var}(b_{1,EC})} = \frac{1}{\text{Var}(b_{1,FE})} + \frac{1}{\text{Var}(b_{1,B})}.$$ 

d. Show equation (7.7). That is, with parts (a)-(c) and the expressions for $b_{1,EC}$ and $\text{Var}(b_{1,EC})$ in Section 7.2.1, show $b_{1,EC} = (1 - \Delta) b_{1,FE} + \Delta b_{1,B}$, where $\Delta = \frac{\text{Var}(b_{1,EC})}{\text{Var}(b_{1,B})}$. 

7.3. Mixed linear model estimation with intercepts

Consider the linear mixed effects model described in Section 3.3 where \( \mathbf{a} \) are treated as random, with mean \( \mathbb{E} \mathbf{a} = \mathbf{a} \) and variance-covariance matrix \( \text{Var} \mathbf{a} = \mathbf{D} \), independent of the error term.

Then, we may re-write the model as

\[
y_i = Z_i \mathbf{a} + X_i \mathbf{b} + \varepsilon_i^*,
\]

where \( \varepsilon_i^* = \varepsilon_i + Z_i (\mathbf{a}_i - \mathbf{a}) \) and \( \text{Var} \varepsilon_i^* = Z_i \mathbf{D} Z_i' + \mathbf{R}_i = \mathbf{V}_i \), a positive definite \( T_i \times T_i \) matrix.

a. Show that we can express the generalized least squares estimator of \( \mathbf{b} \) as

\[
\hat{\mathbf{b}}_{\text{GLS}} = \mathbf{C}^{-1}_{\text{GLS}} \sum_i \mathbf{X}_i' \mathbf{V}_i^{-1} (\sum_i \mathbf{Z}_i' \mathbf{V}_i^{-1} \mathbf{Z}_i)^{-1} \{ \sum_i \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{Z}_i \} \mathbf{y}_i
\]

with

\[
\mathbf{C}_{\text{GLS}} = \left( \sum_i \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{X}_i \right) - \left( \sum_i \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{Z}_i \right) \left( \sum_i \mathbf{Z}_i' \mathbf{V}_i^{-1} \mathbf{Z}_i \right)^{-1} \left( \sum_i \mathbf{Z}_i' \mathbf{V}_i^{-1} \mathbf{X}_i \right).
\]

b. Show that \( \text{Var} \hat{\mathbf{b}}_{\text{GLS}} = \mathbf{C}^{-1}_{\text{RE}} \).

c. Now consider the error components model so that \( q = 1 \), \( \mathbf{D} = \sigma_a^2 \), \( z_it = 1 \) and \( \mathbf{Z}_i = 1_i \). Use part (a) to show that

\[
\hat{\mathbf{b}}_{\text{EC}} = \left( \sum_i (\mathbf{X}_i' \mathbf{Q}_i \mathbf{X}_i + (1 - \zeta_i)T_i (\mathbf{x}_i - \bar{x}_w) (\mathbf{x}_i - \bar{x}_w)') \right)^{-1} \sum_i \mathbf{X}_i' \mathbf{Q}_i \mathbf{y}_i + (1 - \zeta_i)T_i (\mathbf{x}_i - \bar{x}_w) (\bar{y}_i - \bar{y}_w)
\]

where \( \mathbf{Q}_i = \mathbf{1}_i - \frac{1}{T_i} \mathbf{J} \), \( \bar{x}_w = \sum_{i=1}^n \zeta_i \mathbf{x}_i \) and \( \bar{y}_w = \sum_{i=1}^n \zeta_i \mathbf{y}_i \).

d. Consider part (c) and assume in addition that \( K = 1 \). Show that

\[
\hat{\mathbf{b}}_{\text{EC}} = \frac{1}{\sum_{i=1}^n \left( \sum_{t=1}^{T_i} (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i) + (1 - \zeta_i)T_i (\mathbf{x}_i - \bar{x}_w) (\bar{y}_i - \bar{y}_w) \right)} \sum_{i=1}^n \left( \sum_{t=1}^{T_i} (x_{it} - \bar{x}_i)^2 + (1 - \zeta_i)T_i (\mathbf{x}_i - \bar{x}_w)^2 \right)
\]

e. As in part (a), show that the generalized least squares estimator of \( \mathbf{a} \) is

\[
\hat{\mathbf{a}}_{\text{GLS}} = \left( \sum_i \mathbf{Z}_i' \mathbf{V}_i^{-1} \mathbf{Z}_i \right)^{-1} \left( \left( \sum_i \mathbf{Z}_i' \mathbf{V}_i^{-1} \mathbf{y}_i \right) - \left( \sum_i \mathbf{Z}_i' \mathbf{V}_i^{-1} \mathbf{x}_i \right) \hat{\mathbf{b}}_{\text{GLS}} \right).
\]

f. Show that, for the case considered in part (c) with \( q = 1 \), \( \mathbf{D} = \sigma_a^2 \), \( z_it = 1 \) and \( \mathbf{Z}_i = 1_i \), that

\[
\hat{\mathbf{a}}_{\text{EC}} = \bar{y}_w - \mathbf{x}_w \hat{\mathbf{b}}_{\text{EC}},
\]

where \( \hat{\mathbf{b}}_{\text{EC}} \) is given in part (c).

7.4. Robust estimation

Consider the linear mixed effects model described in Problem 7.3. Let

\[
\mathbf{C}_{\text{FE}} = \sum_{i=1}^n X_i' \mathbf{R}_i^{-1/2} Q_{Z,i} \mathbf{R}_i^{-1/2} X_i \quad \text{where} \quad Q_{Z,i} = \mathbf{I}_i - \mathbf{R}_i^{-1/2} Z_i Z_i' \mathbf{R}_i^{-1/2}. \quad \text{Recall that}
\]

\[
\hat{\mathbf{b}}_{\text{FE}} = \mathbf{C}_{\text{FE}}^{-1} \sum_{i=1}^n X_i' \mathbf{R}_i^{-1/2} Q_{Z,i} \mathbf{R}_i^{-1/2} \mathbf{y}_i.
\]

a. Show that \( \mathbb{E} \hat{\mathbf{b}}_{\text{FE}} = \mathbf{b} \).
b. Show that \( \text{Var} b_{FE} = C_{FE}^{-1} \).

### 7.5. Decomposing the random effects estimator

Consider the linear mixed effects model described in Problem 7.3. An alternative estimator of \( \beta \) is the so-called “between-groups” estimator, given as

\[
\mathbf{b}_B = C_B^{-1} \sum_{i=1}^{n} \left( \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{Z}_i \left( \mathbf{Z}_i' \mathbf{V}_i^{-1} \mathbf{Z}_i \right)^{-1} \mathbf{Z}_i' \mathbf{V}_i^{-1} - \left( \sum_{i=1}^{n} \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{Z}_i \right) \left( \sum_{i=1}^{n} \mathbf{Z}_i' \mathbf{V}_i^{-1} \mathbf{Z}_i \right)^{-1} \right) \mathbf{y}_i,
\]

where

\[
C_B = \sum_{i=1}^{n} \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{Z}_i \left( \mathbf{Z}_i' \mathbf{V}_i^{-1} \mathbf{Z}_i \right)^{-1} \mathbf{Z}_i' \mathbf{V}_i^{-1} - \left( \sum_{i=1}^{n} \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{Z}_i \right) \left( \sum_{i=1}^{n} \mathbf{Z}_i' \mathbf{V}_i^{-1} \mathbf{Z}_i \right)^{-1} \left( \sum_{i=1}^{n} \mathbf{Z}_i' \mathbf{V}_i^{-1} \mathbf{X}_i \right).
\]

a. Show that \( \text{Var} \mathbf{b}_B = C_B^{-1} \).

b. Now consider the error components model so that \( q = 1, \mathbf{D} = \sigma_{\alpha}^2, z_{\alpha} = 1 \) and \( \mathbf{Z}_i = 1 \). Use part (a) to show that

\[
\mathbf{b}_B = \left( \sum_{i=1}^{n} \mathbf{T}_i (1 - \zeta_i) \left( \bar{\mathbf{x}}_i - \bar{\mathbf{x}}_w \right) \left( \bar{\mathbf{y}}_i - \bar{\mathbf{y}}_w \right) \right)^{-1} \sum_{i=1}^{n} \mathbf{T}_i (1 - \zeta_i) \left( \bar{\mathbf{x}}_i - \bar{\mathbf{x}}_w \right) \left( \bar{\mathbf{y}}_i - \bar{\mathbf{y}}_w \right).
\]

c. Show that an alternative form for \( \mathbf{b}_B \) is

\[
\mathbf{b}_B = \left( \sum_{i=1}^{n} \zeta_i \left( \bar{\mathbf{x}}_i - \bar{\mathbf{x}}_w \right) \left( \bar{\mathbf{y}}_i - \bar{\mathbf{y}}_w \right) \right)^{-1} \sum_{i=1}^{n} \zeta_i \left( \bar{\mathbf{x}}_i - \bar{\mathbf{x}}_w \right) \left( \bar{\mathbf{y}}_i - \bar{\mathbf{y}}_w \right).
\]

d. Use equation (A.4) of Appendix A.5 to establish

\[
\mathbf{Z}_i' \mathbf{V}_i^{-1} \mathbf{Z}_i = \left( \mathbf{D} + \left( \mathbf{Z}_i' \mathbf{R}_i^{-1} \mathbf{Z}_i \right)^{-1} \right)^{-1}.
\]

e. Use part (d) to establish

\[
\mathbf{V}_i^{-1} - \mathbf{V}_i^{-1} \mathbf{Z}_i \left( \mathbf{Z}_i' \mathbf{V}_i^{-1} \mathbf{Z}_i \right)^{-1} \mathbf{Z}_i' \mathbf{V}_i^{-1} = \mathbf{R}_i^{-1} - \mathbf{R}_i^{-1} \mathbf{Z}_i \left( \mathbf{Z}_i' \mathbf{R}_i^{-1} \mathbf{Z}_i \right)^{-1} \mathbf{Z}_i' \mathbf{R}_i^{-1}.
\]

f. Use Problem 7.3(a), 7.4 and parts (a) –(e) to show that \( C_B + C_{FE} = C_{GLS} \), that is, show that

\[
(\text{Var} \mathbf{b}_B)^{-1} + (\text{Var} b_{FE})^{-1} = (\text{Var} \mathbf{b}_{GLS})^{-1}.
\]

g. Prove Maddala’s decomposition:

\[
\mathbf{b}_{GLS} = (\mathbf{I} - \Delta) \mathbf{b}_{FE} + \Delta \mathbf{b}_B,
\]

where \( \Delta = (\text{Var} \mathbf{b}_{GLS}) (\text{Var} \mathbf{b}_B)^{-1} \).

### 7.6. Omitted variable test

Consider the linear mixed effects model described in Problems 7.3, 7.4 and 7.5.

a. Show that \( \text{Cov}(\mathbf{b}_{GLS}, \mathbf{b}_{FE}) = \text{Var}(\mathbf{b}_{GLS}) \).

b. Use part (a) to show that

\[
\text{Var}(\mathbf{b}_{FE} - \mathbf{b}_{GLS}) = \text{Var}(\mathbf{b}_{FE}) - \text{Var}(\mathbf{b}_{GLS})\.
\]

c. Show that

\[
\chi^2_{FE} = (\mathbf{b}_{FE} - \mathbf{b}_{GLS})' \left( \text{Var}(\mathbf{b}_{FE}) - \text{Var}(\mathbf{b}_{GLS}) \right)^{-1} (\mathbf{b}_{FE} - \mathbf{b}_{GLS})
\]

has an asymptotic (as \( n \to \infty \)) chi-square distribution with \( K \) degrees of freedom.
Chapter 8. Dynamic Models

Abstract. This chapter considers models of longitudinal data sets with longer time dimensions than were considered in earlier chapters. With many observations per subject, analysts have several options for introducing more complex dynamic model features that address questions of interest or that represent important tendencies of the data (or both). One option is based on the serial correlation structure; this chapter extends the basic structures that were introduced in Chapter 2. Another dynamic option is to allow parameters to vary over time. Moreover, for a data set with a long time dimension relative to the number of subjects, we have an opportunity to model the cross-sectional correlation, an important issue in many studies. The chapter also considers the Kalman filter approach that allows the analyst to incorporate many of these features simultaneously. Throughout, the assumption of exogeneity of the explanatory variables is maintained. Chapter 6 considered lagged dependent variables as explanatory variables, another way of introducing dynamic features into the model.

8.1 Introduction

Because longitudinal data vary over time as well as in the cross-section, we have opportunities to model the dynamic, or temporal, patterns in the data. For the data analyst, when is it important to consider dynamic aspects of a problem?

Part of the answer to this question rests on the purpose of the analysis. If the main inferential task is forecasting of future observations as introduced in Chapter 4, then the dynamic aspect is critical. In this instance, every opportunity for understanding dynamic aspects should be explored. In contrast, in other problems the focus is on understanding relations among variables. Here, the dynamic aspects may be less critical. This is because many models still provide the basis for constructing unbiased estimators and reliable testing procedures when dynamic aspects are ignored, at the price of efficiency. To illustrate, for problems with large sample sizes (in the cross-section), efficiency may not be an important issue. Nonetheless, understanding the dynamic correlation structure is important for achieving efficient parameter estimators; this aspect can be vital, especially for data sets with many observations over time.

The importance of dynamics is influenced by the size of the data set, both through

- the choice of the statistical model and
- the type of approximations used to establish properties of parameter estimators.

For many longitudinal data sets, the number of subjects \( (n) \) is large relative to the number of observations per subject \( (T) \). This suggests the use of regression analysis techniques; these methods are designed to understand relationships among variables, observed and unobserved, and to account for subject-level heterogeneity. In contrast, for other problems, \( T \) is large relative to \( n \). This suggests borrowing from other statistical methodologies, such as multivariate time series. Here, although relationships among variables are important, understanding temporal patterns is the focus of this methodology. We remark that the modeling techniques presented in Chapters 1-5
are based on the linear model. In contrast, Section 8.5 presents a modeling technique from the multivariate time series literature, the Kalman filter.

The sample size also influences the properties of our estimators. For longitudinal data sets where \( n \) is large compared to \( T \), this suggests the use of asymptotic approximations where \( T \) is bounded and \( n \) tends to infinity. However, for other data sets, we may achieve more reliable approximations by considering instances where \( n \) and \( T \) approach infinity together or where \( n \) is bounded and \( T \) tends to infinity. For many models, this distinction is not an important one for applications. However, for some models, such as the fixed effects lagged dependent variable model in Section 6.3, the difference is critical. There, the approach where \( T \) is bounded and \( n \) tends to infinity leads to biased parameter estimators.

This chapter deals with problems where the dynamic aspect is important, either because of the inferential purposes underlying the problem or the nature of the data set. We now outline several approaches that are available for incorporating dynamic aspects into a longitudinal data model.

Perhaps the easiest way for handling dynamics is to let one of the explanatory variables be a proxy for time. For example, we might use \( x_{it} = t \), for a linear trend in time model. Another technique is to use “time dummy variables,” that is, binary variables that indicate the presence or absence of a period effect. To illustrate, in Chapter 2, we introduced the two-way model

\[
y_{it} = \alpha_i + \lambda_t + x_{it}' \beta + \epsilon_{it}.
\]  

(8.1)

Here, the parameters \( \{ \lambda_t \} \) are time-specific quantities that do not depend on subjects. Chapter 2 considered the case where \( \{ \lambda_t \} \) were fixed parameters.

In Chapter 3, we allowed \( \{ \lambda_t \} \) to be random. Section 8.3 extends this idea by allowing several parameters in the longitudinal data model to vary with time. To illustrate, one example that we will consider is

\[
y_{it} = x_{it}' \beta_t + \epsilon_{it},
\]

that is, where regression parameters \( \beta \) vary over time.

Unlike cross-sectional data, with longitudinal data we also have the ability to accommodate temporal trends by looking at changes in either the response or the explanatory variables. This technique is straightforward and natural in some areas of application. To illustrate, when examining stock prices, because of financial economics theory, we examine proportional changes in prices, which are simply returns. As another example, we may wish to analyze the model

\[
\Delta y_{it} = \alpha_t + x_{it}' \beta + \epsilon_{it}.
\]  

(8.2)

where \( \Delta y_{it} = y_{it} - y_{i,t-1} \) is the change, or difference, in \( y_{it} \). In general, one must be wary of this approach because you lose \( n \) (initial) observations when differencing.

Re-writing equation (8.2), we have

\[
y_{it} = \alpha_t + y_{i,t-1} + x_{it}' \beta + \epsilon_{it}.
\]

A generalization of this is

\[
y_{it} = \alpha_t + \gamma y_{i,t-1} + x_{it}' \beta + \epsilon_{it},
\]  

(8.3)

where \( \gamma \) is a parameter to be estimated. If \( \gamma = 1 \), then the model in equation (8.3) reduces to the model in equation (8.2). If \( \gamma = 0 \), then the model in equation (8.3) reduces to our “usual” one-way model. Thus, the parameter \( \gamma \) is a measure of the relationship between \( y_{it} \) and \( y_{i,t-1} \). Because it measures the regression of \( y_{i,t-1} \) on \( y_{it} \), it is called an autoregressive parameter. The model in equation (8.3) is an example of a lagged dependent variable model that was introduced in Section 6.3.
Another way to formulate an autoregressive model is

$$y_{it} = \alpha_i + x_{it}' \beta + \epsilon_{it}, \quad (8.4)$$

where $\epsilon_{it} = \rho \epsilon_{i,t-1} + \eta_{it}$. Here, the autoregression is on the disturbance term, not the response. The models in equations (8.3) and (8.4) are similar, yet they differ in some important aspects. To see this, use equation (8.4) twice, to get

$$y_{it} - \rho y_{i,t-1} = (\alpha_i + x_{it}' \beta + \epsilon_{it}) - \rho (\alpha_i + x_{i,t-1}' \beta + \epsilon_{i,t-1}) = \alpha^* + (x_{it} - \rho x_{i,t-1})' \beta + \eta_{it}$$

where $\alpha^* = \alpha(1-\rho)$. Thus, equation (8.4) is similar to equation (8.3) with $\gamma = \rho$; the difference lies in the variable associated with $\beta$. Section 8.2 explores further the modeling strategy of assuming serial correlation directly on the disturbance terms in lieu of the response. There, Section 8.2 notes that because of the assumption of bounded $T$, one need not assume stationarity of errors. This strategy was used implicitly in Chapters 1-5 for handling the dynamics of longitudinal data.

Finally, Section 8.5 shows how to adapt the Kalman filter technique to longitudinal data analysis. The Kalman filter approach is a flexible technique that allows analysts to incorporate time-varying parameters and broad patterns of serial correlation structures into the model. Further, we will show how to use this technique to simultaneously model cross-sectional, heterogeneity, temporal aspects as well as spatial patterns.

### 8.2 Serial correlation models

One approach for handling the dynamics is through the specification of the covariance structure of the disturbance term, $\epsilon$. This section examines stationary and non-stationary specifications of the correlation structure for equally spaced data and then introduces options for data that may not be equally spaced.

#### 8.2.1 Covariance structures

Recall from Section 2.5.1 that $R = \text{Var}(\epsilon)$ is a $T \times T$ temporal variance-covariance matrix. Here, the element in the $r$th row and $s$th column is denoted by $R_{rs}$. For the $i$th subject, we define $\text{Var}(\epsilon_i) = R_{i}(\tau)$, a $T_i \times T_i$ submatrix of $R$ that can be determined by removing the rows and columns of $R$ that correspond to responses not observed. We denote this dependence of $R$ on parameters using $R(\tau)$. Here, $\tau$ is the vector of unknown parameters, called variance components. Section 2.5.1 introduced four specifications of $R$: (i) no correlation, (ii) compound symmetry, (iii) autoregressive of order one and (iv) unstructured.

The autoregressive model of order one is a standard representation used in time series analysis. This field of study also suggests alternative correlation structures. For example, one could entertain autoregressive models of higher order. Further, moving average models suggest the “Toeplitz” specification of $R$:

- $R_{rs} = \sigma_{|r-s|}$. This defines elements of a Toeplitz matrix.
- $R_{rs} = \sigma_{|r-s|}$ for $|r-s| < \text{band}$ and $R_{rs} = 0$ for $|r-s| \geq \text{band}$. This is the banded Toeplitz matrix.

When the band is “$q$”+1, this Toeplitz specification corresponds to a moving average model of order $q$, also known as an $MA(q)$ structure. More complex autoregressive, moving average models may be handled in a similar fashion, see, for example, Jennrich and Schlucuter (1986B).

The Toeplitz specification suggests a general linear variance structure of the form

$$R = \tau_1 R_1 + \tau_2 R_2 + \ldots + \tau_{\dim(\tau)} R_{\dim(\tau)}$$
where \( \text{dim}(\tau) \) is the dimension of \( \tau \) and \( R_1, R_2, \ldots, R_{\text{dim}(\tau)} \) are known matrices. As pointed out in Section 3.5.3 on MIVQUE estimation, this general structure accommodates many, although not all (such as autoregressive) covariance structures.

Another broad covariance structure suggested by the multivariate analysis literature is the factor-analytic structure of the form \( R = \Lambda \Lambda' + \Psi \), where \( \Lambda \) is a matrix of unknown factor loadings and \( \Psi \) is an unknown diagonal matrix. An important advantage of the factor analytic specification is that it easily allows the data analyst to ensure that the estimated variance matrix will be positive (or non-negative) definite, which can be important in some applications.

The covariance structures were described in the context of specification of \( R \), although they also apply to specification of \( \text{Var} \alpha_i = D \).

### 8.2.2 Nonstationary structures

With large \( n \) and a bounded \( T \), we need not restrict ourselves to stationary models. For example, we have already considered the unstructured model for \( R \). Making this specification imposes no additional restrictions on \( R \), including stationarity.

The primary advantage of stationary models is that they provide parsimonious representations for the correlation structure. However, parsimonious nonstationary models are also possible. To illustrate, suppose that the subject-level dynamics are specified through a random walk model \( \varepsilon_{it} = \varepsilon_{i,t-1} + \eta_{it} \). Here, \( \{ \eta_{it} \} \) is an i.i.d. sequence with \( \text{Var} \eta_{it} = \sigma^2_\eta \), which is independent of \( \{ \varepsilon_{i0} \} \). With the notation \( \text{Var} \varepsilon_{i0} = \sigma^2_0 \), we have \( \text{Var} \varepsilon_{it} = \sigma^2_0 + t \sigma^2_\eta \) and \( \text{Cov} (\varepsilon_{ir}, \varepsilon_{is}) = \text{Var} \varepsilon_{ir} = \sigma^2_0 + r \sigma^2_\eta \), for \( r < s \). This yields \( R = \sigma^2_0 \mathbf{J} + \sigma^2_\eta R_{RW} \), where

\[
R_{RW} = \begin{pmatrix}
1 & 1 & 1 & \cdots & 1 \\
1 & 2 & 2 & \cdots & 2 \\
1 & 2 & 3 & \cdots & 3 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 2 & 3 & \cdots & T
\end{pmatrix}
\]

Note that \( R \) is a function of only two unknown parameters. Further, this representation allows us to specify a nonstationary model without differencing the data (and thus without losing the initial set of observations). As shown in Exercise 4.6, this matrix has a simple inverse that can speed computations when \( T \) is large.

More generally, consider \( \varepsilon_{it} = \rho \varepsilon_{i,t-1} + \eta_{it} \) which is similar to the AR(1) specification except that we no longer require stationarity so that we may have \( |\rho| \geq 1 \). To specify covariances, we first define the function

\[
S(\rho) = 1 + \rho^2 + \cdots + \rho^{2(t-1)} = \begin{cases} 
\frac{t}{1-\rho^2} & \text{if } |\rho| < 1 \\
\frac{1-\rho^{2t}}{1-\rho^2} & \text{if } |\rho| \neq 1
\end{cases}
\]

Pleasant calculations show that \( \text{Var} \varepsilon_{it} = \sigma^2_0 + \sigma^2_\eta S(\rho) \) and \( \text{Cov} (\varepsilon_{ir}, \varepsilon_{is}) = \rho^{s-r} \text{Var} \varepsilon_{ir} \), for \( r < s \). This yields \( R = \sigma^2_0 R_{AR}(\rho) + \sigma^2_\eta R_{RW}(\rho) \), where

\[
R_{AR}(\rho) = \begin{pmatrix}
1 & \rho & \rho^2 & \cdots & \rho^{r-1} \\
\rho & 1 & \rho & \cdots & \rho^{r-2} \\
\rho^2 & \rho & 1 & \cdots & \rho^{r-3} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\rho^{r-1} & \rho^{r-2} & \rho^{r-3} & \cdots & 1
\end{pmatrix}
\]

and
A simpler expression assumes that $\varepsilon_i$ is a constant, either known or a parameter to be estimated (Section 8.5 will discuss an estimation method using the Kalman filter). In this case, we have $\sigma_0^2 = 0$ and $\mathbf{R}^{-1} = \frac{1}{\sigma_0^2} \mathbf{R}^{-1}_{RW}(\rho)$. It is easy to see that the Cholesky square root of $\mathbf{R}^{-1}_{RW}(\rho)$ is

$$
\mathbf{R}^{-1/2}_{RW}(\rho) = \begin{pmatrix}
1 & 0 & 0 & \ldots & 0 & 0 \\
-\rho & 1 & 0 & \ldots & 0 & 0 \\
0 & \rho & 1 & \ldots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \ldots & 1 & 0 \\
0 & 0 & 0 & \ldots & -\rho & 1
\end{pmatrix}
$$

This suggests using the transformation

$$
\begin{pmatrix}
y_{i1}^* \\
y_{i2}^* \\
\vdots \\
y_{iT_i}^*
\end{pmatrix} = \mathbf{R}^{-1/2}_{RW}(\rho) \begin{pmatrix}
y_{i1} \\
y_{i2} \\
\vdots \\
y_{iT_i}
\end{pmatrix} = \begin{pmatrix}
y_{i1} \\
y_{i2} - \rho y_{i1} \\
\vdots \\
y_{iT_i} - \rho y_{iT_i-1}
\end{pmatrix},
$$

which is the same as the Prais-Winston transform except for the first row. The Prais-Winston transform is the usual one for a stationary specification. The point of this example is that we do not require $|\rho| < 1$ and thus do not require stationarity.

### 8.2.3 Continuous time correlation models

When data are not equally spaced in time, a natural formulation is to still consider subjects drawn from a population, yet with responses as realizations of a continuous-time stochastic process. The continuous-time stochastic process setting is natural in the context of biomedical applications where, for example, we can envision patients arriving at a clinic for testing at irregularly spaced intervals. Specifically, for each subject $i$, we denote the set of responses as $\{y(t), t \in \mathbb{R}\}$. Here, $t$ denotes that time of the observation that we allow to extend over the entire real line, $\mathbb{R}$. In this context, it is convenient to use the subscript “$j$” for the order of observations within a subject while still using “$i$” for the subject. Observations of the $i$th subject are taken at time $t_{ij}$ so that $y_{ij} = y(i(t_{ij}))$ denotes the $j$th response of the $i$th subject. Similarly, let $\mathbf{x}_i$ and $\mathbf{z}_i$ denote sets of explanatory variables at time $t_{ij}$. We may then model the disturbance terms, $\varepsilon_{ij} = y_{ij} - (\mathbf{z}_{ij}'\mathbf{a}_i + \mathbf{x}_{ij}'\mathbf{b})$ as realizations from a continuous time stochastic process $\varepsilon_i(t)$. In many applications, this is assumed to be a mean zero Gaussian process, which permits a wide choice of correlation structures.

Particularly for unequally spaced data, a parametric formulation for the correlation structure is useful. In this setting, we have $\mathbf{R}_\varepsilon = \text{Cov}(\varepsilon_{ir}, \varepsilon_{is}) = \sigma^2 \rho(\|t_r - t_s\|)$, where $\rho$ is the correlation function of $\{\varepsilon_i(t)\}$. Two widely used choices include the \textit{exponential correlation model}

$$
\rho(u) = \exp(-\phi u), \text{ for } \phi > 0 \quad (8.5)
$$

and the \textit{Gaussian correlation model}

$$
\rho(u) = \exp(-\phi u^2), \text{ for } \phi > 0. \quad (8.6)
$$
Diggle, Heagerty, Liang and Zeger (2002S) provide additional details regarding the continuous-time model.

Another advantage of continuous time stochastic process models is that they easily permit indexing by orderings other than time. By far, the most interesting ordering other than time is a spatial ordering. Spatial orderings are of interest when we wish to model phenomena where responses that are close to one another are related to one another.

For some applications, it is straightforward to incorporate spatial correlations into our models. This is done by allowing $d_{ij}$ to be some measure of spatial or geographical location of the $j$th observation of the $i$th subject. Then, using a measure such as Euclidean distance, we interpret $|d_{ij} - d_{ik}|$ to be the distance between the $jth$ and $k$th observations of the $i$th subject. One could use the correlation structure in either of the equations (8.5) or (8.6).

Another straightforward approach that handles other applications is to reverse the role of $i$ and $j$, allowing $i$ to represent the time period (or replication) and $j$ to represent the subject. To illustrate, suppose that we consider observing purchases of insurance in each of the fifty states in the US over ten years. Suppose that most of the heterogeneity is due to the period effects, that is, changes in insurance purchases are influenced by changes in the country-wide economy. Because insurance is regulated at the state level, we expect each state to have different experiences due to local regulation. Further, we may be concerned that states close to one another share similar economic environments and thus will be more related to one another than states that are geographically distant.

With this reversal of notation, the vector $y_i$ represents all the subjects in the $i$th time period and the term $\alpha_i$ represents temporal heterogeneity. However, in the basic linear mixed effects model, this approach essentially ignores cross-sectional heterogeneity and treats the model as successive independent cross-sections. More details on this approach are in Section 8.2.

To see how to allow for cross-sectional heterogeneity, temporal dynamics and spatial correlations simultaneously, consider a basic two-way model

$$y_{it} = \alpha_i + \lambda_t + x_{it}' \beta + \epsilon_{it}$$

where, for simplicity, we assume balanced data. Stacking over $i$, we have

$$\begin{pmatrix} y_{1t} \\ y_{2t} \\ \vdots \\ y_{nt} \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix} + I_n \lambda_t + \begin{pmatrix} x_{1t}' \\ x_{2t}' \\ \vdots \\ x_{nt}' \end{pmatrix} \beta + \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \\ \vdots \\ \epsilon_{nt} \end{pmatrix},$$

where $I_n$ is a $n \times 1$ vector of ones. We re-write this as

$$y_i = \alpha + I_n \lambda_t + X_i \beta + \epsilon_i .$$

Define $H = \text{Var} \, \epsilon_i$ to be the spatial variance matrix, which we assume does not vary over time. Specifically, the $j$th element of $H$ is $H_{ij} = \text{Cov} (\epsilon_{i0}, \epsilon_{j0}) = \sigma^2 \rho (|d_i - d_j|)$, where $d_i$ is a measure of geographic location. Assuming that $\{\lambda_t\}$ is i.i.d. with variance $\sigma^2$, we have

$$\text{Var} \, y_i = \text{Var} \, \alpha + \sigma^2 I_n + \text{Var} \, \epsilon_i + \sigma^2 J_s + H = \sigma^2 I_n + V_H .$$

Stacking over $t$, we may express equation (8.7) as a special case of the mixed linear model, with $y = (y_1', \ldots, y_T')'$. Because $\text{Cov} (y_s, y_r) = \sigma^2 I_s$ for $s \neq r$, we have $V = \text{Var} \, y = \sigma^2 I_s \otimes J_T + V_H \otimes I_T$. It is easy to verify that $V^{-1} = (\sigma^2 I_s + T V_H^{-1} - T V_H^{-1}) \otimes J_T + V_H^{-1} \otimes I_T$.

Thus, it is straightforward to compute the regression coefficient estimator and the likelihood, as in equation (3.20). For example, with $X = (X_1', \ldots, X_T')'$, the generalized least squares estimator of $\beta$ is
Chapter 8. Dynamic Models

8-7

\[
\mathbf{b}_{GLS} = (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V}^{-1} \mathbf{y}
\]

\[
= \left( \sum_{t=1}^{T} \sum_{s=1}^{T} \mathbf{X}'_t \left( (\sigma^2 \mathbf{I}_n + T \mathbf{V}_H)^{-1} - T \mathbf{V}_H^{-1} \right) \mathbf{X}_s + \sum_{t=1}^{T} \mathbf{X}'_t \mathbf{V}_H^{-1} \mathbf{X}_t \right)^{-1}
\]

\[
\left( \sum_{t=1}^{T} \sum_{s=1}^{T} \mathbf{X}'_t \left( (\sigma^2 \mathbf{I}_n + T \mathbf{V}_H)^{-1} - T \mathbf{V}_H^{-1} \right) \mathbf{y}_s + \sum_{t=1}^{T} \mathbf{X}'_t \mathbf{V}_H^{-1} \mathbf{y}_t \right).
\]

Returning to the simpler case of no subject heterogeneity, suppose that \( \sigma^2 = 0 \). In this case, we have

\[
\mathbf{b}_{GLS} = \left( \sum_{t=1}^{T} \mathbf{X}'_t \mathbf{V}_H^{-1} \mathbf{X}_t \right)^{-1} \left( \sum_{t=1}^{T} \mathbf{X}'_t \mathbf{V}_H^{-1} \mathbf{y}_t \right).
\]

8.3 Cross-sectional correlations and time-series cross-section models

Cross-sectional correlations are particularly important in studies of governmental units, such as states or nations. In some fields, such as political science, when \( T \) is large relative to \( n \) the data are referred to as time-series cross-section data. This nomenclature distinguishes this set-up from the panel context, where \( n \) is large relative to \( T \). To illustrate, according to Beck and Katz (1995O), time-series cross-section data would typically range from 10 to 100 subjects with each subject observed over a long period, perhaps 20 to 50 years; many cross-national studies have ratios of \( n \) to \( T \) that are close to 1. Such studies involve economic, social or political comparisons of countries or states; because of the linkages among governmental units, the interest is in models that permit substantial contemporaneous correlation.

Following the political science literature, we consider a time-series cross-section (TSCS) model of the form

\[
y_i = \mathbf{X}_i \beta + \epsilon_i, \quad i = 1, \ldots, n,
\]

that summarizes \( T_i \) responses over time. Unlike prior chapters, we allow for correlation across different subjects through the notation \( \text{Cov}(\epsilon_i, \epsilon_j) = \mathbf{V}_{ij} \). Because \( n \) is not large relative to \( T \), fixed effects heterogeneity terms could easily be incorporated into the regression coefficients \( \beta \) by using binary indicator (dummy) variables. Incorporation of random effect heterogeneity terms would involve an extension of the current discussion; we follow the literature and ignore this aspect, for now. Stimson (1985O) surveys a range of models of interest in political science.

To complete the specification of the TSCS model, we need to make an assumption about the form of \( \mathbf{V}_{ij} \). Four basic specifications of cross-sectional covariances are:

- \( \mathbf{V}_{ij} = \begin{cases} \sigma^2 \mathbf{I}_m & i = j \\ 0 & i \neq j \end{cases} \). This is the traditional model set-up in which ordinary least squares is efficient.

- \( \mathbf{V}_{ij} = \begin{cases} \sigma^2 \mathbf{I}_m & i = j \\ 0 & i \neq j \end{cases} \). This specification permits heterogeneity across subjects.

- \( \text{Cov}(\epsilon_i, \epsilon_j) = \begin{cases} \sigma_{ts} & t = s \\ 0 & t \neq s \end{cases} \). This specification permits cross-sectional correlations across subjects. However, observations from different time points are uncorrelated.
- Cov(\(\varepsilon_{it}\), \(\varepsilon_{js}\)) = \(\sigma_{ij}\) for \(t = s\) and \(\varepsilon_{i,t} = \rho_i \varepsilon_{i,t-1} + \eta_i\).

This specification permits contemporaneous cross-correlations as well as intra-subject serial correlation through an \(AR(1)\) model. Moreover, with some mild additional assumptions, the model has an easy to interpret cross-lag correlation function of the form \(\text{Cov}(\varepsilon_{it}, \varepsilon_{jt}) = \sigma_{ij} \rho_j^{t-s}\) for \(s < t\).

The TSCS model is estimated using feasible generalized least squares procedures. At the first stage, ordinary least square residuals are used to estimate the variance parameters. One can think of the model as \(n\) separate regression equations and use seemingly unrelated regression techniques, described in Section 6.4.2, to compute estimators. It was in the context of seemingly unrelated regressions that Parks (1967S) proposed the contemporaneous cross-correlation with intra-subject serial \(AR(1)\) correlation model.

Generalized least square (GLS) estimation in a regression context has drawbacks that are well documented, see, for example, Carroll and Ruppert (1988G). That is, GLS estimators are more efficient than ordinary least squares (OLS) estimators when the variance parameters are known. However, because variance parameters are rarely known, one must use instead feasible GLS estimators. Asymptotically, feasible GLS are just as efficient as GLS estimators. However, in finite samples, feasible GLS may be more or less efficient than OLS estimators, depending on the regression design and distribution of disturbances. For the TSCS model that allows for cross-sectional covariances, there are \(n(n+1)/2\) variance parameters. Moreover, for the Parks model, there are additional \(n\) serial correlation parameters. As documented by Beck and Katz (1995O) in the TSCS context, having this many variance parameters means that feasible GLS estimators are inefficient in regression designs that are typically of interest in political science applications.

Thus, Beck and Katz (1995O) recommend using OLS estimators of regression coefficients. To account for the cross-sectional correlations, they recommend using standard errors that are robust to the presence of cross-sectional correlations that they call panel-corrected standard errors. In our notation, this is equivalent to the robust standard errors introduced in Section 2.5.3 without the subject-specific fixed effects yet reversing the roles of \(i\) and \(t\). That is, for the asymptotic theory, we now require independence over time yet allow for (cross-sectional) correlation across subjects.

Specifically, for balanced data, one computes panel-corrected standard errors as:

**Procedure for computing panel-corrected standard errors**

1. Calculate OLS estimators of \(\beta\), \(b_{OLS}\), and the corresponding residuals, \(e_{it} = y_{it} - x_{it}' b_{OLS}\).
2. Define the estimator of the \((ij)\)th cross-sectional covariance to be \(\hat{\sigma}_{ij} = T^{-1} \sum_{t=1}^{T} e_{it} e_{jt}\).
3. Estimate the variance of \(b_{OLS}\) using \(\left(\sum_{t=1}^{n} X_i' X_i\right)^{-1} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \hat{\sigma}_{ij} X_i' X_j\right) \left(\sum_{t=1}^{n} X_i' X_i\right)^{-1}\).

For unbalanced data, steps 2 and 3 need to be modified to align data from the same time periods.

Beck and Katz (1995O) provide simulation studies that establish that the robust \(t\)-statistics resulting from the use of panel-corrected standard errors are preferable to the ordinary \(t\)-statistics, using either OLS or feasible GLS. They also argue that this procedure can be used with serial \(AR(1)\) correlation, by first applying a (Prais-Winston) transformation to the data to induce independence over time. Using simulation, they demonstrate that this procedure is superior to the feasible GLS estimator using the Parks model. For general applications, we caution the reader...
that by reversing the roles of \( i \) and \( t \), one now relies heavily on the independence over time (instead of subjects). The presence of even mild serial correlation means that the usual same asymptotic approximations are no longer valid. Thus, although panel-corrected standard errors are indeed robust to presence of cross-sectional correlations, to use these procedures one must be especially careful about modeling the dynamic aspects of the data.

### 8.4 Time-varying coefficients

#### 8.4.1 The model

Beginning with the basic two-way model in equation (8.1), more generally, we use subject-varying terms

\[
\alpha_{i,t} + \ldots + \alpha_{i,t,q} = z_{\alpha,i} \alpha_i
\]

and time-varying terms

\[
\lambda_{i,t} + \ldots + \lambda_{i,t,r} = z_{\lambda,i} \lambda_i.
\]

With these terms, we define the longitudinal data mixed model with time-varying coefficients as

\[
y_{it} = z_{\alpha,i} \alpha_i + z_{\lambda,i} \lambda_i + x_{it} \beta + \varepsilon_t, \quad t = 1, \ldots, T_i, \quad i = 1, \ldots, n. \tag{8.9}
\]

Here, \( \alpha_i = (\alpha_{i,1}, \ldots, \alpha_{i,q})' \) is a \( q \times 1 \) vector of subject-specific terms and \( z_{\alpha,i} = (z_{\alpha,i,1}, \ldots, z_{\alpha,i,q})' \) is the corresponding vector of covariates. Similarly, \( \lambda_i = (\lambda_{i,1}, \ldots, \lambda_{i,r})' \) is a \( r \times 1 \) vector of time-specific terms and \( z_{\lambda,i} = (z_{\lambda,i,1}, \ldots, z_{\lambda,i,r})' \) is the corresponding vector of covariates. We use the notation \( t = 1, \ldots, T_i \) to indicate the unbalanced nature of the data.

A more compact form of equation (8.9) can be given by stacking over \( t \). This yields a matrix form of the longitudinal data mixed model

\[
y_i = Z_{\alpha,i} \alpha_i + Z_{\lambda,i} \lambda_i + X_i \beta + \varepsilon_i, \quad i = 1, \ldots, n. \tag{8.10}
\]

This expression uses matrices of covariates \( X_i = \left( x_{i1}, x_{i2}, \ldots, x_{iT_i} \right)' \), of dimension \( T_i \times K \), \( Z_{\alpha,i} = (z_{\alpha,i,1}, z_{\alpha,i,2}, \ldots, z_{\alpha,i,T_i})' \), of dimension \( T_i \times q \) matrix and

\[
Z_{\lambda,i} = \begin{pmatrix}
z_{\lambda,i,1}' & 0 & \ldots & 0 \\
0 & z_{\lambda,i,2}' & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & z_{\lambda,i,T_i}'
\end{pmatrix} : 0_{j}
\]

of dimension \( T_i \times rT \), where \( 0_{j} \) is a \( T_i \times r(T-T_i) \) zero matrix. Finally, \( \lambda = (\lambda_1, \ldots, \lambda_r)' \) is the \( rT \times 1 \) vector of time-specific coefficients.

We assume that sources of variability, \( \{ \varepsilon_i \} \), \( \{ \alpha_i \} \) and \( \{ \lambda_i \} \), are mutually independent and mean zero. The non-zero means are accounted for in the “\( \beta \)” parameters. The disturbances are independent between subjects, yet we allow for serial correlation and heteroscedasticity through the notation \( \text{Var} \varepsilon_i = \sigma^2 R_i \). Further, we assume that the subject-specific effects \( \{ \alpha_i \} \) are random with variance-covariance matrix \( \sigma^2 D \), a \( q \times q \) positive definite matrix. Time-specific effects \( \lambda \) have variance-covariance matrix \( \sigma^2 \Sigma_\lambda \), a \( rT \times rT \) positive definite matrix. For each variance component, we separate out the scale parameter \( \sigma^2 \) to simplify the estimation calculations described in Appendix 8A.2. With this notation, we may express the variance of each subject as

\[
\text{Var} y_i = \sigma^2 (V_{\alpha,i} + Z_{\lambda,i} \Sigma_\lambda Z_{\lambda,i}') \text{ where } V_{\alpha,i} = Z_{\alpha,i} D Z_{\alpha,i}' + R_i.
\]

To see how the model in equation (8.10) is a special case of the mixed linear model, take

\[
y = (y_1', y_2', \ldots, y_n')', \quad \varepsilon = (\varepsilon_1', \varepsilon_2', \ldots, \varepsilon_n')', \quad \alpha = (\alpha_1', \alpha_2', \ldots, \alpha_n')',
\]
\[
X = \begin{pmatrix}
X_1 \\
X_2 \\
\vdots \\
X_n
\end{pmatrix}, \quad Z_{\lambda} = \begin{pmatrix}
Z_{\lambda,1} \\
Z_{\lambda,2} \\
\vdots \\
Z_{\lambda,n}
\end{pmatrix} \quad \text{and} \quad Z_u = \begin{pmatrix}
Z_{u,1} & 0 & 0 & \cdots & 0 \\
0 & Z_{u,2} & 0 & \cdots & 0 \\
0 & 0 & Z_{u,3} & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & Z_{u,n}
\end{pmatrix}.
\]

With these choices, we can express the model in equation (8.10) as a mixed linear model, given by
\[
y = Z_{\alpha} \alpha + Z_{\lambda} \lambda + X \beta + \varepsilon.
\]

### 8.4.2 Estimation

By writing equation (8.10) as a mixed linear model, we may appeal to the many estimation results for this latter class of models. To illustrate, for known variance parameters, direct calculations show that the generalized least squares (GLS) estimator of \( \beta \) is
\[
b_{GLS} = (X'V^{-1}X)^{-1}X'V^{-1}y
\]

where \( \text{Var} \ y = \sigma^2 V \). Further, there is a variety of ways of calculating estimators of variance components. These include maximum likelihood, restricted maximum likelihood and several unbiased estimation techniques. For smaller data sets, one may use mixed linear model software directly. For larger data sets, direct appeal to such software may be computationally burdensome. This is because the time-varying random variables \( \lambda_t \) are common to all subjects, obliterating the independence among subjects. However, computational shortcuts are available and are described in detail in Appendix 8A.2.

---

**Example – Forecasting Wisconsin lottery sales – continued**

Table 8.1 reports the estimation results from fitting the two-way error components model in equation (8.1), with and without an AR(1) term. For comparison purposes, the fitted coefficients for the one-way model with an AR(1) term are also presented in this table. As in Table 4.3, we see that the goodness of fit statistic, \( AIC \), indicates that the more complex two-way models provide an improved fit compared to the one-way models. As with the one-way models, the autocorrelation coefficient is statistically significant even with the time-varying parameter \( \lambda_t \). In each of the three models in Table 8.1, only the population size (POP) and education levels (MEDSCHYR) have a significant effect on lottery sales.
**Table 8.1 Lottery Model Coefficient Estimates**

Based on in-sample data of \( n = 50 \) ZIP codes and \( T = 35 \) weeks.

The response is (natural) logarithmic sales.

<table>
<thead>
<tr>
<th>Variable</th>
<th>One-way Error components model with ( AR(1) ) term</th>
<th>Two-way Error components model</th>
<th>Two-way Error components model with ( AR(1) ) term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>13.821 2.18</td>
<td>16.477 2.39</td>
<td>15.897 2.31</td>
</tr>
<tr>
<td>PERPERHH</td>
<td>-1.085 -1.36</td>
<td>-1.210 -1.43</td>
<td>-1.180 -1.40</td>
</tr>
<tr>
<td>MEDSCHYR</td>
<td>-0.821 -2.53</td>
<td>-0.981 -2.79</td>
<td>-0.948 -2.70</td>
</tr>
<tr>
<td>MEDHVL</td>
<td>0.014 0.81</td>
<td>0.001 0.71</td>
<td>0.001 0.75</td>
</tr>
<tr>
<td>PRCRENT</td>
<td>0.032 1.53</td>
<td>0.028 1.44</td>
<td>0.029 1.49</td>
</tr>
<tr>
<td>PRC55P</td>
<td>-0.070 -1.01</td>
<td>-0.071 -1.00</td>
<td>-0.072 -1.02</td>
</tr>
<tr>
<td>HHMEDAGE</td>
<td>0.118 1.09</td>
<td>0.118 1.06</td>
<td>0.120 1.08</td>
</tr>
<tr>
<td>MEDINC</td>
<td>0.043 1.58</td>
<td>0.004 1.59</td>
<td>0.004 1.59</td>
</tr>
<tr>
<td>POPULATN</td>
<td>0.057 2.73</td>
<td>0.001 5.45</td>
<td>0.001 4.26</td>
</tr>
<tr>
<td>NRETAIL</td>
<td>0.021 0.20</td>
<td>-0.009 -1.07</td>
<td>-0.003 -0.26</td>
</tr>
<tr>
<td>( \text{Var } \alpha \left( \sigma^2_{\alpha} \right) )</td>
<td>0.528 0.564</td>
<td>0.554</td>
<td></td>
</tr>
<tr>
<td>( \text{Var } \varepsilon \left( \sigma^2_{\varepsilon} \right) )</td>
<td>0.279 0.022</td>
<td>0.024</td>
<td></td>
</tr>
<tr>
<td>( \text{Var } \lambda \left( \sigma^2_{\lambda} \right) )</td>
<td>0.241 0.241</td>
<td>0.241</td>
<td></td>
</tr>
<tr>
<td>( AR(1) \text{ corr } (\rho) )</td>
<td>0.555 25.88</td>
<td>0.518 25.54</td>
<td></td>
</tr>
<tr>
<td>( AIC )</td>
<td>2,270.97 -1,109.61</td>
<td>-1,574.02</td>
<td></td>
</tr>
</tbody>
</table>

### 8.4.3 Forecasting

For forecasting, we wish to predict

\[
y_{t,L+L} = z'_{t,L+L} \alpha_i + \lambda_{t,L+L} + x'_{t,L+L} \beta + \varepsilon_{t,L+L},
\]

(8.11)

for \( L \) lead time units in the future. We use Chapter 4 results for best linear unbiased prediction (BLUP). To calculate these predictors, we use the sum of squares \( S_{ZZ} = \sum_{i=1}^{n} z'_{i} V a_{i} z_{i} \). The details of the derivation of BLUPs are in Appendix 8A.3.

As intermediate results, it is useful to provide the BLUPs for \( \lambda_i, \varepsilon_i \), and \( \alpha_i \). The BLUP of \( \lambda_i \) turns out to be

\[
\lambda_{BLUP} = (S_{ZZ} + \Sigma^{-1}_{\lambda})^{-1} \sum_{i=1}^{n} z'_{i} V a_{i} e_{i,GLS},
\]

(8.12)

where we use the vector of residuals \( e_{i,GLS} = y_i - X_i b_{GLS} \). The BLUP of \( \varepsilon_i \) turns out to be

\[
e_{i,BLUP} = R_i V a_{i} (e_{i,GLS} - Z_{i} \lambda_{BLUP})
\]

(8.13)

and the BLUP of \( \alpha_i \) is

\[
a_{i,BLUP} = D Z_{i} a_{i} V a_{i} (e_{i,GLS} - Z_{i} \lambda_{BLUP}).
\]

(8.14)

We remark that the BLUP of \( \varepsilon_i \) can also be expressed as

\[
e_{i,BLUP} = y_i - (Z_{i} \alpha_i a_{i,BLUP} + Z_{i} \lambda_{BLUP} + X_i b_{GLS}).
\]
With these quantities, the BLUP forecast of \( y_{i,T_i+L} \) is
\[
\hat{y}_{i,T_i+L} = x'_{i,T_i+L} b_{GLS} + z'_{a,i,T_i+L} a_{i, BLUP} + z'_{k,i,T_i+L} \text{Cov}(\lambda_{T_i+L}, \lambda_i) \left( \sigma^2 \Sigma_\lambda \right)^{-1} \lambda_{BLUP} + \text{Cov}(\epsilon_{i,T_i+L}, \epsilon_i) \left( \sigma^2 R_j \right)^{-1} e_{i, BLUP}.
\]
(8.15)

An expression for the variance of the forecast error, \( \text{Var}(\hat{y}_{i,T_i+L} - y_{i,T_i+L}) \), is given in Appendix 8A.3, equation (8A.20).

Equations (8.12) - (8.15) provide sufficient structure to calculate forecasts for a wide variety of models. Additional computational details appear in Appendix 8A.3. Still, it is instructive to interpret the BLUP forecast in a number of special cases. We first consider the case of independently distributed time-specific components \( \{ \lambda_t \} \).

**Example 8.4.1 – Independent time-specific components**

We consider the special case where \( \{ \lambda_t \} \) are independent and assume that \( T_i + L > T \), so that \( \text{Cov}(\lambda_{T_i+L}, \lambda_i) = 0 \). Thus, from equation (8.14), we have the BLUP forecast of \( y_{i,T_i+L} \) is
\[
\hat{y}_{i,T_i+L} = x'_{i,T_i+L} b_{GLS} + z'_{a,i,T_i+L} a_{i, BLUP} + \text{Cov}(\epsilon_{i,T_i+L}, \epsilon_i) \left( \sigma^2 R_j \right)^{-1} e_{i, BLUP}.
\]
(8.16)

This is similar in appearance to the forecast formula in Chapter 4 equation (4.14). However, note that even when \( \{ \lambda_t \} \) are independent, the time-specific components appear in \( b_{GLS} \), \( e_{i, BLUP} \) and \( a_{i, BLUP} \). Thus, the presence of \( \{ \lambda_t \} \) influences the forecasts.

**Example 8.4.2 – Time-varying coefficients**

Suppose that the model is
\[
y_{it} = x_{it}' \beta_t + \epsilon_{it},
\]
where \( \{ \beta_t \} \) are i.i.d. We can re-write this as:
\[
y_{it} = z_{\lambda,i,t}' \lambda_i + x_{it}' \beta + \epsilon_{it},
\]
where \( E \beta_t = \beta, \lambda_i = \beta_t - \beta \) and \( z_{\lambda,i,t} = x_{i,t} \). With this notation and equation (8.16), the forecast of \( y_{i,T_i+L} \) is
\[
\hat{y}_{i,T_i+L} = x'_{i,T_i+L} b_{GLS}.
\]

**Example 8.4.3 – Two-way error components model**

Consider the basic two-way model given in equation (8.1), with \( \{ \lambda_t \} \) being independent and identically distributed. Here, we have that \( q = r = 1 \) and \( D = \sigma_\alpha^2 / \sigma^2 \), \( z_{a,i,T_i+L} = 1, Z_{a, i} = 1_i \). Further, \( Z_{\lambda,i} = (I_i \ 0_i) \), where \( 0_i \) is a \( T_i \times (T-T_i) \) matrix of zeroes, and \( \Sigma_\lambda = (\sigma_\lambda^2 / \sigma^2) I_T \). Thus, from equation (8.16), we have that the BLUP forecast of \( y_{i,T_i+L} \) is
\[
\hat{y}_{i,T_i+L} = a_{i, BLUP} + x'_{i,T_i+L} b_{GLS}.
\]

Here, from equation (8.12), we have
\[
\lambda_{BLUP} = \left( \sum_{i=1}^n Z_{\lambda,i}' \Sigma^{-1}_{\lambda,i} Z_{\lambda,i} + \frac{\sigma_\alpha^2}{\sigma^2} I_T \right)^{-1} \sum_{i=1}^n Z_{\lambda,i}' \Sigma^{-1}_{\lambda,i} e_{i,GLS},
\]
where \( Z_{\lambda,i} \) is given above, \( \Sigma^{-1}_{\lambda,i} = I_i - \frac{\zeta_i}{T_i} J_i \) and \( \zeta_i = \frac{T_i \sigma_\alpha^2}{\sigma^2 + T_i \sigma_\alpha^2} \). Further, equation (8.16) yields
8. Dynamic Models

For additional interpretation, we assume balanced data so that $T_i = T$; see Baltagi (1995E, page 38). To ease notation, recall $\zeta = \frac{T\sigma^2_\alpha}{\sigma^2 + T\sigma^2_\alpha}$. Here, we have

$$\hat{y}_{i,T+t} = x_{i,T+t}' b_{GLS} + \zeta \left( (\bar{y}_i - \bar{x}_i) b_{GLS} \right) - \frac{n(1 - \zeta)\sigma^2_\lambda}{\sigma^2 + n(1 - \zeta)\sigma^2_\lambda} \left( \bar{y}_i - \bar{x}_i b_{GLS} \right).$$

Example 8.4.4 – Random walk model

Through minor modifications, other temporal patterns of common, yet unobserved, components can be easily included. For this example, we assume that $r = 1$, $\{\lambda_t\}$ are identically and independently distributed, so that the partial sum process $\{\lambda_1 + \lambda_2 + \ldots + \lambda_t\}$ is a random walk process. Thus, the model is

$$y_{it} = z_{a,i} + \sum_{t=1}^T \lambda_t x_i' \beta + \epsilon_{it}, \quad t = 1, \ldots, T, \quad i = 1, \ldots, n. \quad (8.17)$$

Stacking over $t$, this can be expressed in matrix form as equation (8.10) where the $T_i \times T$ matrix $Z_{a,i}$ is a lower triangular matrix of 1’s for the first $T_i$ rows, and zero elsewhere. That is,

$$Z_{a,i} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 1 \\ 1 & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 & 2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 1 & 1 & 1 & \cdots & 0 & 0 & \cdots & 0 & T_i \end{bmatrix}.$$

Thus, it can be shown that

$$\hat{y}_{i,T+t} = x_{i,T+t}' b_{GLS} + \sum_{t=1}^T \lambda_t x_i' b_{GLS} + \sum_{t=1}^T z_{a,i} \lambda_t x_i' + \text{Cov}(\epsilon_{i,T+t}, \epsilon_i)(\sigma^2 R_i)^{-1} \epsilon_{i,BLUP}.$$

8.5 Kalman filter approach

The Kalman filter approach originated in time series analysis. It is a technique for estimating parameters from complex, recursively specified, systems. The essential idea is to use techniques from multivariate normal distributions to express the likelihood recursively, in an easily computable fashion. Then, parameters may be derived using maximum or restricted maximum likelihood. If this is your first exposure to Kalman filters, please skip ahead to the example in Section 8.6 and the introduction of the basic algorithm in Appendix D.

We now consider a class of models known as state space models. These models are well known in the time series literature for their flexibility in describing broad categories of dynamic structures. As we will see, they can be readily fit using the Kalman fit algorithm. These models have been explored in the longitudinal data analysis literature extensively by Jones (1993S). We use recent modifications introduced by Tsimikas and Ledolter (1998S) of this structure for linear mixed effects models. Specifically, we consider equation (8.9), which, in the time series literature, is called the observation equation. The time-specific quantities of equation (8.9) are $\lambda_t = (\lambda_t, \ldots, \lambda_T)$; this vector is our primary mechanism for specifying the dynamics. It is updated recursively through the transition equation,

$$\lambda_t = \Phi_{tt} \lambda_{t-1} + \eta_t. \quad (8.18)$$

Here, $\{\eta_t\}$ are identically and independently distributed, mean zero, random vectors. With state space models, it is also possible to incorporate a dynamic error structure such as an $\text{AR}(p)$ model. The autoregressive of order $p$ (AR($p$)) model for the disturbances $\{\epsilon_t\}$ has the form
where \( \{z_{i,t}\} \) are initially assumed to be identically and independently distributed, mean zero, random variables. Harvey (1989S) illustrates the wide range of choices of dynamic error structures. Further, we shall see that state space models readily accommodate spatial correlations among the responses.

Both the linear mixed effects models and the state space models are useful for forecasting. Because there is an underlying continuous stochastic process for the disturbances, both allow for unequally spaced (in time) observations. Furthermore, both accommodate missing data. Both classes of models can be represented as special cases of the linear mixed model. For state space models, the relationship to linear mixed models has been emphasized by Tsimikas and Ledolter (1994S, 1997S, 1998S).

Perhaps because of their longer history, we find that the linear mixed effects models are easier to implement. These models are certainly adequate for data sets with shorter time dimensions. However, for longer time dimensions, the additional flexibility provided by the newer state space models leads to improve model fitting and forecasting. We first express the longitudinal data model in equations (8.9) and (8.10) as a special case of a more general state space model. To this end, this section considers the transition equations, the set of observations available and the measurement equation. It then describes how to calculate the likelihood associated with this general state space model.

The Kalman filter algorithm is a method for efficiently calculating the likelihood of complex time series using conditioning arguments. Appendix D introduces the general idea of the algorithm, as well as extensions to include fixed and random effects. This section presents only the computational aspects of the algorithm.

### 8.5.1 Transition equations

We first collect the two sources of dynamic behavior, \( \varepsilon \) and \( \lambda \), into a single transition equation. Equation (8.18) specifies the behavior of the latent, time-varying, variable \( \lambda_t \). For the measurement error \( \varepsilon \), we assume that it is governed by a Markovian structure. Specifically, we use an \( AR(p) \) structure as specified in equation (8.19).

To this end, define the \( p \times 1 \) vector \( \xi_{i,t} = (\varepsilon_{i,t}, \varepsilon_{i,t-1}, \ldots, \varepsilon_{i,t-p+1})' \) so that we may write

\[
\begin{pmatrix}
\xi_{i,t} \\
\varepsilon_{i,t}
\end{pmatrix} = \Phi \begin{pmatrix}
\xi_{i,t-1} \\
\varepsilon_{i,t-1}
\end{pmatrix} + \begin{pmatrix}
\eta_{i,t} \\
0
\end{pmatrix}
\]

The first row is the \( AR(p) \) model in equation (8.19). Stacking this over \( i=1, \ldots, n \) yields

\[
\xi_t = \begin{pmatrix}
\xi_{1,t} \\
\vdots \\
\xi_{n,t}
\end{pmatrix} = \left( \Phi \otimes \Phi \right) \xi_{t-1} + \begin{pmatrix}
\eta_{1,t} \\
\vdots \\
\eta_{n,t}
\end{pmatrix} = \left( I_n \otimes \Phi \right) \xi_{t-1} + \eta_{2t}.
\]

Here, \( \xi_t \) is an \( np \times 1 \) vector, \( I_n \) is an \( n \times n \) identity matrix and \( \otimes \) is a Kronecker (direct) product (see Appendix A.6). We assume that \( \{z_{i,t}\} \) are identically distributed with mean zero and variance \( \sigma^2 \). The spatial correlation matrix is defined as \( H_n = \text{Var}(\xi_{1,t}, \ldots, \xi_{n,t}) / \sigma^2 \), for all \( t \). We assume no cross-temporal spatial correlation so that \( \text{Cov}(\xi_{i,t}, \xi_{j,s}) = 0 \) for \( s \neq t \). Thus,
}\), where \( \mathbf{0}_{p-1} \) is a \((p-1) \times (p-1)\) zero matrix. To initialize the recursion, we use \( \zeta_{i,0} = 0 \).

We may now collect the two sources of dynamic behavior, \( \varepsilon \) and \( \lambda \), into a single transition equation. From equations (8.18) and (8.19), we have

\[
\begin{bmatrix}
\mathbf{\Lambda}_t \\
\mathbf{\Omega}_t
\end{bmatrix} 
= 
\begin{bmatrix}
\mathbf{\Phi}_1 \mathbf{\Lambda}_{t-1} \\
\mathbf{\Phi}_2 \mathbf{\Omega}_{t-1}
\end{bmatrix} + 
\begin{bmatrix}
\mathbf{\eta}_t \\
\mathbf{\xi}_t
\end{bmatrix} + 
\begin{bmatrix}
\mathbf{\lambda}_t \\
\mathbf{\xi}_t
\end{bmatrix}
= 
\begin{bmatrix}
\mathbf{T} \delta_{t-1} + \mathbf{\eta}_t \\
\mathbf{\xi}_t
\end{bmatrix}.
\]

We assume that \( \{\lambda_i\} \) and \( \{\xi_i\} \) are independent stochastic processes and express the variance using

\[
Q_t = \text{Var} \mathbf{\eta}_t = \begin{bmatrix}
\text{Var} \mathbf{\eta}_t \\
0
\end{bmatrix} = \sigma^2 \begin{bmatrix}
\mathbf{H}_n \\
0
\end{bmatrix} \begin{bmatrix}
1 & 0 \\
0 & 0
\end{bmatrix} \begin{bmatrix}
\mathbf{H}_n \\
0
\end{bmatrix} = \sigma^2 \mathbf{Q}^*_t.
\]

Finally, to initialize the recursion, we assume that \( \delta_0 \) is a vector of parameters to be estimated.

### 8.5.2 Observation set

To allow for unbalanced data, we use notation analogous to that introduced in Section 5.4.2. Specifically, let \( \{i_1, \ldots, i_n\} \) denote the set of subjects that are available at time \( t \), where \( \{i_1, \ldots, i_n\} \subseteq \{1, 2, \ldots, n\} \). Further, define \( \mathbf{M}_t \) to be the \( n \times n \) design matrix that has a “1” in the \( i_j \)th column and zero otherwise, \( j = 1, \ldots, n \). With this design matrix, we have

\[
\begin{bmatrix}
\mathbf{\alpha}_{i_1} \\
\mathbf{\alpha}_{i_2} \\
\vdots \\
\mathbf{\alpha}_{i_n}
\end{bmatrix} = \begin{bmatrix}
\mathbf{M}_t \otimes \mathbf{I}_q
\end{bmatrix} \begin{bmatrix}
\mathbf{a}_1 \\
\mathbf{a}_2 \\
\vdots \\
\mathbf{a}_n
\end{bmatrix}.
\]

Similarly, with \( \varepsilon_{i,t} = (1 \ 0 \ \ldots \ 0) \mathbf{\xi}_{i,t} \), we have

\[
\begin{bmatrix}
\varepsilon_{i_1,t} \\
\varepsilon_{i_2,t} \\
\vdots \\
\varepsilon_{i_n,t}
\end{bmatrix} = \begin{bmatrix}
\mathbf{I}_n \otimes (1 \ 0 \ \ldots \ 0)
\end{bmatrix} \begin{bmatrix}
\mathbf{\xi}_{i_1,t} \\
\mathbf{\xi}_{i_2,t} \\
\vdots \\
\mathbf{\xi}_{i_n,t}
\end{bmatrix}.
\]

With this observation set for the \( t \)th time period and equation (8.9), we may write

\[
\mathbf{\xi}_{i,t} = \begin{bmatrix}
(\mathbf{M}_t \otimes \mathbf{I}_p)
\end{bmatrix} \begin{bmatrix}
\mathbf{\xi}_{i_1,t} \\
\mathbf{\xi}_{i_2,t} \\
\vdots \\
\mathbf{\xi}_{i_n,t}
\end{bmatrix}.
\]
With equations (8.22) and (8.23), we have
\[
\begin{align*}
y_t &= X_t \beta + Z_{a,t} \alpha + Z_{k,t} \lambda_t + W_{it} \xi_t, \\
&= X_t \beta + Z_t \alpha + W_t \delta_t,
\end{align*}
\] (8.25)

where
\[
X_t = \begin{pmatrix} x'_{i,t} \\ x'_{i,t} \\ \vdots \\ x'_{i,n,t} \end{pmatrix}, \quad Z_t = Z_{a,t} = \begin{pmatrix} z'_{a,i,t} \\ 0 \\ \vdots \\ 0 \end{pmatrix}
\]
and
\[
W_{it} = \begin{pmatrix} M_r & W_{it} \end{pmatrix} \begin{pmatrix} \lambda_t \\ \xi_t \end{pmatrix} = Z_{k,t} \lambda_t + W_{it} \xi_t.
\]

Equation (8.25) collects the time 't' observations into a single expression. To complete the model specification, we assume that \( \{a_i\} \) are identically and independently distribution with mean zero and \( D = \sigma^2 \text{Var } a_i \). Thus, \( \text{Var } a_i = \sigma^2 (I_n \otimes D) \). Here, we write the variance of \( a_i \) as a matrix times a constant \( \sigma^2 \) so that we may concentrate out the constant in the likelihood equations.

### 8.5.4 Initial conditions

We first re-write the measurement and observation equations so that the initial unobserved state vector, \( \delta_0 \), is zero. To this end, define
\[
\delta_t^* = \delta_t - \left( \prod_{r=1}^{t} T_r \right) \delta_0.
\]

With these new variables, we may express equations (8.25) and (8.20) as
\[
\begin{align*}
y_t &= X_t \beta + W_{it} \left( \prod_{r=1}^{t} T_r \right) \delta_t^* + Z_t \alpha + W_t \delta_t^* + \epsilon_t, \\
&= X_t \beta + Z_t \alpha + W_t \delta_t^* + \epsilon_t
\end{align*}
\] (8.26)

and
\[
\delta_t^* = T_t \delta_{t-1}^* + \eta_t
\] (8.27)

where \( \delta_0^* = 0 \).

With equation (8.26), we may consider the initial state variable \( \delta_0 \) to be either fixed, random or a combination of the two. With our assumptions of \( \xi_{t,0} = 0 \) and \( \lambda_0 \) as fixed, we may re-write equation (8.26) as:
\[
\begin{align*}
y_t &= \left( X_t \otimes W_{it} \left( \prod_{r=1}^{t} \Phi_{ir} \right) \right) \beta + Z_t \alpha + W_t \delta_t^*.
\end{align*}
\] (8.28)
Hence, the new vector of fixed parameters to be estimated is \((\beta', \lambda_0')'\) with corresponding \(n_t \times (K+r)\) matrix of covariates \(X_t : W_t \left( \prod_{r=1}^t \Phi_{tr} \right)\). Thus, with this reparameterization, we henceforth consider the state space model with the assumption that \(\delta_0 = 0\).

### 8.5.5 The Kalman filter algorithm

We define the transformed variables \(y^*, X^*\) and \(Z^*\), as follows. Recursively calculate:

\[
d_{t+1}(y) = T_{t+1}d_{t+1}(y) + K_t(y_t - W_t d_{t+1}(y)) \tag{8.29a}
\]
\[
d_{t+1}(X) = T_{t+1}d_{t+1}(X) + K_t(X_t - W_t d_{t+1}(X)), \tag{8.29b}
\]
\[
d_{t+1}(Z) = T_{t+1}d_{t+1}(Z) + K_t(Z_t - W_t d_{t+1}(Z)) \tag{8.29c}
\]

and

\[
P_{t+1} = T_{t+1} \left\{ P_{t+1} - P_{t+1} W_{t} F^{-1}_{t} W_{t} P_{t+1} \right\} T_{t+1}' + Q_{t+1} \tag{8.30a}
\]
\[
F_{t+1} = W_{t+1} T_{t+2} P_{t+1} W_{t+1}' F^{-1}_{t+1} \tag{8.30b}
\]
\[
K_{t+1} = T_{t+2} P_{t+1} W_{t+1}' F^{-1}_{t+1} \tag{8.30c}
\]

We begin the recursions in equations (8.29a-c) with \(d_{1|0}(y) = 0\), \(d_{1|0}(X) = 0\) and \(d_{1|0}(Z) = 0\). Also, for equation (8.30a), use \(P_{1|0} = Q_1\). The \(i\)th component of each transformed variable is

\[
y^*_i = y_i - W_i d_{t+1}(y) \tag{8.31a}
\]
\[
X^*_i = X_i - W_i d_{t+1}(X) \tag{8.31b}
\]
\[
Z^*_i = Z_i - W_i d_{t+1}(Z). \tag{8.31c}
\]

From equations (8.29)-(8.31), note that the calculation of the transformed variables are unaffected by scale changes in \(\{Q_t\}\). Thus, using the sequence \(\{Q_t^*\}\) defined in equation (8.21) in the Kalman filter algorithm yields the same transformed variables and rescaled conditional variances \(F_t^* = \sigma^{-2} F_t\).

### Likelihood equations

To calculate parameter estimators and the likelihood, we use the following sums of squares:

\[
S_{XX,F} = \sum_{t=1}^T X_t'^{-1} X_t', \quad S_{XZ,F} = \sum_{t=1}^T X_t'^{-1} Z_t', \quad S_{ZZ,F} = \sum_{t=1}^T Z_t'^{-1} Z_t', \quad S_{XY,F} = \sum_{t=1}^T X_t'^{-1} Y_t', \quad S_{ZY,F} = \sum_{t=1}^T Z_t'^{-1} Y_t'. \tag{8.32}
\]

With this notation, the generalized least square estimator of \(\beta\) is:

\[
b_{GLS} = \left( S_{XX,F} - S_{XZ,F} \left( S_{ZZ,F} + I_n \otimes D^{-1} \right)^{-1} S_{XZ,F} \right)^{-1} \left( S_{XY,F} - S_{ZY,F} \left( S_{ZZ,F} + I_n \otimes D^{-1} \right)^{-1} S_{XZ,F} \right)'. \tag{8.32}
\]

Let \(\tau\) denote the vector of the other variance components so that \((\sigma^2, \tau)\) represent all variance components. We may express the concentrated logarithmic likelihood as:

\[
L(\sigma^2, \tau) = -\frac{1}{2} \left\{ N \ln 2 \pi + N \ln \sigma^2 + \sigma^2 \text{Error SS} + \sum_{t=1}^T \ln \det F_t^* + n \ln \det D + \ln \det \left( S_{ZZ,F} + I_n \otimes D^{-1} \right) \right\}. \tag{8.33}
\]
where \( Error SS = (S_{yy,F} - S_{zy,F} (S_{zz,F} + I_u \otimes D^{-1})^{-1} S_{zy,F}) - (S_{yx,F}' - S_{zy,F}' (S_{zz,F} + I_u \otimes D^{-1})^{-1} S_{zy,F}) b_{GLS} \). (8.34)

The restricted logarithmic likelihood is:

\[
L_{REML}(\sigma^2, \tau) = -\frac{1}{2} \ln \det \left( S_{xx,F} - S_{xz,F} (S_{zz,F} + I_u \otimes D^{-1})^{-1} S_{zx,F} \right) - K \ln \sigma^2 + L(\sigma^2, \tau),
\]

up to an additive constant. Estimates of the variance components, \( \sigma^2 \) and \( \tau \), may be determined either by maximizing (8.33) or (8.35). This text uses (8.35), which yield the \( REML \) estimators. The restricted maximum likelihood estimator of \( \sigma^2 \) is:

\[
s_{REML}^2 = Error SS / (N-K).
\]

With equation (8.35), the concentrated restricted log likelihood is:

\[
L_{REML}(\tau) = -\frac{1}{2} \ln \det \left( S_{xx,F} - S_{xz,F} (S_{zz,F} + I_u \otimes D^{-1})^{-1} S_{zx,F} \right) - K \ln s_{REML}^2 + L(s_{REML}^2, \tau).\]

Maximizing \( L_{REML}(\tau) \) over \( \tau \) yields the \( REML \) estimator of \( \tau \).

### 8.6 Example: Capital asset pricing model

The capital asset pricing model (CAPM) is a representation that is widely used in financial economics. An intuitively appealing idea, and one of the basic characteristics of the CAPM model, is that there should be a relationship between the performance of a security and the performance of the market. One rationale is simply that if economic forces are such that the market improves, then those same forces should act upon an individual stock, suggesting that it also improve. We measure performance of a security through the return. To measure performance of the market, several market indices exist for each exchange. As an illustration, below we use the return from the “value weighted” index of the market created by the Center for Research in Securities Prices (CRSP). The value weighted index is defined by assuming a portfolio is created when investing an amount of money in proportion to the market value (at a certain date) of firms listed on the New York Stock Exchange, the American Stock Exchange and the Nasdaq Stock Market.

Another rationale for a relationship between security and market returns comes from financial economics theory. This is the CAPM theory, attributed to Sharpe (1964O) and Lintner (1965O) and based on the portfolio diversification ideas of Markowitz (1952O). Other things equal, investors would like to select a return with a high expected value and low standard deviation, the latter being a measure of riskiness. One of the desirable properties about using standard deviations as a measure of riskiness is that it is straightforward to calculate the standard deviation of a portfolio, a combination of securities. One only needs to know the standard deviation of each security and the correlations among securities. A notable security is a risk-free one, that is, a security that theoretically has a zero standard deviation. Investors often use a 30-day U.S. Treasury bill as an approximation of a risk-free security, arguing that the probability of default of the U.S. government within 30 days is negligible. Positing the existence of a risk-free asset and some other mild conditions, under the CAPM theory there exists an efficient frontier called the securities market line. This frontier specifies the minimum expected return that investors should demand for a specified level of risk. To estimate this line, we use the equation

\[
y_{it} = \beta_{0i} + \beta_{1i} x_{mt} + e_{it},
\]

where \( y \) is the security return in excess of the risk-free rate, \( x_{mt} \) is the market return in excess of the risk-free rate. We interpret \( \beta_{1i} \) as a measure of the amount of the \( i \)th security’s return that is attributed to the behavior of the market. According to the CAPM theory, the intercept \( \beta_{0i} \) is zero but we include it to study the robustness of the model.
To assess the empirical performance of the CAPM model, we study security returns from CRSP. We consider \( n = 90 \) firms from the insurance carriers that were listed on the CRSP files as at December 31, 1999. (The “insurance carriers” consists of those firms with standard industrial classification, SIC, codes ranging from 6310 through 6331, inclusive.) For each firm, we used sixty months of data ranging from January 1995 through December 1999.

Table 8.2 summarizes the performance of the market through the return from the value weighted index, VWRETD, and risk free instrument, RISKFREE. We also consider the difference between the two, VWFREE, and interpret this to be the return from the market in excess of the risk-free rate.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Median</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>VWRETD (Value weighted index)</td>
<td>2.091</td>
<td>2.946</td>
<td>-15.677</td>
<td>8.305</td>
<td>4.133</td>
</tr>
<tr>
<td>RISKFREE (Risk free)</td>
<td>0.408</td>
<td>0.415</td>
<td>0.296</td>
<td>0.483</td>
<td>0.035</td>
</tr>
<tr>
<td>VWFREE (Value weighted in excess of risk free)</td>
<td>1.684</td>
<td>2.517</td>
<td>-16.068</td>
<td>7.880</td>
<td>4.134</td>
</tr>
</tbody>
</table>

Source: Center for Research in Securities Prices

Table 8.3 summarizes the performance of individual securities through the monthly return, RET. These summary statistics are based on 5,400 monthly observations taken from 90 firms. The difference between the return and the corresponding risk free instrument is RETFREE.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Median</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>RET (Individual security return)</td>
<td>1.052</td>
<td>0.745</td>
<td>-66.197</td>
<td>102.500</td>
<td>10.038</td>
</tr>
<tr>
<td>RETFREE (Individual security return in excess of risk free)</td>
<td>0.645</td>
<td>0.340</td>
<td>-66.579</td>
<td>102.085</td>
<td>10.036</td>
</tr>
</tbody>
</table>

To examine the relationship between market and individual firm returns, a trellis plot is given in Figure 8.1. Here, only a subset of 18 randomly selected firms is presented; the subset allows one to see important patterns. Each panel in the figure represents a firm’s experience; thus, the market returns (on the horizontal axis) are common to all firms. In particular, note the influential point on the left-hand side of each panel, corresponding to a August, 1998 monthly return of –15.7%. So that this point would not dominate, a nonparametric line was fit for each panel. The lines superimposed show a positive relationship between the market and individual firm returns although the noise about each line is substantial.
Figure 8.1. Trellis Plot of Returns versus Market Return. A random sample of 18 firms are plotted, each panel represents a firm. Within each panel, firm returns versus market returns are plotted. A nonparametric line is superimposed to provide a visual impression of the relationship between the market return and individual firm’s return.

Several fixed effects models were fit using equation (8.38) as a framework. Table 8.4 summarizes the fit of each model. Based on these fits, we will use the variable slopes with an AR(1) error term model as the baseline for investigating time varying coefficients.

<table>
<thead>
<tr>
<th>Summary measure</th>
<th>Homogeneous model</th>
<th>Variable intercepts model</th>
<th>Variable slopes model</th>
<th>Variable intercepts and slopes model</th>
<th>Variable slopes model with AR(1) term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residual std deviation (s)</td>
<td>9.59</td>
<td>9.62</td>
<td>9.53</td>
<td>9.54</td>
<td>9.53</td>
</tr>
<tr>
<td>-2 ln Likelihood</td>
<td>39,751.2</td>
<td>39,488.6</td>
<td>39,646.5</td>
<td>39,350.6</td>
<td>39,610.9</td>
</tr>
<tr>
<td>AIC</td>
<td>39,753.2</td>
<td>39,490.6</td>
<td>39,648.5</td>
<td>39,352.6</td>
<td>39,614.9</td>
</tr>
<tr>
<td>AR(1) corr ($\rho$)</td>
<td>-0.084</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t-statistic for $\rho$</td>
<td>-5.98</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For time-varying coefficients, we investigate models of the form:

$$y_{it} = \beta_0 + \beta_{1,i} x_{m,t} + e_{it},$$

(8.39)

where

$$e_{it} = \rho e_{it-1} + \eta_{1,it},$$

(8.40)

and

$$\beta_{1,i,t} - \beta_{1,i} = \rho \beta (\beta_{1,i,t-1} - \beta_{1,i}) + \eta_{2,it}.$$  

(8.41)
Here, \( \{\eta_{1,lt}\} \) are i.i.d. noise terms. These are independent of \( \{\eta_{2,lt}\} \), that are mutually independent and identical for each firm \( i \). For equations (8.40) and (8.41), we assume that \( \{\varepsilon_{lt}\} \) and \( \{\beta_{1,lt}\} \) are stationary \( AR(1) \) processes. The slope coefficient, \( \beta_{1,lt} \), is allowed to vary by both firm \( i \) and time \( t \). We assume that each firm has its own stationary mean, \( \beta_{1,i} \) and variance \( \text{Var} \beta_{1,lt} \). It is possible to investigate the model in equation (8.39)-(8.41) for each firm \( i \). However, by considering all firms simultaneously, we allow for efficient estimation of common parameters \( \beta_{0}, \rho_{e}, \rho_{\beta}, \) and \( \sigma_{\varepsilon}^{2} = \text{Var} \varepsilon_{lt} \).

To express this model formulation in the notation of Section 8.3, first define \( j_{n,i} \) to be an \( n \times 1 \) vector, with a “one” in the \( i \)th row and zeroes elsewhere. Further define

\[
\begin{pmatrix}
\beta_{0,1}
\beta_{1,1}
\vdots
\beta_{1,n}
\end{pmatrix},
\]

\[
x_{it} = \left[ \begin{array}{c} j_{n,i} \cdot x_{mt} \end{array} \right],
\]

\[
z_{\lambda,lt} = j_{n,i} \cdot x_{mt} \text{ and } \lambda_{t} = \begin{pmatrix}
\beta_{1,lt} - \beta_{1,1}
\vdots
\beta_{1,m} - \beta_{1,n}
\end{pmatrix}.
\]

Thus, with this notation, we have

\[
y_{it} = \beta_{0} + \beta_{1,lt} \cdot x_{mt} + \varepsilon_{lt} = z_{\lambda,lt}' \lambda_{t} + x_{it}' \beta + \varepsilon_{lt}.
\]

This expresses the model as a special case of equation (8.8), ignoring the time-invariant random effects portion and using \( r = n \) time-varying coefficients.

An important component of model estimation routines is \( \text{Var} \lambda = \sigma^{2} \Sigma_{\lambda} \). Straightforward calculations show that this matrix may be expressed as \( \text{Var} \lambda = R_{AR} \otimes \Sigma_{\beta} \), where \( \Sigma_{\beta} = \sigma_{\beta}^{2} \text{I}_{n} \) and \( R_{AR} \) is defined in Section 8.2.2. Thus, this matrix is highly structured and easily invertible. However, it has dimension \( nT \times nT \) which is large. Special routines must take advantage of the structure to make the estimation computationally feasible. The estimation procedure in Appendix 8A.2 assumes that \( r \), the number of time-varying coefficients, is small. (See, for example, equation 8A.5.) Thus, we look to the Kalman filter algorithm for this application.

To apply the Kalman filter algorithm, we use the following conventions. For the updating matrix for time-varying coefficients in equation (8.18), we use \( \Phi_{1} = \text{I}_{n} \rho_{\beta} \). For the error structure in equation (8.19), we use an \( AR(1) \) structure so that \( p = 1 \) and \( \Phi_{2} = \rho_{e} \). Thus, we have

\[
\delta_{t} = \begin{pmatrix}
\lambda_{t}
\varepsilon_{lt}
\end{pmatrix}
\]

\[
T_{t} = \begin{pmatrix}
\text{I}_{n,1} \rho_{\beta}
0
\end{pmatrix}
\]

\[
\otimes \text{I}_{n},
\]

for the vector of time-varying parameters and updating matrix. As in Section 8.5.1, we assume that \( \{\lambda_{t}\} \) and \( \{\varepsilon_{lt}\} \) are independent stochastic processes and express the variance using

\[
Q_{t} = \begin{pmatrix}
\text{Var} \eta_{lt} & 0 \\
0 & \text{Var} \eta_{2lt}
\end{pmatrix} = \begin{pmatrix}
(1 - \rho_{\beta}^{2}) \sigma_{\beta}^{2} \text{I}_{n} & 0 \\
0 & (1 - \rho_{e}^{2}) \sigma_{e}^{2} \text{I}_{n}
\end{pmatrix}.
\]

To reduce the complexity, we assume that the initial vector is zero so that \( \delta_{0} = 0 \).

For the measurement equations, we have
Further, we have $W_{it} = M_t$ and thus, $W_t = (M_t \ x_{mt} : M_t)$. For parameter estimation, we have not specified any time-invariant random effects. Thus, we need only use parts (a) and (b) of equations (8.29) and (8.31), as well as all of equation (8.30). To calculate parameter estimators and the likelihood, we use the following sums of squares:

$S_{XX,F} = \sum_{t=1}^{T} X'_t F_t^{-1} X'_t$, $S_{XY,F} = \sum_{t=1}^{T} X'_t F_t^{-1} y'_t$ and $S_{YY,F} = \sum_{t=1}^{T} y'_t F_t^{-1} y'_t$. With this notation, the generalized least square estimator of $\beta$ is $b_{GLS} = S_{XX,F}^{-1} S_{XY,F}$. We may express the concentrated logarithmic likelihood as:

$L(\sigma^2, \tau) = -\frac{1}{2} \{ N \ln 2 \pi + N \ln \sigma^2 + \sigma^{-2} \text{Error SS} + \sum_{t=1}^{T} \ln \det F_t^* \}$

where $\text{Error SS} = S_{YY,F} - S_{XY,F} \cdot b_{GLS}$. The restricted logarithmic likelihood is

$L_{REML}(\sigma^2, \tau) = -\frac{1}{2} \{ \ln \det (S_{XX,F}) - K \ln \sigma^2 \} + L(\sigma^2, \tau)$

up to an additive constant.

For prediction, we may again use best linear unbiased predictors (BLUPs) introduced in Chapter 4 and extended in Section 8.3.3. Pleasant calculations show that the BLUP of $\beta_{i,t}$ is

$b_{i,t, BLUP} = b_{i,t, GLS} + \sigma^2 \left( \rho_{\beta} \var{X_m} \cdots \rho_{\beta} \var{X_m} \right) \left( \text{Var} \ y_i \right)^{-1} \left( y_i - b_{0,GLS} 1_i - h_{i,t,GLS} x_m \right), \quad (8.42)$

where $x_m = (x_{m,1} \cdots x_{m,T})$, $\text{Var} \ y_i = \sigma^2 \var{X_m} R_{AR}(\rho_{\beta}) X_m + \sigma^2 \var{R_{AR}(\rho_{e})}$ and

$X_m = \text{diag}(x_{m,1} \cdots x_{m,T})$.

Table 8.5 summarizes the fit of the time-varying CAPM model, based on equations (8.39)-(8.41) and the CRSP data. When fitting the model with both autoregressive processes in equations (8.40) and (8.41), it can be difficult to separate the dynamic sources, thus flattening out the likelihood surface. When the likelihood surface is flat, it is difficult to obtain convergence of the likelihood maximization routine. Figure 8.2 shows that the likelihood function is less responsive to changes in the $\rho_{\beta}$ parameter compared to the $\rho_{e}$ parameter.
Figure 8.2. Logarithmic likelihood as a function of the correlation parameters.
The left-hand panel shows the log likelihood as a function of $\rho_\varepsilon$, holding the other parameters fixed at their maximum likelihood values. The right-hand panel shows log likelihood as a function of $\rho_\beta$. The likelihood surface is flatter in the direction of $\rho_\beta$ than $\rho_\varepsilon$.

Table 8.5 Time-Varying CAPM Models

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\sigma$</th>
<th>$\rho_\varepsilon$</th>
<th>$\rho_\beta$</th>
<th>$\sigma_\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model fit with $\rho_\varepsilon$ parameter</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimate</td>
<td>9.527</td>
<td>-0.084</td>
<td>-0.186</td>
<td>0.864</td>
</tr>
<tr>
<td>Standard Error</td>
<td>0.141</td>
<td>0.019</td>
<td>0.140</td>
<td>0.069</td>
</tr>
<tr>
<td>Model fit without $\rho_\varepsilon$ parameter</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimate</td>
<td>9.527</td>
<td>-0.265</td>
<td>0.903</td>
<td></td>
</tr>
<tr>
<td>Standard Error</td>
<td>0.141</td>
<td>0.116</td>
<td>0.068</td>
<td></td>
</tr>
</tbody>
</table>

Because of the interest in the changes of the slope parameter, the model was then re-fit without the correlation parameter for the noise process, $\rho_\varepsilon$. The likelihood surface was much steeper for this reduced model and resulting standard errors are much sharper, as seen in Table 8.5. An alternative model would be to consider $\rho_\beta$ equal to zero yet retain $\rho_\varepsilon$. We leave this as an exercise for the reader.

With the fitted model parameter estimates in Table 8.5, beta prediction and forecasting is possible. For illustration purposes, we calculated the predictions of the slope for each time point using equation (8.42). Figure 8.2 summarizes these calculations for the Lincoln National Corporation. For reference, it turns out that the generalized least square estimator of $\beta_{1,\text{LINCOLN}}$ for this time period is $b_{1,\text{LINCOLN}} = 0.599$.

The upper panel of Figure 8.3 shows the time series of the time-varying efficient predictors of the slope. The lower panel of Figure 8.3 shows the time series of Lincoln returns over the same time period. Here, we see the influence of the firm’s returns on the efficient predictor of $\beta_{1,\text{LINCOLN}}$. For example, we see that the large drop in Lincoln’s return for September of 1999 leads to a corresponding drop in the predictor of the slope.
Figure 8.3. Time series plot of BLUP predictors of the slope associated with the market returns and returns for the Lincoln National Corporation. The upper panel shows the BLUP predictors of the slopes. The lower panel shows the monthly returns.
Appendix 8A. Inference for the Time-varying Coefficient Model

Appendix 8A.1 The Model

To allow for unbalanced data, recall the design matrix $M_i$ specified in equation (7.22). To allow for the observation set described in Section 7.4.1, we may use the matrix form of the linear mixed effects model in equation (8.9) with one exception. That exception is to expand the definition $Z_{\lambda,i}$, a $T_i \times T_r$ matrix of explanatory variables. With the notation in equation (7.22), we have

$$\begin{pmatrix}
Z'_{\lambda,i,1} & 0 & \cdots & 0 \\
0 & Z'_{\lambda,i,2} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & Z'_{\lambda,i,T_r}
\end{pmatrix}
(M_i \otimes I_r) = \text{block diag}(Z'_{\lambda,i,1}, \ldots, Z'_{\lambda,i,T_r})(M_i \otimes I_r). \quad (8A.1)$$

Thus, to complete the specification of equation (8.9), we write

$$Z'_{\lambda,i} = \begin{pmatrix}
Z'_{\lambda,i,1} & 0 & \cdots & 0 \\
0 & Z'_{\lambda,i,2} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & Z'_{\lambda,i,T_r}
\end{pmatrix}
(M_i \otimes I_r) = \text{block diag}(Z'_{\lambda,i,1}, \ldots, Z'_{\lambda,i,T_r})(M_i \otimes I_r). \quad (8A.2)$$

To express the model more compactly, we use the mixed linear model specification. Further, we also use the notation $\text{Var } \varepsilon = \sigma^2 = \text{blockdiag}(R_1, \ldots, R_n)$ and note that $\text{Var } \alpha = \sigma^2 (I_n \otimes D)$. With this notation, we may express the variance-covariance matrix of $y$ as $\text{Var } y = \sigma^2 V$ where

$$V = Z_{\alpha} (I_n \otimes D) Z_{\alpha}' + Z_{\lambda} \Sigma_{\lambda} Z_{\lambda}' + R. \quad (8A.3)$$

Appendix 8A.2 Estimation

For known variances, the usual generalized least squares (GLS) estimator of $\beta$ is $b_{\text{GLS}} = (X' V^{-1} X)^{-1} X' V^{-1} y$ (the scale parameter $\sigma^2$ drops out). To simplify calculations, we note that both $R$ and $Z_{\alpha} (I_n \otimes D)$ $Z_{\alpha}'$ are block diagonal matrices and thus have readily computable inverses. Thus, we define

$$V_{\alpha} = R + Z_{\alpha} (I_n \otimes D) Z_{\alpha}' = \text{block diag}(V_{\alpha,1}, \ldots, V_{\alpha,n})$$

where $V_{\alpha,i}$ is defined in Section 8.3.1. With this notation, we use equation (A.4) of Appendix A.5 to write

$$V^{-1} = (V_{\alpha} + Z_{\lambda} \Sigma_{\lambda} Z_{\lambda}')^{-1} = V_{\alpha}^{-1} - V_{\alpha}^{-1} Z_{\lambda} (Z_{\lambda}' V_{\alpha}^{-1} Z_{\lambda} + \Sigma_{\lambda}^{-1})^{-1} Z_{\lambda}' V_{\alpha}^{-1}. \quad (8A.4)$$

In equation (8A.4), only the block diagonal matrix $V_{\alpha}$ and the $rT \times rT$ matrix $Z_{\lambda}' V_{\alpha}^{-1} Z_{\lambda} + \Sigma_{\lambda}^{-1}$ require inversion, not the $N \times N$ matrix $V$. Define the following sums of squares: $S_{XX} = \sum_{i=1}^{n} X_i' V_{\alpha,j}^{-1} X_i$, $S_{XZ} = \sum_{i=1}^{n} X_i' V_{\alpha,j}^{-1} Z_{\lambda,i}$, $S_{ZZ} = \sum_{i=1}^{n} Z_{\lambda,i}' V_{\alpha,j}^{-1} Z_{\lambda,i}$, $S_{y} = \sum_{i=1}^{n} Z_{\lambda,i}' V_{\alpha,j}^{-1} y_i$ and $S_{XY} = \sum_{i=1}^{n} X_i' V_{\alpha,j}^{-1} y_i$. With this notation and equation (8A.4), we may express the GLS estimator of $\beta$ as
\[ b_{GLS} = (S_{xx} - S_{xz} (S_{zz} + \Sigma_{\lambda}^{-1})^{-1} S_{xz}')^{-1}(S_{xy} - S_{xz} (S_{zz} + \Sigma_{\lambda}^{-1})^{-1} S_{zy}). \] (8A.5)

**Likelihood equations**

We use the notation \( \tau \) to denote the remaining parameters so that \{\( \sigma^2, \tau \)\} represent the variance components. From standard normal theory (see Appendix B), the logarithmic likelihood is

\[ L(\beta, \sigma^2, \tau) = -\frac{1}{2} \{N \ln 2 \pi + N \ln \sigma^2 + \sigma^{-2} (y'X\beta) V^{-1} (y'X\beta) + \ln \det V \} . \] (8A.6)

The corresponding restricted log likelihood is

\[ L_R(\beta, \sigma^2, \tau) = -\frac{1}{2} \{\ln \det (X'V^{-1}X) - K \ln \sigma^2 \} + L(\beta, \sigma^2, \tau) + \text{constant}. \] (8A.7)

Either (8A.6) or (8A.7) can be maximized to determine an estimator of \( \beta \). The result is also the generalized least squares estimator \( b_{GLS} \), given in equation (8A.5). Using \( b_{GLS} \) for \( \beta \) in equations (8A.6) and (8A.7) yields concentrated likelihoods. To determine the REML estimator of \( \sigma^2 \), we maximize \( L_R(b_{GLS}, \sigma^2, \tau) \) (holding \( \tau \) fixed), to get

\[ \sigma^2_{REML} = (N-K)^{-1} (y - X b_{GLS})' V^{-1} (y - X b_{GLS}). \] (8A.8)

Thus, the log-likelihood evaluated at these parameters is

\[ L(b_{GLS}, \sigma^2_{REML}, \tau) = -\frac{1}{2} \{N \ln 2 \pi + N \ln \sigma^2_{REML} + N-K + \ln \det V \} . \] (8A.9)

The corresponding restricted log-likelihood is

\[ L_R = -\frac{1}{2} \{\ln \det (X'V^{-1}X) - K \ln \sigma^2_{REML} \} + L(b_{GLS}, \sigma^2_{REML}, \tau) + \text{constant}. \] (8A.10)

The likelihood expressions in equations (8A.9) and (8A.10) are intuitively straightforward. However, because of the number of dimensions, they can be difficult to compute. We now provide alternative expressions that, although more complex in appearance, are simpler to compute. Using equation (8A.4), we may express

Error SS = (N-K) \( \sigma^2_{REML} \) - y V^{-1} y - y' V^{-1} X b_{GLS}

= S_{yy} - S_{zy}' (S_{zzz} + \Sigma_{\lambda}^{-1})^{-1} S_{zy} - S_{xy}' b_{GLS} + S_{zy}' (S_{zz} + \Sigma_{\lambda}^{-1})^{-1} S_{xz} b_{GLS}. \] (8A.11)

From equation (A.5) of Appendix A.5, we have

\[ \ln \det V = \ln \det V_a + \ln \det \Sigma_{\lambda} + \ln \det (Z_{ij}'V_{a^{-1}}Z_{ij} + \Sigma_{\lambda^{-1}}). \] (8A.12)

Thus, the logarithmic likelihood evaluated at these parameters is

\[ L(b_{GLS}, \sigma^2_{REML}, \tau) = -\frac{1}{2} \{N \ln 2 \pi + N \ln \sigma^2_{REML} + N-K + S_V + \ln \det \Sigma_{\lambda} + \ln \det(S_{zz} + \Sigma_{\lambda}^{-1}) \}, \]

where \( S_V = \sum_{i=1}^{n} \ln \det V_{a_{ij}} \). The corresponding restricted log-likelihood is

\[ L_R(\tau) = -\frac{1}{2} \{ \ln \det (S_{xx} - S_{xz} (S_{zz} + \Sigma_{\lambda}^{-1})^{-1} S_{xz}') - K \ln \sigma^2_{REML} \}

+ L(b_{GLS}, \sigma^2_{REML}, \tau) + \text{constant}. \] (8A.14)

Maximizing \( L_R(\tau) \) over \( \tau \) yields the REML estimator of \( \tau \), say, \( \tau_{REML} \).
Appendix 8A.3 Prediction

To derive the BLUP predictor of \( \lambda \), let \( c_\lambda \) be an arbitrary vector of constants and set \( w = c_\lambda' \lambda \). With this choice, we have \( \mathbb{E}(w) = 0 \). Using equation (4.7), we have

\[
\mathbf{c}_\lambda' \lambda_{\text{BLUP}} = \sigma^{-2} \mathbf{c}_\lambda' \mathbf{Cov}(\lambda, \mathbf{y}) \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X} \mathbf{b}_{\text{GLS}}) = \sigma^{-2} \mathbf{c}_\lambda' \mathbf{Cov}(\lambda, \mathbf{Z}_\lambda \lambda) \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X} \mathbf{b}_{\text{GLS}}).
\]

With Wald’s device, this yields

\[
\lambda_{\text{BLUP}} = \mathbf{Z}_\lambda' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X} \mathbf{b}_{\text{GLS}}) .
\]

Further, using equation (8A.4), we have

\[
\mathbf{Z}_\lambda' \mathbf{V}^{-1} = \mathbf{Z}_\lambda' \mathbf{V}^{-1} (\mathbf{Z}_\lambda' \mathbf{V}^{-1})^{-1} \mathbf{Z}_\lambda' \mathbf{V}^{-1} = \sum_{i=1}^{n} \mathbf{Z}_{i,j} \mathbf{V}^{-1} (\mathbf{y}_i - \mathbf{X}_i \mathbf{b}_{\text{GLS}}).
\]

Thus,

\[
\lambda_{\text{BLUP}} = (\mathbf{S}_{ZZ} + \mathbf{S}_{\lambda\lambda}^{-1})^{-1} \mathbf{Z}_\lambda' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X} \mathbf{b}_{\text{GLS}}).
\]

To simplify this expression, we recall the vector of residuals \( \epsilon_{i,\text{GLS}} = \mathbf{y}_i - \mathbf{X}_i \mathbf{b}_{\text{GLS}} \). This yields

\[
\lambda_{\text{BLUP}} = (\mathbf{S}_{ZZ} + \mathbf{S}_{\lambda\lambda}^{-1})^{-1} \sum_{i=1}^{n} \mathbf{Z}_{i,j} \mathbf{V}^{-1} (\mathbf{y}_i - \mathbf{X}_i \mathbf{b}_{\text{GLS}}) = (\mathbf{S}_{ZZ} + \mathbf{S}_{\lambda\lambda}^{-1})^{-1} \sum_{i=1}^{n} \mathbf{Z}_{i,j} \mathbf{V}^{-1} (\mathbf{y}_i - \mathbf{X}_i \mathbf{b}_{\text{GLS}}).
\]

as in equation (8.12).

We now consider predicting a linear combination of residuals, \( w = c_\epsilon' \epsilon \), where \( c_\epsilon \) is a vector of constants. With this choice, we have \( \mathbb{E}(w) = 0 \). Further, we have

\[
\mathbf{c}_\epsilon' \epsilon_{i,\text{BLUP}} = \sigma^{-2} \mathbf{c}_\epsilon' \mathbf{Cov}(\epsilon_i, \mathbf{y}) \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X} \mathbf{b}_{\text{GLS}})
\]

\[
= \sigma^{-2} \left( \mathbf{c}_\epsilon' \mathbf{Cov}(\epsilon_i, \mathbf{y}) \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X} \mathbf{b}_{\text{GLS}}) - \mathbf{c}_\epsilon' \mathbf{Cov}(\epsilon_i, \mathbf{y}) \mathbf{V}^{-1} \mathbf{Z}_\lambda (\mathbf{S}_{ZZ} + \mathbf{S}_{\lambda\lambda}^{-1})^{-1} (\mathbf{S}_{Zy} - \mathbf{S}_{ZX} \mathbf{b}_{\text{GLS}}) \right)
\]

\[
= \mathbf{c}_\epsilon' \mathbf{R}_i^{-1} \mathbf{V}^{-1} (\mathbf{y}_i - \mathbf{X}_i \mathbf{b}_{\text{GLS}}) - \mathbf{c}_\epsilon' \mathbf{R}_i^{-1} \mathbf{Z}_\lambda (\mathbf{S}_{ZZ} + \mathbf{S}_{\lambda\lambda}^{-1})^{-1} (\mathbf{S}_{Zy} - \mathbf{S}_{ZX} \mathbf{b}_{\text{GLS}})
\]

This yields equation (8.13).

Similarly, we derive the BLUP predictor of \( \mathbf{a} \). Let \( c_\alpha \) be an arbitrary vector of constants and set \( w = c_\alpha' \alpha \). For this choice of \( w \), we have \( \mathbb{E}(w) = 0 \). Further, we have

\[
\mathbf{c}_\alpha' \alpha_{i,\text{BLUP}} = \sigma^{-2} \mathbf{c}_\alpha' \mathbf{Cov}(\alpha_i, \mathbf{y}) \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X} \mathbf{b}_{\text{GLS}})
\]

\[
= \sigma^{-2} \left( \mathbf{c}_\alpha' \mathbf{Cov}(\alpha_i, \mathbf{y}) \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X} \mathbf{b}_{\text{GLS}}) - \mathbf{c}_\alpha' \mathbf{Cov}(\alpha_i, \mathbf{y}) \mathbf{V}^{-1} \mathbf{Z}_\lambda (\mathbf{S}_{ZZ} + \mathbf{S}_{\lambda\lambda}^{-1})^{-1} (\mathbf{S}_{Zy} - \mathbf{S}_{ZX} \mathbf{b}_{\text{GLS}}) \right)
\]

This yields equation (8.13).
Using Wald’s device, we have the BLUP of \( \alpha_i \), given in equation (8.14).

**Forecasting**

First note, from the calculation of BLUPs in equation (4.7), that the BLUP projection is linear. That is, consider estimating the sum of two random variables, \( w_1 + w_2 \). Then, it is immediate that

\[
BLUP(w_1 + w_2) = BLUP(w_1) + BLUP(w_2).
\]

With this and equation (8.9), we have

\[
\hat{y}_{i,T_t+L} = BLUP(y_{i,T_t+L}) = BLUP(z'_{a,i,T_t+L}a_i) + BLUP(z'_{k,i,T_t+L}k_{T_t+L}) + BLUP(x'_{i,T_t+L}\beta) + BLUP(\epsilon_{i,T_t+L})
\]

\[
= z'_{a,i,T_t+L}a_{i,\text{BLUP}} + z'_{k,i,T_t+L}BLUP(k_{T_t+L}) + x'_{i,T_t+L}b_{\text{GLS}} + BLUP(\epsilon_{i,T_t+L}).
\]

From equation (4.7) and the expression of \( \lambda_{BLUP} \), we have

\[
BLUP(k_{T_t+L}) = \sigma^{-2} \text{Cov}(k_{T_t+L}, y)V^{-1}(y - Xb_{\text{GLS}}) = \sigma^{-2} \text{Cov}(k_{T_t+L}, \lambda)Z'_{\lambda}V^{-1}(y - Xb_{\text{GLS}})
\]

\[
= \sigma^{-2} \text{Cov}(k_{T_t+L}, \lambda)\Sigma_{\lambda}^{-1}k_{BLUP}.
\]

From equation (4.7) and the calculation of \( \epsilon_{i,\text{BLUP}} \), we have

\[
BLUP(\epsilon_{i,T_t+L}) = \sigma^{-2} \text{Cov}(\epsilon_{i,T_t+L}, y)V^{-1}(y - Xb_{\text{GLS}})
\]

\[
= \sigma^{-2} \text{Cov}(\epsilon_{i,T_t+L}, \epsilon_i)V^{-1}(y - Xb_{\text{GLS}}) - Z_{\lambda,i}(S_{ZZ} + \Sigma_{\lambda}^{-1})^{-1}(S_{ZY} - S_{ZX}b_{\text{GLS}})
\]

\[
= \sigma^{-2} \text{Cov}(\epsilon_{i,T_t+L}, \epsilon_i)V^{-1}(y - Xb_{\text{GLS}}) - Z_{\lambda,i}\lambda_{BLUP}) = \sigma^{-2} \text{Cov}(\epsilon_{i,T_t+L}, \epsilon_i)R_{\epsilon}^{-1}\epsilon_{i,\text{BLUP}}.
\]

Thus, the BLUP forecast of \( y_{i,T_t+L} \) is

\[
\hat{y}_{i,T_t+L} = x'_{i,T_t+L}b_{\text{GLS}} + \sigma^{-2}z'_{k,i,T_t+L}\text{Cov}(k_{T_t+L}, \lambda)\Sigma_{\lambda}^{-1}k_{BLUP} + z'_{a,i,T_t+L}a_{i,\text{BLUP}}
\]

\[
+ \sigma^{-2} \text{Cov}(\epsilon_{i,T_t+L}, \epsilon_i)R_{\epsilon}^{-1}\epsilon_{i,\text{BLUP}},
\]

as in equation (8.16).

For forecasting, we wish to predict \( w = y_{i,T_t+L} \), given in equation (8.11). It is easy to see that

\[
\text{Var}(y_{i,T_t+L}) = \sigma^2 (z'_{a,i,T_t+L} Dz_{a,i,T_t+L} + z'_{k,i,T_t+L} \text{Cov}(k_{T_t+L})z_{k,i,T_t+L} + \text{Var}(\epsilon_{i,T_t+L}). \quad (8A.16)
\]

To calculate the variance of the forecast error, we use equation (4.9). First, note that

\[
X'V^{-1}X = S_{XX} - S_{XZ}(S_{ZZ} + \Sigma_{\lambda}^{-1})^{-1}S_{ZX}.' \quad (8A.17)
\]

Next, we have

\[
\text{Cov}(y_{i,T_t+L}, y) = \sum_{j=1}^{n} \text{Cov}(y_{i,T_t+L}, y_j)X^{-1}\text{Cov}(y_{i,T_t+L}, y_j)X^{-1}\text{Cov}(y_{i,T_t+L}, y_j)X^{-1}\]

\[
= \sum_{j=1}^{n} \text{Cov}(y_{i,T_t+L}, y_j)X^{-1}(S_{ZZ} + \Sigma_{\lambda}^{-1})^{-1}S_{ZX}. \quad (8A.18)
\]
Similarly, we have

\[
\operatorname{Cov}(y_{i,T+L}, y)\mathbf{V}^{-1} \operatorname{Cov}(y_{i,T+L}, y) = \sum_{j=1}^{n} \operatorname{Cov}(y_{i,T+L}, y_j)\mathbf{V}^{-1}_a \operatorname{Cov}(y_{i,T+L}, y_j)
- \left( \sum_{j=1}^{n} \operatorname{Cov}(y_{i,T+L}, y_j)\mathbf{V}^{-1}_a \mathbf{Z}_{k,j} \right) \left( \mathbf{S}_{\mathbf{Z}} + \Sigma_\lambda^{-1} \right)^{-1} \left( \sum_{j=1}^{n} \operatorname{Cov}(y_{i,T+L}, y_j)\mathbf{V}^{-1}_a \mathbf{Z}_{k,j} \right)^\prime.
\]  

(8A.19)

Thus, using equation (4.8), the variance of the forecast error is

\[
\operatorname{Var}(\hat{y}_{i,T+L} - y_{i,T+L}) = \left( \mathbf{x}_{i,T+L}' - \operatorname{Cov}(y_{i,T+L}, y)\mathbf{V}^{-1} \mathbf{x} \right) \left( \mathbf{x}' \mathbf{V}^{-1} \mathbf{x} \right)^{-1} \left( \mathbf{x}_{i,T+L}' - \operatorname{Cov}(y_{i,T+L}, y)\mathbf{V}^{-1} \mathbf{x} \right)^\prime - \operatorname{Cov}(y_{i,T+L}, y)\mathbf{V}^{-1} \operatorname{Cov}(y_{i,T+L}, y)' + \operatorname{Var}(y_{i,T+L}).
\]  

(8A.20)

where \(\operatorname{Cov}(y_{i,T+L}, y)\mathbf{V}^{-1} \mathbf{x}\) is specified in equation (8A.18), \(\mathbf{x}' \mathbf{V}^{-1} \mathbf{x}\) is specified in equation (8A.17), \(\operatorname{Var}(y_{i,T+L})\) is specified in equation (8A.15) and \(\operatorname{Cov}(y_{i,T+L}, y)\mathbf{V}^{-1} \operatorname{Cov}(y_{i,T+L}, y)\)' is specified in equation (8A.19).
Chapter 9. Binary Dependent Variables

Abstract. This chapter considers situations where the response of interest, $y$, takes on values 0 or 1, a binary dependent variable. To illustrate, one could use $y$ to indicate whether or not a subject possesses an attribute or to indicate a choice made; for example, whether or not a taxpayer employs a professional tax preparer to file income tax returns.

Regression models that describe the behavior of binary dependent variables are more complex than linear regression models. Thus, Section 9.1 reviews basic modeling and inferential techniques without the heterogeneity components. Recall that we refer to models without heterogeneity components as homogeneous models. Sections 9.2 and 9.3 include heterogeneity components by describing random and fixed effects models. Section 9.4 introduces a broader class of models known as marginal models that can be estimated using a moment-based procedure known as generalized estimating equations.

9.1 Homogeneous models

To introduce some of the complexities encountered with binary dependent variables, denote the probability that the response equals 1 by $p_{it} = \text{Prob}(y_{it} = 1)$. Then, we may interpret the mean response to be the probability that the response equals 1, that is, $E y_{it} = 0 \times \text{Prob}(y_{it} = 0) + 1 \times \text{Prob}(y_{it} = 1) = p_{it}$. Further, straightforward calculations show that the variance is related to the mean through the expression $\text{Var} y_{it} = p_{it} (1 - p_{it})$.

Linear probability models

Without heterogeneity terms, we begin by considering a linear model of the form

$$y_{it} = x_{it}' \beta + \epsilon_{it}, \quad (9.1)$$

known as a linear probability model. Assuming $E \epsilon_{it} = 0$, we have that $E y_{it} = p_{it} = x_{it}' \beta$ and $\text{Var} y_{it} = x_{it}' \beta (1 - x_{it}' \beta)$. Linear probability models are widely applied because of the ease of parameter interpretations. For large data sets, the computational simplicity of ordinary least squares estimators is attractive when compared to some complex alternative nonlinear models introduced below. Further, ordinary least squares estimators for $\beta$ have desirable properties. It is straightforward to check that they are consistent and asymptotically normal under mild conditions on the explanatory variables $\{x_{it}\}$.

However, linear probability models have several drawbacks that are serious for many applications. These drawbacks include:

- The expected response is a probability and thus must vary between 0 and 1. However, the linear combination, $x_{it}' \beta$, can vary between negative and positive infinity. This mismatch implies, for example, that fitted values may be unreasonable.
- Linear models assume homoscedasticity (constant variance) yet the variance of the response depends on the mean that varies over observations. The problem of varying variability is known as heteroscedasticity.
- The response must be either a 0 or 1 although the regression models typically regards distribution of the error term as continuous. This mismatch implies, for example, that the usual residual analysis in regression modeling is meaningless.
To handle the heteroscedasticity problem, a (two-stage) weighted least squares procedure is possible. That is, in the first stage, one uses ordinary least squares to compute estimates of $\beta$. With this estimate, an estimated variance for each subject can be computed using the relation $\text{Var}(y_{it}) = x_{it}' \beta (1 - x_{it}' \beta)$. At the second stage, a weighted least squares is performed using the inverse of the estimated variances as weights to arrive at new estimates of $\beta$. It is possible to iterate this procedure, although studies have shown that there are few advantages in doing so (see Carroll and Rupert, 1988). Alternatively, one can use ordinary least square estimators of $\beta$ with standard errors that are robust to heteroscedasticity.

9.1.1 Logistic and probit regression models

Using nonlinear functions of explanatory variables

To circumvent the drawbacks of linear probability models, we consider alternative models in which we express the expectation of the response as a function of explanatory variables, $p_{it} = \pi(x_{it}' \beta) = \text{Prob}(y_{it} = 1 | x_{it})$. We focus on two special cases of the function $\pi(.)$:

- $\pi(z) = \frac{1}{1 + e^{-z}}$, the logit case, and
- $\pi(z) = \Phi(z)$, the probit case.

Here, $\Phi(.)$ is the standard normal distribution function. Note that the choice of the identity function (a special kind of linear function), $\pi(z) = z$, yields the linear probability model. Thus, we focus on nonlinear choices of $\pi$. The inverse of the function, $\pi^{-1}$, specifies the form of the probability that is linear in the explanatory variables, that is, $\pi^{-1}(p_{it}) = x_{it}' \beta$. In Chapter 10, we will refer to this inverse as the link function.

These two functions are similar in that they are almost linearly related over the interval $0.1 \leq \pi \leq 0.9$, (see McCullagh and Nelder, 1989, page 109). This similarity means that it will be difficult to distinguish between the two specifications with most data sets. Thus, to a large extent, the function choice is dependent on the preferences of the analyst.

Threshold interpretation

Both the logit and probit cases can be justified by appealing to the following “threshold” interpretation of the model. To this end, suppose that there exists an underlying linear model, $y_{it}^* = x_{it}' \beta + \epsilon_{it}^*$. Here, we do not observe the response $y_{it}^*$ yet interpret it to be the “propensity” to possess a characteristic. For example, we might think about the speed of a horse as a measure of its propensity to win a race. Under the threshold interpretation, we do not observe the propensity but we do observe when the propensity crosses a threshold. It is customary to assume that this threshold is 0, for simplicity. Thus, we observe $y_{it} = \begin{cases} 0 & y_{it}^* \leq 0 \\ 1 & y_{it}^* > 0 \end{cases}$.

To see how the logit case can be derived from the threshold model, we assume a logit distribution function for the disturbances, so that $\text{Prob}(\epsilon_{it}^* \leq a) = \frac{1}{1 + \exp(-a)}$. Because the logit distribution is symmetric about zero, we have that $\text{Prob}(\epsilon_{it}^* \leq a) = \text{Prob}(-\epsilon_{it}^* \leq a)$. Thus,

$$p_{it} = \text{Prob}(y_{it} = 1) = \text{Prob}(y_{it}^* > 0) = \text{Prob}(\epsilon_{it}^* \leq x_{it}' \beta) = \frac{1}{1 + \exp(-x_{it}' \beta)} = \pi(x_{it}' \beta).$$

This establishes the threshold interpretation for the logit case. The development for the probit case is similar, and is omitted.
Random utility interpretation

Both the logit and probit cases can also be justified by appealing to the following “random utility” interpretation of the model. In economic applications, we think of an individual as selecting between two choices. To illustrate, in Section 9.1.3 we will consider whether or not a taxpayer chooses to employ a professional tax preparer to assist in filing an income tax return. Here, preferences among choices are indexed by an unobserved utility function; individuals select the choice that provides the greater utility.

For the \( it \)th individual at the \( t \)th time period, we use the notation \( u_{it} \) for this utility. We model utility as a function of an underlying value plus random noise, that is, \( U_{it} = u_{it}(V_{it} + \varepsilon_{it}) \), where \( j \) may be 0 or 1, corresponding to the choice. To illustrate, we assume that the individual chooses the category corresponding to \( j = 1 \) if \( U_{it1} > U_{it0} \) and denote this choice as \( y_{it} = 1 \). Assuming that \( u_{it} \) is a strictly increasing function, we have

\[
\Pr(y_{it} = 1) = \Pr(U_{it0} < U_{it1}) = \Pr(u_{it}(V_{it0} + \varepsilon_{it0}) < u_{it}(V_{it1} + \varepsilon_{it1})) = \Pr(\varepsilon_{it0} - \varepsilon_{it1} < V_{it1} - V_{it0}).
\]

To parameterize the problem, assume that the value function is an unknown linear combination of explanatory variables. Specifically, we take \( V_{it0} = 0 \) and \( V_{it1} = x_{it}'\beta \). We may take the difference in the errors, \( \varepsilon_{it0} - \varepsilon_{it1} \), to be normal or logistic, corresponding to the probit and logit cases, respectively. In Section 11.1, we will show that the logistic distribution is satisfied if the errors are assumed to have an extreme-value, or Gumbel, distribution. In Section 9.1.3, linear combinations of taxpayer characteristics will allow us to model the choice of using a professional tax preparer. The analysis allows for taxpayer preferences to vary by subject and over time.

Example 9.1 Job security

Velletta (1999E) studied declining job security using the PSID (Panel Survey of Income Dynamics) database (see Appendix F). We consider here one of the regressions presented by Valetta, based on a sample of male household heads that consists of \( N = 24,168 \) observations over the years 1976-1992, inclusive. The PSID survey records reasons why men left their most recent employment, including plant closures, “quit” and changed jobs for other reasons. However, Valetta focused on dismissals (“laid off” or “fired”) because involuntary separations are associated with job insecurity. Chapter 11 will expand this discussion to consider the other sources of job turnover.

Table 9.1 presents a probit regression model run by Valetta (1999E), using dismissals as the dependent variable. In addition to the explanatory variables listed in Table 9.1, other variables controlled for consisted of education, marital status, number of children, race, years of full-time work experience and its square, union membership, government employment, logarithmic wage, the U.S. employment rate and location as measured through the Metropolitan Statistical Area residence. In Table 9.1, tenure is years employed at the current firm. Further, sector employment was measured by examining CPS (Consumer Price Survey) employment in 387 sectors of the economy, based on 43 industry categories and nine regions of the country.

On the one hand, the tenure coefficient reveals that more experienced workers are less likely to be dismissed. On the other hand, the coefficient associated with the interaction between tenure and time trend reveals an increasing dismissal rate for experienced workers.

The interpretation of the sector employment coefficients is also of interest. With an average tenure of about 7.8 years in the sample, we see the low tenure men are relatively unaffected by changes in sector employment. However, for more experienced men, there is an increasing probability of dismissal associated with sectors of the economy where growth declines.

Valetta also fit a random effects model that will be described in Section 9.2; the results were qualitatively similar to those presented here.
Table 9.1 Dismissal Probit Regression Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Parameter estimate</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tenure</td>
<td>-0.084</td>
<td>0.010</td>
</tr>
<tr>
<td>Time Trend</td>
<td>-0.002</td>
<td>0.005</td>
</tr>
<tr>
<td>Tenure*(Time Trend)</td>
<td>0.003</td>
<td>0.001</td>
</tr>
<tr>
<td>Change in Logarithmic Sector Employment</td>
<td>0.094</td>
<td>0.057</td>
</tr>
<tr>
<td>Tenure*(Change in Logarithmic Sector Employment)</td>
<td>-0.020</td>
<td>0.009</td>
</tr>
</tbody>
</table>

-2 Log Likelihood                     | 7,027.8            |

Pseudo-$R^2$                           | 0.097              |

Logistic regression

An advantage of the logit case is that it permits closed-form expressions, unlike the normal distribution function. Logistic regression is another phrase used to describe the logit case. Using $p = \pi(z)$, the inverse of $\pi$ can be calculated as $z = \pi'(p) = \ln(p/(1-p))$. To simplify future presentations, we define $\text{logit}(p) = \ln(p/(1-p))$ to be the logit function. With logistic regression model, we represent the linear combination of explanatory variables as the logit of the success probability, that is, $x_i' \beta = \text{logit}(p_i)$. 

Odds ratio interpretation

When the response $y$ is binary, knowing only the probability of $y = 1$, $p$, summarizes the distribution. In some applications, a simple transformation of $p$ has an important interpretation. The lead example of this is the odds ratio, given by $p/(1-p)$. For example, suppose that $y$ indicates whether or not a horse wins a race, that is, $y = 1$ if the horse wins and $y = 0$ if the horse does not. Interpret $p$ to be the probability of the horse winning the race and, as an example, suppose that $p = 0.25$. Then, the odds of the horse winning the race is $0.25/(1.00-0.25) = 0.3333$. We might say that the odds of winning are $0.3333$ to $1$, or one to three. Equivalently, we can say that the probability of not winning is $1 - p = 0.75$. Thus, the odds of the horse not winning is $0.75/(1 - 0.75) = 3$. We interpret this to mean the odds against the horse are three to one.

Odds have a useful interpretation from a betting standpoint. Suppose that we are playing a fair game and that we place a bet of $1$ with odds of one to three. If the horse wins, then we get our $1$ back plus winnings of $3$. If the horse loses, then we lose our bet of $1$. It is a fair game in the sense that the expected value of the game is zero because we win $3$ with probability $p = 0.25$ and lose $1$ with probability $1 - p = 0.75$. From an economic standpoint, the odds provide the important numbers (bet of $1$ and winnings of $3$), not the probabilities. Of course, if we know $p$, then we can always calculate the odds. Similarly, if we know the odds, we can always calculate the probability $p$.

The logit is the logarithmic odds function, also known as the log odds.

Logistic regression parameter interpretation

To interpret the regression coefficients in the logistic regression model, $\beta = (\beta_1, \beta_2, \ldots, \beta_k)'$, we begin by assuming that $j$th explanatory variable, $x_{ij}$, is either 0 or 1. Then, with the notation $x_i = (x_{i1} \cdots x_{ij} \cdots x_{ik})'$, we may interpret

$$
\beta_j = (x_{i1} \cdots 1 \cdots x_{ik})' \beta - (x_{i1} \cdots 0 \cdots x_{ik})' \beta
$$

$$
= \ln \left( \frac{\text{Prob}(y_{it} = 1 | x_{ij} = 1)}{1 - \text{Prob}(y_{it} = 1 | x_{ij} = 1)} \right) - \ln \left( \frac{\text{Prob}(y_{it} = 1 | x_{ij} = 0)}{1 - \text{Prob}(y_{it} = 1 | x_{ij} = 0)} \right).
$$
Thus,

\[ e^{\beta_j} = \frac{\text{Prob}(y_{it} = 1|x_{ij} = 1)/(1 - \text{Prob}(y_{it} = 1|x_{ij} = 1))}{\text{Prob}(y_{it} = 1|x_{ij} = 0)/(1 - \text{Prob}(y_{it} = 1|x_{ij} = 0))}. \]

We note that the numerator of this expression is the odds when \( x_{ij} = 1 \), whereas the denominator is the odds when \( x_{ij} = 0 \). Thus, we can say that the odds when \( x_{ij} = 1 \) are \( \exp(\beta_j) \) times as large as the odds when \( x_{ij} = 0 \). To illustrate, if \( \beta_j = 0.693 \), then \( \exp(\beta_j) = 2 \). From this, we say that the odds (for \( y = 1 \)) are twice as great for \( x_j = 1 \) as \( x_j = 0 \).

Similarly, assuming that \( j \)th explanatory variable is continuous (differentiable), we have

\[ \beta_j = \frac{\partial}{\partial x_{ij}} \ln \left( \frac{\text{Prob}(y_{it} = 1|x_{ij})}{1 - \text{Prob}(y_{it} = 1|x_{ij})} \right) = \frac{\partial}{\partial x_{ij}} \left( \frac{\text{Prob}(y_{it} = 1|x_{ij})}{(1 - \text{Prob}(y_{it} = 1|x_{ij}))} \right). \]

Thus, we may interpret \( \beta_j \) as the proportional change in the odds ratio, known as an \textit{elasticity} in economics.

### 9.1.2 Inference for logistic and probit regression models

#### Parameter estimation

The customary method of estimation for homogenous models is maximum likelihood, described in further detail in Appendix C. To provide intuition, we outline the ideas in the context of binary dependent variable regression models.

The \textit{likelihood} is the observed value of the density or mass function. For a single observation, the likelihood is

\[ \begin{cases} 1 - p_{it} & \text{if } y_{it} = 0, \\ p_{it} & \text{if } y_{it} = 1. \end{cases} \]

The objective of maximum likelihood estimation is to find the parameter values that produce the largest likelihood. Finding the maximum of the logarithmic function often yields the same solution as finding the maximum of the corresponding function. Because it is generally computationally simpler, we consider the logarithmic (log-) likelihood, written as

\[ \ln(1 - p_{it}) \text{ if } y_{it} = 0 \]
\[ \ln p_{it} \text{ if } y_{it} = 1. \]

More compactly, the log-likelihood of a single observation is

\[ y_{it} \ln \pi(x_{it}', \beta) + (1 - y_{it}) \ln(1 - \pi(x_{it}', \beta)), \]

where \( p_{it} = \pi(x_{it}', \beta) \). Assuming independence among observations, the likelihood of the data set is a product of likelihoods of each observation. Thus, taking logarithms, the log-likelihood of the data set is the sum of log-likelihoods of single observations. The log-likelihood of the data set is

\[ L(\beta) = \sum_{it} \left\{ y_{it} \ln \pi(x_{it}', \beta) + (1 - y_{it}) \ln(1 - \pi(x_{it}', \beta)) \right\}, \tag{9.1} \]

where the sum ranges over \( \{i = 1, \ldots, T_i, t = 1, \ldots, n\} \). The (log) likelihood is viewed as a function of the parameters, with the data held fixed. In contrast, the joint probability mass (density) function is viewed as a function of the realized data, with the parameters held fixed.

The method of maximum likelihood means finding the values of \( \beta \) that maximize the log-likelihood. The customary method of finding the maximum is taking partial derivatives with respect to the parameters of interest and finding roots of these equations. In this case, taking partial derivatives with respect to \( \beta \) yields the \textit{score equations}
The solution of these equations, say \( b_{\text{MLE}} \), is the maximum likelihood estimator. To illustrate, for the logit function, the score equations in equation (9.2) reduce to

\[
\sum_{it} x_{it} (y_{it} - \pi(x_{it}' \beta)) = 0,
\]

(9.3)

where \( \pi(z) = (1 + \exp(-z))^{-1} \). We note that the solution depends on the responses \( y_{it} \) only through the statistics \( \Sigma_{it} x_{it} y_{it} \). This property, known as sufficiency, will be important in Section 9.3.

An alternative expression for the score equations in equation (9.2) is

\[
\sum_{it} \frac{\partial}{\partial \beta} (E y_{it} | (\text{Var } y_{it})^{-1} (y_{it} - E y_{it}) = 0,
\]

(9.4)

where \( E y_{it} = \pi(x_{it}' \beta) \), \( \frac{\partial}{\partial \beta} E y_{it} = x_{it} \pi'(x_{it}' \beta) \) and \( \text{Var } y_{it} = \pi(x_{it}' \beta)(1 - \pi(x_{it}' \beta)) \). The expression in equation (9.4) is an example of a generalized estimating equation that will be introduced formally in Section 9.4.

An estimator of the asymptotic variance of \( \beta \) may be calculated taking partial derivatives of the score equations. Specifically, the term

\[
\frac{\partial^2}{\partial \beta \partial \beta'} L(\beta)_{b_{\text{MLE}}} \left. \right|_{b_{\text{MLE}}}
\]

is the information matrix evaluated at \( b_{\text{MLE}} \). To illustrate, using the logit function, straightforward calculations show that the information matrix is

\[
\sum_{it} x_{it} x_{it}' \pi(x_{it}' \beta)(1 - \pi(x_{it}' \beta))
\]

The square root of the \( j \)th diagonal element of this matrix yields the standard error for the \( j \)th row of \( b_{\text{MLE}} \), which we denote as \( \text{se}(b_{\text{MLE}}) \).

To assess the overall model fit, it is customary to cite likelihood ratio test statistics in nonlinear regression models. To test the overall model adequacy \( H_0: \beta = 0 \), we use the statistic

\[
LRT = 2 \times (L(b_{\text{MLE}}) - L_0),
\]

where \( L_0 \) is the maximized log-likelihood with only an intercept term. Under the null hypothesis \( H_0 \), this statistic has a chi-square distribution with \( K \) degrees of freedom. Appendix C.8 describes likelihood ratio test statistics in greater detail.

As described in Appendix C.9, measures of goodness of fit are difficult to interpret in nonlinear models. One measure is the so-called max-scaled \( R^2 \), defined as \( R^2_{\text{max}} = \frac{R^2_{\text{max}}}{R^2_{\text{max}}} \), where

\[
R^2 = 1 - \left( \frac{\exp(L_0/N)}{\exp(L(b_{\text{MLE}})/N)} \right)^2 \quad \text{and} \quad R^2_{\text{max}} = 1 - \exp(L_0/N)^2. \]

Here, \( L_0/N \) represents the average value of this log-likelihood.
9.1.3 Example: Income tax payments and tax preparers

To illustrate the methods described in this section, we return to the Income tax payments example introduced in Section 3.2. For this chapter, we now will use the demographic and economic characteristics of a taxpayer described in Table 3.1 to model the choice of using a professional tax preparer, denoted by PREP. The one exception is that we will not consider the variable LNTAX, the tax paid in logarithmic dollars. Although tax paid is clearly related to the choice of a professional tax preparer, it is not clear that this can serve as an “independent” explanatory variable. In econometric terminology, this is considered an endogenous variable (see Chapter 6).

Many summary statistics of the data were discussed in Sections 3.2 and 7.2.3. Tables 9.1 and 9.2 show additional statistics, by level of PREP. Table 9.1 shows that those taxpayers using a professional tax preparer (PREP=1) were more likely to be married, not the head of a household, age 65 and over and self-employed. Table 9.2 shows that those taxpayers using a professional tax preparer had more dependents, larger income and were in a higher tax bracket.

| Table 9.1 Averages of Indicator Variables by Level of PREP |
|-----------------|--------|--------|--------|--------|--------|
| PREP  | Number  | MS     | HH     | AGE    | EMP    |
| 0    | 671     | 0.542  | 0.106  | 0.072  | 0.092  |
| 1    | 619     | 0.709  | 0.066  | 0.165  | 0.212  |

| Table 9.2 Summary Statistics for Other Variables by Level of PREP |
|-----------------|--------|--------|--------|--------|
| Variable | PREP  | Mean   | Median | Minimum | Maximum | Standard deviation |
| DEPEND     | 0     | 2.267  | 2      | 0       | 6       | 1.301               |
|            | 1     | 2.585  | 2      | 0       | 6       | 1.358               |
| LNTAX      | 0     | 9.732  | 9.921  | -0.128  | 12.043  | 1.089               |
|            | 1     | 10.059 | 10.178 | -0.092  | 13.222  | 1.220               |
| MR         | 0     | 21.987 | 21     | 0       | 50      | 11.168              |
|            | 1     | 25.188 | 25     | 0       | 50      | 11.536              |

Table 9.3 provides additional information about the relation between EMP and PREP. To illustrate, for those self-employed individuals (EMP=1), 67.9% (=131/193) of the time they chose to use a tax preparer compared to 44.5% (= 488/1,097) for those not self-employed. Put another way, the odds of self-employed using a preparer are 2.11 (= 0.679/(1-0.679)) compared to 0.801 (=0.445/(1-0.445)) for those not self-employed.

| Table 9.3 Counts of Taxpayers by Levels of PREP and EMP |
|-----------------|--------|--------|
| EMP  | 0    | 1    | Total |
| PREP  | 609  | 62  | 671  |
| 1    | 488  | 131 | 619  |
| Total| 1,097| 193 | 1290 |

Display 9.1 shows a fitted logistic regression model, using LNTAX, MR and EMP as explanatory variables. The calculations were done using SAS PROC LOGISTIC. To interpret this output, we first note that the likelihood ratio test statistic for checking model adequacy is $LRT = 67.2422 = 2 \times (L(b_{MLE}) - 2 \times L_0) = 1786.223 - 1718.981$. Compared to a chi-square with $K=3$ degrees of freedom, this indicates that at least one of the variables LNTAX, MR and EMP is a statistically significant predictors of PREP. Additional model fit statistics, including Akaike’s information criterion (AIC) and Schwarz’s criterion (SC) are described in Appendix C.9.
We interpret the $R^2$ statistic to mean that there is substantial information regarding PREP that is not explained by LNTPI, MR and EMP. It is useful to confirm the calculation of this statistic, this being

$$R^2 = 1 - \frac{\exp(2 \times L_0 / N)}{\exp(2 \times L_\text{MLE} / N)} = 1 - \frac{\exp(2 \times 1786.223/1290)}{\exp(2 \times 1718.981/1290)} = 0.05079.$$ 

Display 9.1 Selected SAS Output

The LOGISTIC Procedure

Model Information
Response Variable PREP
Number of Response Levels 2
Number of Observations 1290
Link Function Logit
Optimization Technique Fisher's scoring

Model Fit Statistics

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Intercept and Covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>Only</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Square 0.0508 Max-rescaled R-Square 0.0678

Testing Global Null Hypothesis: BETA=0

<table>
<thead>
<tr>
<th>Test</th>
<th>Chi-Square</th>
<th>DF</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood Ratio</td>
<td>67.2422</td>
<td>3</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Score</td>
<td>65.0775</td>
<td>3</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Wald</td>
<td>60.5549</td>
<td>3</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Analysis of Maximum Likelihood Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Error</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-2.3447</td>
<td>0.7754</td>
<td>9.1430</td>
<td>0.0025</td>
</tr>
<tr>
<td>LNTPI</td>
<td>1</td>
<td>0.1881</td>
<td>0.0940</td>
<td>4.0017</td>
<td>0.0455</td>
</tr>
<tr>
<td>MR</td>
<td>1</td>
<td>0.0108</td>
<td>0.00884</td>
<td>1.4964</td>
<td>0.2212</td>
</tr>
<tr>
<td>EMP</td>
<td>1</td>
<td>1.0091</td>
<td>0.1693</td>
<td>35.5329</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Odds Ratio Estimates

<table>
<thead>
<tr>
<th>Effect</th>
<th>Point Estimate</th>
<th>95% Wald Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>LNTPI</td>
<td>1.207</td>
<td>1.004, 1.451</td>
</tr>
<tr>
<td>MR</td>
<td>1.011</td>
<td>0.994, 1.029</td>
</tr>
<tr>
<td>EMP</td>
<td>2.743</td>
<td>1.969, 3.822</td>
</tr>
</tbody>
</table>
For parameter interpretation, we note that the coefficient associated with EMP is $b_{EMP} = 1.0091$. Thus, we interpret the odds associated with this estimator, $\exp(1.0091) = 2.743$, to mean that self-employed taxpayers (EMP=1) are 2.743 times more likely to employ a professional tax preparer compared to taxpayers that are not self-employed.

9.2 Random effects models

This section introduces models that use random effects to accommodate heterogeneity. Section 9.3 follows up with the corresponding fixed effect formulation. In contrast, in the linear models portion of the text, we first introduced fixed effects (in Chapter 2) followed by random effects (in Chapter 3). The consistency between these seemingly different approaches is that the text approaches data modeling from an applications orientation. Specifically, for estimation and ease of explanations with users, typically the fixed effects formulation is simpler than the random effects alternative in linear models. This is because fixed effects models are simply special cases of analysis of covariance models, representations that are familiar from applied regression analysis. In contrast, in nonlinear cases such as models with binary dependent variables, random effects models are simpler than corresponding fixed effects alternatives. Here, this is in part computational because random effects summary statistics are easier to calculate. Further, as we will see in Section 9.3, standard estimation routines, such as maximum likelihood, yield fixed effects estimators that do not have the usual desirable asymptotic properties. Thus, the fixed effects formulation requires specialized estimators that can be cumbersome to compute and explain to users.

As in Section 9.1, we expressed the probability of a response equal to one as a function of linear combinations of explanatory variables. To accommodate heterogeneity, we incorporate subject-specific variables of the form $\text{Prob}(y_{it} = 1 | \alpha_i) = \pi(\alpha_i + x_{it}' \beta)$. Here, the subject-specific effects account only for the intercepts and do not include other variables. Chapter 10 will introduce extensions to variable slope models. We assume that $\{\alpha_i\}$ are random effects in this section.

To motivate the random effects formulation, we may assume the two-stage sampling scheme introduced in Section 3.1.

**Stage 1.** Draw a random sample of $n$ subjects from a population. The subject-specific parameter $\alpha_i$ is associated with the $i$th subject.

**Stage 2.** Conditional on $\alpha_i$, draw realizations of $\{y_{it}, x_{it}\}$, for $t = 1, \ldots, T_i$ for the $i$th subject.

In the first stage, one draws subject-specific effects $\{\alpha_i\}$ from a population. In the second stage, for each subject $i$, one draws a random sample of $T_i$ responses $y_{it}$, $t = 1, \ldots, T_i$ and also observes the explanatory variables $\{x_{it}\}$.

**Random effects likelihood**

To develop the likelihood, first note that from the second sampling stage, conditional on $\alpha_i$, the likelihood for the $i$th subject at the $t$th observation is

$$p(y_{it} ; \beta | \alpha_i) = \begin{cases} \pi(\alpha_i + x_{it}' \beta) & \text{if } y_{it} = 1 \\ 1 - \pi(\alpha_i + x_{it}' \beta) & \text{if } y_{it} = 0 \end{cases}.$$  

We summarize this as

$$p(y_{it} ; \beta | \alpha_i) = \pi(\alpha_i + x_{it}' \beta)^{y_{it}} (1 - \pi(\alpha_i + x_{it}' \beta))^{1-y_{it}}.$$  

Because of the independence among responses for a subject conditional on $\alpha_i$, the conditional likelihood for the $i$th subject is
\[ p(y_i; \beta, \alpha_i) = \prod_{t=1}^{T_i} \pi(\alpha_i + x_i' \beta)^{y_i} (1 - \pi(\alpha_i + x_i' \beta))^{1-y_i}. \]

Taking expectations over \( \alpha_i \) yields the unconditional likelihood. Thus, the (unconditional) likelihood for the \( i \)th subject is

\[ p(y_i; \beta, \tau) = \int \left\{ \prod_{t=1}^{T_i} \pi(a + x_i' \beta)^{y_i} (1 - \pi(a + x_i' \beta))^{1-y_i} \right\} dF_{\alpha}(a). \]  (9.5)

In equation (9.5), \( \tau \) is a parameter of the distribution of \( \alpha_i, F_{\alpha}(\cdot) \). Although not necessary, it is customary to use a normal distribution for \( F_{\alpha}(\cdot) \). In this case, \( \tau \) represents the variance of this mean zero distribution. With a specification for \( F_{\alpha} \), the log-likelihood for the data set is

\[ L(\beta, \tau) = \sum_{i=1}^{n} \ln p(y_i; \beta, \tau). \]

To determine maximum likelihood estimators, one maximizes the log-likelihood \( L(\beta, \tau) \) as a function of \( \beta \) and \( \tau \). Closed form analytical solutions for this maximization problem do not exist in general, although numerical solutions are feasible with modern computing equipment. The maximum likelihood estimators can be determined by solving for the roots of the \( K+1 \) score equations

\[ \frac{\partial}{\partial \beta} L(\beta, \tau) = 0 \quad \text{and} \quad \frac{\partial}{\partial \tau} L(\beta, \tau) = 0. \]

Further, asymptotic variances can be computed by taking the matrix of second derivatives of the log-likelihood, known as the information matrix. Appendix C provides additional details.

There are two commonly used specifications of the conditional distribution in the random effects model.

- A logit model for the conditional distribution of a response. That is,

\[ \text{Prob}(y_{it} = 1 | \alpha_i) = \pi(\alpha_i + x_i' \beta) = \frac{1}{1 + \exp(- (\alpha_i + x_i' \beta))}. \]

- A probit model for the conditional distribution of a response. That is,

\[ \text{Prob}(y_{it} = 1 | \alpha_i) = \Phi(\alpha_i + x_i' \beta), \] where \( \Phi \) is the standard normal distribution function.

There are no important advantages or disadvantages when choosing the conditional probability \( \pi \) to be either a logit or a probit. The likelihood involves roughly the same amount of work to evaluate and maximize, although the logit function is slightly easier to evaluate than the standard normal distribution function. The probit model can be easier to interpret because unconditional probabilities can be expressed in terms of the standard normal distribution function. That is, assuming normality for \( \alpha_i \), we have

\[ \text{Prob}(y_{it} = 1) = E \Phi(\alpha_i + x_i' \beta) = \Phi \left( \frac{x_i' \beta}{\sqrt{1 + \tau}} \right). \]

**Example - Income tax payments and tax preparers - Continued**

To see how a random effects dependent variable model works with a data set, we return to the Section 9.1.3 example. Display 9.2 shows a fitted model, using LNTPI, MR and EMP as explanatory variables. The calculations were done using the SAS procedure NLMIXED. This procedure uses a numerical integration technique for mixed effect models, called *adaptive Gaussian quadrature* (see Pinheiro and Bates, 2000, for a description). Display 9.2 shows that this random effects specification is not a desirable model for this data set. By conditioning on the random effects, the parameter estimates turn out to be highly correlated with one another.
We saw that there were a sufficient number of applications to devote an entire chapter (5) to a multilevel framework for linear models. However, extensions to nonlinear models such as binary dependent variable models are only now coming into regular use in the applied sciences. This subsection presents the development of three-level extensions of probit and logit models due to Gibbons and Hedeker (1997B).

We use a set of notation similar to that developed in Section 5.1.2. Let there be \( i = 1, \ldots, n \) subjects (schools, for example), \( j = 1, \ldots, J_i \) clusters within each subject (classrooms, for example), and \( t = 1, \ldots, T_{ij} \) observations within each cluster (over time, or students with a classroom). Combining the three levels, Gibbons and Hedeker considered

\[
y_{ijt}^* = \alpha_i + z_{ijt}^* \mathbf{a}_g + x_{ijt}^* \mathbf{b} + \epsilon_{ijt}.
\]

Here, \( \alpha_i \) represents a level-three heterogeneity term, \( \mathbf{a}_g \) represents a level-two vector of heterogeneity terms and \( \epsilon_{ijt} \) represents the level-one disturbance term. All three terms have zero mean; the means of each level are already accounted for in \( x_{ijt}^* \mathbf{b} \). The left-hand variable of equation (9.6) is latent; we actually observe \( y_{ijt} \) that is a binary variable, corresponding to whether the latent variable \( y_{ijt}^* \) crosses a threshold. The conditional probability of this event is

\[
\text{Prob}(y_{ijt} = 1 | \alpha_i, \mathbf{a}_g) = \pi(\alpha_i + z_{ijt}^* \mathbf{a}_g + x_{ijt}^* \mathbf{b}),
\]

where \( \pi(\cdot) \) is either logit or probit. To complete the model specification, we assume that \( \{\alpha_i\} \) and \( \{\mathbf{a}_g\} \) are each i.i.d. and independent of one another.
The model parameters can be estimated via maximum likelihood. As pointed out by Gibbons and Hedeker, the main computational technique is to take advantage of the independence that we typically assume among levels in multilevel modeling. That is, defining $y_{ij} = (y_{ij1}, ..., y_{ijT_{ij}})'$, the conditional probability mass function is

$$p(y_{ij} | \beta, \alpha_{ij}, a_{ij}) = \prod_{t=1}^{T_{ij}} \pi(\alpha_{ij} + z_{ijt}^{'}, a_{ij} + x_{ijt}^{'}, \beta)^{y_{ijt}} \left(1 - \pi(\alpha_{ij} + z_{ijt}^{'}, a_{ij} + x_{ijt}^{'}, \beta)\right)^{1-y_{ijt}}.$$

Integrating over the level-two heterogeneity effects, we have

$$p(y_{ij} ; \beta, \Sigma_2 | \alpha_{ij}) = \int_a p(y_{ij} ; \beta, a | \alpha_{ij}) d F_{a,2}(a). \quad (9.7)$$

Here, $F_{a,2}(.)$ is the distribution function of $\{a_{ij}\}$ and $\Sigma_2$ are the parameters associated with it. Following Gibbons and Hedeker, we assume that $a_{ij}$ is normally distributed with mean zero and variance-covariance $\Sigma_2$.

Integrating over the level-three heterogeneity effects, we have

$$p(y_{ij1}, ..., y_{ijJ_i} ; \beta, \Sigma_2, \sigma_3^2) = \int_a \prod_{j=1}^{J_i} p(y_{ij} ; \beta, \Sigma_2 | a) d F_{a,3}(a). \quad (9.8)$$

Here, $F_{a,3}(.)$ is the distribution function of $\{a_{ij}\}$ and $\sigma_3^2$ is the parameter associated with it (typically, normal). With equations (9.7) and (9.8), the log-likelihood is

$$L(\beta, \Sigma_2, \sigma_3^2) = \sum_{i=1}^n \ln p(y_{ij1}, ..., y_{ijJ_i} ; \beta, \Sigma_2, \sigma_3^2). \quad (9.9)$$

Computation of the log-likelihood requires numerical integration. However, the integral in equation (9.8) is only 1-dimensional and the integral in equation (9.7) depends on the dimension of $\{a_{ij}\}$, say $q$, that is typically only 1 or 2. This is in contrast to a more direct approach that combines the heterogeneity terms ($\alpha_i, a_{ij}, ..., a_{ijJ_i}$). The dimension of this vector is $(1+qJ_i) \times 1$; using this directly in equation (9.5) is much more computationally intense.

**Example 9.2. Family smoking prevention**

To illustrate their procedures, Gibbons and Hedeker (1997B) considered data from the Television School and Family Smoking Prevention and Cessation Project. In their report of this study, Gibbons and Hedeker considered 1,600 seventh grade students, from 135 classrooms within 28 schools. The dataset was unbalanced; there were between 1 and 13 classrooms from each school and between 2 and 28 students from each classroom. The schools were randomly assigned to one of four study conditions:

- a social resistance classroom in which a school-based curriculum was used to promote tobacco use prevention and cessation,
- a television based curriculum,
- a combination of social resistance and television based curricula and
- no treatments (control).

A tobacco and health scale was used to classify each student as knowledgeable or not, both before and after the intervention.

Table 9.4 provides empirical probabilities of students’ knowledge after the interventions, by type of intervention. This table suggests that social resistance classroom curricula are effective in promoting tobacco prevention awareness.
Table 9.4 Tobacco and Health Scale Post-intervention Performance

<table>
<thead>
<tr>
<th>Social resistance classroom</th>
<th>Television based curriculum</th>
<th>Count</th>
<th>Knowledgeable</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>No</td>
<td>421</td>
<td>41.6</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>416</td>
<td>48.3</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>380</td>
<td>63.2</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>383</td>
<td>60.3</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1,600</td>
<td>52.9</td>
</tr>
</tbody>
</table>

Gibbons and Hedeker estimated both logit and probit models using type of intervention and performance of the pre-intervention test as explanatory variables. They considered a model with random effects as well as a model with classroom as the second level and school as the third level (as well as two two-level models for robustness purposes). For both models, the social resistance classroom curriculum was statistically significant. However, they also found that the model without random effects indicated that the television based intervention was statistically significant whereas the three-level model did not reveal such a strong effect.

9.3 Fixed effects models

As in Section 9.1, we express the probability of the response being a one as a nonlinear function of linear combinations of explanatory variables. To accommodate heterogeneity, we incorporate subject-specific variables of the form

\[ p_{it} = \pi(\alpha_i + x_{it}' \beta). \]

Here, the subject-specific effects account only for the intercepts and do not include other variables. Extensions to variable slope models are possible but, as we will see, even variable intercept models are difficult to estimate. We assume that \( \{\alpha_i\} \) are fixed effects in this section.

Maximum likelihood estimation

Similar to equation (9.1), the log-likelihood of the data set is

\[ L = \sum \left\{ y_{it} \ln \pi(\alpha_i + x_{it}' \beta) + (1 - y_{it}) \ln(1 - \pi(\alpha_i + x_{it}' \beta)) \right\}. \]  

(9.10)

This log-likelihood can be maximized to yield maximum likelihood estimators of \( \alpha_i \) and \( \beta \) that we denote as \( a_{i,MLE} \) and \( b_{MLE} \). Note that there are \( n + K \) parameters to be estimated simultaneously. As in Section 9.1, we consider the logit specification of \( \pi \), so that

\[ p_{it} = \pi(\alpha_i + x_{it}' \beta) = \frac{1}{1 + \exp(\left( \alpha_i + x_{it}' \beta \right))}. \]  

(9.11)

Because \( \ln \left( \pi(x)/(1-\pi(x)) \right) = x \), we have that the log-likelihood in equation (9.10) is

\[ L = \sum \left\{ y_{it} \ln(1 - \pi(\alpha_i + x_{it}' \beta)) + y_{it} \ln\left( \frac{\pi(\alpha_i + x_{it}' \beta)}{1 - \pi(\alpha_i + x_{it}' \beta)} \right) \right\} \]

\[ = \sum \left\{ y_{it} \ln(1 - \pi(\alpha_i + x_{it}' \beta)) + y_{it}(\alpha_i + x_{it}' \beta) \right\}. \]  

(9.12)

Straightforward calculations show that the score equations are:
\[
\frac{\partial L}{\partial \alpha_i} = \sum_i (y_{it} - \pi(\alpha_i + x_{it}'\beta)) \quad \text{and} \quad \frac{\partial L}{\partial \beta} = \sum_i x_{it}(y_{it} - \pi(\alpha_i + x_{it}'\beta)).
\] (9.13)

Finding the roots of these equations yield our maximum likelihood estimators.

**Example - Income tax payments and tax preparers - Continued**

To see how maximum likelihood works with a data set, we return to the Section 9.1.3 example. For this data set, we have \( n = 258 \) taxpayers and consider \( K=3 \) explanatory variables, LNTPI, MR and EMP. Fitting this model yields \(-2 \times \log\text{-likelihood} = 416.024 \) and \( R^2 = 0.6543 \). According to standard likelihood ratio tests, the additional intercept terms are highly statistically significant. That is, the likelihood ratio test statistic for assessing the null hypothesis \( H_0: \alpha_1 = \ldots = \alpha_{558} \) is

\[
LRT = 1,718.981 - 416.024 = 1,302.957.
\]

The null hypothesis is rejected based on a comparison of this statistic with a chi-square distribution with 258 degrees of freedom.

Unfortunately, the above analysis is based on approximations that are known to be unreliable. The difficulty is that, as the subject size \( n \) tends to infinity, the number of parameters also tends to infinity. It turns out that our ability to estimate \( \beta \) is corrupted by our inability to estimate consistently the subject-specific effects \( \{\alpha_i\} \). In contrast, in the linear case, maximum likelihood estimators are equivalent to the least squares estimators that are consistent. The least squares procedure “sweeps out” intercept estimates when producing estimates of \( \beta \). This is not the case in nonlinear regression models.

To get a better feel for the types of things that can go wrong, suppose that we have no explanatory variables. In this case, from display (9.13), the root of the score equation is

\[
\frac{\partial L}{\partial \alpha_i} = \sum_i \left\{ - \frac{\exp(\alpha_i)}{1 + \exp(\alpha_i)} + y_{it} \right\} = 0.
\]

The solution \( a_{i,MLE} \) is

\[
\bar{y}_i = \frac{\exp(a_{i,MLE})}{1 + \exp(a_{i,MLE})} \quad \text{or} \quad a_{i,MLE} = \logit(\bar{y}_i).
\]

Thus, if \( \bar{y}_i = 1 \), then \( a_{i,MLE} = \infty \) and if \( \bar{y}_i = 0 \), then \( a_{i,MLE} = -\infty \). Thus, intercept estimators are unreliable in these circumstances. An examination of the score functions in display (9.13) shows that similar phenomena also occurs even in the presence of explanatory variables. To illustrate, in Example 9.1 we have 97 taxpayers who do not use a professional tax preparer in any of the five years under considerations (\( \bar{y}_i=0 \)) whereas 89 taxpayers always use a tax preparer (\( \bar{y}_i=1 \)).

Even when the intercept estimators are finite, maximum likelihood estimators of global parameters \( \beta \) are inconsistent in fixed effects binary dependent variable models. See Example 9.2.

**Illustration 9.1 - Inconsistency of maximum likelihood estimates (Chamberlain, 1978, Hsiao 1986).**

Now, as a special case, let \( T_i = 2, K=1 \) and \( x_{i1} = 0 \) and \( x_{i2}=1 \). Using the score equations in display (9.12), Appendix 9A shows how to calculate directly the maximum likelihood estimator of \( \beta, b_{MLE} \), for this special case. Further, Appendix 9A.1 argues that the probability limit of \( b_{MLE} \) is 2 \( \beta \). Hence, it is an inconsistent estimator of \( \beta \).
Conditional maximum likelihood estimation

To circumvent the problem of the intercept estimators corrupting the estimator of $\beta$, we use the conditional maximum likelihood estimator. This estimation technique is due to Chamberlain (1980E) in the context of fixed effects binary dependent variable models. We consider the logit specification of $\pi$ as in equation (9.12). With this specification, it turns out that $\Sigma_t y_{it}$ is a sufficient statistic for $\alpha_i$. The idea of sufficiency is reviewed in Appendix 10A.2. In this context, it means that if we condition on $\Sigma_t y_{it}$, then the distribution of the responses will not depend on $\alpha_i$.

Illustration 9.2 - Sufficiency

Continuing with the set-up of Illustration 9.1, we now illustrate how to separate the intercept from the slope effects. Here, we only assume that $T_i = 2$, not that $K = 1$.

To show that the conditional distribution of the responses do not depend on $\alpha_i$, begin by supposing that the sum, $\Sigma_t y_{it} = y_{i1} + y_{i2}$, equals either 0 or 2. Consider three cases. For the first case, assume that sum equals 0. Then, the conditional distribution of the responses is $\text{Prob}(y_{i1} = 0, y_{i2} = 0 | y_{i1} + y_{i2} = \text{sum}) = 1$; this clearly does not depend on $\alpha_i$. For the second case, assume that sum equals 2. Then, the conditional distribution of the responses is $\text{Prob}(y_{i1} = 1, y_{i2} = 1 | y_{i1} + y_{i2} = \text{sum}) = 1$, which also does not depend on $\alpha_i$. Stated another way, if $y_{ij}$ is either 0 or 1, then the statistic that is being conditioned on determines all the responses, resulting in no contribution to a conditional likelihood.

Now consider the third case, where the sum equals 1. Basic probability calculations establish that

$$\text{Prob}(y_{i1} + y_{i2} = 1) = \frac{\text{Prob}(y_{i1} = 0)\text{Prob}(y_{i2} = 1) + \text{Prob}(y_{i1} = 1)\text{Prob}(y_{i2} = 0)}{(1 + \exp(\alpha_i + x'_{i1}\beta))(1 + \exp(\alpha_i + x'_{i2}\beta))}$$

Thus, if sum equals 1, then

$$\text{Prob}(y_{i1} = 0, y_{i2} = 1 | y_{i1} + y_{i2} = 1) = \frac{\text{Prob}(y_{i1} = 0)\text{Prob}(y_{i2} = 1)}{\text{Prob}(y_{i1} + y_{i2} = 1)} = \frac{\exp(\alpha_i + x'_{i1}\beta)}{\exp(\alpha_i + x'_{i2}\beta) + \exp(\alpha_i + x'_{i2}\beta)} = \frac{\exp(x'_{i1}\beta)}{\exp(x'_{i1}\beta) + \exp(x'_{i2}\beta)}.$$  

Thus, the conditional distribution of the responses does not depend on $\alpha_i$. We also note that if an explanatory variable $x_j$ is time-constant ($x_{t2j} = x_{t1j}$), then the corresponding parameter $\beta_j$ disappears from the conditional likelihood.
Conditional likelihood estimation

To define the conditional likelihood, let \( S_i \) be the random variable representing \( \sum y_{it} \) and let \( \text{sum}_i \) be the realization of \( \sum y_{it} \). With this notation, the conditional likelihood of the data set is

\[
\prod_{i=1}^{n} \left\{ \frac{1}{\text{Prob}(S_i = \text{sum}_i)} \prod_{t=1}^{T_i} p^{y_{it}} \left( 1 - p^{1-y_{it}} \right) \right\}.
\]

Note that the ratio within the curly brackets equals one when \( \text{sum}_i \) equal 0 or \( T_i \). Taking the log of the function and then finding values of \( \beta \) that maximize it yields \( b_{\text{CMLE}} \), the conditional maximum likelihood estimator. We remark that this can be computationally difficult. That is, the distribution of \( S_i \) is messy and is difficult to compute for moderate size data sets with \( T \) more than 10. Appendix 9A.2 provides details.

Illustration 9.3 - Conditional maximum likelihood estimator

To see that the conditional maximum likelihood estimator is consistent in a case where the maximum likelihood is not, we continue with Illustration 9.1. As argued in Illustration 9.2, we need only be concerned with the case \( y_{i1} + y_{i2} = 1 \). The conditional likelihood is

\[
\prod_{i=1}^{n} \left\{ \frac{1}{\text{Prob}(y_{i1} + y_{i2} = 1)} \prod_{t=1}^{2} p^{y_{it}} \left( 1 - p^{1-y_{it}} \right) \right\} = \prod_{i=1}^{n} \frac{y_{i1} \exp(x_{i1}' \beta) + y_{i2} \exp(x_{i2}' \beta)}{\exp(x_{i1}' \beta) + \exp(x_{i2}' \beta)}.
\]  (9.14)

As in Example 9.2, take \( K=1 \) and \( x_{i1} = 0 \) and \( x_{i2} = 1 \). Then, by taking the derivative with respect to \( \beta \) of the log of the conditional likelihood and setting this equal to zero, one can determine explicitly the conditional maximum likelihood estimator, denoted as \( b_{\text{CMLE}} \). Straightforward limit theory shows this to be a consistent estimator of \( \beta \). Appendix 9A.1 provides details.

A note on terminology - conditional maximum likelihood estimation for the logit model differs from the conditional logit model that we will introduce in Section 11.1.

9.4 Marginal models and GEE

For marginal models, we require only the specification of the first two moments, specifically the mean and variance of a response as well as the covariances among responses. This is much less information than the entire distribution, as required by the likelihood based approaches in Sections 9.2 and 9.3. Of course, if the entire distribution is assumed known, then we can always calculate the first two moments. Thus, the estimation techniques applicable to marginal models can also be used when the entire distribution is specified.

Marginal models are estimated using a special type of moment estimation known as the generalized estimating equations, or GEE, approach in the biological sciences. In the social sciences, this approach is part of the generalized method of moments, or GMM. For the applications that we have in mind, it is most useful to develop the estimation approach using the GEE notation. However, analysts should keep in mind that this estimator is really just another type of GMM estimator.

To describe GEE estimators, one must specify a mean, variance and covariance structure. To illustrate the development, we begin by assuming that the Section 9.2 random effects model is valid and we wish to estimate parameters of this distribution. The general GEE procedure is described in Appendix C.6 and will be further developed in Chapter 10.
### GEE estimators for the random effects binary dependent variable model

From Section 9.2, we have that the conditional first moment of the response is \( E(y_{it} \mid \alpha_i) = \pi(\alpha_i + x_{it}' \beta) \). Thus, the mean may be expressed as

\[
\mu_{it} = \mu_{it}(\beta, \tau) = \int \pi(a + x_{it}' \beta) d F_\alpha(a). \tag{9.15}
\]

Recall that \( \tau \) is a parameter of the distribution function \( F_\alpha(.) \). For example, if \( F_\alpha(.) \) represents a normal distribution, then \( \tau \) represents the variance. Occasionally, it is useful to use the notation \( \mu_{it}(\beta, \tau) \) to remind ourselves that the mean function \( \mu_{it} \) depends on the parameters \( \beta \) and \( \tau \). Let \( \mu_i = (\mu_{i1}, \ldots, \mu_{iT_i})' \) denote the \( T_i \times 1 \) vector of means.

For this model, straightforward calculations show that the variance can be expressed as

\[
\text{Var}(y_{it}) = \mu_{it}(1 - \mu_{it}).
\]

Regarding covariances, for \( r \neq s \), we have

\[
\text{Cov}(y_{ir}, y_{is}) = E(y_{ir} y_{is}) - \mu_{ir} \mu_{is} = E \left( E(y_{ir} y_{is} \mid \alpha_i) \right) - \mu_{ir} \mu_{is} = E \left( \pi(\alpha_i + x_{ir}' \beta) \pi(\alpha_i + x_{is}' \beta) \right) - \mu_{ir} \mu_{is}. \tag{9.16}
\]

Let \( V_i = V_i(\beta, \tau) \) be the usual \( T_i \times T_i \) variance-covariance matrix for the \( i \)th subject; that is, the \( r \)th diagonal element of \( V_i \) is \( \text{Var}(y_{it}) \) whereas for non-diagonal elements, the \( r \)th row and \( s \)th column of \( V_i \) is given by \( \text{Cov}(y_{ir}, y_{is}) \).

For GEE, we also require derivatives of certain moments. For the mean function, from equation (9.15), we will use

\[
\frac{\partial}{\partial \beta} \mu_{it} = x_{it} \int \pi'(a + x_{it}' \beta) d F_\alpha(a).
\]

As is customary in applied data analysis, this calculation assumes a sufficient amount of regularity of the distribution function \( F_\alpha(.) \) so that we may interchange the order of differentiation and integration. In general, we will use the notation

\[
G_\mu(\beta, \tau) = \begin{pmatrix}
\frac{\partial \mu_{i1}}{\partial \beta} & \cdots & \frac{\partial \mu_{iT_i}}{\partial \beta}
\end{pmatrix},
\]

a \( K \times T_i \) matrix of derivatives.

### GEE estimation procedure

The GEE estimators are computed according to the following general recursion. Begin with initial estimators of \( (\beta, \tau) \), say \( (b_{0,EE}, \tau_{0,EE}) \). Typically, initial estimators \( b_{0,EE} \) are calculated assuming zero covariances among responses. Initial estimators \( \tau_{0,EE} \) are computed using residuals based on the \( b_{0,EE} \) estimate. Then, at the \( (n+1) \)st stage, recursively:

1. Use \( \tau_{n,EE} \) and the solution of the equation

\[
0_K = \sum_{i=1}^n G_\mu(b, \tau_{n,EE}) (V_i(b, \tau_{n,EE}))^{-1} (y_i - \mu_i(b, \tau_{n,EE}))
\]  \( \tag{9.17} \)

   to determine an updated estimator of \( \beta \), say \( b_{n+1,EE} \).

2. Use the residuals \( \{y_{it} - \mu_{it}(b_{n+1,EE}, \tau_{n,EE})\} \) to determine an updated estimator of \( \tau \), say \( \tau_{n+1,EE} \).

3. Repeat steps 1 and 2 until convergence.
Let $b_{EE}$ and $\tau_{EE}$ denote the resulting estimators of $\beta$ and $\tau$. Under broad conditions, $b_{EE}$ is consistent and asymptotically normal with asymptotic variance

$$
\left( \sum_{i=1}^{n} G_{\mu}(b_{EE}, \tau_{EE}) \left( V_i(b_{EE}, \tau_{EE}) \right)^{-1} G_{\mu}(b_{EE}, \tau_{EE})' \right)^{-1}.
$$

(9.18)

The solution $b_{EE}$ in equation (9.17) can be computed quickly using iterated reweighted least squares, a procedure described in Appendix C.3. However, the specified estimation procedure is still tedious because it relies on the numerical integration computations in calculating $\mu_\alpha$ in equation (9.15) and $\text{Cov}(y_{ir}, y_{is})$ in equation (9.16). Now, in Section 9.2, we saw that numerical integration in (9.15) could be avoided by specifying normal distributions for $\pi$ and $F_{\alpha}$. However, even with this specification, we would still require numerical integration to calculate $\text{Cov}(y_{ir}, y_{is})$ in equation (9.16). A single numerical integration is straightforward in modern-day computing environment. However, evaluation of $V_i$ would require $T_i (T_i - 1)/2$ numerical integrations for the covariance terms. Thus, each evaluation of equation (9.17) would require $\Sigma_i \{ T_i (T_i - 1)/2 \}$ numerical integrations (this is $258 \times 5 \times (5-1)/2 = 2,580$ for the Section 9.1.3 example). Many evaluations of equation (9.17) would be required prior to successful convergence of the recursive procedure. In summary, this approach is often computationally prohibitive.

To reduce these computational complexities, the focus of marginal models is the representation of the first two moments directly, with or without reference to underlying probability distributions. By focusing directly on the first two moments, we may keep the specification simple and computationally feasible.

To illustrate, we may choose to specify the mean function as $\mu_\alpha = \Phi(\frac{x_{it}' \beta}{\sqrt{1 + \tau}})$. This is certainly plausible under the random effects binary dependent variable model. For the variance function, we consider $\text{Var}(y_\alpha) = \phi \mu_\alpha (1 - \mu_\alpha)$. Here, $\phi$ is an overdispersion parameter that we may either assume to be 1 or to be estimated from the data. Finally, it is customery in the literature to specify correlations in lieu of covariances. Use the notation $\text{Corr}(y_{ir}, y_{is})$ to denote the correlation between $y_{ir}$ and $y_{is}$. To illustrate, it is common to use the exchangeable correlation structure specified as

$$
\text{Corr}(y_{ir}, y_{is}) = \begin{cases} 
1 & \text{for } r = s \\
\rho & \text{for } r \neq s.
\end{cases}
$$

Here, the motivation is that the latent variable $\alpha_i$ is common to all observations within a subject, thus inducing a common correlation. For this illustration, the parameters $\tau = (\phi, \rho)'$ constitute the variance components.

Estimation may then proceed as described in the recursion beginning with equation (9.17). However, as with linear models, the second moments may be misspecified. For this reason, the correlation specification is commonly known as a working correlation. For linear models, weighted least squares provides estimators with desirable properties. Although not optimal compared to generalized least squares, weighted least squares estimators are typically consistent and asymptotically normal. In the same fashion, GEE estimators based on “working correlations” have desirable properties, even when the correlation structure is not perfectly specified.

However, if the correlation structure is not valid, then the asymptotic standard errors provided through the asymptotic variance in equation (9.17) are not valid. Instead, empirical standard errors may be calculated using the following estimator of the asymptotic variance of $b_{EE}$.


Specifically, the standard error of the $j$th component of $b_{EE}$, $se(b_{j,EE})$, is defined to be the square root of the $j$th diagonal element of the variance-covariance matrix in display (9.19).

**Example - Income tax payments and tax preparers - Continued**

To see how a marginal model works with a data set, we return to the Section 9.1.3 example. Table 9.5 shows the fit of two models, each using LNTPI, MR and EMP as explanatory variables. The calculations were done using the SAS procedure GENMOD. For the first model, the exchangeable working correlation structure was used. Parameter estimates, as well as equation (9.19) empirical standard errors and equation (9.18) model-based standard errors, appear in Table 9.5. The estimated correlation parameter turned out to be $\hat{\rho} = 0.712$. For the second model, an unstructured working correlation matrix was used. (Table 2.5.1 provides an example of an unstructured covariance matrix.) Table 9.6 provides these estimated correlations.

We may interpret the ratio of the estimate to standard error as a $t$-statistic and use this to assess the statistical significance of a variable. Examining Table 9.5, we see that LNTPI and MR are not statistically significant, using either type of standard error or correlation structure. The variable EMP ranges from being strongly statistically significant, for the case with model-based standard errors and an exchangeable working correlation, to being not statistically significant, for the case of empirical standard errors and an unstructured working correlation. Overall, the GEE estimates of the marginal model provide dramatically different results when compared to either the Section 9.1 homogeneous model or the Section 9.2 random effects results.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Exchangeable Working Correlation</th>
<th>Unstructured Working Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>Empirical Standard Error</td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.9684</td>
<td>0.7010</td>
</tr>
<tr>
<td>LNTPI</td>
<td>0.0764</td>
<td>0.0801</td>
</tr>
<tr>
<td>MR</td>
<td>0.0024</td>
<td>0.0083</td>
</tr>
<tr>
<td>EMP</td>
<td>0.5096</td>
<td>0.2676</td>
</tr>
</tbody>
</table>

**Table 9.6 Estimate of Unstructured Correlation Matrix**

<table>
<thead>
<tr>
<th>Time</th>
<th>Time = 1</th>
<th>Time = 2</th>
<th>Time = 3</th>
<th>Time = 4</th>
<th>Time = 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time = 1</td>
<td>1.0000</td>
<td>0.8663</td>
<td>0.7072</td>
<td>0.6048</td>
<td>0.4360</td>
</tr>
<tr>
<td>Time = 2</td>
<td>0.8663</td>
<td>1.0000</td>
<td>0.8408</td>
<td>0.7398</td>
<td>0.5723</td>
</tr>
<tr>
<td>Time = 3</td>
<td>0.7072</td>
<td>0.8408</td>
<td>1.0000</td>
<td>0.9032</td>
<td>0.7376</td>
</tr>
<tr>
<td>Time = 4</td>
<td>0.6048</td>
<td>0.7398</td>
<td>0.9032</td>
<td>1.0000</td>
<td>0.8577</td>
</tr>
<tr>
<td>Time = 5</td>
<td>0.4360</td>
<td>0.5723</td>
<td>0.7376</td>
<td>0.8577</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
Further reading

More extensive introductions to (homogeneous) binary dependent variable models are available in Agresti (2002G) and Hosmer and Lemshow (2000G). For an econometrics perspective, see Cameron and Trivedi (1998E).

For discussions of binary dependent models with endogenous explanatory variables, see Wooldridge (2002E) and Arellano and Honoré (2001E).

For models of binary dependent variables with random intercepts, maximum likelihood estimators can be computed using numerical integration techniques to approximate the likelihood. McCulloch and Searle (2001G) discuss numerical integration for mixed effect models. Pinheiro and Bates (2000S) describe the adaptive Gaussian quadrature method that is used in SAS PROC NLMIXED.

Appendix 9. Likelihood calculations

Appendix 9A.1. Consistency of Likelihood Estimators

Illustration 9.1 - Inconsistency of maximum likelihood estimates – Continued

Recall that $T_i = 2$, $K = 1$ and $x_{i1} = 0$ and $x_{i2} = 1$. Thus, from equation (9.13), we have

$$\frac{\partial L}{\partial \alpha_i} = y_{i1} + y_{i2} - \frac{e^{\alpha_i}}{1 + e^{\alpha_i}} - \frac{e^{\alpha_i + \beta}}{1 + e^{\alpha_i + \beta}} = 0 \quad (9A.1)$$

and

$$\frac{\partial L}{\partial \beta} = \sum_i \left( y_{i2} - \frac{e^{\alpha_i + \beta}}{1 + e^{\alpha_i + \beta}} \right) = 0. \quad (9A.2)$$

From equation (9A.1), it is easy to see that if $y_{i1} + y_{i2} = 0$, then $a_{i,mle} = -\infty$. Further, if $y_{i1} + y_{i2} = 2$, then $a_{i,mle} = \infty$. For both cases, the contribution to the sum in equation (9A.2) is zero. Thus, we consider the case $y_{i1} + y_{i2} = 1$ and let $d_i$ be the indicator variable that $y_{i1} + y_{i2} = 1$. In this case, we have that $a_{i,mle} = -b_{mle}/2$ from equation (9A.1). Putting this into equation (9A.2) yields

$$\sum_i d_i y_{i2} = \sum_i d_i \frac{\exp(a_{i,mle} + b_{mle})}{1 + \exp(a_{i,mle} + b_{mle})} = \sum_i d_i \frac{\exp(b_{mle}/2)}{1 + \exp(b_{mle}/2)} = n_1 \frac{\exp(b_{mle}/2)}{1 + \exp(b_{mle}/2)}.$$

where $n_1 = \Sigma_i d_i$ is the number of subjects where $y_{i1} + y_{i2} = 1$. Thus, with the notation $\bar{y}_2^+ = \frac{1}{n_1} \sum_i d_i y_{i2}$, we have $b_{mle} = 2 \ln \frac{\bar{y}_2^+}{1 - \bar{y}_2^+}$.

To establish the inconsistency, straightforward weak law of large numbers can be used to show that the probability limit of $\bar{y}_2^+$ is $\frac{e^\beta}{1 + e^\beta}$. Thus, the probability limit of $b_{mle}$ is $2 \beta$, and hence it is an inconsistent estimator of $\beta$.

Illustration 9.3 - Conditional maximum likelihood estimator– Continued

Recall that $K = 1$, $x_{i1} = 0$ and $x_{i2} = 1$. Then, from equation (9.14), the conditional likelihood is
Chapter 9. Binary Dependent Variables / 9-21

\[ \prod_{i=1}^{n} \left( \frac{y_{i1} + y_{i2} \exp(\beta)}{1 + \exp(\beta)} \right)^{d_i} = \prod_{i=1}^{n} \left( \frac{\exp(\beta y_{i2})}{1 + \exp(\beta)} \right)^{d_i}, \]

because \( y_{i1} + y_{i2} = 1 \) and \( d_i \) is a variable to indicate \( y_{i1} + y_{i2} = 1 \). Thus, the conditional log-likelihood is

\[ L_{c}(\beta) = \sum_{i=1}^{n} d_i \left\{ \beta y_{i2} - \ln(1 + e^{\beta}) \right\}. \]

To find the conditional maximum likelihood estimator, we have

\[ \frac{\partial L_{c}(\beta)}{\partial \beta} = \sum_{i=1}^{n} d_i \left\{ y_{i2} - \ln(1 + e^{\beta}) \right\} = \sum_{i=1}^{n} d_i \left\{ y_{i2} - \frac{e^{\beta}}{1 + e^{\beta}} \right\} = 0. \]

The root of this is \( b_{CMLE} = \ln \frac{\bar{y}_{2}^{+}}{1 - \bar{y}_{2}^{+}} \). In Example 9.2, we used the fact that the probability limit of \( y_{i} \) is \( \beta \). Thus, the probability limit of \( b_{CMLE} \) is \( \beta \) and hence is consistent.

Appendix 9A.2. Computing Conditional Maximum Likelihood Estimators

Computing the Distribution of Sums of Nonidentically, Independently Distributed, Bernoulli Random Variables

We begin by presenting an algorithm for the computation of the distribution of sums of nonidentically, independently distributed, Bernoulli random variables. Thus, we take \( y_{it} \) to be independent Bernoulli random variables with \( \text{Prob}(y_{it} = 1) = \pi(x_{it}' \beta) \). For convenience, we use the logit form of \( \pi \). Define the sum random variable \( S_{iT} = y_{i1} + y_{i2} + \ldots + y_{iT} \). We wish to evaluate \( \text{Prob}(S_{iT} = s) \), for \( s = 0, \ldots, T \). For notational convenience, we omit the \( i \) subscript on \( T \).

We first note that it is straightforward to compute

\[ \text{Prob}(S_{iT} = 0) = \prod_{t=1}^{T} \left\{ 1 - \pi(x_{it}' \beta) \right\} = \prod_{t=1}^{T} \left\{ 1 + \exp(x_{it}' \beta) \right\}^{-1}, \]

using a logit form for \( \pi \). Continuing, we have

\[ \text{Prob}(S_{iT} = 1) = \sum_{i=1}^{T} \pi(x_{i}' \beta) \prod_{r=1, r \neq i}^{T} \left\{ 1 - \pi(x_{r}' \beta) \right\} = \sum_{i=1}^{T} \frac{\pi(x_{i}' \beta)}{1 - \pi(x_{i}' \beta)} \prod_{r=1}^{T} \left\{ 1 - \pi(x_{r}' \beta) \right\}. \]

Using a logit form for \( \pi \), we have \( \pi(z)/(1-\pi(z)) = e^{z} \). Thus, with this notation, we have

\[ \text{Prob}(S_{iT} = 1) = \text{Prob}(S_{iT} = 0) \left( \sum_{i=1}^{T} \exp(x_{i}' \beta) \right). \]

Let \( \{j_1, j_2, \ldots, j_s\} \) be a subset of \( \{1, 2, \ldots, T\} \) and \( \Sigma_{s,T} \) be the sum over all such subsets. Thus, for the next step in the iteration, we have

\[ \text{Prob}(S_{iT} = 2) = \sum_{2, r} \pi(x_{i,j_1}' \beta) \pi(x_{i,j_2}' \beta) \prod_{r=1, r \neq j_1, r \neq j_2}^{T} \left\{ 1 - \pi(x_{r}' \beta) \right\} \]

\[ = \text{Prob}(S_{iT} = 0) \sum_{2, r} \frac{\pi(x_{i,j_1}' \beta)}{1 - \pi(x_{i,j_1}' \beta)} \frac{\pi(x_{i,j_2}' \beta)}{1 - \pi(x_{i,j_2}' \beta)}. \]

Continuing this, in general we have
\[
\text{Prob}(S_{iT} = s) = \text{Prob}(S_{iT} = 0) \sum_{s,T} \pi(x_{i,j}^t \beta) \cdot \frac{\pi(x_{i,j}^t \beta)}{1 - \pi(x_{i,j}^t \beta)}.
\]

Using a logit form for \(\pi\), we may express the distribution as
\[
\text{Prob}(S_{iT} = s) = \text{Prob}(S_{iT} = 0) \sum_{s,T} \exp \left( x_{i,j} + \ldots + x_{i,j}^t \beta \right).
\]  

(9A.3)

Thus, even with the logit form for \(\pi\), we see the difficulty in computing \(\text{Prob}(S_{iT} = s)\) is that it involves the sum over \(\sum_{s,T}\) quantities in \(\Sigma_{s} T\). This expresses the distribution in terms of \(\text{Prob}(S_{iT} = 0)\). It is also possible to derive a similar expression in terms of \(\text{Prob}(S_{iT} = T)\); this alternative expression is more computationally useful than equation (9A.3) for evaluating the distribution at large values of \(s\).

**Computing the Conditional Maximum Likelihood Estimator**

From Section 9.3, the logarithmic conditional likelihood is
\[
\ln CL = \sum_{i=1}^{n} \left[ \sum_{t=1}^{T_i} y_{it} \ln \pi(x_{it}^t \beta) + (1 - y_{it}) \ln(1 - \pi(x_{it}^t \beta)) \right] - \ln \text{Prob}(S_{iT_i} = \text{sum}_{iT_i})
\]

where we have taken \(\alpha_i\) to be zero, without loss of generality. As remarked in Section 9.3, when summing over all subjects \(i = 1, \ldots, n\), we need not consider those subjects where \(\sum_{t=1}^{T_i} y_{it} = 0\) or \(T_i\) because the conditional likelihood is identically equal to one.

To find those values of \(\beta\) that maximize \(\ln CL\), one could use the Newton-Raphson recursive algorithm (see Appendix C.2). To this end, we require the vector of partial derivatives
\[
\frac{\partial}{\partial \beta} \ln CL = \sum_{i=1}^{n} \left[ \left( \sum_{t=1}^{T_i} x_{it} (y_{it} - \pi(x_{it}^t \beta)) \right) - \frac{\partial}{\partial \beta} \ln \text{Prob}(S_{iT_i} = \text{sum}_{iT_i}) \right].
\]

The Newton-Raphson algorithm also requires the matrix of second derivatives but computational considerations of this matrix are similar to the ones for the first derivative, and are omitted.

From the form of the vector of partial derivatives, we see that the main task is to compute the gradient of \(\ln \text{Prob}(S_{iT} = s)\). Using a logit form for \(\pi\) and equation (9A.3) (dropping the \(i\) subscript on \(T\)) we have
\[
\frac{\partial}{\partial \beta} \ln \text{Prob}(S_{iT} = s) = \frac{\partial}{\partial \beta} \ln \text{Prob}(S_{iT} = 0) + \frac{\partial}{\partial \beta} \ln \sum_{s,T} \exp \left( x_{i,j} + \ldots + x_{i,j}^t \beta \right)
\]
\[
= \frac{\sum_{s,T} \left( x_{i,j} + \ldots + x_{i,j}^t \right) \exp \left( x_{i,j} + \ldots + x_{i,j}^t \beta \right)}{\sum_{s,T} \exp \left( x_{i,j} + \ldots + x_{i,j}^t \beta \right)} - \sum_{t=1}^{T} x_{it} \exp(x_{it}^t \beta).
\]

As with the probability in equation (9A.3), this is easy to compute for values of \(s\) that are close to \(0\) or \(T\). However, in general the calculation requires sum over \(\sum_{s,T}\) quantities in \(\Sigma_{s,T}\), both in the numerator and denominator. Moreover, this is required for each subject \(i\) at each stage of the
Chapter 9. Binary Dependent Variables / 9-23

9. Exercises and Extensions

Section 9.1

9.1. Threshold interpretation of the probit regression model

Consider an underlying linear model, $y_{it}^* = \mathbf{x}_{it}' \beta + \epsilon_{it}^*$, where $\epsilon_{it}^*$ is normally distributed with mean zero and variance $\sigma^2$. Define $y_{it} = \begin{cases} 0 & y_{it}^* \leq 0 \\ 1 & y_{it}^* > 0 \end{cases}$. Show that $p_{it} = \Phi\left(\frac{\mathbf{x}_{it}' \beta}{\sigma}\right)$, where $\Phi$ is the standard normal distribution function.

9.2. Random utility interpretation of the logistic regression model

Under the random utility interpretation, an individual with utility $U_{ij} = u_{ij}(V_{ij} + \epsilon_{ij})$, where $j$ may be 0 or 1, selects category corresponding to $j = 1$ with probability $p_{it} = \text{Prob}(y_{it} = 1) = \text{Prob}(U_{i0} < U_{i1})$

$= \text{Prob}(u_{it}(V_{i0} + \epsilon_{i0}) < u_{it}(V_{i1} + \epsilon_{i1})) = \text{Prob}(\epsilon_{i0} - \epsilon_{i1} < V_{i1} - V_{i0}).$

Suppose that the errors are from an extreme value distribution of the form $\text{Prob}(\epsilon_{ij} < a) = \exp(-e^{-a})$.

Show that the choice probability $p_{it}$ has a logit form. That is, show $p_{it} = \frac{1}{1 + \exp(-\mathbf{x}_{it}' \beta)}$.

9.3. Marginal distribution of the probit random effects model

Consider a normal model for the conditional distribution of a response. That is, $\text{Prob}(y_{it} = 1|\alpha_i) = \Phi(\alpha_i + \mathbf{x}_{it}' \beta)$, where $\Phi$ is the standard normal distribution function.

Assume further that $\alpha_i$ is normally distributed with mean zero and variance $\tau$. Show that $p_{it} = \Phi\left(\frac{\mathbf{x}_{it}' \beta}{\sqrt{1 + \tau}}\right)$.

Empirical Exercise

9.4. Choice of yogurt brands

These data are known as scanner data because they are obtained from optical scanning of grocery purchases at check-out. The subjects consist of $n = 100$ households in Springfield, Missouri. The response of interest is the type of yogurt purchased. For this exercise, we consider only the brand “Yoplait” or another choice. The households were monitored over a two-year period with the number of purchases ranging from 4 to 185; the total number of purchases is $N = 2,412$. More extensive motivation is provided in Section 11.1.

The two marketing variables of interest are price and features. We use two price variables for this study, PY, the price of Yoplait, and PRICEOTHER, the lowest price of the other brands. For features, these are binary variables, defined to be one if there was a newspaper feature advertising the brand at time of purchase, and zero otherwise. We use two feature variables, FY,
the features associated with Yoplait, and FEATOTHER, a variable to indicate if any other brand has a featured displayed.

a. Basic summary statistics
   Create the basic summary statistics for Yoplait and the four explanatory variables.
   i. What are the odds of purchasing Yoplait?
   ii. Determine the odds ratio when assessing whether or not Yoplait is featured. Interpret your result.

b. Logistic regression models
   Run a logistic regression model with the two explanatory variables, PY and FY. Further, run a second logistic regression model with four explanatory variables, PY, FY PRICEOTHER and FEATOTHER. Compare these two models and say which you prefer. Justify your choice by appealing to standard statistical hypothesis tests.

c. Random effects model
   Fit a random effects model with four explanatory variables. Interpret the regression coefficients of this model.

d. GEE model
   Fit a GEE model with four explanatory variables.
   i. Give a brief description of the theory behind this model.
   ii. Compare the part d(i) description to the random effects model in part (c). In particular, how does each model address the heterogeneity?
   iii. For this data set, describe whether or not your model choice is robust to any important interpretations about the regression coefficients.
Chapter 10. Generalized Linear Models

Abstract. This chapter extends the linear model introduced in Part I and the binary dependent variable model in Chapter 9 to the generalized linear model formulation. Generalized linear models, often known by the acronym GLM, represent an important class of nonlinear regression models that have found extensive use in practice. In addition to the normal and Bernoulli distributions, these models include the binomial, Poisson and Gamma families as distributions for dependent variables.

Section 10.1 begins this chapter with a review of homogeneous GLM models, that is, GLM models that do not incorporate heterogeneity. The Section 10.2 example reinforces this review. Section 10.3 then describes marginal models and generalized estimating equations, a widely applied framework for incorporating heterogeneity. Then, Sections 10.4 and 10.5 allow for heterogeneity by modeling subject-specific quantities as random and fixed effects models, respectively. Section 10.6 ties together fixed and random effects under the umbrella of Bayesian inference.

10.1 Homogeneous models

This section introduces the generalized linear model (GLM); a more extensive treatment may be found in the classic work by McCullagh and Nelder (1989G). The GLM framework generalizes linear models in the following sense. Linear model theory provides a platform for choosing appropriate linear combinations of explanatory variables to predict a response. In Chapter 9, we saw how to use nonlinear functions of these linear combinations to provide better predictors, at least for responses with Bernoulli (binary) outcomes. In GLM, we widen the class of distributions to allow us to handle other types of non-normal outcomes. This broad class includes as special cases the normal, Bernoulli and Poisson distributions. As we will see, the normal distribution corresponds to linear models, the Bernoulli to the Chapter 9 binary response models. To motivate GLM, we focus on the Poisson distribution; this distribution allows us to readily model count data. Our treatment follows the structure introduced in Chapter 9; we first introduce the homogenous version of the GLM framework so that this section does not use any subject-specific parameters nor does it introduce terms to account for serial correlation. Subsequent sections will introduce techniques for handling heterogeneity in longitudinal and panel data.

The section begins with an introduction of the response distribution in Subsection 10.1.1 and then shows how to link the distribution’s parameters to regression variables in Subsection 10.1.2. Subsection 10.1.3 then describes estimation principles.

10.1.1 Linear exponential families of distributions

This chapter considers the linear exponential family of the form

\[ p(y, \theta, \phi) = \exp \left( \frac{y \theta - b(\theta)}{\phi} + S(y, \phi) \right). \tag{10.1} \]

Here, \( y \) is a dependent variable and \( \theta \) is the parameter of interest. The quantity \( \phi \) is a scale parameter that we often will assume is known. The term \( b(\theta) \) depends only on the parameter
\[ \theta, \text{ not the dependent variable. The statistic } S(y, \phi) \text{ is a function of the dependent variable and the scale parameter, not the parameter } \theta \]

The dependent variable \( y \) may be discrete, continuous or a mixture. Thus, \( p(.) \) may be interpreted to be a density or mass function, depending on the application. Table 10A.1 provides several examples, including the normal, binomial and Poisson distributions. To illustrate, consider a normal distribution with a probability density function of the form

\[
f(y, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi \sigma}} \exp \left( -\frac{(y - \mu)^2}{2\sigma^2} \right) = \exp \left( \frac{(y\mu - \mu^2 / 2)}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2} \ln(2\pi\sigma^2) \right).
\]

With the choices \( \theta = \mu, \phi = \sigma^2, b(\theta) = \theta^2/2 \) and \( S(y, \phi) = - (y^2 / (2\phi) + \ln(2\pi\phi))/2 \), we see that the normal probability density function can be expressed as in equation (10.1).

For the function in equation (10.1), some straightforward calculations show that

- \( \text{E} y = b'(\theta) \) and
- \( \text{Var} y = \phi b''(\theta) \).

For reference, these calculations appear in Appendix 10A.1. To illustrate, in the context of the normal distribution example above, it is easy to check that \( \text{E} y = b'(\theta) = \theta = \mu \) and \( \text{Var} y = \sigma^2 b''(\mu) = \sigma^2 \), as anticipated.

10.1.2 Link functions

In regression modeling situations, the distribution of \( y_{it} \) varies by observation through the subscripts “it.” Specifically, we allow the distribution’s parameters to vary by observation through the notation \( \theta_{it} \) and \( \phi_{it} \). For our applications, the variation of the scale parameter is due to known weight factors. Specifically, when the scale parameter varies by observation, it is according to \( \phi_{it} = \phi / w_{it} \), that is, a constant divided by a known weight \( w_{it} \). With the relation \( \text{Var} y_{it} = \phi_{it} b''(\theta) = \phi b''(\theta)/w_{it} \), we have that a larger weight implies a smaller variance, other things being equal.

In regression situations, we wish to understand the impact of a linear combination of explanatory variables on the distribution. In the GLM context, it is customary to call \( \eta_{it} = x_{it}' \beta \) the systematic component of the model. This systematic component is related to the mean through the expression

\[
\eta_{it} = g(\mu_{it}) .
\]

Here, \( g(.) \) is known as the link function. As we saw in the prior subsection, we can express the mean of \( y_{it} \) as \( \text{E} y_{it} = \mu_{it} = b'(\theta_{it}) \). Thus, equation (10.2) serves to “link” the systematic component to the parameter \( \theta_{it} \). It is possible to use the identity function for \( g(.) \) so that \( \mu_{it} = x_{it}' \beta \). Indeed, this is the usual case in linear regression. However, linear combinations of explanatory variables, \( x_{it}' \beta \), may vary between negative and positive infinity whereas means are often restricted to smaller range. For example, Poisson means vary between zero and infinity. The link function serves to map the domain of the mean function onto the whole real line.

Special case: Links for the Bernoulli distribution

Bernoulli means are probabilities and thus vary between zero and one. For this case, it is useful to choose a link function that maps the unit interval \((0,1)\) onto the whole real line. The following are three important examples of link functions for the Bernoulli distribution:

- Logit: \( g(\mu) = \text{logit}(\mu) = \ln (\mu / (1 - \mu)) \).
- Probit: \( g(\mu) = \Phi^{-1}(\mu) \), where \( \Phi^{-1} \) is the inverse of the standard normal distribution function.
- Complementary log-log: \( g(\mu) = \ln (-\ln(1 - \mu)) \).
This example demonstrates that there may be several link functions that are suitable for a particular distribution. To help with the selection, an intuitively appealing case occurs when the systematic component equals the parameter of interest \( (\eta = \theta) \). To see this, first recall that \( \eta = g(\mu) \) and \( \mu = b'(\theta) \), dropping the “\( i \)th” subscripts for the moment. Then, it is easy to see that if \( g^{-1} = b' \), then \( \eta = g(b'(\theta)) = \theta \). The choice of \( g \) that is the inverse of \( b' \) is called the canonical link.

Table 10.1 shows the mean function and corresponding canonical link for several important distributions.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Mean function ((b'(\theta)))</th>
<th>Canonical link ((g(\theta)))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>( \theta )</td>
<td>( \theta )</td>
</tr>
<tr>
<td>Bernoulli</td>
<td>( e^{\theta}/(1 + e^{\theta}) )</td>
<td>( \text{logit}(\theta) )</td>
</tr>
<tr>
<td>Poisson</td>
<td>( e^{\theta} )</td>
<td>( \ln \theta )</td>
</tr>
<tr>
<td>Gamma</td>
<td>(-1/\theta)</td>
<td>(-\theta)</td>
</tr>
</tbody>
</table>

10.1.3 Estimation

This section presents maximum likelihood, the customary form of estimation. To provide intuition, we begin with the simpler case of canonical links and then extend the results to more general links.

Maximum likelihood estimation for canonical links

From equation (10.1) and the independence among observations, the log-likelihood is

\[
\ln p(y) = \sum_\mathbf{y} \left\{ y_{it} \theta_{it} - b(\theta_{it}) \right\} + S(y_{it}, \phi_{it}).
\] (10.3)

Recall that for canonical links, we have equality between the distribution’s parameter and the systematic component, so that \( \theta_{it} = \eta_{it} = x_{it}' \beta \). Thus, the log-likelihood is

\[
\ln p(y) = \sum_\mathbf{y} \left\{ y_{it} x_{it}' \beta - b(x_{it}' \beta) \right\} + S(y_{it}, \phi_{it}).
\] (10.4)

Taking the partial derivative with respect to \( \beta \) yields the score function

\[
\frac{\partial}{\partial \beta} \ln p(y) = \sum_\mathbf{y} \left\{ x_{it} y_{it} - b'(x_{it}' \beta) \right\} / \phi_{it}.
\]

Because \( \mu_{it} = b'(\theta_{it}) = b'(x_{it}' \beta) \) and \( \phi_{it} = \phi / w_{it} \), we can solve for the maximum likelihood estimators of \( \beta \), \( \beta_{MLE} \), through the “normal equations”

\[
0 = \sum_\mathbf{y} w_{it} x_{it} (y_{it} - \mu_{it}).
\] (10.5)

One reason for the widespread use of GLM methods is that the maximum likelihood estimators can be computed quickly through a technique known as iterated reweighted least squares, described in Appendix C.3.

Note that, like ordinary linear regression normal equations, we do not need to consider estimation of the variance scale parameter \( \phi \) at this stage. That is, we can first compute \( \beta_{MLE} \) and then estimate \( \phi \).
Maximum likelihood estimation for general links

For general links, we no longer assume the relationship \( \theta_{it} = x_{it}' \beta \) but assume that \( \beta \) is related to \( \theta_{it} \) through the relation \( \mu_{it} = b'(\theta_{it}) \) and \( x_{it}' \beta = g(\mu_{it}) \). Using equation (10.3), we have that the \( j \)th element of the score function is

\[
\sum \frac{1}{\phi_{it}} \int \frac{\partial \ln p(y)}{\partial \beta_{j}} \int \left( \frac{\partial \theta_{it}}{\beta_{j}} - \frac{\partial y_{it}}{\partial \beta_{j}} \right) d\theta_{it}.
\]

because \( b'(\theta_{it}) = \mu_{it} \). Now, use the chain rule and the relation \( \text{Var} y_{it} = \phi_{it} b''(\theta_{it}) \) to get

\[
\frac{\partial \mu_{it}}{\partial \beta_{j}} = \frac{\partial b'(\theta_{it})}{\partial \beta_{j}} = b'(\theta_{it}) \frac{\partial \theta_{it}}{\partial \beta_{j}} = \text{Var} y_{it} \frac{\partial \theta_{it}}{\partial \beta_{j}}.
\]

Thus, we have

\[
\frac{\partial \theta_{it}}{\partial \beta_{j}} \frac{1}{\phi_{it}} = \frac{\partial \mu_{it}}{\partial \beta_{j}} \frac{1}{\text{Var} y_{it}}.
\]

This yields

\[
\frac{1}{\phi_{it}} \int \frac{\partial \ln p(y)}{\partial \beta_{j}} \int \left( \frac{\partial \theta_{it}}{\partial \beta_{j}} \left( \text{Var} y_{it} \right)^{-1} (y_{it} - \mu_{it}) \right) d\theta_{it},
\]

which is known as the generalized estimating equations form. This is the topic of Section 10.3.

Standard errors for regression coefficient estimators

As described in Appendix C.2, maximum likelihood estimators are consistent and asymptotically normally distributed under broad conditions. To illustrate, consider the maximum likelihood estimator determined from equation (10.4) using the canonical link. The asymptotic variance-covariance matrix of \( b_{MLE} \) is the inverse of the information matrix that, from equation (C.4), is

\[
\sum \frac{b'(x_{it}' b_{MLE})}{\phi_{it}} x_{it} x_{it}' \phi_{it}.
\]

The square root of the \( j \)th diagonal element of the inverse of this matrix yields the standard error for the \( j \)th row of \( b_{MLE} \), which we denote as \( se(b_{MLE}) \). Extensions to general links are similar.

Overdispersion

An important feature of several members of the linear exponential family of distributions, such as the Bernoulli and the Poisson distributions, is that the variance is determined by the mean. In contrast, the normal distribution has a separate parameter for the variance, or dispersion. When fitting models to data with binary or count dependent variables, it is common to observe that the variance exceeds that anticipated by the fit of the mean parameters. This phenomenon is known as overdispersion. Several alternative probabilistic models may be available to explain this phenomenon, depending on the application at hand. See Section 10.3 for an example and McCullagh and Nelder (1989) for a more detailed inventory.

Although arriving at a satisfactory probabilistic model is the most desirable route, in many situations analysts are content to postulate an approximate model through the relation

\[
\text{Var} y_{it} = \sigma^2 \frac{b''(x_{it}' \beta)}{w_{it}}.
\]

The scale parameter \( \phi \) is specified through the choice of the distribution whereas the scale parameter \( \sigma^2 \) allows for extra variability. For example, Table 10A.1 shows that by specifying either the Bernoulli or Poisson distribution, we have \( \phi = 1 \). Although the scale parameter \( \sigma^2 \)
allows for extra variability, it may also accommodate situations in which the variability is smaller than specified by the distributional form (although this situation is less common). Finally, note that for some distributions such as the normal distribution, the extra term is already incorporated in the \( \phi \) parameter and thus serves no useful purpose.

When the additional scale parameter \( \sigma^2 \) is included, it is customary to estimate it by Pearson’s chi-square statistic divided by the error degrees of freedom. That is,

\[
\sigma^2 = \frac{1}{N - K} \sum_{i} w_i \left( \frac{y_{it} - b'(x_{it}'b_{MLE})}{\phi b''(x_{it}'b_{MLE})} \right)^2.
\]

### 10.2 Example: Tort filings

There is a widespread belief that, in the United States, contentious parties have become increasingly willing to go to the judicial system to settle disputes. This is particularly true when one party is from the insurance industry, an industry designed to spread risk among individuals. Litigation in the insurance industry arises from two types of disagreement among parties, breach of faith and tort. A breach of faith is a failure by a party to the contract to perform according to its terms. A tort action is a civil wrong, other than breach of contract, for which the court will provide a remedy in the form of action for damages. A civil wrong may include malice, wantonness oppression or capricious behavior by a party. Generally, large damages can be collected for tort actions because the award may be large enough to “sting” the guilty party. Because large insurance companies are viewed as having “deep pockets,” these awards can be quite large.

The response that we consider is the number of filings, NUMFILE, of tort actions against insurance companies \((y)\). For each of six years, 1984-1989, the data were obtained from 19 states with two observations unavailable for a total of \(N = 112\) observations. The issue is to understand ways in which state legal, economic and demographic characteristics affect the number of filings. Table 10.2 describes these characteristics. More extensive motivation is provided in Lee (1994O) and Lee, Browne and Schmit (1994O).

<table>
<thead>
<tr>
<th>Table 10.2. State Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent Variable</strong></td>
</tr>
<tr>
<td>NUMFILE</td>
</tr>
<tr>
<td><strong>State Legal Characteristics</strong></td>
</tr>
<tr>
<td>JSLIAB</td>
</tr>
<tr>
<td>COLLRULE</td>
</tr>
<tr>
<td>CAPS</td>
</tr>
<tr>
<td>PUNITIVE</td>
</tr>
<tr>
<td><strong>State Economic and Demographic Characteristics</strong></td>
</tr>
<tr>
<td>POP</td>
</tr>
<tr>
<td>POPLAWYR</td>
</tr>
<tr>
<td>VEHCMILE</td>
</tr>
<tr>
<td>POPDENSY</td>
</tr>
<tr>
<td>WCMPMAX</td>
</tr>
<tr>
<td>URBAN</td>
</tr>
<tr>
<td>UNEMPLOY</td>
</tr>
</tbody>
</table>

*Source: An Empirical Study of the Effects of Tort Reforms on the Rate of Tort Filings, unpublished Ph.D. Dissertation, Han-Duck Lee, University of Wisconsin (1994O).*

Tables 10.3 and 10.4 summarize the state legal, economic and demographic characteristics. To illustrate, in Table 10.3 we see that 23.2 percent of the 112 state-year observations were under limits (caps) on non-economic reform. Those observations not under
limits on non-economic reforms had a larger average number of filings. The correlations in Table 10.4 show that several of the economic and demographic variables appear to be related to the number of filings. In particular, we note that the number of filings is highly related to the state population.

Table 10.3 Averages with Explanatory Indicator Variables

<table>
<thead>
<tr>
<th>Explanatory Variable</th>
<th>JSLIAB</th>
<th>COLLRULE</th>
<th>CAPS</th>
<th>PUNITIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Explanatory Variable</td>
<td>0.491</td>
<td>0.304</td>
<td>0.232</td>
<td>0.321</td>
</tr>
<tr>
<td>Average NUMFILE When Explanatory Variable = 0</td>
<td>15,530</td>
<td>20,727</td>
<td>24,682</td>
<td>17,693</td>
</tr>
<tr>
<td>When Explanatory Variable = 1</td>
<td>25,967</td>
<td>20,027</td>
<td>6,727</td>
<td>26,469</td>
</tr>
</tbody>
</table>

Table 10.4. Summary Statistics for Other Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Median</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Standard deviation</th>
<th>Correlation with NUMFILE</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUMFILE</td>
<td>20514</td>
<td>9085</td>
<td>512</td>
<td>137455</td>
<td>29039</td>
<td>1.0</td>
</tr>
<tr>
<td>POP</td>
<td>6.7</td>
<td>3.4</td>
<td>0.5</td>
<td>29.0</td>
<td>7.2</td>
<td>0.902</td>
</tr>
<tr>
<td>POPLAWYR</td>
<td>377.3</td>
<td>382.5</td>
<td>211.0</td>
<td>537.0</td>
<td>75.7</td>
<td>-0.378</td>
</tr>
<tr>
<td>VEHCMILE</td>
<td>654.8</td>
<td>510.5</td>
<td>63.0</td>
<td>1899.0</td>
<td>515.4</td>
<td>0.518</td>
</tr>
<tr>
<td>POPDENS</td>
<td>168.2</td>
<td>63.9</td>
<td>0.9</td>
<td>1043.0</td>
<td>243.9</td>
<td>0.368</td>
</tr>
<tr>
<td>WCMPMAX</td>
<td>350.0</td>
<td>319.0</td>
<td>203.0</td>
<td>1140.0</td>
<td>151.7</td>
<td>-0.265</td>
</tr>
<tr>
<td>URBAN</td>
<td>69.4</td>
<td>78.9</td>
<td>18.9</td>
<td>100.0</td>
<td>24.8</td>
<td>0.550</td>
</tr>
<tr>
<td>UNEMPLOY</td>
<td>6.2</td>
<td>6.0</td>
<td>2.6</td>
<td>10.8</td>
<td>1.6</td>
<td>0.008</td>
</tr>
</tbody>
</table>

Figure 10.1 Multiple time series plot of NUMFILE.
Figure 10.1 is a multiple time series plot of the number of filings. The state with the largest number of filings is California. This plot shows the state level heterogeneity of filings.

When data represent counts such as the number of tort filings, it is customary to consider the Poisson model to represent the distribution of responses. From mathematical statistics theory, it is known that the sums of independent Poisson random variables have a Poisson distribution. Thus, if \( \lambda_1, \ldots, \lambda_n \) represent independent Poisson random variables, each with parameter \( E \lambda_i = \mu_i \), then \( \lambda_1 + \ldots + \lambda_n \) is Poisson distributed with parameter \( \mu = (\mu_1 + \ldots + \mu_n)/n \). We assume that \( y_{it} \) is Poisson distribution with parameter \( \text{POP}_{it} \exp(x_{it}' \beta) \), where \( \text{POP}_{it} \) is the population of the \( i \)th state at time \( t \). To account for this “known” relationship with population, we assume that the (natural) logarithmic population is one of the explanatory variables, yet has a known regression coefficient equal to one. In GLM terminology, such as variable is known as an offset. Thus, our Poisson parameter for \( y_{it} \) is

\[
\text{exp}\left[\ln \text{POP}_{it} + x_{it,1} \beta_1 + \cdots + x_{it,K} \beta_K \right] = \exp\left(\ln \text{POP}_{it} + x_{it}' \beta\right) = \text{POP}_{it} \exp(x_{it}' \beta).
\]

An alternative approach is to use the average number of tort filings as the response and assume approximate normality. This was the approach taken by Lee et al. (1994O); the reader has an opportunity to practice this approach in Exercises 2.18 and 3.12. Note that in the Poisson model above the expectation of the average response is \( E(y_{it} / \text{POP}_{it}) = \exp(x_{it}' \beta) \) whereas the variance is \( \text{Var}(y_{it} / \text{POP}_{it}) = \exp(x_{it}' \beta)/\text{POP}_{it} \). Thus, to make these two approaches compatible, one must use weighted regression, using estimated reciprocal variances as the weights.

Table 10.5 summarizes the fit of three Poisson models. With the basic homogeneous Poisson model, all explanatory variables turn out to be statistically significant, as evidenced by the small \( p \)-values. However, the Poisson model assumes that the variance equals the mean; this is often a restrictive assumption for empirical work. Thus, Table 10.5 also summarizes a homogenous Poisson model with an estimated scale parameter, to account for potential overdispersion. Table 10.5 emphasizes that although the regression coefficient estimates do not change with the introduction of the scale parameter, estimated standard errors and thus \( p \)-values do change. Many variables, such as CAPS, turn out to be statistically insignificant predictors of the number of filings when a more flexible model for the variance is introduced. Subsequent sections will introduce models of the state level heterogeneity. Although not as important for this data set, it is still easy to examine temporal heterogeneity in this context through year binary (dummy) variables. The goodness of fit statistics, the deviance and Pearson chi-square, favor including the time categorical variable (see Appendix C.8 for definitions of these goodness of fit statistics).
Table 10.5 Tort Filings Model Coefficient Estimates
Based on \( N = 112 \) observations from \( n = 19 \) states and \( T = 6 \) years.

Logarithmic population is used as an offset.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Parameter estimate</th>
<th>p-values</th>
<th>Parameter estimate</th>
<th>p-values</th>
<th>Parameter estimate</th>
<th>p-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-7.943</td>
<td>&lt;.0001</td>
<td>-7.943</td>
<td>&lt;.0001</td>
<td>2.123</td>
<td>0.0004</td>
</tr>
<tr>
<td>POPLAWYR/1000</td>
<td>2.163</td>
<td>&lt;.0001</td>
<td>2.163</td>
<td>0.0002</td>
<td>0.856</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>VEHCMILE/1000</td>
<td>0.862</td>
<td>&lt;.0001</td>
<td>0.862</td>
<td>&lt;.0001</td>
<td>0.384</td>
<td>0.0067</td>
</tr>
<tr>
<td>POPDENSY/1000</td>
<td>0.392</td>
<td>&lt;.0001</td>
<td>0.392</td>
<td>0.0038</td>
<td>-0.826</td>
<td>0.1523</td>
</tr>
<tr>
<td>WCMPMAX/1000</td>
<td>-0.802</td>
<td>&lt;.0001</td>
<td>-0.802</td>
<td>0.1226</td>
<td>-0.826</td>
<td>0.1523</td>
</tr>
<tr>
<td>URBAN/1000</td>
<td>0.892</td>
<td>&lt;.0001</td>
<td>0.892</td>
<td>0.8187</td>
<td>0.977</td>
<td>0.8059</td>
</tr>
<tr>
<td>UNEMPLOY</td>
<td>0.087</td>
<td>&lt;.0001</td>
<td>0.087</td>
<td>0.0005</td>
<td>0.086</td>
<td>0.0024</td>
</tr>
<tr>
<td>JSLIAB</td>
<td>0.177</td>
<td>&lt;.0001</td>
<td>0.177</td>
<td>0.0292</td>
<td>0.130</td>
<td>0.2705</td>
</tr>
<tr>
<td>COLLRULE</td>
<td>-0.030</td>
<td>&lt;.0001</td>
<td>-0.030</td>
<td>0.7444</td>
<td>-0.023</td>
<td>0.8053</td>
</tr>
<tr>
<td>CAPS</td>
<td>-0.032</td>
<td>&lt;.0001</td>
<td>-0.032</td>
<td>0.7457</td>
<td>-0.056</td>
<td>0.6008</td>
</tr>
<tr>
<td>PUNITIVE</td>
<td>0.030</td>
<td>&lt;.0001</td>
<td>0.030</td>
<td>0.6623</td>
<td>0.053</td>
<td>0.4986</td>
</tr>
<tr>
<td>Scale</td>
<td>1.000</td>
<td></td>
<td>35.857</td>
<td></td>
<td>36.383</td>
<td></td>
</tr>
<tr>
<td>Deviance</td>
<td>118,309.0</td>
<td></td>
<td>118,309.0</td>
<td></td>
<td>115,496.4</td>
<td></td>
</tr>
<tr>
<td>Pearson Chi-Square</td>
<td>129,855.7</td>
<td></td>
<td>129,855.7</td>
<td></td>
<td>127,073.9</td>
<td></td>
</tr>
</tbody>
</table>

10.3 Marginal models and GEE

Marginal models underpin a widely applied framework for handling heterogeneity in longitudinal data. As introduced in Section 9.4, marginal models are semiparametric. Specifically, we model the first and second moments as functions of unknown parameters without specifying a full parametric distribution for the responses. For these models, parameters can be estimated using extensions of the method of moment estimation technique known as generalized estimating equations, or GEE.

Marginal models

To re-introduce marginal models, again use \( \beta \) to be a \( K \times 1 \) vector of parameters associated with the explanatory variables \( x_{it} \). Further, let \( \tau \) be a \( q \times 1 \) vector of variance component parameters. Assume that the mean of each response is

\[ E(y_{it}) = \mu_{it} = \mu(a(\beta, \tau)), \]

where \( \mu_{it} \) is a known function. Further, denote the variance of each response as

\[ \text{Var}(y_{it}) = v_{it} = v(a(\beta, \tau)), \]

where \( v_{it} \) is a known function.

To complete the second moment specification, we need assumptions on the covariances among responses. We assume that observations among different subjects are independent, and thus have zero covariance. For observations within a subject, we assume that the correlation between two observations within the same subject is a known function. Specifically, we will denote

\[ \text{corr}(y_{ir}, y_{is}) = \rho_{rs}(\mu_{ir}, \mu_{is}, \tau), \]
where \( \rho_{rs}(.) \) is a known function. As in Chapter 3, marginal models use correlations among observations within a subject to represent heterogeneity, that is, the tendency for observations from the same subject to be related.

**Special case – Generalized linear model with canonical links**

As we saw in Section 10.1, in the context of the generalized linear model with canonical links, we have \( \mu_{it} = E y_{it} = b'(x_{it}' \beta) \) and \( v_{it} = \text{Var } y_{it} = \phi b''(x_{it}' \beta) / w_{it} \). Here, \( b(.) \) is a known function that depends on the choice of the distribution of responses. There is only one variance component, so \( q = 1 \) and \( \tau = \phi \). For the cases of independent observations among subjects discussed in Section 10.1, the correlation function \( \rho_{rs} \) is 1 if \( r=s \) and 0 otherwise. For this example, note that the mean function depends on \( \beta \) but does not depend on the variance component \( \tau \).

**Special case – Error components model**

For the error components model introduced in Section 3.1, we may use a generalized linear model with a normal distribution for responses, so that \( b(\theta) = \theta^2 / 2 \). Thus, from the above example, we have \( \mu_{it} = E y_{it} = x_{it}' \beta \) and \( v_{it} = \text{Var } y_{it} = \phi = \sigma_a^2 + \sigma_e^2 \). Unlike the above example, observations within subjects are not independent but have an exchangeable correlation structure. Specifically, in Section 3.1 we saw that, for \( r \neq s \),

\[
\text{corr}(y_{ir}, y_{is}) = \rho = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}.
\]

With this specification, the vector of variance components is \( \tau = (\phi, \sigma_a^2)' \), where \( \phi = \sigma_a^2 + \sigma_e^2 \).

In Section 10.1, the parametric form of the first two moments was a consequence of the assumption of exponential family distribution. With marginal models, the parametric form is now a basic assumption. We now give matrix notation to express these assumptions more compactly. To this end, define \( \mu_i(\beta, \tau) = \mu_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{iT_i})' \) to be the vector of means for the \( i \)th subject. Let \( V_i \) be the \( T_i \times T_i \) variance covariance matrix, \( \text{Var } y_i \), where the \( rs \)th element of \( V_i \) is given by

\[
\text{Cov}(y_{ir}, y_{is}) = \text{corr}(y_{ir}, y_{is}) \sqrt{v_{ir} v_{is}}.
\]

As in Section 9.4, for estimation purposes we will also require the \( K \times T_i \) gradient matrix,

\[
G_{\mu}(\beta, \tau) = \left( \frac{\partial \mu_{i1}}{\partial \beta} \quad \frac{\partial \mu_{i2}}{\partial \beta} \quad \ldots \quad \frac{\partial \mu_{iT_i}}{\partial \beta} \right).
\]

**Generalized estimating equations**

Marginal models provide sufficient structure to allow for a method of moments estimation procedure called generalized estimating equations (GEE), a type of generalized method of moments (GMM) estimation. As with generalized least squares, we first describe this estimation method assuming the variance component parameters are known. We then describe extensions to incorporate variance component estimation. The general GEE procedure is described in Appendix C.6; recall that an introduction was provided in Section 9.4. Assuming that the variance components in \( \tau \) are known, the GEE estimator of \( \beta \) is the solution of
\[0_K = \sum_{i=1}^{n} G_{\mu}(b) V_i^{-1}(b)(y_i - \mu_i(b)). \] (10.8)

These are the \textit{generalized estimating equations}. We will denote this solution as \(b_{EE}\). Under mild regularity conditions, this estimator is consistent and asymptotically normal with variance-covariance matrix

\[
\text{Var } b_{EE} = \left( \sum_{i=1}^{n} G_{\mu}(b) V_i^{-1}(b) G_{\mu}^T(b) \right)^{-1}. \tag{10.9}
\]

\textbf{Special case – Generalized linear model with canonical links - Continued}

Because \(\mu_i = b'(x_i' \beta)\), it is easy to see that that \(\frac{\partial \mu_i}{\partial \beta} = x_{ii} b'(x_{ii}' \beta) = w_{ii} x_{ii} v_{ii}/\phi\). Thus, \(G_{\mu}(\beta, \tau) = \frac{1}{\phi} \left( w_{i1} x_{i1} v_{i1} \cdots w_{iT_i} x_{iT_i} v_{iT_i} \right)\) is our \(K \times T_i\) matrix of derivatives. Assuming independence among observations within a subject we have \(V_i = \text{diag}(v_{i1}, \ldots, v_{iT_i})\). Thus, we may express the generalized estimating equations as

\[0_K = \sum_{i=1}^{n} G_{\mu}(b) V_i^{-1}(b)(y_i - \mu_i(b)) - \frac{1}{\phi} \sum_{i=1}^{n} \sum_{t=1}^{T_i} w_{it} x_{it} (y_{it} - \mu_{it}(b)). \]

This yields the same solution as the maximum likelihood “normal equations” in equation (10.5). Thus, the GEE estimators are equal to the maximum likelihood estimators for this special case. Note that this solution does not depend on knowledge of the variance component, \(\phi\).

\textbf{GEEs with unknown variance components}

To determine GEE estimates, the first task is to determine an initial estimator of \(\beta\), say \(b_{0,EE}\). To illustrate, one might use the GLM model with independence among observations, as above, to get an initial estimator. Next, we use the initial estimator \(b_{0,EE}\) to compute residuals and determine an initial estimator of the variance components, say \(\tau_{0,EE}\). Then, at the \((n+1)\)st stage, recursively:

1. Use \(\tau_{n,EE}\) and the solution of the equation

\[0_K = \sum_{i=1}^{n} G_{\mu}(b, \tau_{n,EE}) V_i^{-1}(b, \tau_{n,EE})(y_i - \mu_i(b, \tau_{n,EE}))\]

   to determine an updated estimator of \(\beta\), say \(b_{n+1,EE}\).

2. Use the residuals \(\{y_{it} - \mu_i(b_{n+1,EE}, \tau_{n,EE})\}\) to determine an updated estimator of \(\tau\), say \(\tau_{n+1,EE}\).

3. Repeat steps 1 and 2 until convergence.

Alternatively, for the first stage, we may update estimators using a one-step procedure such as a Newton-Raphson iteration. As another example, the statistical package SAS uses a Fisher scoring type update of the form:

\[b_{n+1,EE} = b_{n,EE} + \left( \sum_{i=1}^{n} G_{\mu}(b_{n,EE}) V_i^{-1}(b_{n,EE}) G_{\mu}^T(b_{n,EE}) \right)^{-1} \left( \sum_{i=1}^{n} G_{\mu}(b_{n,EE}) V_i^{-1}(b_{n,EE})(y_i - \mu_i(b_{n,EE})) \right)\]
For GEEs in more complex problems with unknown variance components, Prentice (1988B) suggests using a second estimating equation of the form:

$$\sum_i \left( \frac{\partial \mathbf{E} y_i^*}{\partial \tau} \right) \mathbf{W}_i^{-1} (y_i^* - \mathbf{E} y_i^*) = 0.$$  

Here, $y_i^*$ is a vector of squares and cross-products of observations within a subject of the form:

$$y_i^* = \left( y_{i1}^2 \quad y_{i1} y_{i2} \quad \cdots \quad y_{i1} y_{iT_i} \quad \cdots \quad y_{iT_i}^2 \quad y_{i2}^2 \quad \cdots \quad y_{iT_i}^2 \right).$$

For the variance of $y_i^*$, Diggle et al. 2002S) suggest using the identity matrix for $\mathbf{W}_i$ with most discrete data. However, for binary responses, they note that the last $T_i$ observations are redundant because $y_{iT_i} = y_{iT_i}^2$. These should be ignored; Diggle et al. recommend using $\mathbf{W}_i = \text{diag}(\text{Var}(y_{i1}^2) \quad \cdots \quad \text{Var}(y_{iT_i}^2)).$

### Robust estimation of standard errors

As discussed in Sections 2.5.3 and 3.4 for linear models, one must be concerned with unsuspected serial correlation and heteroscedasticity. This is particularly true with marginal models, where the specification is based only on the first two moments and not a full probability distribution. Because the specification of the correlation structure may be suspect, these are often called working correlations. Standard errors that rely on the correlation structure specification from the relation in equation (10.9) are known as model-based standard errors. In contrast, empirical standard errors may be calculated using the following estimator of the asymptotic variance of $\mathbf{b}_{EE}$:

$$\left( \sum_{i=1}^{n} \mathbf{G}_\mu^{-1} \mathbf{V}_i^{-1} \mathbf{G}_\mu' \right)^{-1} \left( \sum_{i=1}^{n} \mathbf{G}_\mu^{-1} \mathbf{V}_i^{-1} (y_i - \mu_i)(y_i - \mu_i)' \mathbf{V}_i^{-1} \mathbf{G}_\mu' \right) \left( \sum_{i=1}^{n} \mathbf{G}_\mu^{-1} \mathbf{G}_\mu' \right)^{-1}.$$

Specifically, the standard error of the $j$th component of $\mathbf{b}_{EE}$, $\text{se}(\mathbf{b}_{EE,j})$, is defined to be the square root of the $j$th diagonal element of the variance-covariance matrix in the above display.

### Example: Tort filings – continued

To illustrate marginal models and GEE estimators, we return to the Section 10.2 Tort filing example. The fitted models, which appear in Table 10.6, were estimated using the SAS procedure GENMOD. Assuming an independent working correlation, we arrive at the same parameter estimators as in Table 10.5, under the homogenous Poisson model with an estimated scale parameter. Further, although not displayed, the fits provide the same model-based standard errors. Compared to the empirical standard errors, we see that almost all empirical standard errors are larger than the corresponding model-based standard errors (the exception is CAPS).

To test the robustness of this model fit, we fit the same model with an autoregressive model of order one, $AR(1)$, working correlation. This model fit also appears in Table 10.6. Because we use the working correlation as a weight matrix, many of the coefficients have different values than the corresponding fit with the independent working correlation matrix. Note that the asterisk indicates when a estimate exceeds twice the empirical standard error; this gives a rough idea of the statistical significance of the variable. Under the $AR(1)$ working correlation, variables that are statistically significant using empirical standard errors are also statistically significant using model-based standard errors. (Again, CAPS is a borderline exception.) Comparing the independent to the $AR(1)$ working correlation results, we see that VEHCMILE, UNEMPLOY and JSLIAB are consistently statistically significant whereas POPLAWYR, POPDENS, WCMPMAX, URBAN, COLLRULE and PUNITIVE are consistently statistically insignificant. However, the judgment of statistical significance depends on the model selection for CAPS.
### Table 10.6 Comparison of GEE Estimators.

All models use an estimated scale parameter. Logarithmic population is used as an offset.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Independent Working Correlation</th>
<th>AR(1) Working Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>Empirical Standard Error</td>
</tr>
<tr>
<td>Intercept</td>
<td>-7.943*</td>
<td>0.612</td>
</tr>
<tr>
<td>POPLAWYR/1000</td>
<td>2.163</td>
<td>1.101</td>
</tr>
<tr>
<td>VEHCMILE/1000</td>
<td>0.862*</td>
<td>0.265</td>
</tr>
<tr>
<td>POPDENSY/1000</td>
<td>0.392*</td>
<td>0.175</td>
</tr>
<tr>
<td>WCMPMAX/1000</td>
<td>-0.802</td>
<td>0.895</td>
</tr>
<tr>
<td>URBAN/1000</td>
<td>0.892</td>
<td>5.367</td>
</tr>
<tr>
<td>UNEMPLOY</td>
<td>0.087*</td>
<td>0.042</td>
</tr>
<tr>
<td>JSLIAB</td>
<td>0.177*</td>
<td>0.089</td>
</tr>
<tr>
<td>COLLRULE</td>
<td>-0.030</td>
<td>0.120</td>
</tr>
<tr>
<td>CAPS</td>
<td>-0.032</td>
<td>0.098</td>
</tr>
<tr>
<td>PUNITIVE</td>
<td>0.030</td>
<td>0.125</td>
</tr>
<tr>
<td>Scale</td>
<td>35.857</td>
<td></td>
</tr>
<tr>
<td>AR(1) Coefficient</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The asterisk (*) indicates that the estimate is more than twice the empirical standard error, in absolute value.

---

### 10.4 Random effects models

An important method for accounting for heterogeneity in longitudinal and panel data is through a random effects formulation. As in Sections 3.1 and 3.3.1 for linear models, this model is easy to introduce and interpret in the following hierarchical fashion:

- **Stage 1.** Subject effects \( \{ \alpha_i \} \) are a random sample from a distribution that is known up to a vector of parameters.
- **Stage 2.** Conditional on \( \{ \alpha_i \} \), the responses \( \{ y_{i1}, y_{i2}, \ldots, y_{iT} \} \) are a random sample from a GLM with systematic component \( \eta_i = z_{it}' \alpha_i + x_{it}' \beta \).

Additional motivation and sampling issues regarding random effects were introduced in Chapter 3 for linear models; they also pertain to generalized linear models.

Thus, the random effects model introduced above is a generalization of:

- The Chapter 3 linear random effects model. Here, we used a normal distribution for the GLM component.
- The Section 9.2 binary dependent variables model with random effects. Here, we used a Bernoulli distribution for the GLM component. (In Section 9.2, we also focused on the case \( z_{it} = 1 \).)

By assuming a distribution for the heterogeneity components \( \alpha_i \), there is typically a much smaller number of parameters to be estimated when compared to the fixed effects panel data models that will be described in Section 10.5. The maximum likelihood method of estimation is available although is computationally intensive. To use this method, the customary practice is to assume normally distributed random effects. Other estimation methods are also available in the literature. For example, the EM (for expectation-maximization) algorithm for estimation is described in Diggle et al. (2002S) and McCulloch and Searle (2001G).
Random effects likelihood
To develop the likelihood, we first note that from the second sampling stage, conditional on \( \alpha \), the likelihood for the \( i \)th subject at the \( t \)th observation is

\[
p(y_{it} \mid \beta, \alpha) = \exp \left( \frac{y_{it} \theta_{it} - b(\theta_{it})}{\phi} + S(y_{it}, \phi) \right),
\]

where \( b' (\theta_{it}) = E (y_{it} \mid \alpha) \) and \( \eta_{it} = z_{it}' \alpha + x_{it}' \beta = g(E (y_{it} \mid \alpha)) \). Because of the independence among responses within a subject conditional on \( \alpha \), the conditional likelihood for the \( i \)th subject is

\[
p(y_i \mid \beta, \alpha) = \exp \left( \sum_t \left( \frac{y_{it} \theta_{it} - b(\theta_{it})}{\phi} + S(y_{it}, \phi) \right) \right),
\]

Taking expectations over \( \alpha \) yields the (unconditional) likelihood. To see this explicitly, we use the canonical link so that \( \theta_{it} = \eta_{it} \). Thus, the (unconditional) likelihood for the \( i \)th subject is

\[
p(y_i \mid \beta) = \exp \left( \sum_t \exp \left( \sum_t \left( \frac{y_{it} (z_{it}' \alpha + x_{it}' \beta) - b(z_{it}' \alpha + x_{it}' \beta)}{\phi} \right) \right) \right) dF_{\alpha}(\alpha). \tag{10.11}
\]

Here, \( F_{\alpha}(\cdot) \) represents the distribution of \( \alpha \), which we will assume to be multivariate normal with mean zero and variance-covariance matrix \( \Sigma_{\alpha} \). Consistent with the notation in Chapter 3, let \( \tau \) denote the vector of parameters associated with scaling in stage 2, \( \phi \), and the stage 1 parameters in the matrix \( D \). With this notation, we may write the total log-likelihood as

\[
\ln p(y, \beta, \tau) = \sum_i \ln p(y_i, \beta).
\]

From equation (10.11), we see that evaluating the log-likelihood requires numerical integration and thus is more difficult to compute than likelihoods for homogeneous models.

Special case – Poisson distribution
To illustrate, assume \( q = 1 \) and \( z_{it} = 1 \) so that only intercepts \( \alpha_i \) vary by subject. Assuming a Poisson distribution for the conditional responses, we have \( \phi = 1 \), \( b(a) = e^a \), and \( S(y, \phi) = -\ln(y!) \). Thus, from equation (10.11), the log-likelihood for the \( i \)th subject is

\[
\ln p(y_i, \beta) = -\sum_t \ln(y_{it}!) + \ln \left( \int \exp \left( \sum_t \left( y_{it} (a + x_{it}' \beta) - \exp(a + x_{it}' \beta) \right) \right) f_{\alpha}(a) da \right)
\]

\[
= -\sum_t \ln(y_{it}!) + \sum_t y_{it} x_{it}' \beta + \ln \left( \int \exp \left( \sum_t \left( y_{it} a - \exp(a + x_{it}' \beta) \right) \right) f_{\alpha}(a) da \right),
\]

where \( f_{\alpha}(\cdot) \) is the probability density function of \( \alpha_i \). As before, evaluating and maximizing the log-likelihood requires numerical integration.
Serial correlation and overdispersion

The random effects GLM model introduces terms that account for heterogeneity in the data. However, the model also introduces certain forms of serial correlation and overdispersion that may or may not be evident in the data. Recall that in Chapter 3 we saw that permitting subject-specific effects, \( \alpha_i \), to be random induced serial correlation in the responses \( \{y_{i1}, y_{i2}, \ldots, y_{iT}\} \). This is also true for the nonlinear GLM models, as shown in the following example.

Special case – Poisson distribution – continued

To illustrate, we use the assumptions of Example 10.1 and recall that, for a canonical link, we have \( E(y_{it} | \alpha_i) = b'(\theta_{it}) = b'(\eta_{it}) = b'(\alpha_i + x_{it}' \beta) \). For the Poisson distribution, we have \( b'(a) = e^a \), so that

\[
\mu_{it} = E(y_{it}) = E(E(y_{it} | \alpha_i)) = E b'(\alpha_i + x_{it}' \beta) = \exp(x_{it}' \beta) E e^{\alpha_i}.
\]

Here, we have dropped the subscript on \( \alpha_i \) because the distribution is identical over \( i \).

To see the serial correlation, we examine the covariance between two observations, for example, \( y_{i1} \) and \( y_{i2} \). By the conditional independence, we have

\[
\text{Cov}(y_{i1}, y_{i2}) = E(\text{Cov}(y_{i1}, y_{i2}) | \alpha_i) + \text{Cov}(E(y_{i1} | \alpha_i), E(y_{i2} | \alpha_i))
\]

\[
= \text{Cov}(b'(\alpha_i + x_{i1}' \beta), b'(\alpha_i + x_{i2}' \beta)) = \text{Cov}(e^{\alpha_i} \exp(x_{i1}' \beta), e^{\alpha_i} \exp(x_{i2}' \beta))
\]

\[
= \exp((x_{i1} + x_{i2})' \beta) \text{Var } e^{\alpha_i}.
\]

This covariance is always nonnegative, indicating that we can anticipate positive serial correlation using this model.

Similar calculations show that

\[
\text{Var } y_{it} = E(\text{Var } (y_{it} | \alpha_i)) + \text{Var}(E(y_{it} | \alpha_i))
\]

\[
= E \phi b''(\alpha_i + x_{it}' \beta) + \text{Var}(b'(\alpha_i + x_{it}' \beta)) = E e^{\alpha_i} \exp(x_{it}' \beta) + \text{Var}(\exp(\alpha_i + x_{it}' \beta))
\]

\[
= \mu_{it} + \exp(2 x_{it}' \beta) \text{Var } e^{\alpha_i}.
\]

Thus, the variance always exceeds the mean. Compared to the usual Poisson models that require equality between the mean and the variance, the random effects specification induces a larger variance. This is a specific example of the phenomenon known as overdispersion.

Example: Tort filings – continued

To illustrate the random effects estimators, we return to the Section 10.2 Tort filing example. Table 10.7 summarizes a random effects model that was fit using the SAS statistical procedure NLMIXED. For comparison, the Table 10.5 fits from the homogeneous Poisson model with an estimated scale parameter are included in Table 10.7. The random effects model assumes a conditional Poisson distributed response, with a scalar homogeneity parameter that has a normal distribution.
### Table 10.7 Tort Filings Model Coefficient Estimates – Random Effects

Logarithmic population is used as an offset.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Homogeneous Model with estimated scale parameter</th>
<th>Random Effects Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Parameter estimate</td>
<td>p-values</td>
</tr>
<tr>
<td>Intercept</td>
<td>-7.943</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>POPLAWYR/1000</td>
<td>2.163</td>
<td>0.0002</td>
</tr>
<tr>
<td>VEHCMILE/1000</td>
<td>0.862</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>POPDENSY/1000</td>
<td>0.392</td>
<td>0.0038</td>
</tr>
<tr>
<td>WCMPMAX/1000</td>
<td>-0.802</td>
<td>0.1226</td>
</tr>
<tr>
<td>URBAN/1000</td>
<td>0.892</td>
<td>0.8187</td>
</tr>
<tr>
<td>UNEMPLOY</td>
<td>0.087</td>
<td>0.0005</td>
</tr>
<tr>
<td>JSLIAB</td>
<td>0.177</td>
<td>0.0292</td>
</tr>
<tr>
<td>COLLRULE</td>
<td>-0.030</td>
<td>0.7444</td>
</tr>
<tr>
<td>CAPS</td>
<td>-0.032</td>
<td>0.7457</td>
</tr>
<tr>
<td>PUNITIVE</td>
<td>0.030</td>
<td>0.6623</td>
</tr>
<tr>
<td>State Variance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-2 Log Likelihood</td>
<td></td>
<td>119,576</td>
</tr>
</tbody>
</table>

### Computational considerations

As is evident from equation (10.11), maximum likelihood estimation of regression coefficients requires one or more \( q \)-dimensional numerical integrations, for each subject and each iteration of an optimization routine. As we have seen in our Chapter 9 and 10 examples, this computational complexity is manageable for random intercept models where \( q = 1 \). According to McCulloch and Searle (2001G), this direct method is also available for applications with \( q = 2 \) or \( 3 \); however, for higher-order models, such as with crossed-random effects, alternative approaches are necessary.

We have already mentioned the **EM** (expectation-maximization) algorithm in Chapter 9 as one alternative, see McCulloch and Searle (2001G) or Diggle et al. (2002S) for more details. Another alternative is to use simulation techniques. McCulloch and Searle (2001G) summarize a Monte Carlo Newton-Raphson approach for approximating the score function, a simulated maximum likelihood approach for approximating the integrated likelihood and a stochastic approximation method for a more efficient and sequential approach of simulation.

The most widely used set of alternatives are based on Taylor-series expansions, generally about the link function or the integrated likelihood. There are several justifications for this set of alternatives. One is that a Taylor-series is used to produce adjusted variables that follow an approximate linear (mixed effects) model. (Appendix C.3 describes this adjustment in the linear case.) Another justification is these methods are determined through a “penalized” quasi-likelihood function, where there is a so-called “penalty” term for the random effects. This set of alternatives is the basis for the SAS macro **GLM800.sas** and, for example, the **S-plus** (a statistical package) procedure **nlme** (for nonlinear mixed effects). The disadvantage of this set of alternatives is that they do not work well for distributions that are “far” from normality, such as Bernoulli distributions (Lin and Breslow, 1996B). The advantage is that the approximation procedures work even for relatively large number of random effects. We refer the reader to McCulloch and Searle (2001G) for further discussion.

We also note that generalized linear models can be expressed as special cases of **nonlinear regression models**. Here, by “nonlinear,” we mean that the regression function need not be a linear function of the predictors but can be expressed as a nonlinear function of the form \( f(x_{it}, \)
The exponential family of distributions provides a special case of nonlinear regression functions. This is relevant to computing considerations because many computational routines have been developed for nonlinear regression and can be used directly in the GLM context. This is also true of extensions to mixed effects models, as suggested by the reference to \textit{nlme} procedure above. For discussions of nonlinear regression models with random effects, we refer the reader to Davidian and Giltinan (1995S), Vonesh and Chinchilli (1997S) and Pinheiro and Bates (2000S).

### 10.5 Fixed effects models

As we have seen, another method for accounting for heterogeneity in longitudinal and panel data is through a fixed effects formulation. Specifically, to incorporate heterogeneity, we allow for a systematic component of the form \( \eta_{it} = z_{it}' \alpha_i + x_{it}' \beta \). Thus, this section extends the Section 9.3 discussion where we only permitted varying intercepts \( \alpha_i \). However, many of the same computational difficulties arise. Specifically, it will turn out that maximum likelihood estimators are generally inconsistent. Conditional maximum likelihood estimators are consistent but difficult to compute for many distributional forms. The two exceptions are the normal and Poisson families, where we show that it is easy to compute conditional maximum likelihood estimators. Because of the computational difficulties, we restrict consideration in this section to canonical links.

#### 10.5.1 Maximum likelihood estimation for canonical links

Assume a canonical link so that \( \theta_{it} = \eta_{it} = z_{it}' \alpha_i + x_{it}' \beta \). Thus, with equation (10.3), we have the log-likelihood

\[
\ln p(y) = \sum_{it} \left\{ y_{it} \frac{z_{it}' (z_{it}' \alpha_i + x_{it}' \beta) - b(z_{it}' \alpha_i + x_{it}' \beta)}{\phi} + S(y_{it}, \phi) \right\}. \tag{10.12}
\]

To determine maximum likelihood estimators of \( \alpha_i \) and \( \beta \), we take derivatives of \( \ln p(y) \), set the derivatives equal to zero and solve for the roots of these equations.

Taking the partial derivative with respect to \( \alpha_i \) yields

\[
0 = \sum_{i} \left\{ z_{it} \left( y_{it} - b(z_{it}' \alpha_i + x_{it}' \beta) \right) \right\} = \sum_{i} \left\{ z_{it} \left( y_{it} - \mu_{it} \right) \right\},
\]

because \( \mu_{it} = b'(\theta_{it}) = b'(z_{it}' \alpha_i + x_{it}' \beta) \). Taking the partial derivative with respect to \( \beta \) yields

\[
0 = \sum_{i} \left\{ x_{it} \left( y_{it} - b(z_{it}' \alpha_i + x_{it}' \beta) \right) \right\} = \sum_{i} \left\{ x_{it} \left( y_{it} - \mu_{it} \right) \right\}.
\]

Thus, we can solve for the maximum likelihood estimators of \( \alpha_i \) and \( \beta \) through the “normal equations”

\[
0 = \sum_{i} z_{it} (y_{it} - \mu_{it}) \quad \text{and} \quad 0 = \sum_{i} x_{it} (y_{it} - \mu_{it}). \tag{10.13}
\]

This is a special case of the method of moments. Unfortunately, as we have seen in Section 9.3, this procedure may produce inconsistent estimates of \( \beta \). The difficulty is that the number of parameter estimators, \( q \times n + K \), grows with the number of subjects, \( n \). Thus, the usual asymptotic theorems that ensure our distributional approximations are no longer valid.
Example: Tort filings – continued

To illustrate the fixed effects estimators, we return to the Section 10.2 Tort filing example. Table 10.8 summarizes the fit of the fixed effects model. For comparison, the Table 10.5 fits from the homogeneous Poisson model with an estimated scale parameter are included in Table 10.8.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Homogeneous Model</th>
<th>Model with state categorical variable</th>
<th>Model with state and time categorical variables</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Parameter estimate</td>
<td>p-values</td>
<td>Parameter estimate</td>
</tr>
<tr>
<td>Intercept</td>
<td>-7.943</td>
<td>&lt;.0001</td>
<td>0.788</td>
</tr>
<tr>
<td>POPLAWYR/1000</td>
<td>2.163</td>
<td>0.0002</td>
<td>0.303</td>
</tr>
<tr>
<td>VEHCMILE/1000</td>
<td>0.862</td>
<td>&lt;.0001</td>
<td>0.2565</td>
</tr>
<tr>
<td>POPDENSY/1000</td>
<td>0.392</td>
<td>0.0038</td>
<td>0.3791</td>
</tr>
<tr>
<td>WCMPMAX/1000</td>
<td>-0.802</td>
<td>0.1226</td>
<td>0.3567</td>
</tr>
<tr>
<td>URBAN/1000</td>
<td>0.892</td>
<td>0.8187</td>
<td>-33.941</td>
</tr>
<tr>
<td>UNEMPLOY</td>
<td>0.087</td>
<td>0.028</td>
<td>0.014</td>
</tr>
<tr>
<td>JSLIAB</td>
<td>0.177</td>
<td>0.0092</td>
<td>0.311</td>
</tr>
<tr>
<td>COLLRULE</td>
<td>-0.032</td>
<td>0.7444</td>
<td>0.653</td>
</tr>
<tr>
<td>CAPS</td>
<td>-0.032</td>
<td>0.7457</td>
<td>0.2053</td>
</tr>
<tr>
<td>PUNITIVE</td>
<td>0.030</td>
<td>0.6623</td>
<td>0.6377</td>
</tr>
<tr>
<td>Scale</td>
<td>35.857</td>
<td>16.779</td>
<td>16.315</td>
</tr>
<tr>
<td>Deviance</td>
<td>118,309.0</td>
<td>22,463.4</td>
<td>19,834.2</td>
</tr>
<tr>
<td>Pearson Chi-Square</td>
<td>129,855.7</td>
<td>23,366.1</td>
<td>20,763.0</td>
</tr>
</tbody>
</table>

10.5.2 Conditional maximum likelihood estimation for canonical links

Using equation (10.12), we see that certain parameters depend on the responses on through certain summary statistics. Specifically, using the factorization theorem described in Appendix 10A.2, we have that the statistics ∑ᵢ yᵢᵣ zᵢᵣ are sufficient for αᵢ and that the statistics ∑ᵢ yᵢᵣ xᵢᵣ are sufficient for β. This convenient property of canonical links is not available for other choices of links. Recall that sufficiency means that the distribution of the responses will not depend on a parameter when conditioned by the corresponding sufficient statistic. Specifically, we now consider the likelihood of the data conditional on { ∑ᵢ yᵢᵣ zᵢᵣ, i = 1, ..., n }, so that the conditional likelihood will not depend on { αᵢ }. By maximizing this conditional likelihood, we achieve consistent estimators of β.

To this end, let Sᵢ be the random variable representing ∑ᵢ zᵢᵣ yᵢᵣ and let sumᵢ be the realization of ∑ᵢ zᵢᵣ yᵢᵣ. The conditional likelihood of the responses is

\[ \prod_{i=1}^{n} \frac{p(y_{i1}, \alpha_i, \beta) p(y_{i2}, \alpha_i, \beta) \cdots p(y_{iT}, \alpha_i, \beta)}{p_{S_i}(sum_i)}, \] (10.14)

where \( p_{S_i}(sum_i) \) is the probability density (or mass) function of \( S_i \) evaluation at \( sum_i \). This likelihood does not depend on \{ αᵢ \}, only on β. Thus, when evaluating it, we can take αᵢ to be a zero vector without loss of generality. Under broad conditions, maximizing equation (10.14) with
respect to $\beta$ yields root-$n$ consistent estimators, see, for example, McCullagh and Nelder (1989G). Still, as in Section 9.3, for most parametric families, it is difficult to compute the distribution of $S_i$. Clearly, the normal distribution is one exception, because if the responses are normal, then the distribution of $S_i$ is also normal. The following subsection describes another important application where the computation is feasible under conditions likely to be encountered in applied data analysis.

**10.5.3 Poisson distribution**

This subsection considers Poisson distributed data, a widely used distribution for count responses. To illustrate, we consider the canonical link with $q = 1$ and $z_{it} = 1$, so that $\theta_{it} = \alpha_i + x_{it}'\beta$. The Poisson distribution is given by

$$p(y_{it}, \alpha_i, \beta) = \frac{(\mu_{it})^{y_{it}} \exp(-\mu_{it})}{y_{it}!}$$

with $\mu_{it} = b'(\theta_{it}) = \exp(\alpha_i + x_{it}'\beta)$. Conversely, because $\ln \mu_{it} = \alpha_i + x_{it}'\beta$, we have that the logarithmic mean is a linear combination of explanatory variables. This is the basis of the so-called “log-linear” model.

As noted in Section 10A.2, $\Sigma_t y_{it}$ is a sufficient statistic for $\alpha_i$. Further, it is easy to check that the distribution of $\Sigma_t y_{it}$ turns out to be Poisson, with mean $\Sigma_t \exp(\alpha_i + x_{it}'\beta)$. In the subsequent development, we will use the ratio of means,

$$\pi_{it}(\beta) = \frac{\mu_{it}}{\sum_t \mu_{it}} = \frac{\exp(x_{it}'\beta)}{\sum_t \exp(x_{it}'\beta)},$$

(10.15)

that does not depend on $\alpha_i$.

Now, using equation (10.14), the conditional likelihood for the $i$th subject is

$$\text{Prob}(S_i = \sum_t y_{it})$$

$$= \frac{\prod_t (\mu_{it})^{y_{it}} \exp(-\mu_{it})}{\prod_t y_{it}!} / \left( \prod_t \frac{\exp(x_{it}'\beta)}{\sum_t \exp(x_{it}'\beta)} \right)$$

$$= \prod_t \pi_{it}(\beta)^{y_{it}}$$

where $\pi_{it}(\beta)$ is given in equation (10.15). This is a multinomial distribution. Thus, the joint distribution of $\{y_{i1}, \ldots, y_{iT_i}\}$ given $\sum_t y_{it}$ has a multinomial distribution.

Using equation (10.14), the conditional likelihood is
Taking partial derivatives of the log conditional likelihood yields

$$\frac{\partial}{\partial \beta} \ln CL = \sum_i y_{it} \ln \pi_{it} (\beta) = \sum_i y_{it} \left( x_{it} - \sum_r x_{ir} \pi_{ir} (\beta) \right).$$

Thus, the conditional maximum likelihood estimate, $b_{\text{CMLE}}$, is the solution of

$$\sum_{i=1}^{n_T} \sum_{r=1}^{R} y_{it} \left( x_{it} - \sum_{r=1}^{R} x_{ir} \pi_{ir} (b_{\text{CMLE}}) \right) = 0.$$
Begin by recalling Bayes’ rule

\[
p(\text{parameters} \mid \text{data}) = \frac{p(\text{data} \mid \text{parameters}) \times p(\text{parameters})}{p(\text{data})}
\]

where

• \( p(\text{parameters}) \) is the distribution of the parameters, known as the prior distribution.
• \( p(\text{data} \mid \text{parameters}) \) is the sampling distribution. In a frequentist context, it is used for making inferences about the parameters and is known as the likelihood.
• \( p(\text{parameters} \mid \text{data}) \) is the distribution of the parameters having observed the data, known as the posterior distribution.
• \( p(\text{data}) \) is the marginal distribution of the data. It is generally obtained by integrating (or summing) the joint distribution of data and parameters over parameter values. This is often the difficult step in Bayesian inference.

In a regression context, we have two types of data, the response variable \( y \) and the set of explanatory variables \( X \). Let \( \theta \) and \( \psi \) denote the sets of parameters that describe the sampling distributions of \( y \) and \( X \), respectively. Moreover, assume that \( \theta \) and \( \psi \) are independent. Then, using Bayes’ rule, the posterior distribution is

\[
p(\theta, \psi \mid y, X) \propto \frac{p(y, X \mid \theta, \psi) \times p(\theta, \psi)}{p(y, X)} \propto \{p(y \mid X, \theta) \times p(\theta)\} \times \{p(X \mid \psi) \times p(\psi)\}
\]

\[
\propto p(\theta \mid y, X) \times p(\psi \mid X).
\]

Here, the symbol “\( \propto \)” means “is proportional to.” Thus, the joint posterior distribution of the parameters can be factored into two pieces, one for the responses and one for the explanatory variables. Assuming no dependencies between \( \theta \) and \( \psi \), there is no loss of information in the traditional regression setting by ignoring the distributions associated with the explanatory variables. By the “traditional regression setting,” we mean that one essentially treats the explanatory variables as non-stochastic.

Most statistical inference can be accomplished readily having computed the posterior. With this entire distribution, summarizing likely values of the parameters through confidence intervals or unlikely values through hypothesis tests is straightforward. Bayesian methods are also especially suitable for forecasting. In the regression context, suppose we wish to summarize the distribution of a set of new responses, \( y_{\text{new}} \), given new explanatory variables, \( X_{\text{new}} \), and previously observed data \( y \) and \( X \). This distribution, \( p(y_{\text{new}} \mid X_{\text{new}}, y, X) \) is a type of predictive distribution. We have that

\[
p(y_{\text{new}} \mid y, X, X_{\text{new}}) = \int p(y_{\text{new}} \mid \theta, y, X, X_{\text{new}}) \times p(\theta \mid y, X, X_{\text{new}}) \, d\theta.
\]

This assumes that the parameters \( \theta \) are continuous. Here, \( p(y_{\text{new}} \mid \theta, y, X, X_{\text{new}}) \) is the sampling distribution of \( y_{\text{new}} \) and \( p(\theta \mid y, X, X_{\text{new}}) = p(\theta \mid y, X) \) is the posterior distribution (assuming that values of the new explanatory variables are independent of \( \theta \)). Thus, the predictive distribution can be computed as a weighted average of the sampling distribution, where the weights are given by the posterior.

A difficult aspect of Bayesian inference can be the assignment of priors. Classical assignments of priors are generally either non-informative or conjugate. “Non-informative” priors are distributions that are designed to interject the least amount of information possible. Two
important types of non-informative priors are uniform (also known as “flat”) priors and “Jeffrey’s prior.” A uniform prior is simply a constant value; thus, no value is more likely than any other. A drawback of this type of prior is that not invariant under transformation of the parameters. To illustrate, consider the normal linear regression model, so that \( y \mid X, \beta, \sigma^2 \sim N(X \beta, \sigma^2 I) \). A widely used non-informative prior is a flat prior on \((\beta, \log \sigma^2)\), so that the joint distribution of \((\beta, \sigma^2)\) turns out to be proportional to \(\sigma^{-2}\). Thus, although uniform in \(\log \sigma^2\), this prior gives heavier weight to small values of \(\sigma^2\). Jeffrey’s priors are complex in the case of multidimensional parameters and thus we will not consider them further here.

Conjugacy of a prior is actually a property that depends on both the prior as well as sampling distribution. When the prior and posterior distributions come from the same family of distributions, then the prior is known as a conjugate prior. Appendix 10A.3 gives several examples of the more commonly used conjugate priors.

For longitudinal and panel data, it is convenient to formulate Bayesian models in three stages: one for the parameters, one for the data and one stage for the latent variables used to represent the heterogeneity. Thus, extending Section 7.3.1, we have:

**Stage 0.** (Prior distribution) Draw a realization of a set of parameters from a population. The parameters consist of regression coefficients \(\beta\) and variance components \(\tau\).

**Stage 1.** (Heterogeneity effects distribution) Conditional on the parameters from stage 0, draw a random sample of \(n\) subjects from a population. The vector of subject-specific effects \(\alpha_i\) is associated with the \(i\)th subject.

**Stage 2.** (Conditional sampling distribution) Conditional on \(\beta^*\) and \(\{Z, X\}\), \(\{y_{it}\}\) for \(t = 1, \ldots, T_i\) for the \(i\)th subject. Summarize these draws as \(\{y_i, Z_i, X_i\}\).

A common method of analysis is to combine stages 0 and 1 and to treat \(\beta^* = (\beta', \alpha_1, \ldots, \alpha_n)'\) as the regression parameters of interest. Also common is to use a normal prior distribution for this set of regression parameters. This is the conjugate prior when the sampling distribution is normal. When the sampling distribution is from the GLM family (but not normal), there is no general recipe for conjugate priors. A normal prior is useful because it is flexible and computationally convenient. To be specific, we consider.

**Stage 1*.** (Prior distribution) Assume that

\[
\beta^* = \begin{bmatrix} \beta \\ \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} \sim N\left( \begin{pmatrix} E\beta \\ \Sigma_\beta \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} D \otimes I_n \end{pmatrix} \right).
\]

Thus, both \(\beta\) and \((\alpha_1, \ldots, \alpha_n)\) are normally distributed.

**Stage 2.** (Conditional sampling distribution) Conditional on \(\beta^*\) and \(\{Z, X\}\), \(\{y_{it}\}\) are independent and the distribution of \(y_{it}\) is from a generalized linear model (GLM) family with parameter \(\eta_{it} = z_{it}'\alpha_i + x_{it}'\beta\).

For some types of GLM families (such as the normal), an additional scale parameter is used that is typically included in the prior distribution specification.

With this specification, in principle one simply applies Bayes’ rule to determine the posterior distribution of \(\beta^*\) given the data \(\{y, X, Z\}\). However, as a practical matter, this is difficult to do without conjugate priors. Specifically, to compute the marginal distribution of the data, one must use numerical integration to remove parameter distributions; this is computationally intensive for many problems of interest. To circumvent this difficult, modern
Bayesian analysis regular employs simulation techniques known as Markov Chain Monte Carlo (MCMC) methods and an especially important special case, the Gibbs sampler. MCMC methods produce simulated values of the posterior distribution and are available in many statistical packages, including the shareware that is favored by many Bayesian analysts, BUGS/WINBUGS (available at http://www.mrc-bsu.cam.ac.uk/bugs/). There are many specialized treatments that discuss the theory and applications of this approach; we refer the reader to Gelman et al. (2004G), Gill (2002G) and Congdon (2003G).

For some applications such as prediction, the interest is in the full joint posterior distribution of the global regression parameters and the heterogeneity effects, \( \beta, \alpha_1, \ldots, \alpha_n \). For other applications, the interest is in the posterior distribution of the global regression coefficients, \( \beta \). In this case, one integrates out the heterogeneity effects from the joint posterior distribution of \( \beta^* \).

**Example – Respiratory infections**

Zeger and Karim (1991S) introduced Bayesian inference for GLM longitudinal data. Specifically, Zeger and Karim considered the stage 2 GLM conditional sampling distribution and, for the heterogeneity distribution, assumed that \( \alpha_i \) are conditionally i.i.d. \( N(0, D) \). They allowed for a general form for the prior distribution.

To illustrate this set-up, Zeger and Karim examined infectious disease data on \( n = 250 \) Indonesian children who were observed up to \( T = 6 \) consecutive quarters, for a total of \( N = 1,200 \) observations. The goal was to assess determinants of a binary variable that indicates the presence of a respiratory disease. They used normally distributed random intercepts \( (q = 1 \text{ and } z_{it} = 1) \) with a logistic conditional sampling distribution.

**Example – Patents**

Chib, Greenberg and Winkelman (1998E) discussed parameterization of Bayesian Poisson models. They considered a conditional Poisson model with a canonical link so that \( y_{it} | \alpha_i, \beta \sim \text{Poisson}(\theta_{it}) \), where \( \theta_{it} = E(y_{it} | \alpha_i, \beta) = z_{it}' \alpha_i + x_{it}' \beta \). Unlike prior work, they did not assume that the heterogeneity effects have zero mean but instead used \( \alpha_i \sim N(\eta_\alpha, D) \). With this specification, they did not use the usual convention of assuming that \( z_{it} \) is a subset of \( x_{it} \). This, they argued, leads to more stable convergence algorithms when computing posterior distributions.

To complete the specification, Chib et al. used prior distributions:

\[ \beta \sim N(\mu_\beta, \Sigma_\beta) \text{, } \eta_\alpha \sim N(\eta_{0\alpha}, \Sigma_\eta) \text{ and } D^{-1} \sim \text{Wishart}. \]

The Wishart distribution is a multivariate extension of the chi-square distribution, see, for example, Anderson (1958G).

To illustrate, Chib et al. used patent data first considered by Hausman, Hall and Griliches (1984E). These data include the number of patents received by \( n = 642 \) firms over \( T = 5 \) years, 1975-1979. The explanatory variables included the logarithm of research and development (R&D) expenditures, as well as their 1, 2 and 3 year lags, and time dummy variables. Chib et al. use a variable intercept and a variable slope for the logarithmic R&D expenditures.
Further reading

McCullagh and Nelder (1989S) provide a more extensive introduction to (homogeneous) generalized linear models.

Conditional likelihood testing in connection with exponential families was suggested by Rasch (1961EP) and asymptotic properties were developed by Andersen (1970S), motivated in part by the “presence of infinitely many nuisance parameters” of Neyman and Scott (1948E). Panel data conditional likelihood estimation was introduced by Chamberlain (1980E) for binary logit models and by Hausman, Hall and Griliches (1984E) for Poisson (as well as negative binomial) count models.


For additional discussions on computing aspects, we refer to McCulloch and Searle (2001G) and Pinheiro and Bates (2000S).
Appendix 10A. Exponential families of distributions

The distribution of the random variable $y$ may be discrete, continuous or a mixture. Thus, $p(.)$ in equation (10.1) may be interpreted to be a density or mass function, depending on the application. Table 10A.1 provides several examples, including the normal, binomial and Poisson distributions.

### Table 10A.1 Selected Distributions of the One-Parameter Exponential Family

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Parameters</th>
<th>Density or Mass Function</th>
<th>Components</th>
<th>$E_y$</th>
<th>Var $y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>$\theta$, $\phi$</td>
<td>$\exp\left(\frac{y\theta - b(\theta)}{\phi} + S(y, \phi)\right)$</td>
<td>$\theta$, $\phi$, $b(\theta)$, $S(y, \phi)$, $b'(\theta)$, $b''(\theta)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>$\mu$, $\sigma^2$</td>
<td>$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$</td>
<td>$\mu$, $\sigma^2$, $\frac{\theta^2}{2} - \frac{y^2 + \ln(2\pi\phi)}{2\phi}$</td>
<td>$\theta = \mu$, $\phi = \sigma^2$</td>
<td></td>
</tr>
<tr>
<td>Binomial</td>
<td>$\pi$</td>
<td>$\pi^n (1-\pi)^{n-y} \ln\left(\frac{\pi}{1-\pi}\right)$</td>
<td>$n \ln(1 + e^\phi)$</td>
<td>$n \frac{e^\phi}{1 + e^\phi}$</td>
<td>$n \frac{e^\phi}{1 + e^\phi} n = n \pi = n \pi(1 - \pi)$</td>
</tr>
<tr>
<td>Poisson</td>
<td>$\lambda$</td>
<td>$\lambda^y \exp(-\lambda)$</td>
<td>$\ln \lambda$, $1$, $e^\phi$</td>
<td>$-\ln(y!)$, $e^\phi$, $\lambda$, $e^\phi = \lambda$</td>
<td></td>
</tr>
<tr>
<td>Gamma</td>
<td>$\alpha$, $\beta$</td>
<td>$\frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \exp(-y/\beta)$</td>
<td>$-\frac{\beta}{\alpha}$, $\frac{1}{\alpha}$, $-\ln(-\theta)$, $-\phi^{-1} \ln \phi$</td>
<td>$\frac{1}{\theta}$, $\frac{\phi}{\theta^2}$, $\frac{\alpha}{\beta^2}$, $\frac{1}{\theta} + (\phi^{-1} - 1) \ln y$</td>
<td></td>
</tr>
</tbody>
</table>

### 10A.1 Moment Generating Function

To assess the moments of exponential families, it is convenient to work with the moment generating function. For simplicity, we assume that the random variable $y$ is continuous. Define the moment generating function

$$M(s) = E e^{sy} = \int \exp\left(sy + \frac{y\theta - b(\theta)}{\phi} + S(y, \phi)\right) dy$$

$$= \exp\left(\frac{b(\theta + s\phi) - b(\theta)}{\phi}\right) \int \exp\left(y(\theta + s\phi) - b(\theta + s\phi) + S(y, \phi)\right) dy$$

$$= \exp\left(\frac{b(\theta + s\phi) - b(\theta)}{\phi}\right) \int \exp\left(y(b(\theta + s\phi) - b(\theta) + S(y, \phi)\right) dy = \exp\left(\frac{b(\theta + s\phi) - b(\theta)}{\phi}\right).$$

With this expression, we can generate the moments. Thus, for the mean, we have

$$E_y = M'(0) = \frac{\partial}{\partial s} \exp\left(\frac{b(\theta + s\phi) - b(\theta)}{\phi}\right)\bigg|_{s=0} = \left[b'(\theta + s\phi) \exp\left(\frac{b(\theta + s\phi) - b(\theta)}{\phi}\right)\right]_{s=0} = b'(\theta).$$

Similarly, for the second moment, we have
This yields $E y^2 = M^*(0) = \phi b^*(\theta) + (b'(\theta))^2$ and $\text{Var} y = \phi b^*(\theta)$.

### 10A.2 Sufficiency

In complex situations, it is convenient to be able to decompose the likelihood into several pieces that can be analyzed separately. To accomplish this, the concept of sufficiency is useful. A statistic $T(y_1, \ldots, y_n) = T(y)$ is **sufficient** for a parameter $\theta$ if the distribution of $y_1, \ldots, y_n$ conditional on $T(y)$ does not depend on $\theta$.

When checking whether or not a statistic is sufficient, an important result is the factorization theorem. This result indicates, under certain regularity conditions, that a statistic $T(y)$ is sufficient for $\theta$ if and only if the density (mass) function of $y$ can be decomposed into the product of two components

$$p(y, \theta) = p_1(T(y), \theta) p_2(y).$$  \hspace{1cm} (10A.2)

Here, the first portion, $p_1$, may depend on $\theta$ but depends on the data only through the sufficient statistic $T(y)$. The second portion, $p_2$, may depend on the data but does not depend on the parameter $\theta$. See, for example, Bickel and Doksum (1977G).

To illustrate, if $\{y_1, \ldots, y_n\}$ are independent and follow the distribution in equation (10.1), then the joint distribution is

$$p(y, \theta, \phi) = \exp \left( \frac{\theta \left( \sum_{i=1}^n y_i \right) - n b(\theta)}{\phi} + \sum_{i=1}^n S(y_i, \phi) \right).$$

Thus, with $p_2(y) = \exp \left( \sum_{i=1}^n S(y_i, \phi) \right)$ and $p_1(T(y), \theta) = \exp \left( \frac{\theta T(y) - n b(\theta)}{\phi} \right)$, the statistic $T(y) = \sum_{i=1}^n y_i$ is sufficient for $\theta$.

### 10A.3 Conjugate Distributions

Assume that the parameter $\theta$ is random with distribution $\pi(\theta, \tau)$, where $\tau$ is a vector of parameters that describe the distribution of $\theta$. In Bayesian models, the distribution $\pi$ is known as the prior and reflects our belief or information about $\theta$. The likelihood $p(y, \theta) = p(y | \theta)$ is a probability conditional on $\theta$. The distribution of $\theta$ with knowledge of the random variables, $\pi(\theta, \tau | y)$, is called the posterior distribution. For a given likelihood distribution, priors and posteriors that come from the same parametric family are known as conjugate families of distributions.

For a linear exponential likelihood, there exists a natural conjugate family. For the likelihood in equation (10.1), define the prior distribution

$$\pi(\theta, \tau) = C \exp ( \theta a_1(\tau) - b(\theta) a_2(\tau) ),$$  \hspace{1cm} (10A.2)

where $C$ is a normalizing constant. Here, $a_1(\tau)$ and $a_2(\tau)$ are functions of the parameters $\tau$. The joint distribution of $y$ and $\theta$ is given by $p(y, \theta) = p(y | \theta) \pi(\theta, \tau)$. Using Bayes Theorem, the posterior distribution is
\[ \pi(\theta, \tau \mid y) = C_1 \exp \left( a_1(\tau) + \frac{y}{\phi} \right) \theta - b(\theta) \left( a_2(\tau) + \frac{1}{\phi} \right) \right), \]

where \( C_1 \) is a normalizing constant. Thus, we see that \( \pi(\theta, \tau \mid y) \) has the same form as \( \pi(\theta, \tau) \).

### Special case 10A.1 Normal-Normal Model

Consider a normal likelihood in equation (10.1) so that \( b(\theta) = \theta^2/2 \). Thus, with equation (10A.2), we have

\[ \pi(\theta, \tau) = C \exp \left( a_1(\tau)\theta - \frac{\theta^2}{2} a_2(\tau) \right) = C_1(\tau) \exp \left( -\frac{a_2(\tau)}{2} \left( \theta - \frac{a_1(\tau)}{a_2(\tau)} \right)^2 \right). \]

Thus, the prior distribution of \( \theta \) is normal with mean \( a_1(\tau)/a_2(\tau) \) and variance \( (a_2(\tau))^{-1} \). The posterior distribution of \( \theta \) given \( y \) is normal with mean \( (a_1(\tau)+y/\phi)/(a_2(\tau)+\phi^{-1}) \) and variance \( (a_2(\tau)+\phi^{-1})^{-1} \).

### Special case 10A.2 Binomial-Beta Model

Consider a binomial likelihood in equation (10.1) so that \( b(\theta) = \ln(1+e^\theta) \). Thus, with equation (10A.2), we have

\[ \pi(\theta, \tau) = C \exp \left( a_1(\tau)\theta - n a_2(\tau) \ln(1+e^\theta) \right) = C \left( \frac{e^\theta}{1+e^\theta} \right)^{a_1(\tau)} \left( 1+e^\theta \right)^{-n a_2(\tau)+a_1(\tau)} \]

Thus, we have that prior of logit(\( \theta \)) is a beta distribution with parameters \( a_1(\tau) \) and \( n a_2(\tau)-a_1(\tau) \). The posterior of logit(\( \theta \)) is a beta distribution with parameters \( a_1(\tau)+y/\phi \) and \( n (a_2(\tau)+\phi^{-1}) - (a_1(\tau)+y/\phi) \).

### Special case 10A.3 Poisson-Gamma Model

Consider a Poisson likelihood in equation (10.1) so that \( b(\theta) = e^\theta \). Thus, with equation (10A.2), we have

\[ \pi(\theta, \tau) = C \exp \left( a_1(\tau)\theta - a_2(\tau) e^\theta \right) = C \left( e^\theta \right)^{a_1(\tau)} \exp \left( -e^\theta a_2(\tau) \right). \]

Thus, we have that the prior of \( e^\theta \) is a Gamma distribution with parameters \( a_1(\tau) \) and \( a_2(\tau) \). The posterior of \( e^\theta \) is a Gamma distribution with parameters \( a_1(\tau)+y/\phi \) and \( a_2(\tau)+\phi^{-1} \).

### 10A. 4 Marginal Distributions

In many conjugate families, the marginal distribution of the random variable \( y \) turns out to be from a well-known parametric family. To see this, consider the prior distribution in equation (10A.2). Because densities integrate (sum, in the discrete case) to 1, we may express the constant \( C \) as

\[ C^{-1} = (C(a_1, a_2))^{-1} = \int \exp(\theta a_1 - b(\theta)a_2) d\theta, \]

a function of the parameters \( a_1 = a_1(\tau) \) and \( a_2 = a_2(\tau) \). With this and equation (10A.1), the marginal distribution of \( y \) is

\[ g(y) = \int p(y, \theta) \pi(\theta) d\theta = C(a_1, a_2) \exp(S(y, \phi)) \int \exp \left( \theta \left( \frac{y}{\phi} + a_1 \right) - b(\theta) \left( \frac{1}{\phi} + a_2 \right) \right) d\theta \]
\[ C(a_1, a_2) \exp(S(y, \phi)) = \frac{C(a_1 + \frac{y}{\phi}, a_2 + \frac{1}{\phi})}{C(a_1, a_2)}. \]  
\[ \text{(10A.3)} \]

It is of interest to consider several special cases.

**Special case 10A.1 Normal-Normal Model - Continued**

Using Table 10A.1, we have \( \exp(S(y, \phi)) = \frac{1}{\sqrt{2\pi\phi}} \exp\left(-\frac{y^2}{2\phi}\right) \). Straightforward calculations show that \( C(a_1, a_2) = C_1 \exp\left(-\frac{a_1^2}{2a_2}\right) = \sqrt{\frac{2a_2}{\pi}} \exp\left(-\frac{a_1^2}{2a_2}\right) \). Thus, from equation (10A.3), the marginal distribution is

\[
g(y) = \frac{\sqrt{a_2/(2\pi)} \exp\left(-\frac{a_1^2}{2a_2}\right) \sqrt{1/(2\pi\phi)} \exp\left(-\frac{y^2}{2\phi}\right)}{\sqrt{(a_2 + \phi^{-1})/(2\pi)} \exp\left(-\frac{(a_1 + y/\phi)^2}{2(a_2 + \phi^{-1})}\right)}
\]

\[
= \frac{1}{\sqrt{2\pi(\phi + a_2^{-1})}} \exp\left(-\frac{1}{2} \left(\frac{a_1^2}{a_2} + \frac{y^2}{\phi} - \frac{(a_1 + y/\phi)^2}{a_2 + \phi^{-1}}\right)\right) = \frac{1}{\sqrt{2\pi(\phi + a_2^{-1})}} \exp\left(-\frac{(y - a_1/\phi)^2}{2(a_2 + \phi^{-1})}\right).
\]

Thus, \( y \) is normally distributed with mean \( a_1/a_2 \) and variance \( \phi+a_2^{-1} \).

**Special case 10A.2 Binomial-Beta Model - Continued**

Using Table 10A.1, we have \( \exp(S(y, \phi)) = \binom{n}{y} \) and \( \phi = 1 \). Further, straightforward calculations show that \( C(a_1, a_2) = \frac{\Gamma(na_2)}{\Gamma(a_1)\Gamma(na_2 - a_1)} = \frac{1}{B(a_1, na_2 - a_1)} \), where \( B(.,.) \) is called the Beta function. Thus, from equation (10A.3), we have

\[
g(y) = \binom{n}{y} B(a_1 + y, na_2 + 1) \binom{n}{y} B(a_1, na_2 - a_1).
\]

This is called the beta-binomial distribution.
Special case 10A.3 Poisson-Gamma Model - Continued

Using Table 10A.1, we have \(\exp(S(y,\phi)) = 1/y! = 1/(y+1)\) and \(\phi = 1\). Further, straightforward calculations show that \(C(a_1,a_2) = \frac{(a_2)^{y_i}}{\Gamma(a_1)}\). Thus, from equation (10A.4), we have \(g(y) = \frac{(a_2)^{y_i}}{\Gamma(a_1)} \frac{1}{\Gamma(y+1)} = \frac{\Gamma(a_1+y)(a_2)^{y_i}}{\Gamma(a_1)\Gamma(y+1)(a_2+1)^{y_i+1}} = \frac{\Gamma(a_1+y)}{\Gamma(a_1)\Gamma(y+1)(a_2+1)^{y_i+1}} \left( \frac{a_2}{a_2+1} \right)^y\).

This is a negative binomial distribution.

10. Exercises and Extensions

Poisson and Binomials
10.1 Consider \(y_i\) that has a Poisson distribution with mean parameter \(\mu_i\), \(i=1,2\), and suppose that \(y_1\) is independent of \(y_2\).

a. Show that \(y_1 + y_2\) has a Poisson distribution with mean parameter \(\mu_1 + \mu_2\).

b. Show that \(y_1\) given \(y_1 + y_2\) has a Binomial distribution. Determine the parameters of the Binomial distribution.

c. Suppose that you tell the computer to evaluate the likelihood of \((y_1,y_2)\) assuming that \(y_1\) is independent of \(y_2\), each has a Poisson distribution with mean parameters \(\mu_1\) and \(\mu_2 = 1 - \mu_1\).

i. Evaluate the likelihood of \((y_1,y_2)\).

ii. However, the data are such that \(y_1\) is binary and \(y_2 = 1 - y_1\). Evaluate the likelihood that the computer is evaluating.

iii. Show that this equals the likelihood assuming \(y_1\) has a Bernoulli distribution with mean \(\mu_1\), up to a proportionality constant. Determine the constant.

d. Consider a longitudinal data set \(\{y_{it}, x_{it}\}\) where \(y_{it}\) are binary with mean \(p_{it} = \pi(x_{it}' \beta)\), where \(\pi(.)\) is known. You would like to evaluate the likelihood but can only do so in terms of a package using Poisson distributions. Using the part (c) result, explain how to do so.
Chapter 11. Categorical Dependent Variables and Survival Models

Abstract. Extending Chapter 9, this chapter considers dependent variables having more than two possible categorical alternatives. As in Chapter 9, we often interpret a categorical variable to be an attribute possessed or choice made by an individual, household or firm. By allowing more than two alternatives, we substantially broaden the scope of applications to complex social science problems of interest.

The pedagogic approach of this chapter follows the pattern established in earlier chapters; we begin with homogeneous models in Section 11.1 followed by the Section 11.2 model that incorporates random effects, thus providing a heterogeneity component. Section 11.3 introduces an alternative method for accommodating time patterns through transition, or Markov, models. Although transition models are applicable in the Chapter 10 generalized linear models, they are particularly useful in the context of categorical dependent variables. Many repeated applications of the idea of transitioning gives rise to survival models, another important class of longitudinal models. Section 11.4 develops this link.

11.1 Homogeneous models

We now consider a response that is an unordered categorical variable. We assume that the dependent variable \( y \) may take on values 1, 2, \ldots, \( c \), corresponding to \( c \) categories. We first introduce the homogenous version so that this section does not use any subject-specific parameters nor does it introduce terms to account for serial correlation.

In many social science applications, the response categories correspond to an attribute possessed or choices made by individuals, households or firms. Some applications of categorical dependent variable models include:

- employment choice, such as Valletta (1999E),
- mode of transportation choice, such as the classic work by McFadden (1978E),
- choice of political party affiliation, such as Brader and Tucker (2001O) and
- marketing brand choice, such as Jain et al. (1994O).

Example 11.1 Political party affiliation

Brader and Tucker (2001O) studied the choice made by Russian voters of political parties in 1995-1996 elections. Their interest was in assessing effects of meaningful political party attachments during Russia’s transition to democracy. They examined a \( T = 3 \) wave survey of \( n = 2,841 \) respondents, taken: (1) three to four weeks before the 1995 parliamentary elections, (2) immediately after the parliamentary elections and (3) after the 1996 presidential elections. This survey design was modeled on the American National Election Studies (see Appendix F).

The dependent variable was the political party voted for, consisting of \( c = 10 \) parties including the Liberal Democratic Party of Russia, Communist Party of the Russian Federation, Our Home is Russia, and others. The independent variables included social characteristics such as education, gender, religion, nationality, age, and location (urban versus rural), economic
characteristics such as income and employment status and political attitudinal characteristics such
attitudes toward market transitions and privatization.

11.1.1 Statistical inference

For an observation from subject \( i \) at time \( t \), denote the probability of choosing the \( j \)th
category as \( \pi_{it,j} = \text{Prob}(y_{it} = j) \), so that \( \pi_{it,1} + \ldots + \pi_{it,c} = 1 \). In the homogeneous framework, we
assume that observations are independent from one another.

In general, we will model these probabilities as a (known) function of parameters and use
maximum likelihood estimation for statistical inference. We use the notation \( y_{itj} \) to be an indicator
of the event \( y_{it} = j \). The likelihood for the \( it \)th subject at the \( t \)th time point is:

\[
\prod_{j=1}^{c} \left( \pi_{it,j} \right)^{y_{itj}} = \begin{cases} 
\pi_{it,1} & \text{if } y_{it} = 1 \\
\pi_{it,2} & \text{if } y_{it} = 2 \\
\vdots & \vdots \\
\pi_{it,c} & \text{if } y_{it} = c 
\end{cases}.
\]

Thus, assuming independence among observations, the total log-likelihood is

\[
L = \sum_{i} \sum_{j=1}^{c} y_{itj} \ln \pi_{it,j}.
\]

With this framework, standard maximum likelihood estimation is available. Thus, our main
task is to specify an appropriate form for \( \pi \).

11.1.2 Generalized logit

Like standard linear regression, generalized logit models employ linear combinations of
explanatory variables of the form:

\[
V_{it,j} = \mathbf{x}_{it} \beta_j.
\]  

(11.1)

Because the dependent variables are not numerical, we cannot model the response \( y \) as a linear
combination of explanatory variables plus an error. Instead we use the probabilities

\[
\text{Prob}(y_{it} = j) = \pi_{it,j} = \frac{\exp(V_{it,j})}{\sum_{k=1}^{c} \exp(V_{it,k})}, \quad j = 1, 2, \ldots, c 
\]

(11.2)

Note here that \( \beta_j \) is the corresponding vector of parameters that may depend on the alternative, or
choice, whereas the explanatory variables \( \mathbf{x}_{it} \) do not. So that probabilities sum to one, a
convenient normalization for this model is \( \beta_c = 0 \). With this normalization and the special case of
\( c = 2 \), the generalized logit reduces to the logit model introduced in Section 9.1.

Parameter interpretations

We now describe an interpretation of coefficients in generalized logit models, similar to
Section 9.1.1 for the logistic model. From equations (11.1) and (11.2), we have

\[
\ln \left( \frac{\text{Prob}(y_{it} = j)}{\text{Prob}(y_{it} = c)} \right) = V_{it,j} - V_{it,c} = \mathbf{x}_{it}' \beta_j.
\]

The left-hand side of this equation is interpreted to be the logarithmic odds of choosing choice \( j \)
compared to choice \( c \). Thus, as in Section 9.1.1, we may interpret \( \beta_j \) as the proportional change in
the odds ratio.
Generalized logits have an interesting nested structure that we will explore briefly in Section 9.1.5. That is, it is easy to check that conditional on not choosing the first category, the form of $\text{Prob}(y_{it} = j \mid y_{it} \neq 1)$ has a generalized logit form in equation (11.2). Further, if $j$ and $h$ are different alternatives, we note that

$$
\text{Prob}(y_{it} = j \mid \{y_{it} = j \text{ or } y_{it} = h\}) = \frac{\text{Prob}(y_{it} = j)}{\text{Prob}(y_{it} = j) + \text{Prob}(y_{it} = h)} = \frac{\exp(V_{it,j})}{\exp(V_{it,j}) + \exp(V_{it,h})} = \frac{1}{1 + \exp(x'_it (\beta_j - \beta_h))}.
$$

This has a logit form that was introduced in Section 9.1.1.

**Special case – Intercept only model**

To develop intuition, we now consider the model with only intercepts. Thus, let $x_{it} = 1$ and $\beta_j = \beta_0, j = \alpha_j$. With the convention $\alpha_c = 0$, we have

$$
\text{Prob}(y_{it} = j) = \pi_{it,j} = \frac{e^{\alpha_j}}{e^{\alpha_1} + e^{\alpha_2} + \ldots + e^{\alpha_c} + 1}
$$

and

$$
\ln \left( \frac{\text{Prob}(y_{it} = j)}{\text{Prob}(y_{it} = c)} \right) = \alpha_j.
$$

From the second relation, we may interpret the $j$th intercept $\alpha_j$ to be the logarithmic odds of choosing alternative $j$ compared to alternative $c$.

**Example 11.2 Job security**

This is a continuation of the Example 9.1 on the determinants of job turnover, based on the work of Valetta (1999E). The Chapter 9 analysis of this data considered only the binary dependent variable dismissal, the motivation being that this is the main source of job insecurity. Valetta (1999E) also presented results from a generalized logit model, his primary motivation being that the economic theory describing turnover implies that other reasons for leaving a job may affect dismissal probabilities.

For the generalized logit model, the response variable has $c = 5$ categories: dismissal, left job because of plant closures, “quit,” changed jobs for other reasons and no change in employment. The latter category is the omitted one in Table 11.1. The explanatory variables of the generalized logit are same as the probit regression described in Example 9.1; the estimates summarized in Example 9.1 are reproduced here for convenience.

Table 11.1 shows that turnover declines as tenure increases. To illustrate, consider a typical man in the 1992 sample where we have time = 16 and focus on dismissal probabilities. For this value of time, the coefficient associated with tenure for dismissal is $-0.221 + 16 \times (0.008) = -0.093$ (due to the interaction term). From this, we interpret an additional year of tenure to imply that the dismissal probability is $\exp(-0.093) = 91\%$ of what it would be otherwise, representing a decline of $9\%$.

Table 11.1 also shows that the generalized coefficients associated with dismissal are similar to the probit fits. The standard errors are also qualitatively similar, although higher for the generalized logits when compared to the probit model. In particular, we again see that the
coefficient associated with the interaction between tenure and time trend reveals an increasing dismissal rate for experienced workers. The same is true for the rate of quitting.

Table 11.1 Turnover Generalized Logit and Probit Regression Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Probit Regression Model</th>
<th>Generalized Logit Model</th>
<th>Dismissed</th>
<th>Plant Closed</th>
<th>Other Reason</th>
<th>Quit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tenure</td>
<td>-0.084</td>
<td>-0.221</td>
<td>-0.086</td>
<td>-0.068</td>
<td>-0.127</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.010)</td>
<td>(0.025)</td>
<td>(0.019)</td>
<td>(0.020)</td>
<td>(0.012)</td>
<td></td>
</tr>
<tr>
<td>Time Trend</td>
<td>-0.002</td>
<td>-0.008</td>
<td>-0.024</td>
<td>0.011</td>
<td>-0.022</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.005)</td>
<td>(0.011)</td>
<td>(0.016)</td>
<td>(0.013)</td>
<td>(0.007)</td>
<td></td>
</tr>
<tr>
<td>Tenure × (Time Trend)</td>
<td>0.003</td>
<td>0.008</td>
<td>0.004</td>
<td>-0.005</td>
<td>0.006</td>
<td></td>
</tr>
<tr>
<td>Change in Logarithmic Sector</td>
<td>0.094</td>
<td>-0.286</td>
<td>0.459</td>
<td>-0.022</td>
<td>0.333</td>
<td></td>
</tr>
<tr>
<td>Employment</td>
<td>(0.057)</td>
<td>(0.123)</td>
<td>(0.189)</td>
<td>(0.158)</td>
<td>(0.082)</td>
<td></td>
</tr>
<tr>
<td>Tenure × (Change in Logarithmic</td>
<td>-0.020</td>
<td>-0.061</td>
<td>-0.053</td>
<td>-0.005</td>
<td>-0.027</td>
<td></td>
</tr>
<tr>
<td>Sector Employment</td>
<td>(0.009)</td>
<td>(0.023)</td>
<td>(0.025)</td>
<td>(0.025)</td>
<td>(0.012)</td>
<td></td>
</tr>
</tbody>
</table>

Notes: Standard errors in parentheses. Omitted category is no change in employment. Other variables controlled for consisted of education, marital status, number of children, race, years of full-time work experience and its square, union membership, government employment, logarithmic wage, the U.S. employment rate and location.

11.1.3 Multinomial (conditional) logit

Similar to equation (11.1), an alternative linear combination of explanatory variables is

$$V_{it,j} = x_{it,j}^\prime \beta.$$  \hspace{1cm} (11.3)

Here, $x_{it,j}$ is a vector of explanatory variables that depends on the $j$th alternative, whereas the parameters $\beta$ do not. Using the expression in equation (11.2) is the basis for the multinomial logit model, also known as the conditional logit model. With this specification, the total log-likelihood is

$$L(\beta) = \sum_{it} \left[ \sum_{j=1}^{c} y_{it,j} \ln \pi_{it,j} \right] = \sum_{it} \left[ \sum_{j=1}^{c} y_{it,j} x_{it,j}^\prime \beta - \ln \left( \sum_{k=1}^{c} \exp \left( x_{it,k}^\prime \beta \right) \right) \right].$$  \hspace{1cm} (11.4)

This straightforward expression for the likelihood enables maximum likelihood inference to be easily performed.

The generalized logit model is a special case of the multinomial logit model. To see this, consider explanatory variables $x_{it}$ and parameters $\beta_{i}$, each of dimension $K \times 1$. Define

$$x_{it,j} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ x_{it} \end{pmatrix} \quad \text{and} \quad \beta = \begin{pmatrix} \beta_{1} \\ \beta_{2} \\ \vdots \\ \beta_{c} \end{pmatrix}.$$
Specifically, $x_{it,j}$ is defined as $j-1$ zero vectors (each of dimension $K \times 1$), followed by $x_t$ and then followed by $c-j$ zero vectors. With this specification, we have $x_{it,j}' \beta = x_t' \beta_j$. Thus, a statistical package that performs multinomial logit estimation can also perform generalized logit estimation through the appropriate coding of explanatory variables and parameters. Another consequence of this connection is that some authors use the descriptor multinomial logit when referring to the Section 11.1.2 generalized logit model.

Moreover, through similar coding schemes, multinomial logit models can also handle linear combinations of the form:

$$
V_{it,j} = x_{it,1,j}' \beta + x_{it,2}' \beta_j.
$$

Here, $x_{it,1,j}$ are explanatory variables that depend on the alternative whereas $x_{it,2}$ do not. Similarly, $\beta_j$ are parameters that depend on the alternative whereas $\beta$ do not. This type of linear combination is the basis of a mixed logit model. As with conditional logits, it is customary to choose one set of parameters as the baseline and specify $\beta_c = 0$ to avoid redundancies.

To interpret parameters for the multinomial logit model, we may compare alternatives $h$ and $k$ using equations (11.2) and (11.3), to get

$$
\ln \left( \frac{\text{Prob}(y_{it} = h)}{\text{Prob}(y_{it} = k)} \right) = (x_{it,h} - x_{it,k})' \beta.
$$

Thus, we may interpret $\beta_j$ as the proportional change in the odds ratio, where the change is the value of the $j$th explanatory variable, moving from the $k$th to the $h$th alternative.

With equation (11.2), note that $\pi_{it,1} / \pi_{it,2} = \exp(V_{it,1}) / \exp(V_{it,2})$. This ratio does not depend on the underlying values of the other alternatives, $V_{it,j}$, for $j=3, ..., c$. This feature, called the independence of irrelevant alternatives, can be a drawback of the multinomial logit model for some applications.

**Example 11.3 Choice of yogurt brands**

We now consider a marketing data set introduced by Jain et al. (1994O) that was further analyzed by Chen and Kuo (2001S). These data, obtained from A. C. Nielsen, are known as scanner data because they are obtained from optical scanning of grocery purchases at check-out. The subjects consist of $n=100$ households in Springfield, Missouri. The response of interest is the type of yogurt purchased, consisting of four brands: Yoplait, Dannon, Weight Watchers and Hiland. The households were monitored over a two-year period with the number of purchases ranging from 4 to 185; the total number of purchases is $N=2,412$.

The two marketing variables of interest are PRICE and FEATURES. For the brand purchased, PRICE is recorded as price paid, that is, the shelf price net of the value of coupons redeemed. For other brands, PRICE is the shelf price. FEATURES is a binary variable, defined to be one if there was a newspaper feature advertising the brand at time of purchase, and zero otherwise. Note that the explanatory variables vary by alternative, suggesting the use of a multinomial (conditional) logit model.

Tables 11.2 and 11.3 summarize some important aspects of the data. Table 11.2 shows that Yoplait was the most frequently (33.9%) selected type of yogurt in our sample whereas Hiland was the least frequently selected (2.9%). Yoplait was also the most heavily advertised, appearing in newspaper advertisements 5.6% of the time that the brand was chosen. Table 11.3 shows that Yoplait was also the most expensive, costing 10.6 cents per ounce, on average. Table 11.3 also shows that there are several prices that were far below the average, suggesting some potential influential observations.
TABLE 11.2 Summary Statistics by Choice of Yogurt

<table>
<thead>
<tr>
<th>Summary Statistics</th>
<th>Yoplait</th>
<th>Dannon</th>
<th>Weight Watchers</th>
<th>Hiland</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Choices</td>
<td>818</td>
<td>970</td>
<td>553</td>
<td>71</td>
<td>2,412</td>
</tr>
<tr>
<td>Number of Choices, in Percent</td>
<td>33.9</td>
<td>40.2</td>
<td>22.9</td>
<td>2.9</td>
<td>100.0</td>
</tr>
<tr>
<td>Feature Averages, in Percent</td>
<td>5.6</td>
<td>3.8</td>
<td>3.0</td>
<td>3.7</td>
<td>4.4</td>
</tr>
</tbody>
</table>

Table 11.3 Summary Statistics for Prices

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Median</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yoplait</td>
<td>0.107</td>
<td>0.108</td>
<td>0.003</td>
<td>0.193</td>
<td>0.019</td>
</tr>
<tr>
<td>Dannon</td>
<td>0.082</td>
<td>0.086</td>
<td>0.019</td>
<td>0.111</td>
<td>0.011</td>
</tr>
<tr>
<td>Weight Watchers</td>
<td>0.079</td>
<td>0.079</td>
<td>0.004</td>
<td>0.104</td>
<td>0.008</td>
</tr>
<tr>
<td>Hiland</td>
<td>0.054</td>
<td>0.054</td>
<td>0.025</td>
<td>0.086</td>
<td>0.008</td>
</tr>
</tbody>
</table>

A multinomial logit model was fit to the data, using the following specification for the systematic component

\[ V_{it,j} = \alpha_j + \beta_1 \text{PRICE}_{it} + \beta_2 \text{FEATURE}_{it,j}, \]

using Hiland as the omitted alternative. The results are summarized in Table 11.4. Here, we see that each parameter is statistically significantly different from zero. Thus, the parameter estimates may be useful when predicting the probability of choosing a brand of yogurt. Moreover, in a marketing context, the coefficients have important substantive interpretations. Specifically, we interpret the coefficient associated with FEATURES to suggest that a consumer is \(\exp(0.4914) = 1.634\) times more likely to purchase a product that is featured in a newspaper ad compared to one that is not. For the \(\text{PRICE}\) coefficient, a one cent decrease in price suggests that a consumer is \(\exp(0.3666) = 1.443\) times more likely to purchase a brand of yogurt.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Parameter estimate</th>
<th>t-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yoplait</td>
<td>4.450</td>
<td>23.78</td>
</tr>
<tr>
<td>Dannon</td>
<td>3.716</td>
<td>25.55</td>
</tr>
<tr>
<td>Weight Watchers</td>
<td>3.074</td>
<td>21.15</td>
</tr>
<tr>
<td>FEATURES</td>
<td>0.491</td>
<td>4.09</td>
</tr>
<tr>
<td>PRICE</td>
<td>-36.658</td>
<td>-15.04</td>
</tr>
<tr>
<td>-2 Log Likelihood</td>
<td>10,148</td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>10,138</td>
<td></td>
</tr>
</tbody>
</table>

11.1.4 Random utility interpretation

In economic applications, we think of an individual choosing among \(c\) categories where preferences among categories are determined by an unobserved utility function. For the \(i\)th individual at the \(t\)th time period, use \(U_{it,j}\) for the utility of the \(j\)th choice. To illustrate, assume that the individual chooses the first category if \(U_{it,1} > U_{it,j}\) for \(j=2, \ldots, c\) and denote this choice as \(y_{it} = 1\). We model utility as an underlying value plus random noise, that is, \(U_{it,j} = V_{it,j} + \epsilon_{it,j}\), where \(V_{it,j}\) is specified in equation (11.4). The noise variable is assumed to have an extreme-value distribution of the form

\[ F(a) = \text{Prob}(\epsilon_{it,j} \leq a) = \exp(-e^a). \]
This form is computationally convenient. Omitting the observation-level subscripts \{i,t\} for the moment, we have

\[
\text{Prob}(y = 1) = \text{Prob}(U_1 > U_j \text{ for } j = 2, \ldots, c) = \text{Prob}(\epsilon_j < \epsilon_1 + V_j - V_j \text{ for } j = 2, \ldots, c)
\]

\[
= E\{\text{Prob}(\epsilon_j < \epsilon_1 + V_j - V_j \text{ for } j = 2, \ldots, c | \epsilon_1)\} = E\{F(\epsilon_1 + V_1 - V_j) \cdots F(\epsilon_1 + V_1 - V_c)\}
\]

\[
= E \exp[\exp(- (\epsilon_1 + V_1 - V_j)) + \cdots + \exp(- (\epsilon_1 + V_1 - V_c))]
\]

\[
= E \exp[k \exp(- \epsilon_1)],
\]

where \( k = \sum_{j=2}^{c} \exp(V_j - V_1) \). Now, it is a pleasant exercise in calculus to show, with the distribution function given above, that \( E \exp[k \exp(- \epsilon_1)] = \frac{1}{k+1} \). Thus, we have

\[
\text{Prob}(y = 1) = \frac{1}{1 + \sum_{j=2}^{c} \exp(V_j - V_1) \sum_{j=1}^{c} \exp(V_j)}.
\]

Because this argument is valid for all alternatives \( j = 1, 2, \ldots, c \), the random utility representation yields the multinomial logit model.

### 11.1.5 Nested logit

To mitigate the problem of independence of irrelevant alternatives, we now introduce a type of hierarchical model known as a **nested logit model**. To interpret the nested logit model, in the first stage one chooses an alternative (say the first type) with probability

\[
\pi_{it,1} = \text{Prob}(y_{it} = 1) = \frac{\exp(V_{it,1})}{\sum_{j=2}^{c} \exp(V_{it,j})},
\]

Then, conditional on not choosing the first alternative, the probability of choosing the any one of the other alternatives follows a multinomial logit model with probabilities

\[
\frac{\pi_{it,j}}{1 - \pi_{it,1}} = \text{Prob}(y_{it} = j | y_{it} \neq 1) = \frac{\exp(V_{it,j} / \rho)}{\sum_{k=2}^{c} \exp(V_{it,k} / \rho)}, \quad j = 2, \ldots, c.
\]

In equations (11.7) and (11.8), the parameter \( \rho \) measures the association among the choices \( j = 2, \ldots, c \). The value of \( \rho = 1 \) reduces to the multinomial logit model that we interpret to mean independence of irrelevant alternatives. We also interpret \( \text{Prob}(y_{it} = 1) \) to be a weighted average of values from the first choice and the others. Conditional on not choosing the first category, the form of \( \text{Prob}(y_{it} = j | y_{it} \neq 1) \) has the same form as the multinomial logit.

The advantage of the nested logit is that it generalizes the multinomial logit model in a way such that we no longer have the problem of independence of irrelevant alternatives. A disadvantage, pointed out by McFadden (1981E), is that only one choice is observed; thus, we do not know which category belongs in the first stage of the nesting without additional theory regarding choice behavior. Nonetheless, the nested logit generalizes the multinomial logit by allowing alternative “dependence” structures. That is, one may view the nested logit as a robust alternative to the multinomial logit and examine each one of the categories in the first stage of the nesting.
11.1.6 Generalized extreme value distribution

The nested logit model can also be given a random utility interpretation. To this end, return to the random utility model but assume that the choices are related through a dependence within the error structure. McFadden (1978E) introduced the \textit{generalized extreme-value distribution} of the form:

\[
F(a_1, \ldots, a_c) = \exp\left\{ - G(e^{-a_1}, \ldots, e^{-a_c}) \right\}.
\]

Under regularity conditions on \( G \), McFadden (1978E) showed that this yields

\[
\text{Prob}(y = j) = \text{Prob}(U_j > U_k \text{ for } k = 1, \ldots, c, k \neq j) = \frac{e^{V_j}}{G(e^{-V_1}, \ldots, e^{-V_c})},
\]

where \( G_j(x_1, \ldots, x_c) = \frac{\partial}{\partial x_j} G(x_1, \ldots, x_c) \) is the \( j \)th partial derivative of \( G \).

\section*{Special cases}

1. Let \( G(x_1, \ldots, x_c) = x_1 + \ldots + x_c \). In this case, \( G_j = 1 \) and \( \text{Prob}(y = j) = \frac{\exp(V_j)}{\sum_{k=1}^c \exp(V_k)} \). This is the multinomial logit case.

2. Let \( G(x_1, \ldots, x_c) = x_1 + \left( \sum_{k=2}^c x_k^{1/\rho} \right)^\rho \). In this case, \( G_1 = 1 \) and

\[
\text{Prob}(y = 1) = \frac{\exp(V_1)}{\exp(V_1) + \left( \sum_{k=2}^c \exp(V_k / \rho) \right)^\rho}.
\]

Additional calculations show that

\[
\text{Prob}(y = j | y \neq 1) = \frac{\exp(V_j / \rho)}{\sum_{k=2}^c \exp(V_k / \rho)}.
\]

This is the nested logit case.

Thus, the generalized extreme-value distribution provides a framework that encompasses the multinomial and conditional logit models. Amemiya (1985E) provides background on more complex nested models that utilize the generalized extreme-value distribution.

11.2 Multinomial logit models with random effects

Repeated observations from an individual tend to be similar; in the case of categorical choices, this means that individuals tend to make the same choices from one observation to the next. This section models that similarity through a common heterogeneity term. To this end, we augment our systematic component with a heterogeneity term and, similar to equation (11.4), consider linear combinations of explanatory variables of the form

\[
V_{it,j} = z'_{it,j} \alpha_i + x'_{it,j} \beta_j.
\]
As before, \( \alpha_i \) represents the heterogeneity term that is subject-specific. The form of equation (11.9) is quite general and includes many applications of interest. However, to develop intuition, we focus on the special case

\[
V_{it,j} = \alpha_{ij} + x'_{it,j} \beta.
\]  

(11.10)

Here, intercepts vary by individual and alternative but are common over time. With this specification for the systematic component, the conditional (on the heterogeneity) probability that the \( i \)th subject at time \( t \) chooses the \( j \)th alternative is

\[
\pi_{ij}(a_i) = \frac{\exp(V_{it,j})}{\sum_{k=1}^{c} \exp(V'_{it,k})} = \frac{\exp(\alpha_{ij} + x'_{it,j} \beta)}{\sum_{k=1}^{c} \exp(\alpha_{ik} + x'_{it,k} \beta)}, \quad j = 1, 2, \ldots, c,
\]  

(11.11)

where we now denote the set of heterogeneity terms as \( \alpha_i = (\alpha_{i1}, \ldots, \alpha_{ic})' \). From the form of this equation, we see that a heterogeneity term that is constant over alternatives \( j \) does not affect the conditional probability. To avoid parameter redundancies, a convenient normalization is to take \( \alpha_{ic} = 0 \).

For statistical inference, we begin with likelihood equations. Similar to the development in Section 11.1.1, the conditional likelihood for the \( i \)th subject is

\[
L(y_i | a_i) = \prod_{t=1}^{T_i} \prod_{j=1}^{c} \left( \pi_{it,j}(a_i) \right)^{y_{it,j}} \prod_{t=1}^{T_i} \left[ \frac{\exp\left(\sum_{j=1}^{c} V_{it,j}(a_{ij} + x'_{it,j} \beta)\right)}{\sum_{k=1}^{c} \exp(\alpha_{ik} + x'_{it,k} \beta)} \right].
\]  

(11.12)

We assume that \( \{a_i\} \) is i.i.d. with distribution function \( G_a \), that is typically taken to be multivariate normal. With this convention, the (unconditional) likelihood for the \( i \)th subject is

\[
L(y_i) = \int L(y_i | a) dG_a(a).
\]

Assuming independence among subjects, the total log-likelihood is \( L = \sum_i L(y_i) \).

**Relation with nonlinear random effects Poisson model**

With this framework, standard maximum likelihood estimation is available. However, for applied work, there are relatively few statistical packages available for estimating multinomial logit models with random effects. As an alternative, one can look to properties of the multinomial distribution and link it to other distributions. Chen and Kuo (2001S) recently surveyed these linkages in the context of random effects models and we now present one link to a nonlinear Poisson model. Statistical packages for nonlinear Poisson models are readily available; with this link, they can be used to estimate parameters of the multinomial logit model with random effects.

To this end, an analyst would instruct a statistical package to “assume” that the binary random variables \( y_{it,j} \) are Poisson distributed with conditional means \( \pi_{it,j} \) and, conditional on the heterogeneity terms, are independent. This is a nonlinear Poisson model because, from Section 10.5.3, a linear Poisson model takes the logarithmic (conditional) mean to be a linear function of explanatory variables. In contrast, from equation (11.11), log \( \pi_{it,j} \) is a nonlinear function. Of course, this “assumption” is not valid. Binary random variables have only two outcomes and thus cannot have a Poisson distribution. Moreover, the binary variables must sum to one (that is, \( \sum_j y_{it,j} = 1 \)) and thus are not even conditionally independent. Nonetheless, with this “assumption” and the Poisson distribution (reviewed in Section 10.5.3), the conditional likelihood interpreted by the statistical package is:
Up to the constant, this is the same conditional likelihood as in equation (11.12) (see Exercise 10.1). Thus, a statistical package that performs nonlinear Poisson models with random effects can be used to get maximum likelihood estimates for the multinomial logit model with random effects. See Chen and Kuo (2001S) for a related algorithm based on a linear Poisson model with random effects.

**Example 11.3 Choice of yogurt brands - Continued**

To illustrate, we used a multinomial logit model with random effects on the yogurt data introduced in Example 11.1. Following Chen and Kuo (2001S), random intercepts for Yoplait, Dannon and Weight Watchers were assumed to follow a multivariate normal distribution with an unstructured covariance matrix. Table 11.5 shows results from fitting this model, based on the nonlinear Poisson model link and using SAS PROC NLMIXED. Here, we see that the coefficients for FEATURES and PRICE are qualitatively similar to the model without random effects, reproduced for convenience from Table 11.4. They are qualitatively similar in the sense that they have the same sign and same degree of statistical significance. Overall, the AIC statistic suggests that the model with random effects is the preferred model.

Table 11.5 Yogurt Multinomial Logit Model Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Without Random Effects</th>
<th>With Random Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Parameter estimate</td>
<td>t-statistic</td>
</tr>
<tr>
<td>Yoplait</td>
<td>4.450</td>
<td>23.78</td>
</tr>
<tr>
<td>Dannon</td>
<td>3.716</td>
<td>25.55</td>
</tr>
<tr>
<td>Weight Watchers</td>
<td>3.074</td>
<td>21.15</td>
</tr>
<tr>
<td>FEATURES</td>
<td>0.491</td>
<td>4.09</td>
</tr>
<tr>
<td>PRICE</td>
<td>-36.658</td>
<td>-15.04</td>
</tr>
<tr>
<td>-2 Log Likelihood</td>
<td>10,148</td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>10,138</td>
<td></td>
</tr>
</tbody>
</table>

As with binary dependent variables, conditional maximum likelihood estimators have been proposed; see, for example, Conaway (1989S). Appendix 11A provides a brief introduction to these alternative estimators.

### 11.3 Transition (Markov) Models

Another way of accounting for heterogeneity is to trace the development of a dependent variable over time and represent the distribution of its current value as a function of its history. To this end, define $H_{it}$ to be the history of the $i$th subject up to time $t$. For example, if the explanatory variables are assumed to be non-stochastic, then we might use $H_{it} = \{y_{i1}, \ldots, y_{it}\}$. With this information set, we may partition the likelihood for the $i$th subject is

$$L(y_i | a_i) = \prod_{t=2}^{T_t} \prod_{j=1}^{c} \left( \frac{\pi_{i,t,j}(a_i)^{y_{i,t,j}}}{y_{i,t,j}} \exp\left(-\pi_{i,t,j}(a_i)\right) \right) e^{-1}.$$  

(11.13)
Chapter 11. Categorical Dependent Variables and Survival Models / 11-11

where \( f(y_{it} \mid H_{it}) \) is the conditional distribution of \( y_{it} \) given its history and \( f(y_{i1}) \) is the marginal distribution of \( y_{i1} \). To illustrate, one type of application is through a conditional generalized linear model (GLM) of the form

\[
\begin{align*}
\logit(\pi_{it,j} / \phi) &= \beta_j x_{it} + \sum_k \phi_{j,k} y_{it-k} \\
\pi_{it,j} &= \frac{e^{\beta_j x_{it} + \sum_k \phi_{j,k} y_{it-k}}}{1 + \sum_{k \neq j} e^{\beta_j x_{it} + \sum_k \phi_{j,k} y_{it-k}}} 
\end{align*}
\]

where \( \text{E}(y_{it} \mid H_{it}) = b(\theta_{it}) \) and \( \text{Var}(y_{it} \mid H_{it}) = \phi b^*(\theta_{it}) \). Assuming a canonical link, for the systematic component, one could use

\[
\theta_{it} = g(\text{E}(y_{it} \mid H_{it})) = x_{it}^T \beta + \sum_j \phi_{j} y_{it-j}.
\]

See Diggle, Heagerty, Liang and Zeger (2002S, Chapter 10) for further applications of and references to general transition GLMs. We focus on categorical responses.

**Unordered categorical response**

To simplify our discussion of unordered categorical responses, we also assume discrete unit time intervals. To begin, we consider Markov models of order 1. Thus, the history \( H_{it} \) need only contain \( y_{i,t-1} \). More formally, we assume that

\[
\pi_{it,j} = \text{Prob}(y_{it} = j \mid y_{i,t-1} = \cdot) = \text{Prob}(y_{it} = j \mid \{y_{i,t-1} = \cdot, y_{i,t-2}, \ldots, y_{i,1}\}).
\]

That is, given the information in \( y_{i,t-1} \), there is no additional information content in \( \{y_{i,t-2}, \ldots, y_{i,1}\} \) about the distribution of \( y_{it} \).

Without covariate information, it is customary to organize the set of transition probabilities \( \pi_{it,j} \) as a matrix of the form

\[
\Pi_{it} = \begin{pmatrix}
\pi_{it,11} & \pi_{it,12} & \cdots & \pi_{it,1c} \\
\pi_{it,21} & \pi_{it,22} & \cdots & \pi_{it,2c} \\
\vdots & \vdots & \ddots & \vdots \\
\pi_{it,c1} & \pi_{it,c2} & \cdots & \pi_{it,cc}
\end{pmatrix}
\]

Here, each row sums to one. With covariate information and an initial state distribution \( \text{Prob}(y_{i1}) \), one can trace the history of the process knowing only the transition matrix \( \Pi_{it} \). We call the row identifier, \( j \), the state of origin and the column identifier, \( k \), the destination state.

For complex transition models, it can be useful to graphically summarize the set of feasible transitions under consideration. To illustrate, Figure 11.1 summarizes an employee retirement system with \( c = 4 \) categories. Here,

- 1 denotes active continuation in the pension plan,
- 2 denotes retirement from the pension plan,
- 3 denotes withdrawal from the pension plan and
- 4 denotes death.

For this system, the circles represent the nodes of the graph and correspond to the response categories. The arrows, or arcs, indicate the modes of possible transitions. This graph indicates that movement from state 1 to states 1, 2, 3 or 4 is possible, so that we would assume \( \pi_{ij} \geq 0 \), for \( j = 1, 2, 3, 4 \). However, once an individual is in states 2, 3, or 4, it is not possible to move from those states (known as absorbing states). Thus, we use \( \pi_{ij} = 1 \) for \( j = 2, 3, 4 \) and \( \pi_{jk} = 0 \), for \( j = 2, 3, 4 \) and \( k \neq j \). Note that although death is certainly possible (and even eventually certain) for those in retirement, we assume \( \pi_{24} = 0 \) with the understanding that the plan has paid pension benefits at retirements and need no longer be concerned with additional transitions after exiting.
the plan, regardless of the reason. This is assumption is often convenient because it is difficult to track individuals having left active membership in a benefit plan.

Figure 11.1 Graphical Summary of a Transition Model for a Hypothetical Employment Retirement System.

For another example, consider the modification summarized in Figure 11.2. Here, we see that retirees are now permitted to re-enter the workforce so that $\pi_{21}$ may be positive. Moreover, now the transition from retirement to death is also explicitly accounted for so that $\pi_{24} \geq 0$. This may be of interest in a system that pays retirement benefits as long as a retiree lives. We refer to Haberman and Pitacco (1999O) for many additional examples of Markov transition models that are of interest in employee benefit and other types of actuarial systems.

Figure 11.2 A Modified Transition Model for a Employment Retirement System.

We can parameterize the problem by choosing a multinomial logit, one for each state of origin. Thus, we use

$$
\pi_{it,jk} = \frac{\exp(V_{it,jk})}{\sum_{h=1}^{c} \exp(V_{it,jh})},
$$

where the systematic component $V_{it,jk}$ is given by

$$
V_{it,jk} = x'_{it,jk} \beta_j.
$$
As discussed in the context of employment retirement systems, in a given problem one assumes that a certain subset of transition probabilities are zero, thus constraining the estimation of $\beta_j$.

For estimation, we may proceed as in Section 11.1. Define

$$y_{it,jk} = \begin{cases} 1 & \text{if } y_{it} = k \text{ and } y_{it-1} = j \\ 0 & \text{otherwise} \end{cases}$$

With this notation, the conditional likelihood is

$$f(y_{it} | y_{it-1}) = \prod_{j=1}^c \prod_{k=1}^c (\pi_{it,jk})^{y_{it,jk}}.$$  (11.16)

Here, in the case that $\pi_{it,jk} = 0$ (by assumption), we have that $y_{it,jk} = 0$ and use the convention that $0^0 = 1$.

To simplify matters, we assume that the initial state distribution, Prob($y_{i1}$), is described by a different set of parameters than the transition distribution, $f(y_{it} | y_{it-1})$. Thus, to estimate this latter set of parameters, one only needs to maximize the partial log-likelihood

$$L_p = \sum_{t=2}^T \sum_{i} \ln f(y_{it} | y_{it-1})$$  (11.17)

where $f(y_{it} | y_{it-1})$ is specified in equation (11.16). In some cases, the interesting aspect of the problem is the transition. In this case, one loses little by focusing on the partial likelihood. In other cases, the interesting aspect is the state, such as the proportion of retirements at a certain age. Here, a representation for the initial state distribution takes on greater importance.

In equation (11.15), we specified separate components for each alternative. Assuming no implicit relationship among the components, this specification yields a particularly simple analysis. That is, we may write the partial log-likelihood as

$$L_p = \sum_{j=1}^c L_{p,j}(\beta_j)$$

where, from equations (11.14)-(11.16), we have

$$L_{p,j}(\beta_j) = \sum_{i} \sum_{t=2}^T \ln f(y_{it} | y_{it-1} = j) = \sum_{i} \sum_{t=2}^T \left( \sum_{k=1}^c y_{it,jk} x_{it,jk} \beta_j - \ln \left( \sum_{k=1}^c \exp(x_{it,jk} \beta_j) \right) \right)$$

as in equation (11.4). Thus, we can split up the data according to each (lagged) choice and determine maximum likelihood estimators for each alternative, in isolation of the others.

**Example 11.3 Choice of yogurt brands - Continued**

To illustrate, we return to the Yogurt data set. We now explicitly model the transitions between brand choices, as denoted in Figure 11.3. Here, purchases of yogurt occur intermittently over a two-year period; the data are not observed at discrete time intervals. By ignoring the length of time between purchases, we are using what is sometimes referred to as “operational time.” In effect, we are assuming that one’s most recent choice of a brand of yogurt has the same effect on today’s choice, regardless as to whether the prior choice was one day or one month ago. This assumption suggests future refinements to the transition approach in modeling yogurt choice.
Tables 11.6a and 11.6b show the relation between the current and most recent choice of yogurt brands. Here, we call the most recent choice the “origin state” and the most current choice the “destination state.” Table 11.6a shows that there are only 2,312 observations under consideration; this is because initial values from each of 100 subjects are not available for the transition analysis. For most observation pairs, the current choice of the brand of yogurt is the same as chosen most recently, exhibiting “brand loyalty.” Other observation pairs can be described as “switchers.” Brand loyalty and switching behavior is more apparent in Table 11.6b, where we rescale counts by row totals to give (rough) empirical transition probabilities. Here, we see that customers of Yoplait, Dannon and Weight Watchers exhibit more brand loyalty compared to those of Hiland who are more prone to switching.

<table>
<thead>
<tr>
<th>Destination State</th>
<th>Yoplait</th>
<th>Dannon</th>
<th>Weight Watchers</th>
<th>Hiland</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yoplait</td>
<td>654</td>
<td>65</td>
<td>41</td>
<td>17</td>
<td>777</td>
</tr>
<tr>
<td>Dannon</td>
<td>71</td>
<td>822</td>
<td>19</td>
<td>16</td>
<td>928</td>
</tr>
<tr>
<td>Weight Watchers</td>
<td>44</td>
<td>18</td>
<td>473</td>
<td>5</td>
<td>540</td>
</tr>
<tr>
<td>Hiland</td>
<td>14</td>
<td>17</td>
<td>6</td>
<td>30</td>
<td>67</td>
</tr>
<tr>
<td>Total</td>
<td>783</td>
<td>922</td>
<td>539</td>
<td>68</td>
<td>2,312</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Destination State</th>
<th>Yoplait</th>
<th>Dannon</th>
<th>Weight Watchers</th>
<th>Hiland</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yoplait</td>
<td>84.2</td>
<td>8.4</td>
<td>5.3</td>
<td>2.2</td>
<td>100.0</td>
</tr>
<tr>
<td>Dannon</td>
<td>7.7</td>
<td>88.6</td>
<td>2.0</td>
<td>1.7</td>
<td>100.0</td>
</tr>
<tr>
<td>Weight Watchers</td>
<td>8.1</td>
<td>3.3</td>
<td>87.6</td>
<td>0.9</td>
<td>100.0</td>
</tr>
<tr>
<td>Hiland</td>
<td>20.9</td>
<td>25.4</td>
<td>9.0</td>
<td>44.8</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Of course, Tables 11.6a and 11.6b do not account for changing aspects of price and features. In contrast, these explanatory variables are captured in the multinomial logit fit, displayed in Table 11.7. Table 11.7 shows that purchase probabilities for customers of Dannon, Weight Watchers and Hiland are more responsive to a newspaper ad than Yoplait customers.
Moreover, compared to the other three brands, Hiland customers are not price sensitive in that changes in PRICE have relatively impact on the purchase probability (it is not even statistically significant).

Table 11.6b suggests that prior purchase information is important when estimating purchase probabilities. To test this, it is straightforward use a likelihood ratio test of the null hypothesis $H_0: \beta_j = \beta$, that is, the components do not vary by origin state. Table 11.7 shows that the total (minus two times the partial) log-likelihood is $2,379.8 + \ldots + 281.5 = 6,850.3$. Estimation of this model under the null hypothesis yields a corresponding value of 9,741.2. Thus, the likelihood ratio test statistic is $LRT = 2,890.9$. There are 15 degrees of freedom for this test statistic. Thus, this provides strong evidence for rejecting the null hypothesis, corroborating the intuition that the most recent type of purchase has a strong influence in the current brand choice.

<table>
<thead>
<tr>
<th>Table 11.7 Yogurt Transition Model Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>State of Origin</td>
</tr>
<tr>
<td>Variable</td>
</tr>
<tr>
<td>Yoplait</td>
</tr>
<tr>
<td>Dannon</td>
</tr>
<tr>
<td>Weight Watchers</td>
</tr>
<tr>
<td>FEATURES</td>
</tr>
<tr>
<td>PRICE</td>
</tr>
<tr>
<td>-2 Log Likelihood</td>
</tr>
</tbody>
</table>

Example 11.3 on the choice of yogurt brands illustrated an application of a conditional (multinomial) logit model where the explanatory variables FEATURES and PRICE depend on the alternative. To provide an example where this is not so, we return to Example 9.1.3 on the choice of a professional tax preparer. Here, financial and demographic characteristics do not depend on the alternative and so we apply a straightforward logit model.

**Example 11.4 Income tax payments and tax preparers**

This is a continuation of the Chapter 9 Example 9.1.3 on the choice of whether or not a tax filer will use a professional tax preparer. Our Chapter 9 analysis of this data indicated strong similarities in responses within a subject. In fact, our Section 9.3 fixed effects analysis indicated that 97 tax filers never used a preparer in the five years under consideration and 89 always did, out of 258 subjects. We now model these time transitions explicitly in lieu of using a time constant latent variable $\alpha_i$ to account for this similarity.

Table 11.8 shows the relationship between the current and most recent choice of preparer. Although we began with 1,290 observations, 258 initial observations are not available for the transition analysis, reducing our sample size to $1,290 - 258 = 1,032$. Table 11.8 strongly suggests that the most recent choice is an important predictor of the current choice.

| Table 11.8 Tax Preparers Transition Empirical Probabilities, in Percent |
|-----------------------------|----------------|-------------|
| Origin State               |               | Destination State |
| Count                       | PREP = 0 | PREP = 1 |
| PREP = 0                    | 546      | 89.4       | 10.6       |
| PREP = 1                    | 486      | 8.4        | 91.6       |
Table 11.9 provides a more formal assessment with a fit of a logit transition model. To assess whether or not the transition aspect is an important piece of the model, we can use a likelihood ratio test of the null hypothesis \( \beta_j = \beta \), that is, the coefficients do not vary by origin state. Table 11.9 shows that the total (minus two times the partial) log-likelihood is 361.5 + 264.6 = 626.1. Estimation of this model under the null hypothesis yields a corresponding value of 1,380.3. Thus, the likelihood ratio test statistic is \( LRT = 754.2 \). There are 4 degrees of freedom for this test statistic. Thus, this provides strong evidence for rejecting the null hypothesis, corroborating the intuition that most recent choice is an important predictor of the current choice.

To interpret the regression coefficients in Table 11.9, we use the summary statistics in Section 9.1.3 to describe a “typical” tax filer and assume that LNTP1 = 10, MR = 23 and EMP = 0. If this tax filer had not previously chosen to use a preparer, the estimated systematic component is \( V = -10.704 + 1.024(10) - 0.072(23) + 0.352(0) = -2.12 \). Thus, the estimated probability of choosing to use a preparer is \( \exp(-2.12)/(1+\exp(-2.12)) = 0.107 \). Similar calculations show that, if this tax filer had chosen to use a preparer, then the estimated probability is 0.911. These calculations are in accord with the estimates in Table 11.8 that do not account for the explanatory variables. This illustration points out the importance of the intercept in determining these estimated probabilities.

<table>
<thead>
<tr>
<th>Variable</th>
<th>PREP = 0</th>
<th>PREP = 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-10.704</td>
<td>0.208</td>
</tr>
<tr>
<td>LNTP1</td>
<td>1.024</td>
<td>0.104</td>
</tr>
<tr>
<td>MR</td>
<td>-0.072</td>
<td>0.047</td>
</tr>
<tr>
<td>EMP</td>
<td>0.352</td>
<td>0.85</td>
</tr>
<tr>
<td>-2 Log Likelihood</td>
<td>361.5</td>
<td>264.6</td>
</tr>
</tbody>
</table>

Higher order Markov models

There are strong serial relationships in the Taxpayer data and these may not be completely captured by simply looking at the most recent choice. For example, it may be that a tax filer who uses a preparer for two consecutive periods has a substantially different choice probability than a comparable filer who does not use a preparer in one period but elects to use a preparer in the subsequent period.

It is customary in Markov modeling to simply expand the state space to handle higher order time relationships. To this end, we may define a new categorical response, \( y_{it}^* = \{y_{it}, y_{i,t-1}\} \). With this new response, the order 1 transition probability, \( f(y_{it}^* | y_{i,t-1}^*) \), is equivalent to an order 2 transition probability of the original response, \( f(y_{it} | y_{i,t-1}, y_{i,t-2}) \). This is because the conditioning events are the same, \( y_{i,t-1}^* = \{y_{i,t-1}, y_{i,t-2}\} \), and because \( y_{i,t-1} \) is completely determined by the conditioning event \( y_{i,t-1}^* \). Expansions to higher orders can be readily accomplished in a similar fashion.

To simplify the exposition, we consider only binary outcomes so that \( c = 2 \). Examining the transition probability, we are now conditioning on four states, \( y_{i,t-1}^* = \{y_{i,t-1}, y_{i,t-2}\} = \{(0,0), (0,1), (1,0), (1,1)\} \). As above, one can split up the (partial) likelihood into four components, one for each state. Alternatively, one can write the logit model as

\[
\text{Prob}(y_{it} | y_{i,t-1}, y_{i,t-2}) = \logit \left( V_{it} \right)
\]

with
Chapter 11. Categorical Dependent Variables and Survival Models / 11-17

\[ V_{it} = x_{it}^\prime \beta_1 I(y_{i,t-1} = 0, y_{i,t-2} = 0) + x_{it}^\prime \beta_2 I(y_{i,t-1} = 1, y_{i,t-2} = 0) + x_{it}^\prime \beta_3 I(y_{i,t-1} = 0, y_{i,t-2} = 1) + x_{it}^\prime \beta_4 I(y_{i,t-1} = 1, y_{i,t-2} = 1), \]

where \( I(.) \) is the indicator function of a set. The advantage of running the model in this fashion, as compared to splitting it up into four distinct components, is that one can test directly the equality of parameters and consider a reduced parameter set by combining them. The advantage of the alternative approach is computational convenience; one performs a maximization procedure over a smaller data set and a smaller set of parameters, albeit several times.

Example 11.4 Income tax payments and tax preparers - Continued

To investigate the usefulness of a second order component, \( y_{i,t-2} \), in the transition model, we begin in Table 11.10 with empirical transition probabilities. Table 11.10 suggests that there are important differences in the transition probabilities for each lag 1 origin state \( (y_{i,t-1} = \text{Lag PREP}) \) between levels of the lag two origin state \( (y_{i,t-2} = \text{Lag 2 PREP}) \).

Table 11.11 provides a more formal analysis by incorporating potential explanatory variables. The total (minus two times the partial) log-likelihood is 469.4. Estimation of this model under the null hypothesis yields a corresponding value of 1,067.7. Thus, the likelihood ratio test statistic is \( LRT = 567.3 \). There are 12 degrees of freedom for this test statistic. Thus, this provides strong evidence for rejecting the null hypothesis. With this data set, estimation of the model incorporating lag one differences yields total (partial) (minus two times) log-likelihood of 490.2. Thus, the likelihood ratio test statistic is \( LRT = 20.8 \). With 8 degrees of freedom, comparing this test statistic to a chi-square distribution yields a \( p \)-value of 0.0077. Thus, the lag two component is statistically significant contribution to the model.

| Table 11.10 Tax Preparers Order 2 Transition Empirical Probabilities, in Percent |
|-----------------------------|-------------|-----------------|-----------------|
| Origin State | Destination State |
| Lag 2 PREP | Count | PREP = 0 | PREP = 1 |
| 0 | 0 | 376 | 89.1 | 10.9 |
| 0 | 1 | 28 | 67.9 | 32.1 |
| 1 | 0 | 43 | 25.6 | 74.4 |
| 1 | 1 | 327 | 6.1 | 93.9 |

| Table 11.11 Tax Preparers Order 2 Transition Model Estimates |
|-----------------------------|-------------|-----------------|-----------------|
| State of Origin | |
| Lag PREP = 0, Lag 2 PREP = 0 | Lag PREP = 0, Lag 2 PREP = 1 | Lag PREP = 1, Lag 2 PREP = 0 | Lag PREP = 1, Lag 2 PREP = 1 |
| Variable | Estimate | t-stat | Estimate | t-stat | Estimate | t-stat | Estimate | t-stat |
| Intercept | -9.866 | -2.30 | -7.331 | -0.81 | 1.629 | 0.25 | -0.251 | -0.19 |
| LNTPI | 0.923 | 1.84 | 0.675 | 0.63 | -0.210 | -0.27 | 0.197 | 1.21 |
| MR | -0.066 | -1.79 | -0.001 | -0.01 | 0.065 | 0.89 | 0.040 | 1.42 |
| EMP | 0.406 | 0.84 | 0.050 | 0.04 | NA | NA | 1.406 | 1.69 |
| -2 Log Likelihood | 254.2 | 33.4 | 42.7 | 139.1 |
Just as one can incorporate higher order lags into a Markov structure, it is also possible to bring in the time spent in a state. This may be of interest in a model of health states, where we might wish to accommodate the time spent in a “healthy” state or an “at-risk” state. This phenomenon is known as “lagged duration dependence.” Similarly, the transition probabilities may depend on the number of prior occurrences of an event, known as “occurrence dependence.” For example, when modeling employment, we may wish to allow transition probabilities to depend on the number of previous employment spells. For further considerations of these and other specialized transition models, see Lancaster (1990E) and Haberman and Pitacco (1999O).

11.4 Survival Models

Categorical data transition models, where one models the probability of movement from one state to another, are closely related to survival models. In survival models, the dependent variable is the time until an event of interest. The classic example is time until death (the complement of death being survival). Survival models are now widely applied in many scientific disciplines; other examples of events of interest include the onset of Alzheimer’s disease (biomedical), time until bankruptcy (economics) and time until divorce (sociology).

Like the data studied elsewhere in this text, survival data are longitudinal. The cross-sectional aspect typically consists of multiple subjects, such as persons or firms, under study. There may be only one measurement on each subject but the measurement is taken with respect to time. This combination of cross-sectional and temporal aspects gives survival data their longitudinal flavor. Because of the importance of survival models, it is not uncommon for researchers to equate the phrase “longitudinal data” with survival data.

Some events of interest, such as bankruptcy or divorce, may not happen for a specific subject. It is common that an event of interest may not have yet occurred within the study period so that the data are (right) censored. Thus, the complete observation times may not be available due to the design of the study. Moreover, firms may merge or be acquired by other firms and subjects may move from a geographical area, leaving the study. Thus, the data may be incomplete due to events that are extraneous to the research question under consideration, known as random censoring. Censoring is a regular feature of survival data; large values of a dependent variable are more difficult to observe than small values, other things being equal. In Section 7.4 we introduced mechanisms and models for handling incomplete data. For the repeated cross-sectional data described in this text, models for incompleteness have become available only relatively recently (although researchers have long been aware of these issues, focusing on attrition). In contrast, models of incompleteness have been historically important and one of the distinguishing features of survival data.

Some survival models can be written in terms of the Section 11.3 transition models. To illustrate, suppose that \( Y_i \) is the time until an event of interest and, for simplicity, assume that it is discrete positive integer. From knowledge of \( Y_i \), we may define \( y_{it} \) to be one if \( Y_i = t \) and zero otherwise. With this notation, we may write the likelihood

\[
\text{Prob}(Y_i = n) = \text{Prob}(y_{i1} = 0, \ldots, y_{i,n-1} = 0, y_{in} = 1)
\]

\[
= \text{Prob}(y_{i1} = 0 \prod_{t=2}^{n-1} \text{Prob}(y_{it} = 0 | y_{i,t-1} = 0)) \text{Prob}(y_{in} = 1 | y_{i,n-1} = 0),
\]

in terms of transition probabilities \( \text{Prob}(y_{it} | y_{i,t-1}) \) and the initial state distribution \( \text{Prob}(y_{i1}) \). Note that in Section 11.3 we considered \( n \) to be the non-random number of time units under consideration whereas here it is a realized value of a random variable.
Example 11.5 – Time until bankruptcy

Shumway (2001O) examined the time to bankruptcy for 3,182 firms listed on Compustat Industrial File and the CRSP Daily Stock Return File for the New York Stock Exchange over the period 1962-1992. Several explanatory financial variables were examined, including working capital to total assets, retained to total assets, earnings before interest and taxes to total assets, market equity to total liabilities, sales to total assets, net income to total assets, total liabilities to total assets and current assets to current liabilities. The data set included 300 bankruptcies from 39,745 firm years.

See also Kim et al. (1995O) for a similar study on insurance insolvencies.

Survival models are frequently expressed in terms of continuous dependent random variables. To summarize the distribution of \( Y \), define the hazard function

\[
h(t) = \frac{\text{probability density function}}{\text{survival function}} = \frac{-\frac{\partial}{\partial t} \text{Prob}(Y > t)}{\text{Prob}(Y > t)} = -\frac{\partial}{\partial t} \ln \text{Prob}(Y > t),
\]

the “instantaneous” probability of failure, conditional on survivorship up to time \( t \). This is also known as the force of mortality in actuarial science, as well as the failure rate in engineering. A related quantity of interest is the cumulative hazard function,

\[
H(t) = \int_0^t h(s)ds.
\]

This quantity can also be expressed as the minus the log survival function, and conversely, \( \text{Prob}(Y > t) = \exp(-H(t)) \).

Survival models regularly allow for “non-informative” censoring. Thus, define \( \delta \) to be an indicator function for right-censoring, that is,

\[
\delta = \begin{cases} 
1 & \text{if } Y \text{ is censored} \\
0 & \text{otherwise}
\end{cases}.
\]

Then, the likelihood of a realization of \((Y, \delta)\), say \((y, d)\), can be expressed in terms of the hazard function and cumulative hazard as

\[
\begin{cases} 
\text{Prob}(Y > y) & \text{if } Y \text{ is censored} \\
-\frac{\partial}{\partial y} \text{Prob}(Y > y) & \text{otherwise}
\end{cases}
\]

\[
= (\text{Prob}(Y > y))^d \left(h(y)\text{Prob}(Y > y)\right)^{1-d}
\]

\[
= h(y)^{1-d} \exp(-H(y)).
\]

There are two common methods for introducing regression explanatory variables, one is the accelerated failure time model and the other is the Cox proportional hazard model. Under the former, one essentially assumes a linear model in the logarithmic time to failure. We refer to any standard treatment of survival models for more discussion of this mechanism. Under the latter, one assumes that the hazard function can be written as the product of some “baseline” hazard and a function of a linear combination of explanatory variables. To illustrate, we use

\[
h(t) = h_0(t) \exp (x^T \beta), \quad (11.18)
\]
where \( h_0(t) \) is the baseline hazard. This is known as a proportional hazards model because if one takes the ratio of hazard functions for any two sets of covariates, say \( x_1 \) and \( x_2 \), one gets

\[
\frac{h(t | x_1)}{h(t | x_2)} = \frac{h_0(t) \exp(x_1' \beta)}{h_0(t) \exp(x_2' \beta)} = \exp((x_1 - x_2)' \beta),
\]

that the ratio is independent of time \( t \).

To express the likelihood function for the Cox model, let \( H_0 \) be the cumulative hazard function associated with the baseline hazard function \( h_0 \). Let \((Y_1, \delta_1), \ldots, (Y_n, \delta_n)\) be independent and assume that \( Y_i \) follows a Cox proportional hazard model with regressors \( x_i \). Then the likelihood is

\[
L(\beta, h_0) = \prod_{i=1}^{n} h(Y_i)^{1-\delta_i} \exp(-H(Y_i)) = \prod_{i=1}^{n} (h_0(Y_i) \exp(x_i' \beta))^{1-\delta_i} \exp(-H_0(Y_i) \exp(x_i' \beta)).
\]

Maximizing this in terms of \( h_0 \) yields the partial likelihood

\[
L_p(\beta) = \prod_{i=1}^{n} \left( \frac{\exp(x_i' \beta)}{\sum_{j \in R(Y_i)} \exp(x_j' \beta)} \right)^{1-\delta_i}, \tag{11.19}
\]

where \( R(t) \) is the set of all \( \{Y_1, \ldots, Y_n\} \) such that \( Y_i \geq t \), that is, the set of all subjects still under study at time \( t \).

From equation (11.19), we see that inference for the regression coefficients depends only on the ranks of the dependent variables \( \{Y_1, \ldots, Y_n\} \), not their actual values. Moreover, equation (11.19) suggests (and it is true) that large sample distribution theory has properties similar to the usual desirable (fully) parametric theory. This is mildly surprising because the proportional hazards model is semi-parametric; in equation (11.18) the hazard function has a fully parametric component, \( \exp(x_i' \beta) \), but also contains a nonparametric baseline hazard, \( h_0(t) \). In general, nonparametric models are more flexible than parametric counterparts for model fitting but result in less desirable large sample properties (specifically, slower rates of convergence to an asymptotic distribution).

An important feature of the proportional hazards model is that it can readily handle time dependent covariates of the form \( x_i(t) \). In this case, one can write the partial likelihood as

\[
L_p(\beta) = \prod_{i=1}^{n} \left( \frac{\exp(x_i' \beta)}{\sum_{j \in R(Y_i)} \exp(x_j' \beta)} \right)^{1-\delta_i}.
\]

Maximization of this likelihood is somewhat complex but can be readily accomplished with modern statistical software.

To summarize, there is a large overlap between survival models and the longitudinal and panel data models considered in this text. Survival models are concerned with dependent variables that are time until an event of interest whereas the focus of longitudinal and panel data models is broader. Because they concern time, survival models using conditioning arguments extensively in model specification and estimation. Also because of the time element, survival models heavily involve censoring and truncation of variables (it is often more difficult to observe large values of a time variable, other things being equal). Like longitudinal/panel data models, survival models address repeated observations on a subject. Unlike longitudinal/panel data models, survival models also address repeated occurrences of an event (such as marriage). To track events over time, survival models may be expressed in terms of stochastic processes. This
formulation allows one to model many complex data patterns of interest. There are many excellent applied introductions to survival modeling, see for example, Klein and Moeschberger (1997B) and Singer and Willet (2003EP). For a more technical treatment, see Hougaard (2000B).

Appendix 11A. Conditional likelihood estimation for multinomial logit models with heterogeneity terms

To estimate the parameters $\beta$ in the presence of the heterogeneity terms $\alpha_{ij}$, we may again look to conditional likelihood estimation. The idea is to condition the likelihood using sufficient statistics, introduced in Appendix 10A.2.

From equations (11.4) and (11.11), the log-likelihood for the $i$th subject is

$$\ln L(y_i | \alpha_i) = \sum_t \sum_{j=1}^c y_{ij} \ln \pi_{ij} = \sum_t \left\{ \sum_{j=1}^{c-1} y_{ij} \ln \frac{\pi_{ij}}{\pi_{ic}} + \ln \pi_{ic} \right\}$$

$$= \sum_t \left\{ \sum_{j=1}^{c-1} y_{ij} (\alpha_{ij} + (x_{it,j} - x_{it,c})' \beta) + \ln \pi_{ic} \right\},$$

because $\ln \pi_{ij}/\pi_{ic} = V_{ij} - V_{ic} = \alpha_{ij} + (x_{ij} - x_{ic})' \beta$. Thus, using the factorization theorem in Appendix 10A.2, $\Sigma_t y_{ij}$ is sufficient for $\alpha_{ij}$. We interpret $\Sigma_t y_{ij}$ to be the number of choices of alternative $j$ in $T_i$ time periods.

To calculate the conditional likelihood, we let $S_{ij}$ be the random variable representing $\Sigma_t y_{ij}$ and let $\text{sum}_{ij}$ be the realization of $\Sigma_t y_{ij}$. With this, the distribution of the sufficient statistic is

$$\text{Prob}(S_{ij} = \text{sum}_{ij}) = \sum_{B_i} \prod_{j=1}^c (\sigma_{ij})^{y_{ij}},$$

where $B_i$ is the sum over all sets of the form $\{y_{ij} : \Sigma_t y_{ij} = \text{sum}_{ij}\}$. By sufficiency, we may take to $\alpha_{ij} = 0$ without loss of generality. Thus, the conditional likelihood of the $i$th subject is

$$\frac{L(y_i | \alpha_i)}{\text{Prob}(S_{ij} = \text{sum}_{ij})} = \frac{\exp \left( \sum_t \left\{ \sum_{j=1}^{c-1} y_{ij} (x_{it,j} - x_{it,c})' \beta + \ln \pi_{ic} \right\} \right)}{\sum_{B_i} \prod_{j=1}^c (\sigma_{ij})^{y_{ij}}}$$

$$= \frac{\exp \left( \sum_{j=1}^{c-1} y_{ij} (x_{it,j} - x_{it,c})' \beta + \ln \pi_{ic} \right)}{\sum_{B_i} \exp \left( \sum_t \sum_{j=1}^{c-1} y_{ij} (x_{it,j} - x_{it,c})' \beta + \ln \pi_{ic} \right)}.$$

As in Appendix 9A.2, this can be maximized in $\beta$. However, it is computationally intensive.
Appendices

Appendix A. Elements of Matrix Algebra

A.1 Basic Definitions

- **matrix** - a rectangular array of numbers arranged in rows and columns (the plural of matrix is matrices).
- **dimension** of the matrix - the number of rows and columns of the matrix.
- Consider a matrix \( A \) that has dimension \( m \times k \). Let \( a_{ij} \) be the symbol for the number in the \( i \)th row and \( j \)th column of \( A \). In general, we work with matrices of the form

\[
\begin{pmatrix}
a_{11} & a_{12} & \cdots & a_{1k} \\
a_{21} & a_{22} & \cdots & a_{2k} \\
\vdots & \vdots & \ddots & \vdots \\
a_{m1} & a_{m2} & \cdots & a_{mk}
\end{pmatrix}
\]

- **vector** - a (column) vector is a matrix containing only one row (\( m = 1 \)).
- **row vector** - a matrix containing only one column (\( k = 1 \)).
- **transpose** - transpose of a matrix \( A \) is defined by interchanging the rows and columns and is denoted by \( A' \) (or \( A^T \)). Thus, if \( A \) has dimension \( m \times k \), then \( A' \) has dimension \( k \times m \).
- **square matrix** - a matrix where the number of rows equals the number of columns, that is, \( m = k \).
- **diagonal element** – the number in the \( r \)th row and column of a square matrix, \( r = 1, 2, \ldots \)
- **diagonal matrix** - a square matrix where all non-diagonal numbers are equal to zero.
- **identity matrix** - a diagonal matrix where all the diagonal elements are equal to one and is denoted by \( I \).
- **symmetric matrix** - a square matrix \( A \) such that the matrix remains unchanged if we interchange the roles of the rows and columns, that is, if \( A = A' \). Note that a diagonal matrix is a symmetric matrix.
- **gradient vector** – a vector of partial derivatives. If \( f(.) \) is a function of the vector \( x = (x_1, \ldots, x_m)' \), then the gradient vector is \( \partial f(x)/\partial x \). The \( i \)th row of the gradient vector is \( \partial f(x)/\partial x_i \).
- **Hessian matrix** – a matrix of second derivatives. If \( f(.) \) is a function of the vector \( x = (x_1, \ldots, x_m)' \), then the Hessian matrix is \( \partial^2 f(x)/\partial x \partial x' \). The element in the \( i \)th row and \( j \)th column of the Hessian matrix is \( \partial^2 f(x)/\partial x_i \partial x_j \).

A.2 Basic Operations

- **scalar multiplication.** Let \( c \) be a real number, called a scalar. Multiplying a scalar \( c \) by a matrix \( A \) is denoted by \( cA \) and defined by

\[
\begin{pmatrix}
ca_{11} & ca_{12} & \cdots & ca_{1k} \\
ca_{21} & ca_{22} & \cdots & ca_{2k} \\
\vdots & \vdots & \ddots & \vdots \\
ca_{m1} & ca_{m2} & \cdots & ca_{mk}
\end{pmatrix}
\]

- **matrix addition and subtraction.** Let \( A \) and \( B \) be matrices, each with dimension \( m \times k \). Use \( a_{ij} \) and \( b_{ij} \) to denote the numbers in the \( i \)th row and \( j \)th column of \( A \) and \( B \), respectively. Then, the matrix \( C = A + B \) is defined to be the matrix with the number \( (a_{ij} + b_{ij}) \) to denote the number in the \( i \)th row and \( j \)th column. Similarly, the matrix \( C = A - B \) is defined to be the matrix with the number \( (a_{ij} - b_{ij}) \) to denote the numbers in the \( i \)th row and \( j \)th column.
• matrix multiplication. If \( A \) is a matrix of dimension \( m \times c \) and \( B \) is a matrix of dimension \( c \times k \), then \( C = A \times B \) is a matrix of dimension \( m \times k \). The number in the \( i \)th row and \( j \)th column of \( C \) is \( \sum_{c=1}^{c} a_{ij} b_{ij} \).

• determinant - a function of a square matrix, denoted by \( \text{det}(A) \), or \( |A| \). For a \( 1 \times 1 \) matrix, the determinant is \( \text{det}(A) = a_{11} \). To define determinants for larger matrices, we need two additional concepts. Let \( A_{rs} \) be the \((m-1) \times (m-1)\) submatrix of \( A \) defined by removing the \( r \)th row and \( s \)th column. The determinant of \( A_{rs} \), \( \text{det}(A_{rs}) \), is called the minor of the element \( a_{rs} \). Further, \((-1)^{r+s}\) is called the minor of the element \( a_{rs} \). Recursively, define \( \text{det}(A) = \sum_{r=1}^{m} a_{rs} \text{det}(A_{rs}) \), for any \( r = 1, \ldots, n \). For example, for \( n = 2 \), we have the determinant is \( \text{det}(A) = a_{11} a_{22} - a_{12} a_{21} \).

• matrix inverse. In matrix algebra, there is no concept of “division.” Instead, we extend the concept of “reciprocals” of real numbers. To begin, suppose that \( A \) is a square matrix of dimension \( k \times k \) such that \( \text{det}(A) \neq 0 \). Further, let \( I \) be the \( k \times k \) identity matrix. If there exists a \( k \times k \) matrix \( B \) such that \( A \times B = I = B \times A \), then \( B \) is called the inverse of \( A \) and is written \( B = A^{-1} \).

### A.3 Further Definitions

• linearly dependent vectors – a set of vectors \( e_1, \ldots, e_k \) is said to be linearly dependent if one of the vectors in the set can be written as a linear combination of the others.

• linearly independent vectors – a set of vectors \( e_1, \ldots, e_k \) is said to be linearly independent if they are not linearly dependent. Specifically, a set of vectors \( e_1, \ldots, e_k \) is said to be linearly independent if and only if the only solution of the equation \( x_1 e_1 + \ldots + x_k e_k = 0 \) is \( x_1 = \ldots = x_k = 0 \).

• rank of a matrix – the largest number of linearly independent columns (or rows) of a matrix.

• singular matrix – a square matrix \( A \) such that \( \text{det}(A) = 0 \).

• non-singular matrix – a square matrix \( A \) such that \( \text{det}(A) \neq 0 \).

• positive definite matrix – a symmetric square matrix \( A \) such that \( x' A x > 0 \) for \( x \neq 0 \).

• non-negative definite matrix – a symmetric square matrix \( A \) such that \( x' A x \geq 0 \) for \( x \neq 0 \).

• orthogonal – two matrices \( A \) and \( B \) are orthogonal if \( A' B = 0 \), a zero matrix.

• orthogonal matrix - a matrix \( A \) such that \( A A' = I \).

• idempotent – a square matrix such that \( A A = A \).

• trace – the sum of all diagonal elements of a square matrix.

• eigenvalues – the solutions of the \( n \)th degree polynomial \( \text{det}(A - \lambda I) = 0 \). Also known as characteristic roots and latent roots.

• eigenvector – a vector \( x \) such that \( A x = \lambda x \), where \( \lambda \) is an eigenvalue of \( A \). Also known as a characteristic vector and latent vector.

• generalized inverse - of a matrix \( A \) is a matrix \( B \) such that \( A B A = A \). We use the notation \( A^- \) to denote the generalized inverse of \( A \). In the case that \( A \) is invertible, then \( A^- \) is unique and equals \( A^{-1} \). Although there are several definitions of generalized inverses, the above definition suffices for our purposes. See Searle (1987G) for further discussion of alternative definitions of generalized inverses.

### A.4 Matrix Decompositions

• Let \( A \) be an \( m \times m \) symmetric matrix. Then \( A \) has \( m \) pairs of eigenvalues and eigenvectors \( (\lambda_1, e_1), \ldots, (\lambda_m, e_m) \). The eigenvectors can be chosen to have unit length, \( e'_i e_j = 1 \), and to be orthogonal to one another, \( e'_i e_j = 0 \), for \( i \neq j \). The eigenvectors are unique unless two or more eigenvalues are equal.
• **Spectral decomposition** - Let $A$ be an $m \times m$ symmetric matrix. Then we can write $A = \lambda_1 e_1 e_1' + \ldots + \lambda_m e_m e_m'$, where the eigenvectors are of unit length and are mutually orthogonal.

  Suppose that $A$ is positive definite. Then, each eigenvalue is positive. Similarly, if $A$ is non-negative definite, then each eigenvalue is non-negative.

  Using the spectral decomposition, we may write an $m \times m$ symmetric matrix $A$ as $A = P \Lambda P'$, where $P = [e_1; \ldots; e_m]$ and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_m)$.

• **Square root matrix** – for a non-negative definite matrix $A$, we may define the square root matrix as $A^{1/2} = P \Lambda^{1/2} P'$, where $\Lambda^{1/2} = \text{diag}(\lambda_1^{1/2}, \ldots, \lambda_m^{1/2})$. The matrix $A^{1/2}$ is symmetric and is such that $A^{1/2} A^{1/2} = A$.

• **Matrix power** – for a positive definite matrix $A$, we may define $A^c = P \Lambda^c P'$, where $\Lambda^c = \text{diag}(\lambda_1^c, \ldots, \lambda_m^c)$, for any scalar $c$.

• **Cholesky factorization** - Suppose that $A$ is positive definite. Then, we may write $A = L L'$, where $L$ is a lower triangular matrix ($l_{ij} = 0$ for $i < j$). Further, $L$ is invertible so that $A^{-1} = (L')^{-1} L^{-1}$. $L$ is known as the Cholesky square-root matrix.

  Let $A$ be an $m \times k$ matrix. Then, there exists an $m \times m$ orthogonal matrix $U$ and a $k \times k$ orthogonal matrix $V$ such that $A = U \Lambda V'$. Here, $\Lambda$ is an $m \times k$ matrix such that the $(i, i)$ entry of $\Lambda$ is $\lambda_i \geq 0$ for $i = 1, \ldots, \min(m, k)$ and the other entries are zero. The elements $\lambda_i$ are called the singular values of $A$.

• **Singular value decomposition** – For $\lambda_i > 0$, let $u_i$ and $v_i$ be the corresponding columns of $U$ and $V$, respectively. The singular value decomposition of $A$ is $A = \sum_{i=1}^r \lambda_i u_i v_i'$, where $r$ is the rank of $A$.

• **QR decomposition** – Let $A$ be an $m \times k$ matrix, with $m \geq k$ of rank $k$. Then, there exists an $m \times m$ orthogonal matrix $Q$ and a $k \times k$ upper triangular matrix $R$ such that $A = Q R$.

  We can also write $A = Q_k R$, where $Q_k$ consists of the first $k$ columns of $Q$.

### A.5 Partitioned Matrices

• A standard result on inverses of partitioned matrices is

\[
\begin{pmatrix}
B_{11} & B_{12} \\
B_{21} & B_{22}
\end{pmatrix}^{-1} = \begin{pmatrix}
C_{11} & -B_{11} B_{12} C_{22} \\
-C_{22} B_{21} B_{11} & C_{22}
\end{pmatrix}^{-1}
\]

where $C_{11} = B_{11} - B_{12} B_{22} B_{21}$ and $C_{22} = B_{22} - B_{21} B_{11}^{-1} B_{12}$.

• A related result on determinants of partitioned matrices is

\[
\det \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} = \det B_{11} \det C_{22} = \det B_{11} \det \left( B_{22} - B_{21} B_{11}^{-1} B_{12} \right).
\]

• Another standard result on inverses of partitioned matrices is

\[
\left( B_{11} - B_{12} B_{22}^{-1} B_{21} \right)^{-1} = B_{11}^{-1} + B_{11}^{-1} B_{12} \left( B_{22} - B_{21} B_{11}^{-1} B_{12} \right)^{-1} B_{21} B_{11}^{-1}
\]

• To illustrate, with $R = B_{11}$, $Z = B_{12}$, $Z' = B_{21}$, and $D^{-1} = B_{22}$, we have

\[
V^{-1} = (R + Z D Z')^{-1} = R^{-1} - R^{-1} Z (D^{-1} + Z' R^{-1} Z)^{-1} Z' R^{-1}.
\]

• The related determinant result is

\[
\ln \det V = \ln \det R - \ln \det D^{-1} + \ln \det \left( D^{-1} + Z' R^{-1} Z \right)^{-1}.
\]
• Suppose that $A$ is an invertible, $p \times p$ matrix and $c$, $d$ are a $p \times 1$ vectors. Then, from for example, Graybill (1983G), Theorem 8.9.3, we have

$$(A + c d')^{-1} = A^{-1} - \frac{A^{-1} c d' A^{-1}}{1 + d' A^{-1} c}.$$  

(A.6)

To check this result, simply multiply $A + c d'$ by the right hand side to get $I$, the identity matrix.

• Let $P$, $Q$ be idempotent and orthogonal matrices. Let $a$, $b$ be positive contants. Then,

$$(a P + b Q)' = a^* P + b^* Q,$$  

for scalar $c$.       

(A.7)

(Baltagi, 2001E).

A.6 Kronecker (Direct) Product

Let $A$ be an $m \times n$ matrix and $B$ be an $m_1 \times n_2$ matrix. The direct product is defined to be

$$A \otimes B = \begin{pmatrix}
a_{11}B & a_{12}B & \cdots & a_{1n}B \\
a_{21}B & a_{22}B & \cdots & a_{2n}B \\
\vdots & \vdots & \ddots & \vdots \\
a_{m1}B & a_{m2}B & \cdots & a_{mn}B
\end{pmatrix},$$

an $m m_1 n_2 \times n_2$ matrix.

Some properties of Direct Products.

$$(A \otimes B)(F \otimes G) = (A F ) \otimes (B G)$$

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$$

If $P$, $Q$ are orthogonal matrices, then $P \otimes Q$ is an orthogonal matrix.

See Graybill (1969G, Chapter 8).
Appendix B. Normal Distribution

Univariate normal distribution

Recall that the probability density function of \( y \sim N(\mu, \sigma^2) \) is given by
\[
f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right).
\]
If \( \mu = 0 \) and \( \sigma = 1 \), then \( y \sim N(0, 1) \) is said to be standard normal. The standard normal probability density function is
\[
\varphi(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right).
\]
The corresponding standard normal distribution function is denoted by \( \Phi(y) = \int_{-\infty}^{y} \varphi(z)dz \).

Multivariate normal distribution

A vector of random variables \( y = (y_1, y_2, ..., y_n)' \) is said to be multivariate normal if all linear combinations of the form \( a'y = \sum_i a_i y_i \) are normally distributed, where \( a_i \)'s are constants. In this case, we write \( y \sim N(\mu, V) \), where \( \mu = E(y) \) is the expected value of \( y \) and \( V = \text{Var}(y) \) is the variance-covariance matrix of \( y \). From the definition, we have that \( y \sim N(\mu, V) \) implies that \( a'y \sim N(a'\mu, a'V a) \).

The multivariate probability density function of \( y \sim N(\mu, V) \) is given by
\[
f(y) = f(y_1, ..., y_n) = (2\pi)^{-n/2} (\det V)^{-1/2} \exp\left(-\frac{1}{2} (y-\mu)' V^{-1} (y-\mu)\right).
\]
For mixed linear models, the mean is a function of linear combinations of parameters such that \( \mu = X\beta \). Thus, the probability density function of \( y \sim N(X\beta, V) \) is given by
\[
f(y) = f(y_1, ..., y_n) = (2\pi)^{-n/2} (\det V)^{-1/2} \exp\left(-\frac{1}{2} (y-X\beta)' V^{-1} (y-X\beta)\right).
\]

Normal likelihood

A logarithmic probability density function evaluated using the observations is known as a log-likelihood. Suppose that this density depends on the mean parameters \( \beta \) and variance components \( \tau \). Then, the log-likelihood for the multivariate normal can be expressed as
\[
L(\beta, \tau) = -\frac{1}{2} \left( n \ln(2\pi) + \ln(\det V) + (y-X\beta)' V^{-1} (y-X\beta) \right). \tag{B.1}
\]

Conditional distributions

Suppose that \((y_1', y_2')' \) is a multivariate normally distributed vector such that
\[
\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim N\left( \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{pmatrix} \right).
\]
Then, the conditional distribution of \( y_1 | y_2 \) is also normal. Specifically, we have
\[
y_1 | y_2 \sim N\left( \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (y_2 - \mu_2), \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{12}' \right). \tag{B.2}
\]
Thus, \( E(y_1 | y_2) = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (y_2 - \mu_2) \) and \( \text{Var}(y_1 | y_2) = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{12}' \).
Appendix C. Likelihood-Based Inference

Begin with a random vector \( y \) whose joint distribution is known up to a vector of parameters \( \theta \). This joint probability density (mass) function is denoted as \( p(y; \theta) \). The log-likelihood function is \( \ln p(y; \theta) = L(y; \theta) = \log p(y; \theta) \), when evaluated at a realization of \( y \). That is, the log-likelihood is a function of the parameters with the data fixed rather than a function of the data with the parameters fixed.

C.1 Characteristics of Likelihood Functions

Two basic characteristics of likelihood functions are:

\[
E\left( \frac{\partial}{\partial \theta} L(\theta) \right) = 0
\]

(C.1)

and

\[
E\left( \frac{\partial^2}{\partial \theta \partial \theta'} L(\theta) \right) + E\left( \frac{\partial L(\theta)}{\partial \theta} \frac{\partial L(\theta)}{\partial \theta'} \right) = 0
\]

(C.2)

The derivative of the log-likelihood function, \( \frac{\partial}{\partial \theta} L(\theta) \), is called the score function. From equation (C.1), we see that it has mean zero. To see equation (C.1), under suitable regularity conditions, we have

\[
E\left( \frac{\partial}{\partial \theta} L(\theta) \right) = E\left( \frac{\partial}{\partial \theta} \frac{p(y; \theta)}{p(y; \theta)} \right) = \int \frac{\partial}{\partial \theta} p(y; \theta) \, dy = \frac{\partial}{\partial \theta} \int p(y; \theta) \, dy = \frac{\partial}{\partial \theta} 1 = 0.
\]

For convenience, this demonstration assumes a density for \( y \); extensions to mass and mixtures are straightforward. The proof of equation (C.2) is similar and is omitted. Some “suitable regularity conditions” are required to allow the interchange of the derivative and integral sign.

Using equation (C.2), we may define

\[
I(\theta) = E\left( \frac{\partial^2}{\partial \theta \partial \theta'} L(\theta) \right) = -E\left( \frac{\partial}{\partial \theta} \frac{\partial L(\theta)}{\partial \theta} \right),
\]

the information matrix. This quantity is used in the scoring algorithm for parameter estimation.

Under broad conditions, we have that \( \frac{\partial}{\partial \theta} L(\theta) \) is asymptotically normal with mean \( \theta \) and variance \( I(\theta) \).

C.2 Maximum Likelihood Estimators

Maximum likelihood estimators are values of the parameters \( \theta \) that are “most likely” to have been produced by the data. Consider random variables \((y_1, \ldots, y_n) = y\) with probability function \( p(y; \theta) \). The value of \( \theta \), say \( \theta_{MLE} \), that maximizes \( p(y; \theta) \) is called the maximum likelihood estimator. We may also determine \( \theta_{MLE} \) by maximizing \( L(y; \theta) = L(\theta) \), the log-likelihood function. In many applications, this can be done by finding roots of the score function, \( \frac{\partial}{\partial \theta} L(\theta) \).

Under broad conditions, we have that \( \theta_{MLE} \) is asymptotically normal with mean \( \theta \) and variance \( (I(\theta))^{-1} \). Moreover, maximum likelihood estimators are the most efficient in the following sense. Suppose that \( \hat{\theta} \) is an alternative unbiased estimator. The Cramer-Rao theorem states, under mild regularity conditions, for all vectors \( \mathbf{c} \),

\[
\text{Var}(\mathbf{c}' \theta_{MLE}) \leq \text{Var}(\mathbf{c}' \hat{\theta}),
\]

for sufficiently large \( n \).
We also note that $2(L(\theta_{MLE}) - L(\theta))$ has a chi square distribution with degrees of freedom equal to the dimension of $\theta$.

**Example - One parameter exponential family**

Let $y_1, \ldots, y_n$ be independent draws from a one parameter exponential family distribution as in equation (9.1),

$$p(y, \theta, \phi) = \exp \left( \frac{y \theta - b(\theta)}{\phi} + S(y, \phi) \right).$$

(C.3)

The score function is

$$\frac{\partial L(\theta)}{\partial \theta} = \frac{\partial}{\partial \theta} \sum_{i=1}^n \ln p(y_i, \theta, \phi) = \frac{\partial}{\partial \theta} \sum_{i=1}^n \left( \frac{y_i \theta - b(\theta)}{\phi} + S(y_i, \phi) \right) = \sum_{i=1}^n \left( \frac{y_i - b'(\theta)}{\phi} \right) = \frac{n(\bar{y} - b'(\theta))}{\phi}. $$

Thus, setting this equal zero yields $\bar{y} = b'(\theta_{MLE})$, or $\theta_{MLE} = b^{-1}(\bar{y})$. The information matrix is

$$I(\theta) = -E \left[ \frac{\partial^2}{\partial \theta^2} L(\theta) \right] = \frac{n b^*(\theta)}{\phi}. $$

In general, maximum likelihood estimators are determined iteratively. For general likelihoods, two basic procedures are used:

- **Newton-Raphson** uses the iterative algorithm $\theta_{NEW} = \theta_{OLD} - \left[ \frac{\partial^2 L}{\partial \theta \partial \theta'} \right]^{-1} \frac{\partial L}{\partial \theta}_{\theta = \theta_{OLD}}.$

- **Fisher scoring** uses the iterative algorithm $\theta_{NEW} = \theta_{OLD} + I(\theta_{OLD})^{-1} \left[ \frac{\partial L}{\partial \theta} \right]_{\theta = \theta_{OLD}},$ where $I(\theta)$ is the information matrix.

**Example - Generalized linear model**

Let $y_1, \ldots, y_n$ be independent draws from a one parameter exponential family with distribution in equation (C.3). Suppose that the random variable $y_i$ has mean $\mu_i$ with systematic component $\eta_i = g(\mu_i) = x'_i \beta$ and canonical link so that $\eta_i = \theta_i$. Assume that the scale parameter varies by observation so that $\phi_i = \phi / w_i$, where $w_i$ is a known weight function. Then, the score function is

$$\frac{\partial L(\beta)}{\partial \beta} = \frac{\partial}{\partial \beta} \sum_{i=1}^n \left( \frac{y_i x'_i \beta - b(x'_i \beta)}{\phi_i} + S(y_i, \phi) \right) = \sum_{i=1}^n w_i \left( \frac{y_i - b'(x'_i \beta) x_i}{\phi} \right).$$

The matrix of second derivatives is

$$\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta'} = -\sum_{i=1}^n w_i b^*(x'_i \beta) x_i x'_i. $$

(C.4)

Thus, the Newton-Raphson algorithm is

$$\beta_{NEW} = \beta_{OLD} - \left( \sum_{i=1}^n w_i b^*(x_i \beta_{OLD}) x_i x'_i \right)^{-1} \left( \sum_{i=1}^n w_i (y_i - b'(x_i \beta_{OLD})) x_i \right).$$
Because the matrix of second derivatives is non-stochastic, we have that \( \frac{\partial^2 L(\beta)}{\partial \beta \partial \beta'} = I(\beta) \) and thus the Newton-Raphson is equivalent to the Fisher scoring algorithm.

### C.3 Iterated Reweighted Least Squares

Continue with the prior example concerning the generalized linear model and define an “adjusted dependent variable”

\[
y_i^*(\beta) = x_i'\beta + \frac{y_i - b'(x_i'\beta)}{b''(x_i'\beta)}.
\]

This has variance

\[
\text{Var}[y_i^*(\beta)] = \frac{\text{Var}[y_i]}{(b''(x_i'\beta))^2} = \frac{\phi_i}{w_i} = \frac{\phi}{b''(x_i'\beta)}.
\]

Use the new weight as the reciprocal of the variance, \( w_i(\beta) = w_i b''(x_i'\beta) / \phi \). Then, with the expression

\[
w_i(y_i - b'(x_i'\beta)) = w_i b''(x_i'\beta)(y_i^*(\beta) - x_i'\beta) = \phi w_i(b)(y_i^*(\beta) - x_i'\beta),
\]

from the Newton-Raphson iteration, we have

\[
\beta_{NEW} = \beta_{OLD} - \left( \sum_{i=1}^{n} w_i b''(x_i'\beta_{OLD}) x_i x_i' \right)^{-1} \left( \sum_{i=1}^{n} w_i (y_i - b'(x_i'\beta_{OLD})) x_i \right)
\]

\[
= \beta_{OLD} - \left( \sum_{i=1}^{n} \phi w_i (\beta_{OLD}) x_i x_i' \right)^{-1} \left( \sum_{i=1}^{n} \phi w_i (\beta_{OLD}) (y_i^*(\beta_{OLD}) - x_i'\beta_{OLD}) x_i \right)
\]

\[
= \beta_{OLD} - \left( \sum_{i=1}^{n} w_i (\beta_{OLD}) x_i x_i' \right)^{-1} \left( \sum_{i=1}^{n} w_i (\beta_{OLD}) x_i y_i^*(\beta_{OLD}) - \sum_{i=1}^{n} w_i (\beta_{OLD}) x_i x_i'\beta_{OLD} \right)
\]

\[
= \left( \sum_{i=1}^{n} w_i (\beta_{OLD}) x_i x_i' \right)^{-1} \left( \sum_{i=1}^{n} w_i (\beta_{OLD}) x_i y_i^*(\beta_{OLD}) \right).
\]

Thus, this provides a method for iteration using weighted least squares.

### C.4 Profile Likelihood

Split the vector of parameters into two components, say, \( \theta = (\theta_1, \theta_2)' \). Here, interpret \( \theta_1 \) to be the parameters of interest whereas \( \theta_2 \) are auxiliary, or “nuisance,” parameters, of secondary interest. Let \( \theta_{2,MLE}(\theta_1) \) be the maximum likelihood estimator for a fixed value of \( \theta_1 \). Then, the profile likelihood for \( \theta_1 \) is

\[
p(\theta_1, \theta_{2,MLE} ; y) = \sup_{\theta_2} p(\theta_1, \theta_2, y).
\]

This is not a likelihood in the usual sense. To illustrate, it is straightforward to check that equation (C.1) does not hold for \( p(\theta_1, \theta_{2,MLE} ; y) \).

### C.5 Quasi-Likelihood

Suppose that \( y \) is distributed according to the one parameter exponential family in equation (C.3). Then, \( E y = \mu = b'(\theta) \) and \( \text{Var} y = \phi b''(\theta) \). Thus, \( \frac{\partial \mu}{\partial \theta} = \frac{\partial b'(\theta)}{\partial \theta} = b''(\theta) = \text{Var} y / \phi \).

Using the chain rule, we have

\[
\frac{\partial}{\partial \mu} \ln p(y, \theta, \phi) = \frac{\partial \theta}{\partial \mu} \frac{\partial}{\partial \theta} \ln p(y, \theta, \phi) = \frac{\phi}{\text{Var} y} \frac{y - b'(\theta)}{\phi} = \frac{y - \mu}{\text{Var}(\mu)}.
\]
Here, we have explicitly denoted the variance of $y$ as a function of the mean $\mu$ by using the notation $\text{Var} \, y = V(\mu)$. We write $t(y, \mu) = (y - \mu) / V(\mu)$, a function of $y$ and $\mu$. Similar to the score function, this function has the following properties:

$$
(1) \, \mathbb{E} \, t(y, \mu) = 0 \quad \text{and} \quad (2) \, -\mathbb{E} \frac{\partial}{\partial \mu} t(y, \mu) = \frac{1}{V(\mu)} = \text{Var} \, t(y, \mu).
$$

Since these properties are the ones that make the asymptotics of likelihood analysis go, we may think of $Q(y, \mu) = \int_{y}^{\mu} t(y, s) ds$ as a “quasi” log-likelihood. The parameter $\mu$ is known up to a finite dimension vector of parameters, $\beta$, and thus we write $\mu(\beta)$ for $\mu$. Thus, the quasi-score function is

$$
\sum_{i=1}^{n} \frac{\partial Q(y_i, \mu_i)}{\partial \beta} = \sum_{i=1}^{n} t(y_i, \mu_i(\beta)) \frac{\partial \mu_i}{\partial \beta}.
$$

### C.6 Estimating Equations

An alternative method for parameter estimation is using the notion of an “estimating equation” that extends the idea of moment estimation. In the following, we summarize treatments due to McCullagh and Nelder (1989G, Chapter 9) and Diggle et al. (2002S). From an econometrics perspective where this procedure is known as “generalized method of moments,” see Hayashi (2000E).

An estimating function is a function $g(\cdot)$ of $y$, an $n \times 1$ vector of random variables and $\theta$, a $p \times 1$ vector of parameters such that

$$
\mathbb{E} \, g(y; \theta) = 0_n \quad \text{for all } \theta \text{ in a parameter space of interest}, \quad (C.5)
$$

where $0_n$ is an $n \times 1$ vector of zeroes. For example, in many applications, we take $g(y; \theta) = y - \mu(\theta)$, where $\mathbb{E} \, y = \mu(\theta)$. The choice of the $g$ function is critical in applications. In econometrics, equation (C.5) is known as the “moment condition.”

Let $H$ be an $n \times p$ matrix and define the estimator as the solution of the equation

$$
H' \, g(y; \theta) = 0_p,
$$

denoted as $\theta_{EE}$. What is the best choice of $H$? Using a Taylor-series expansion, we see that

$$
\theta_p = H' \, g(y; \theta_{EE}) \approx H' \{ \, g(y; \theta) + \frac{\partial g(y; \theta)}{\partial \theta} \, (\theta_{EE} - \theta) \} \quad \text{so that}
$$

$$
\theta_{EE} - \theta \approx -H' \left( \frac{\partial g(y; \theta)}{\partial \theta} \right)^{-1} H' \, g(y; \theta).
$$

Thus, the asymptotic variance is

$$
\text{Var} \, \theta_{EE} = \left( H' \frac{\partial g(y; \theta)}{\partial \theta} \right)^{-1} H' \left( \text{Var} \, g(y; \theta) \right) H \left( \frac{\partial g(y; \theta)}{\partial \theta} \right)' \left( \frac{\partial g(y; \theta)}{\partial \theta} \right)^{-1} H'.
$$
The choice of \( H \) that yields the most efficient estimator is 

\[ \text{Var} \theta_{EE} = \left( \frac{\partial g(y; \theta)}{\partial \theta} \right)^{\prime} \left( \text{Var} g(y; \theta) \right)^{-1} \frac{\partial g(y; \theta)}{\partial \theta}. \]

This yields

\[ \text{Var} \theta_{EE} = \left( \frac{\partial \mu(\theta)}{\partial \theta} \right)^{\prime} V^{-1} \frac{\partial \mu(\theta)}{\partial \theta}. \]

For the case \( g(y, \theta) = y - \mu(\theta) \), we have \( H = V^{-1} \frac{\partial \mu(\theta)}{\partial \theta} \) where \( V = \text{Var} y \), and

\[ \text{Var} \theta_{EE} = \left( \frac{\partial \mu(\theta)}{\partial \theta} \right)^{\prime} V^{-1} \frac{\partial \mu(\theta)}{\partial \theta}. \]

In this case, the estimating equations estimator \( \theta_{EE} \) is the solution of the equation

\[ \theta_p = \frac{\partial \mu(\theta)}{\partial \theta} V^{-1} (y - \mu(\theta)). \]

For independent observations, this representation is identical to the quasi-likelihood estimators presented in Appendix C.5.

As another example, suppose that \( (w_i, x_i, y_i) \) are i.i.d. and let \( g(y_i; \theta) = (g_1(y_i; \theta), \ldots, g_n(y_i; \theta))^{\prime} \), where

\[ g_i(y; \theta) = w_i(y_i - x_i^{\prime} \theta), \]

and \( E(y_i | w_i, x_i) = x_i^{\prime} \theta \). Assume that \( \text{Var} g_i(y; \theta) = \text{Var} (w_i(y_i - x_i^{\prime} \theta)) = \sigma^2 E(w_i w_i^{\prime}) = \sigma^2 \Sigma_w \).

Thus,

\[ H = (\text{Var} g(y; \theta))^{-1} \frac{\partial g(y; \theta)}{\partial \theta} = (I_n \otimes (\sigma^2 \Sigma_w)) \begin{pmatrix} -w_1 x_1^{\prime} \\ \vdots \\ -w_n x_n^{\prime} \end{pmatrix} = \sigma^{-2} \begin{pmatrix} -\Sigma_w w_1 x_1^{\prime} \\ \vdots \\ -\Sigma_w w_n x_n^{\prime} \end{pmatrix}. \]

Thus, the estimator is a solution of

\[ \theta = \sigma^{-2} \begin{pmatrix} -x_1 w_1^{\prime} \Sigma_w^{-1} \cdots x_n w_n^{\prime} \Sigma_w^{-1} \\ \vdots \\ w_n (y_n - x_n^{\prime} \theta) \end{pmatrix} = -\sigma^{-2} \sum_{i=1}^{n} x_i w_i^{\prime} \Sigma_w^{-1} w_i (y_i - x_i^{\prime} \theta). \]

This yields

\[ \theta_{EE} = \left( \sum_{i=1}^{n} x_i w_i^{\prime} \Sigma_w^{-1} w_i x_i^{\prime} \right)^{-1} \sum_{i=1}^{n} x_i w_i^{\prime} \Sigma_w^{-1} w_i y_i. \]

Using \( n^{-1} \sum_{i=1}^{n} w_i w_i^{\prime} \) in place of \( \Sigma_w \) yields the instrumental variable estimator.

For the case of longitudinal data mixed models, we will assume that the data vector can be decomposed as \( y = (y_1, \ldots, y_n)^{\prime} \) where \( E y_i = \mu_i(\theta) \) and \( \text{Var} y_i = V_i = V_i(\theta, \tau) \). Here, the \( r \times 1 \) vector \( \tau \) is our vector of variance components. Assuming independence among subjects, we consider

\[ G_{\theta}(\theta, \tau) = \sum_{i=1}^{n} \frac{\partial \mu_i(\theta)}{\partial \theta} V_i^{-1} (y_i - \mu_i(\theta)). \]  

The “estimating equations” estimator of \( \theta \), denoted by \( \theta_{EE} \), is the solution of the equation \( \theta_p = G_{\theta}(\theta, \tau) \), where \( \theta_p \) is a \( p \times 1 \) vector of zeroes.
To compute $\theta_{EE}$, we require estimators of the variance components $\tau$. For second moments, we will use $y_i, y'_i$ as our primary data source. For notation, let $\text{vech}(M)$ denote the column vector created by stacking the columns of the matrix $M$. Thus, for example, $\text{vech}(y, y'_i) = (y_{1i}, y_{2i}, \ldots, y_{1iT_{r,1}}, y_{1T_{r,1}}, y_{21}, y_{2T_{r,1}}, \ldots, y_{1iT_{r,1}}, y_{1T_{r,1}})'$. Thus, we use $h_i = (\text{vech}(y_i, y'_i))$ as our data vector and let $\eta_i = \text{E} h_i$. Thus, the estimating equation for $\tau$ is

$$G_{\tau}(\theta, \tau) = \sum_{i=1}^{n} \frac{\partial h'_i}{\partial \tau} (h_i - \eta_i).$$

(C.7)

The “estimating equations” estimator of $\tau$, denoted by $\tau_{EE}$, is the solution of the equation $G_{\tau}(\theta, \tau_{EE}) = 0$. To summarize, we first compute initial estimators of $(\theta, \tau)$, say $(\theta_{0,EE}, \tau_{0,EE})$, typically using basic moment conditions. Then, at the $n$th stage, recursively:

1. Use $\tau_{n,EE}$ and equation (C.6) to update the estimator of $\theta$; that is, $\theta_{n+1,EE}$ is the solution of the equation $G_{\theta}(\theta, \tau_{n,EE}) = 0$
2. Use $\theta_{n+1,EE}$ and equation (C.7) to update the estimator of $\tau$; that is, $\tau_{n+1,EE}$ is the solution of the equation $G_{\tau}(\theta_{n+1,EE}, \tau) = 0$
3. Repeat steps 1 and 2 until convergence.

Under mild regularity conditions, $(\theta_{EE}, \tau_{EE})$ is consistent and asymptotically normal; see for example, Diggle et al. (2002S). Under mild regularity conditions, Gourieroux, Monfort and Trognon (1984E) show that the estimator $\theta_{EE}$, calculated using the estimated $\tau$ is just as efficient asymptotically as if $\tau$ were known. Liang and Zeger (1986B) also provide the following estimator of the asymptotic variance-covariance matrix of $(\theta_{EE}, \tau_{EE})$.

$$
\left( \sum_{i=1}^{n} G_{ii} G_{2i} \right)^{-1} \left( \sum_{i=1}^{n} G_{ii} \left( y_i - \mu_i \right) \left( y_i - \mu_i \right)' G_{ii}' \right) \left( \sum_{i=1}^{n} G_{2i} G_{2i}' \right)^{-1},
$$

(C.8)

where $G_{ii} = \begin{pmatrix} \frac{\partial \mu_i}{\partial \theta} & \frac{\partial \mu_i}{\partial \tau} \\ \frac{\partial \mu_i}{\partial \theta} & \frac{\partial \mu_i}{\partial \tau} \end{pmatrix}$ and $G_{2i} = \begin{pmatrix} \frac{\partial \mu_i}{\partial \theta} & \frac{\partial \mu_i}{\partial \tau} \\ \frac{\partial \mu_i}{\partial \theta} & \frac{\partial \mu_i}{\partial \tau} \end{pmatrix}$.

### C.7 Hypothesis Tests

We consider testing the null hypothesis $H_0: h(\theta) = d$, where $d$ is a known vector of dimension $r \times 1$ and $h(.)$ is known and differentiable.

There are three widely used approaches for testing the null hypothesis, called the likelihood ratio, Wald and Rao tests. The Wald approach evaluates a function of the likelihood at $\theta_{MLE}$. The likelihood ratio approach uses $\theta_{MLE}$ and $\theta_{Reduced}$. Here, $\theta_{Reduced}$ is the value of $\theta$ that maximizes $L(\theta)$ under the restriction that $h(\theta) = d$. The Rao approach also uses $\theta_{Reduced}$ but determines it by maximizing $L(\theta) - \lambda'(h(\theta) - d)$, where $\lambda$ is a vector of Lagrange multipliers. Hence, Rao's test is also called the Lagrange multiplier test. The three statistics are:

- LRT = $2 \left( L(\theta_{MLE}) - L(\theta_{Reduced}) \right)$
- Wald = $TS_{\theta}(\theta_{MLE})$, where $TS_{\theta}(\theta) = (h(\theta) - d)' \frac{\nabla h(\theta)'}{-I(\theta)}^{-1} \nabla h(\theta)'^{-1} (h(\theta) - d)$.
- Rao = $TS_{\theta}(\theta_{Reduced})$, where $TS_{\theta}(\theta) = \nabla L(\theta)' (-I(\theta))^{-1} \nabla L(\theta)$.

Here, $\nabla h(\theta) = \partial h(\theta)/\partial \theta$ is the gradient of $h(\theta)$ and $\nabla L(\theta) = \partial L(\theta)/\partial \theta$ is the gradient of $L(\theta)$, the score function.
The main advantage of the Wald statistic is that it only requires computation of $\theta_{\text{MLE}}$ and not $\theta_{\text{Reduced}}$. Similarly, the main advantage of the Rao statistic is that it only requires computation of $\theta_{\text{Reduced}}$ and not $\theta_{\text{MLE}}$. In many applications, computation of $\theta_{\text{MLE}}$ is onerous.

Under broad conditions, all three test statistics are asymptotically chi-square with $r$ degrees of freedom under $H_0$. All asymptotic methods work well when the number of parameters is finite dimensional and the null hypothesis specifies that $\theta$ is on the interior of the parameter space.

In the usual fixed effects model, the number of individual-specific parameters is the same order as the number of subjects. Here, the number of parameters tends to infinity as the number of subjects tends to infinity and the usual asymptotic approximations are not valid. Instead, special conditional maximum likelihood estimators enjoy the asymptotic properties similar to maximum likelihood estimators.

When a hypothesis specifies that $\theta$ is on the boundary, then the asymptotic distribution is no longer valid without corrections. An example is $H_0$: $\theta = \sigma^2 = 0$. Here, the parameter space is $[0, \infty)$. By specifying the null hypothesis at 0, we are on the boundary. Self and Liang (1987S) provide some corrections that improve the asymptotic approximation.

**C.8 Goodness of Fit Statistics**

In linear regression models, the most widely cited goodness of fit statistic is the $R^2$ measure that is based on the decomposition

$$
\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2 + 2\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}).
$$

In the language of Section (2.3), this decomposition is:

$$
\text{Total SS} = \text{Error SS} + \text{Regression SS} + 2 \times \text{Sum of Cross-Products}.
$$

The difficulty with nonlinear models is that the \textit{Sum of Cross-Products} term rarely equals zero. Thus, one gets different statistics when defining $R^2$ as $\left(\frac{\text{Regression SS}}{\text{Total SS}}\right)$ as compared to $\left(1 - \frac{\text{Error SS}}{\text{Total SS}}\right)$.

An alternative widely cited goodness of fit measure is the \textit{Pearson chi-square} statistic. To define this statistic, suppose that $E y_i = \mu_i$, $\text{Var} y_i = V(\mu_i)$ for some function $V(.)$ and that $\hat{\mu}_i$ is an estimator of $\mu_i$. Then, the Pearson chi-square statistic is defined as $\sum \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$. For Poisson models of count data, this formulation reduces to the form $\sum \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$.

In the context of generalized linear models, a goodness of fit measure is the \textit{deviance statistic}. To define this statistic, suppose $E y = \mu = \mu(\theta)$ and write $L(\hat{\mu})$ for the log-likelihood evaluated at $\hat{\mu} = \mu(\hat{\theta})$. The \textit{scaled} deviance statistic is defined as $D^*(y, \hat{\mu}) = 2 (L(y) - L(\hat{\mu}))$. In linear exponential families, we multiply by the scaling factor $\phi$ to define the deviance statistic, $D(y, \hat{\mu}) = \phi D^*(y, \hat{\mu})$. This multiplication actually removes the variance scaling factor from the definition of the statistic.

Using Appendix 9A, it is straightforward to check that the deviance statistic reduces to the following forms for three important distributions:

- **Normal:** $D(y, \hat{\mu}) = \sum (y_i - \hat{\mu}_i)^2$
- **Bernoulli:** $D(y, \hat{\pi}) = \sum \left\{ y_i \ln \frac{y_i}{\hat{\pi}_i} + (1 - y_i) \ln \frac{1 - y_i}{1 - \hat{\pi}_i} \right\}$
• Poisson: \[ D(y, \hat{\mu}) = \sum_i \left( y_i \ln \frac{y_i}{\hat{\mu}_i} + (y_i - \hat{\mu}_i) \right). \]

Here, we use the convention that \( y \ln y = 0 \) when \( y = 0 \).

**C.9 Information Criteria**

Likelihood ratio tests are useful for choosing between two models that are *nested*, that is, where one model is a subset of the other. How do we compare models when they are not nested? One way is to use the following information criteria.

The distance between two probability distributions given by probability density function \( g \) and \( f_\theta \) can be summarized by

\[
KL(g, f_\theta) = \mathbb{E}_g \ln \frac{g(y)}{f_\theta(y)}.
\]

This is the *Kullback-Leibler distance* that turns out to be nonnegative. Here, we have indexed \( f \) by a vector of parameters \( \theta \). Picking the function \( g \) to be \( f_{\theta_0} \), then minimizing \( KL(f_{\theta_0}, f_\theta) \) is equivalent to the maximum likelihood principle.

In general, we have to estimate \( g(\cdot) \). Akaike showed that a reasonable alternative is to minimize

\[
AIC = -2 \ln(\hat{\theta}_{MLE}) + 2 \text{(number of parameters)},
\]

known as *Akaike’s Information Criterion*. This statistic is used by when comparing several alternative (nonnested) models. One picks the model that minimizes \( AIC \). If the models under consideration have the same number of parameters, this is equivalent to choosing the model that maximizes the log-likelihood. We remark that, in time series analysis, the \( AIC \) is rescaled by the number of parameters.

The statistic \( AIC \) is also useful in that it reduces to the \( C_p \), a statistic that is widely used in regression analysis. This statistic minimizes a bias and variance trade-off when selecting among linear regression models.

Schwarz derived an alternative criterion using Bayesian methods. His measure is known as the *Bayesian Information Criterion*, defined as

\[
BIC = -2 \ln(\hat{\theta}_{MLE}) + \ln \text{(number of parameters)}.
\]

This measure gives greater weight to the number of parameters. That is, other things being equal, \( BIC \) will suggest a more parsimonious model than \( AIC \).
Appendix D. State Space Model and the Kalman Filter

D.1 Basic State Space Model
Consider the observation equation
\[ y_t = W_t \delta_t + \epsilon_t \quad t = 1, \ldots, T \]  
(D.1)
where \( y_t \) is an \( n_t \times 1 \) vector and \( \delta_t \) is a \( m \times 1 \) vector. The transition equation is
\[ \delta_t = T_t \delta_{t-1} + \eta_t. \]  
(D.2)
Together, equations (D.1) and (D.2) define the state space model. To complete the specification, define \( \text{Var}_{t-1} \epsilon_t = H_t \) and \( \text{Var}_{t-1} \eta_t = Q_t \), where \( \text{Var} \) is a variance conditional on information up to and including time \( t \), that is, \( \{ y_1, \ldots, y_t \} \). Similarly, let \( E_t \) denote the conditional expectation and assume that \( \text{E}_{t-1} \epsilon_t = 0 \) and \( \text{E}_{t-1} \eta_t = 0 \). Further define \( d_0 = E \delta_0 \), \( P_0 = \text{Var} \delta_0 \) and \( P_t = \text{Var} \delta_t \).
Assume that \( \{ \epsilon_t \} \) and \( \{ \eta_t \} \) are mutually independent.

In subsequent sections, it will be useful to summarize equation (D.1). Thus, we define
\[ y = \begin{pmatrix} y_1 \\ \vdots \\ y_T \end{pmatrix} = \begin{pmatrix} W_1 \delta_1 \\ \vdots \\ W_T \delta_T \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_T \end{pmatrix} = \begin{pmatrix} W_1 & 0 & \cdots & 0 \\ 0 & W_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & W_T \end{pmatrix} \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_T \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_T \end{pmatrix} = W \delta + \epsilon. \]  
(D.3)
With the notation \( N = \sum_{t=1}^T n_t \), we have that \( y \) is a \( N \times 1 \) vector of random variables, \( \delta \) is a \( Tm \times 1 \) vector of state variables, \( W \) is a \( N \times Tm \) matrix of known variables and \( \epsilon \) is a \( N \times 1 \) vector of disturbance terms.

D.2 Kalman Filter Algorithm
Taking a conditional expectation and variance of equation (D.2) yields the “prediction equations”
\[ d_{t-1} = \text{E}_{t-1} \delta_t = T_t d_{t-1} \]  
(D.4a)
and
\[ P_{t-1} = \text{Var}_{t-1} \delta_t = T_t P_{t-1} T_t' + Q_t. \]  
(D.4b)
Taking a conditional expectation and variance of equation (D.1) yields
\[ \text{E}_{t-1} y_t = W_t d_{t-1} \]  
(D.5a)
and
\[ F_t = \text{Var}_{t-1} y_t = W_t P_{t-1} W_t' + H_t. \]  
(D.5b)
The “updating equations” are
\[ d_t = d_{t-1} + P_{t-1} W_t' F_t^{-1} (y_t - W_t d_{t-1}) \]  
(D.6a)
and
\[ P_t = P_{t-1} - P_{t-1} W_t' F_t^{-1} W_t P_{t-1}. \]  
(D.6b)
The updating equations can be motivated by assuming that \( \delta_t \) and \( y_t \) are jointly normally distributed. With this assumption, and equation (B.2) of Appendix B, we have
\[ \text{E}_t \delta_t = \text{E}_{t-1} \delta_t + \text{Cov}_{t-1} (\delta_t, y_t) (\text{Var}_{t-1} y_t)^{-1} (y_t - \text{E}_{t-1} y_t) \]  
and
\[ \text{Var}_t \delta_t = \text{Var}_{t-1} \delta_t - \text{Cov}_{t-1} (\delta_t, y_t) (\text{Var}_{t-1} y_t)^{-1} \text{Cov}_{t-1} (\delta_t, y_t). \]
These expressions yield the updating equations immediately.
For computational convenience, the Kalman filter algorithm in equations (D.4)-(D.6) can be expressed more compactly as

\[
d_{t+1} = T_{t+1} d_t + K_t (y_t - W_t d_{t-1})
\]

(D.7)

and

\[
P_{t+1} = T_{t+1} (P_{t+1-1} W_t F_t^{-1} W_t' P_{t+1-1}) T_{t+1}' + Q_{t+1},
\]

(D.8)

where \(K_t = T_{t+1} P_{t+1-1} W_t F_t^{-1}\) is known as the gain matrix. To start these recursions, from (D.4) we have

\[
d_{1/0} = T_1 d_0
\]

and

\[
P_{1/0} = T_1 P_0 T_1' + Q_1.
\]

D.3 Likelihood Equations

Assuming the conditional variances, \(Q_t\) and \(H_t\), and initial conditions, \(d_0\) and \(P_0\), are known, equations (D.7) and (D.8) all allow one to recursively compute \(E_{t-1}y_t\) and \(\text{Var}_{t-1}y_t = F_t\). These quantities are important because they allow us to directly evaluate the likelihood of \(\{y_1, \ldots, y_T\}\). That is, assume that \(\{y_1, \ldots, y_T\}\) are jointly multivariate normal and let “f” be used for the joint and conditional density. Then, with equations (B.1) and (B.2) of Appendix B, the logarithmic likelihood is

\[
L = \ln f(y_1, \ldots, y_T) = \ln f(y_1) + \sum_{t=2}^{T} \ln f(y_t | y_1, \ldots, y_{t-1})
\]

\[
= -\frac{1}{2} \left[ N \ln 2\pi + \sum_{t=1}^{T} \ln \det(F_t) + \sum_{t=1}^{T} (y_t - E_{t-1}y_t)' F_t^{-1} (y_t - E_{t-1}y_t) \right].
\]

(D.9)

From the Kalman filter algorithm in equations (D.4)-(D.6), we see that \(E_{t-1}y_t\) is a linear combination of \(\{y_1, \ldots, y_{t-1}\}\). Thus, we may write

\[
Ly = L \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix} = \begin{pmatrix} y_1 - E_0 y_1 \\ y_2 - E_1 y_2 \\ \vdots \\ y_T - E_{T-1} y_T \end{pmatrix},
\]

(D.10)

where \(L\) is a \(N \times N\) lower triangular matrix with one’s on the diagonal. Elements of the matrix \(L\) do not depend on the random variables. The advantages of this transformation are that the components of the right hand side of equation (D.10) are mean zero and are mutually uncorrelated.

D.4 Extended State Space Model and Mixed Linear Models

To handle the longitudinal data model described in Section 6.5, we now extend the state space model to mixed linear models, so that it handles fixed and random effects. By incorporating these effects, we will also be able to introduce initial conditions that are estimable.

Specifically, consider the observation equation

\[
y_t = X_t \beta + Z_t \alpha + W_t \delta_t + \epsilon_t.
\]

(D.11)

Here, \(\beta\) is a \(K \times 1\) vector of unknown parameters, called “fixed effects.” Further, \(\alpha\) is a \(q^* \times 1\) random vector, known as “random effects.” We assume that \(\alpha\) has mean \(\theta\), variance-covariance matrix \(\sigma^2 B = \text{Var} \alpha\), and is independent of \(\{\delta_t\}\) and \(\{\epsilon_t\}\). The transition equation is as in equation (D.2), that is, we assume

\[
\delta_t = T_t \delta_{t-1} + \eta_t.
\]

(D.12)

Similar to equation (D.3), we summarize equation (D.11) as
Appendices

y = X \beta + Z \alpha + W \delta + \epsilon. \tag{D.13}

Here, y, W, \delta, and \epsilon are defined in equation (D.3), and X = (X_1', X_2', \ldots, X_T')' and Z = (Z_1', Z_2', \ldots, Z_T')'.

### D.5 Likelihood Equations for Mixed Linear Models

To handle fixed and random effects, begin with the transformation matrix \(L\) defined in equation (D.10). In developing the likelihood, the important point to observe is that the transformed sequence of random variables, \(y - (X \beta + Z \alpha)\), has the same properties as the basic set of random variables \(y\) in Appendix D.1. Specifically, with the transformation matrix \(L\), the \(T\) components of the vector

\[
L \left( y - (X \beta + Z \alpha) \right) = v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_T \end{pmatrix}
\]

are mean zero and mutually uncorrelated. Further, conditional on \(\alpha\) and \(\{y_1, \ldots, y_{t-1}\}\), the \(t\)th component of this matrix, \(v_t\), has variance \(F_t\).

With the Kalman filter algorithm in (D.7) and (D.8), we define the transformed variables \(y^* = L y, X^* = L X\) and \(Z^* = L Z\). To illustrate, for the \(j\)th column of \(X\), say \(X_j\), one would recursively calculate

\[
d_{t+1/0}(X_j^*) = T_{t+1} d_{t/0}(X_j^*) + K_t (X_t^* - W_t d_{t/0}(X_j^*)) \tag{D.7*}
\]

in place of equation (D.7). We begin the recursion in equation (D.7*) with \(d_{1/0}(X_j^*) = 0\). Equation (D.8) remains unchanged. Then, analogous to expression (D.5a), the \(t\)th component of \(X^*\) is

\[
X_t^* = X_t - W_t d_{t/0}(X_j^*).
\]

With these transformed variables, we may express the transformed random variables as

\[
y^* = L y = v + X^* \beta + Z^* \alpha.
\]

Recall \(\sigma^2 B = \text{Var} \alpha\) and note that

\[
\text{Var} v = \text{Var} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_T \end{pmatrix} = \begin{pmatrix} F_1 & 0 & \cdots & 0 \\ 0 & F_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & F_T \end{pmatrix} = \sigma^2 \Lambda. \tag{D.14}
\]

This yields \(E y^* = X^* \beta\) and \(\text{Var} y^* = \sigma^2 (A + Z^* B Z^*) = \sigma^2 \Lambda\). We use \(\tau\) to denote the vector of (unknown) quantities that parameterized \(V\).

From equation (B.1) of Appendix B, the logarithmic likelihood is

\[
L(\beta, \sigma^2, \tau) = -\frac{1}{2} \{N \ln 2 \pi + N \ln \sigma^2 + \sigma^2 (y^* - X^* \beta)' V^{-1} (y^* - X^* \beta) + \ln \det V \}. \tag{D.15}
\]

The corresponding restricted log-likelihood is

\[
L_R(\beta, \sigma^2, \tau) = -\frac{1}{2} \{\ln \det (X^* V^{-1} X^*) - K \ln \sigma^2 \} + L(\beta, \sigma^2, \tau) + \text{constant}. \tag{D.16}
\]

Either (D.15) or (D.16) can be maximized to determine an estimator of \(\beta\). The result is equivalent to the generalized least squares estimator

\[
b_{GLS} = (X^* V^{-1} X^*)^{-1} X^* V^{-1} y^*. \tag{D.17}
\]
Using $b_{GLS}$ for $\beta$ in equations (D.16 and (D.17) yields concentrated likelihoods. To determine the REML estimator of $\sigma^2$, we maximize $L_R(b_{GLS}, \sigma^2, \tau)$ (holding $\tau$ fixed), to get

$$s^2_{REML} = (N-K)^{-1} (y^* - X^* b_{GLS})' V^{-1} (y^* - X^* b_{GLS}).$$

(D.18)

Thus, the logarithmic likelihood evaluated at these parameters is

$$L(b_{GLS}, s^2_{REML}, \tau) = -\frac{1}{2} \{N \ln 2 + N \ln s^2_{REML} + N-K + \ln \det V \}.$$ 

(D.19)

The corresponding restricted logarithmic likelihood is

$$L_{REML} = -\frac{1}{2} \{\ln \det (X^* V^{-1} X^*) - K \ln s^2_{REML} \} + L(b_{GLS}, s^2_{REML}, \tau) + constant.$$ 

(D.20)

The likelihood expressions in equations (D.19) and (D.20) are intuitively straightforward. However, because of the number of dimensions, they can be difficult to compute. We now provide alternative expressions that, although more complex, are simpler to compute with the Kalman filter algorithm. From equations (A.3) and (A.5) of Appendix A, we have

$$V^{-1} = \Lambda^{-1} - \Lambda^{-1} Z^* (B^{-1} + Z^* \Lambda^{-1} Z^*)^{-1} Z^* \Lambda^{-1}$$

(D.21)

and

$$\ln \det V = \ln \det \Lambda - \ln \det B^{-1} + \ln \det(B^{-1} + Z^* \Lambda^{-1} Z^*).$$

(D.22)

With equation (D.21), we immediately have the expression for $b_{GLS}$ in equation (6.32). From equation (D.18), the restricted maximum likelihood estimator of $\sigma^2$ can be expressed as

$$s^2_{REML} = (N-K)^{-1} \{y^* V^{-1} y^* - y^* V^{-1} X^* b_{GLS}\}$$

which is sufficient for equation (6.34). This, equation (D.20) and (D.22) are sufficient for the equation (6.35).
## Appendix E. Symbols and Notation

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Chapter defined</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i, t$</td>
<td>indices for the $i$th subject, $t$th time period</td>
<td>1</td>
</tr>
<tr>
<td>$T_i$</td>
<td>number of observations for the $i$th subject</td>
<td>1</td>
</tr>
<tr>
<td>$n$</td>
<td>number of subjects</td>
<td>1</td>
</tr>
<tr>
<td>$N$</td>
<td>total number of observations, $N = T_1 + T_2 + \ldots + T_n$</td>
<td>2</td>
</tr>
<tr>
<td>$y_{it}$</td>
<td>response for the $i$th subject, $t$th time period</td>
<td>1</td>
</tr>
<tr>
<td>$y_i$</td>
<td>$T_i \times 1$ vector of responses for the $i$th subject, $y_i = (y_{i1}, y_{i2}, \ldots, y_{iT})'$.</td>
<td>2</td>
</tr>
<tr>
<td>$x_{it}$</td>
<td>$j$th explanatory variable associated with global parameters, for the $i$th subject, $t$th time period</td>
<td>2</td>
</tr>
<tr>
<td>$K$</td>
<td>number of explanatory variables associated with global parameters</td>
<td>2</td>
</tr>
<tr>
<td>$x_{it}$</td>
<td>$K \times 1$ vector of explanatory variables associated with global parameters for the $i$th subject, $t$th time period, $x_{it} = (x_{it,1}, x_{it,2}, \ldots, x_{it,K})'$</td>
<td>2</td>
</tr>
<tr>
<td>$X_i$</td>
<td>$T_i \times K$ matrix of explanatory variables associated with global parameters for the $i$th subject, $X_i = (x_{i1}, x_{i2}, \ldots, x_{iT})'$</td>
<td>2</td>
</tr>
<tr>
<td>$z_{it}$</td>
<td>$j$th explanatory variable associated with subject-specific parameters, for the $i$th subject, $t$th time period</td>
<td>2</td>
</tr>
<tr>
<td>$q$</td>
<td>number of explanatory variables associated with subject-specific parameters</td>
<td>2</td>
</tr>
<tr>
<td>$z_{it}$</td>
<td>$q \times 1$ vector of explanatory variables associated with subject-specific parameters for the $i$th subject, $t$th time period, $z_{it} = (z_{it,1}, z_{it,2}, \ldots, z_{it,q})'$</td>
<td>2</td>
</tr>
<tr>
<td>$Z_i$</td>
<td>$T_i \times q$ matrix of explanatory variables associated with subject-specific parameters for the $i$th subject, $Z_i = (z_{i1}, z_{i2}, \ldots, z_{iT})'$</td>
<td>2</td>
</tr>
<tr>
<td>$\varepsilon_{it}$</td>
<td>error term for the $i$th subject, $t$th time period</td>
<td>2</td>
</tr>
<tr>
<td>$\varepsilon_i$</td>
<td>$T_i \times 1$ vector of error terms for the $i$th subject, $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT})'$.</td>
<td>2</td>
</tr>
<tr>
<td>$\beta_j$</td>
<td>$j$th global parameter, associated with $x_{it}$</td>
<td>2</td>
</tr>
<tr>
<td>$\beta$</td>
<td>$K \times 1$ vector of global parameters $\beta = (\beta_1, \beta_2, \ldots, \beta_K)'$</td>
<td>2</td>
</tr>
<tr>
<td>$\alpha_i$</td>
<td>subject-specific intercept parameter for the $i$th subject</td>
<td>2</td>
</tr>
<tr>
<td>$\alpha_{ij}$</td>
<td>$j$th subject-specific parameter for the $i$th subject, associated with $z_{it}$</td>
<td>2</td>
</tr>
<tr>
<td>$\alpha_i$</td>
<td>$q \times 1$ vector of subject-specific parameter for the $i$th subject, $\alpha_i = (\alpha_{i1}, \ldots, \alpha_{iq})'$</td>
<td>2</td>
</tr>
<tr>
<td>$\lambda_t$</td>
<td>time-specific parameter</td>
<td>2</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>variance of the error term under the homoscedastic model, $\sigma^2 = \text{Var} \varepsilon_{it}$</td>
<td>2</td>
</tr>
<tr>
<td>$\rho$</td>
<td>correlation between two error terms, $\rho = \text{corr}(\varepsilon_{it}, \varepsilon_{is})$</td>
<td>2</td>
</tr>
<tr>
<td>$b_j$, $b$</td>
<td>estimators of $\beta_j$, $\beta$</td>
<td>2</td>
</tr>
<tr>
<td>$a_i$, $a_{ij}$, $a_i$</td>
<td>estimators of $\alpha_i$, $\alpha_{ij}$, $\alpha_i$</td>
<td>2</td>
</tr>
<tr>
<td>$s^2$</td>
<td>unbiased estimator of $\sigma^2$</td>
<td>2</td>
</tr>
<tr>
<td>$e_{it}$</td>
<td>residual for the $i$th subject, $t$th time period</td>
<td>2</td>
</tr>
<tr>
<td>$e_i$</td>
<td>$T_i \times 1$ vector of residuals for the $i$th subject, $e_i = (e_{i1}, e_{i2}, \ldots, e_{iT})'$.</td>
<td>2</td>
</tr>
<tr>
<td>Symbol</td>
<td>Description</td>
<td>Chapter defined</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td>----------------</td>
</tr>
<tr>
<td>$\bar{y}_i$</td>
<td>average, over time, of the responses from the $i$th subject, $\bar{y}<em>i = T_i^{-1} \sum</em>{t=1}^{T_i} y_{it}$</td>
<td>2</td>
</tr>
<tr>
<td>$x_i$</td>
<td>$K \times 1$ vector of averages, over time, of the explanatory variables associated with the global parameters from the $i$th subject, $x_i = T_i^{-1} \sum_{t=1}^{T_i} x_{it}$</td>
<td>2</td>
</tr>
<tr>
<td>$W_i$</td>
<td>a $K \times K$ weight matrix for the $i$th subject, $W_i = \sum_{t=1}^{T_i} (x_{it} - \bar{x}<em>i)(x</em>{it} - \bar{x}_i)'$</td>
<td>2</td>
</tr>
<tr>
<td>$r_{it}$</td>
<td>the rank of the $i$th residual $e_{it}$ from the vector of residuals ${e_{i1}, \ldots, e_{iT}}$. Spearman's rank correlation coefficient between the $i$th and $j$th subjects, $sr_{ij} = \frac{\sum_{t=1}^{T_i} (r_{i,t} - (T+1)/2)(r_{j,t} - (T+1)/2)}{\sum_{t=1}^{T_i} (r_{i,t} - (T+1)/2)^2}$</td>
<td>2</td>
</tr>
<tr>
<td>$R_{AVE}$</td>
<td>the average of Spearman's rank correlations, $R_{AVE} = \frac{1}{n(n-1)/2} \sum_{i&lt;j} sr_{ij}$</td>
<td>2</td>
</tr>
<tr>
<td>$R^2_{AVE}$</td>
<td>the average of squared Spearman's rank correlations, $R^2_{AVE} = \frac{1}{n(n-1)/2} \sum_{i&lt;j} (sr_{ij})^2$</td>
<td>2</td>
</tr>
<tr>
<td>$R$</td>
<td>$T \times T$ variance-covariance matrix, $R = \text{Var } \varepsilon$</td>
<td>2</td>
</tr>
<tr>
<td>$R_{rs}$</td>
<td>the element in the $r$th row and $s$th column of $R$, $R_{rs} = \text{Cov } (\varepsilon_r, \varepsilon_s)$</td>
<td>2</td>
</tr>
<tr>
<td>$R_i$</td>
<td>$T_i \times T_i$ matrix variance-covariance for the $i$th subject, $R_i = \text{Var } \varepsilon_i$</td>
<td>2</td>
</tr>
<tr>
<td>$I_i$</td>
<td>$T_i \times 1$ vector of ones</td>
<td>2</td>
</tr>
<tr>
<td>$I$</td>
<td>Identity matrix, generally of dimension $T \times T$</td>
<td>2</td>
</tr>
<tr>
<td>$I_i$</td>
<td>$T_i \times T_i$ identity matrix</td>
<td>2</td>
</tr>
<tr>
<td>$J$</td>
<td>matrix of ones, generally of dimension $T \times T$</td>
<td>2</td>
</tr>
<tr>
<td>$J_i$</td>
<td>$T_i \times T_i$ matrix of ones</td>
<td>2</td>
</tr>
<tr>
<td>$\tau$</td>
<td>vector of variance components</td>
<td>2</td>
</tr>
<tr>
<td>$Q_i$</td>
<td>matrix that projects a vector of responses to OLS residuals, $Q_i = I_i - Z_i(Z_i'Z_i)^{-1}Z_i'$</td>
<td>2</td>
</tr>
<tr>
<td>$Q_{Z,i}$</td>
<td>matrix that projects a vector of responses to GLS residuals, $Q_{Z,i} = I_i - R_i^{-1/2}Z_i(Z_i'[R_i^{-1}Z_i]'Z_i'R_i^{-1/2})^{-1}Z_i'R_i^{-1/2}$</td>
<td>2</td>
</tr>
<tr>
<td>$\sigma^2_\alpha$</td>
<td>variance of the subject-specific intercept in the one-way error components model, $\text{Var } \alpha_t$</td>
<td>3</td>
</tr>
<tr>
<td>$s^2_\alpha$</td>
<td>unbiased estimator of $\sigma^2_\alpha$</td>
<td>3</td>
</tr>
<tr>
<td>$b_{EC}$</td>
<td>generalized least squares (GLS) estimator of $\beta$ in the error components model</td>
<td>3</td>
</tr>
<tr>
<td>$TS$</td>
<td>a test statistic for assessing homogeneity in the error components model</td>
<td>3</td>
</tr>
<tr>
<td>$\delta$</td>
<td>a group effect variable</td>
<td>3</td>
</tr>
<tr>
<td>Symbol</td>
<td>Description</td>
<td>Chapter defined</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>(\mathbf{D})</td>
<td>variance-covariance matrix of subject-specific effects in the longitudinal data mixed model, (\text{Var} , \mathbf{\alpha}_i)</td>
<td>3</td>
</tr>
<tr>
<td>(\mathbf{V}_i)</td>
<td>variance-covariance matrix of the (i)th subject in the longitudinal data mixed model, (\mathbf{V}_i = \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i' + \mathbf{R}_i)</td>
<td>3</td>
</tr>
<tr>
<td>(\mathbf{X})</td>
<td>(N \times K) matrix of explanatory variables associated with fixed effects in the mixed linear model</td>
<td>3</td>
</tr>
<tr>
<td>(\mathbf{\beta})</td>
<td>(K \times 1) vector of fixed effects in the mixed linear model</td>
<td>3</td>
</tr>
<tr>
<td>(\mathbf{Z})</td>
<td>(N \times q) matrix of explanatory variables associated with random effects in the mixed linear model</td>
<td>3</td>
</tr>
<tr>
<td>(\mathbf{a})</td>
<td>(q \times 1) vector of random effects in the mixed linear model</td>
<td>3</td>
</tr>
<tr>
<td>(\mathbf{y})</td>
<td>(N \times 1) vector of responses in the mixed linear model</td>
<td>3</td>
</tr>
<tr>
<td>(\mathbf{e})</td>
<td>(N \times 1) vector of disturbances in the mixed linear model</td>
<td>3</td>
</tr>
<tr>
<td>(\mathbf{b}_{\text{GLS}})</td>
<td>GLS estimator of (\mathbf{\beta}) in the mixed linear model</td>
<td>3</td>
</tr>
<tr>
<td>(\mathbf{b}_{\text{MLE}})</td>
<td>maximum likelihood estimator (MLE) of (\mathbf{\beta}) in the mixed linear model</td>
<td>3</td>
</tr>
<tr>
<td>(l_i(.))</td>
<td>logarithmic likelihood of the (i)th subject</td>
<td>3</td>
</tr>
<tr>
<td>(L(.))</td>
<td>logarithmic likelihood of the entire data set</td>
<td>3</td>
</tr>
<tr>
<td>(\mathbf{b}_W)</td>
<td>weighted least square estimator of (\mathbf{\beta}) in the mixed linear model</td>
<td>3</td>
</tr>
<tr>
<td>(\text{se}(\mathbf{b}_W))</td>
<td>robust standard error of (\mathbf{b}_W)</td>
<td>3</td>
</tr>
<tr>
<td>(L_R)</td>
<td>logarithmic restricted maximum likelihood of the entire data set</td>
<td>3</td>
</tr>
<tr>
<td>(\mathbf{b}_{i,OLS})</td>
<td>ordinary least squares estimator of (\mathbf{\alpha}_i + \mathbf{\beta}) in the random coefficients model</td>
<td>3</td>
</tr>
<tr>
<td>(\mathbf{D}_{\text{SWAMY}})</td>
<td>Swamy's estimator of (\mathbf{D})</td>
<td>3</td>
</tr>
<tr>
<td>(\tilde{\mathbf{y}}_{i,s})</td>
<td>shrinkage estimator of (\mu + \mathbf{\alpha}_i) in the one-way random effects ANOVA model</td>
<td>4</td>
</tr>
<tr>
<td>(\zeta_i)</td>
<td>weighting (credibility) factor used to compute the shrinkage estimator</td>
<td>4</td>
</tr>
<tr>
<td>(\mathbf{y}_{i,\text{BLUP}})</td>
<td>best linear unbiased predictor (BLUP) of (\mu + \mathbf{\alpha}_i) in the one-way random effects ANOVA model</td>
<td>4</td>
</tr>
<tr>
<td>(m_{\mu,\text{GLS}})</td>
<td>GLS estimator of (\mu) in the one-way random effects ANOVA model</td>
<td>4</td>
</tr>
<tr>
<td>(e_{i,\text{BLUP}})</td>
<td>BLUP residual, predictor of (\mathbf{e}_i)</td>
<td>4</td>
</tr>
<tr>
<td>(w)</td>
<td>generic random variable to be predicted</td>
<td>4</td>
</tr>
<tr>
<td>(w_{\text{BLUP}})</td>
<td>BLUP predictor of (w)</td>
<td>4</td>
</tr>
<tr>
<td>(d_{i,\text{BLUP}})</td>
<td>BLUP predictor of (\mathbf{\alpha}_i) in the one-way error components model</td>
<td>4</td>
</tr>
<tr>
<td>(\mathbf{\beta}_{i,j})</td>
<td>level 2 dependent variable in a three-level model</td>
<td>5</td>
</tr>
<tr>
<td>(y_i)</td>
<td>level 3 dependent variable in a three-level model</td>
<td>5</td>
</tr>
<tr>
<td>(\chi^2(q))</td>
<td>chi–square random variable with (q) degrees of freedom</td>
<td>5</td>
</tr>
<tr>
<td>(y_{i_1,i_2,\ldots,i_k})</td>
<td>A typical dependent variable in a (k)-level model</td>
<td>5</td>
</tr>
<tr>
<td>(i(k))</td>
<td>Index set. The set of all indices ((i_1, i_2, \ldots, i_k)) such that (y_{i_1,i_2,\ldots,i_k}) is observed.</td>
<td>5</td>
</tr>
<tr>
<td>(i(k))</td>
<td>A typical element of (i(k)) of the form ((i_1, i_2, \ldots, i_k)).</td>
<td>5</td>
</tr>
<tr>
<td>(i(k-s))</td>
<td>Index set. The set of all indices ((i_1, \ldots, i_{k-s})) such that (y_{i_1,i_2,\ldots,i_{k-s},\ldots,i_k}) is observed for some ({j_{k-s+1}, \ldots, j_k}).</td>
<td>5</td>
</tr>
<tr>
<td>(i(k-s))</td>
<td>A typical element of (i(k-s)).</td>
<td>5</td>
</tr>
<tr>
<td>Symbol</td>
<td>Description</td>
<td>Chapter defined</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>( Z_{(g)}^{(k+1-g)} )</td>
<td>Level ( g ) covariates matrices in the high order multilevel model, analogous to Chapter 2 and 3 ( Z ) and ( X ) matrices.</td>
<td>5</td>
</tr>
<tr>
<td>( X_{(g)}^{(k+1-g)} )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta_{(k-g)}^{(g)} )</td>
<td>Level ( g ) parameter matrices in the high order multilevel model, analogous to Chapter 2 and 3 ( \alpha ) and ( \beta ).</td>
<td>5</td>
</tr>
<tr>
<td>( \beta_g )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta_{(g-1)}^{(k+1-g)} )</td>
<td>Level ( g ) response and disturbance terms in the high order multilevel model, analogous to Chapter 2 and 3 ( \gamma ) and ( \epsilon ).</td>
<td>5</td>
</tr>
<tr>
<td>( \epsilon_{(g)}^{(k+1-g)} )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( L(\epsilon_i</td>
<td>x_i) )</td>
<td>A linear projection of ( \epsilon_i ) on ( x_i ).</td>
</tr>
<tr>
<td>( f(y_1, ..., y_T, x_1, ..., x_T) )</td>
<td>A generic joint probability density (or mass) function { ( y ), ..., ( y ), ( x ), ..., ( x ) }.</td>
<td>6</td>
</tr>
<tr>
<td>( \theta, \psi )</td>
<td>Vectors of parameters for ( f(\cdot) ).</td>
<td>6</td>
</tr>
<tr>
<td>( w_i )</td>
<td>A set of predetermined variables.</td>
<td>6</td>
</tr>
<tr>
<td>( W )</td>
<td>A matrix of instrumental variables.</td>
<td>6</td>
</tr>
<tr>
<td>( P_W )</td>
<td>A projection matrix, ( P_W = W(W'W)^{-1}W' ).</td>
<td>6</td>
</tr>
<tr>
<td>( X^* )</td>
<td>A collection of all observed explanatory variables, ( X^* = { X_1, Z_1, ..., X_n, Z_n } ).</td>
<td>6</td>
</tr>
<tr>
<td>( o_{it} ), ( o_i )</td>
<td>( o_{it} ) is a ((q+K) \times 1) vector of observed effects ( o_{it} = (z_{it}', x_{it}')' ), ( o_i = (o_{i1}', ..., o_{iT}')' ).</td>
<td>6</td>
</tr>
<tr>
<td>( \Delta )</td>
<td>First difference operator, for example, ( \Delta y_{it} = y_{it} - y_{i,t-1} ).</td>
<td>6</td>
</tr>
<tr>
<td>( K )</td>
<td>A ((T-1) \times T) upper triangular matrix such that ( K I = 0 ).</td>
<td>6</td>
</tr>
<tr>
<td>( b_{IV} )</td>
<td>Instrumental variable estimator of ( \beta ).</td>
<td>6</td>
</tr>
<tr>
<td>( \Sigma_{IV} )</td>
<td>Variance-covariance matrix used to compute the variance of ( b_{IV} ).</td>
<td>6</td>
</tr>
<tr>
<td>( \Sigma_{IV} = E(W_i'K \epsilon_i \epsilon_i'K'W_i) ).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( Y, X, \Gamma )</td>
<td>Responses, explanatory variables and parameters in the multivariate regression model.</td>
<td>6</td>
</tr>
<tr>
<td>( \Sigma )</td>
<td>Variance of the response in the multivariate regression model.</td>
<td>6</td>
</tr>
<tr>
<td>( G_{OLS} )</td>
<td>Ordinary least squares estimator of ( \Gamma ) in the multivariate regression model.</td>
<td>6</td>
</tr>
<tr>
<td>( B, \Gamma )</td>
<td>Regression coefficients for endogenous and exogenous regressors, respectively, in the simultaneous equations model.</td>
<td>6</td>
</tr>
<tr>
<td>( \Pi )</td>
<td>Reduced form regression coefficients in the simultaneous equations model, ( \Pi = (I-B)^{-1} \Gamma ).</td>
<td>6</td>
</tr>
<tr>
<td>( \eta_i )</td>
<td>Reduced form disturbance term in the simultaneous equations model, ( \eta_i = (I-B)^{-1} \epsilon_i ).</td>
<td>6</td>
</tr>
<tr>
<td>( \Omega )</td>
<td>Variance of the reduced form disturbance term in the simultaneous equations model.</td>
<td>6</td>
</tr>
<tr>
<td>Symbol</td>
<td>Description</td>
<td>Chapter defined</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>$\tau_x, \Lambda_x$</td>
<td>regression coefficients in the $x$-measurement equation of simultaneous equations</td>
<td>6</td>
</tr>
<tr>
<td>$\xi_i, \delta_i$</td>
<td>latent explanatory variable and disturbance term, respectively, in the $x$-measurement equation of simultaneous equations</td>
<td>6</td>
</tr>
<tr>
<td>$\mu_\xi$</td>
<td>expected value of $\xi_i$, $E(\xi_i) = \mu_\xi$</td>
<td>6</td>
</tr>
<tr>
<td>$\Phi, \Theta_\delta$</td>
<td>variances of $\xi_i$ and $\delta_i$, respectively. $Var(\xi_i) = \Phi$ and $Var(\delta_i) = \Theta_\delta$.</td>
<td>6</td>
</tr>
<tr>
<td>$\tau_y, \Lambda_y$</td>
<td>regression coefficients in the $y$-measurement equation of simultaneous equations</td>
<td>6</td>
</tr>
<tr>
<td>$\eta_i, \varepsilon_i$</td>
<td>latent explanatory variable and disturbance term, respectively, in the $y$-measurement equation of simultaneous equations</td>
<td>6</td>
</tr>
<tr>
<td>$\Theta_\varepsilon$</td>
<td>variances $\varepsilon_i$, $Var(\varepsilon_i) = \Theta_\varepsilon$</td>
<td>6</td>
</tr>
<tr>
<td>$\tau_{\eta}, B, \Gamma$</td>
<td>regression coefficients in the structural equation of simultaneous equations</td>
<td>6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Chapter defined</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi(.)$</td>
<td>Standard normal probability density function</td>
<td>7</td>
</tr>
<tr>
<td>$\Phi(.)$</td>
<td>Standard normal probability distribution function</td>
<td>7</td>
</tr>
<tr>
<td>$o_{it}, o_i$</td>
<td>$o_{it}$ is a $(q+K) \times 1$ vector of observed effects $o_{it} = (z_{it}', x_{it}')'$, $o_i = (o_{i1}', \ldots, o_{iT_i}')'$.</td>
<td>7</td>
</tr>
<tr>
<td>$U_i$</td>
<td>a $T_i \times g$ matrix of unobserved, independent variables</td>
<td>7</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>a $g \times 1$ vector of parameters corresponding to $U_i$</td>
<td>7</td>
</tr>
<tr>
<td>$\eta_i$</td>
<td>a $T_i \times 1$ i.i.d. noise vector</td>
<td>7</td>
</tr>
<tr>
<td>$G_i$</td>
<td>a $T_i \times g$ matrix of “augmented,” independent variables</td>
<td>7</td>
</tr>
<tr>
<td>$M_i$</td>
<td>a $T_i \times T$ matrix used to specify the availability of observations</td>
<td>7</td>
</tr>
<tr>
<td>$r_{ij}, r$</td>
<td>$r_{ij}$ is an indicator of the $ij$th observation (a “1” indicates the observation is available), $r = (r_{11}, \ldots, r_{1T_i}, \ldots, r_{n1}, \ldots, r_{nT})'$</td>
<td>7</td>
</tr>
<tr>
<td>$Y$</td>
<td>$Y$ is the vector of all potentially observed responses, $Y = (y_{11}, \ldots, y_{1T_i}, \ldots, y_{n1}, \ldots, y_{nT})'$</td>
<td>7</td>
</tr>
<tr>
<td>Symbol</td>
<td>Description</td>
<td>Chapter defined</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td>----------------</td>
</tr>
<tr>
<td>$R_{AR}(\rho)$</td>
<td>Correlation matrix corresponding to an $AR$ (1) process</td>
<td>8</td>
</tr>
<tr>
<td>$R_{RW}(\rho)$</td>
<td>Correlation matrix corresponding to a (generalized) random walk process</td>
<td>8</td>
</tr>
<tr>
<td>$M_i$</td>
<td>$T_i \times T_i$ design matrix that describes missing observations</td>
<td>8</td>
</tr>
<tr>
<td>$\rho(u)$</td>
<td>correlation function for spatial data</td>
<td>8</td>
</tr>
<tr>
<td>$H$</td>
<td>spatial variance matrix, $= \text{Var } \varepsilon_t$</td>
<td>8</td>
</tr>
<tr>
<td>$V_H$</td>
<td>$\sigma^2_J + H$</td>
<td>8</td>
</tr>
<tr>
<td>$r$</td>
<td>number of time-varying coefficients per time period</td>
<td>8</td>
</tr>
<tr>
<td>$\lambda$, $\lambda_1$, $\ldots$, $\lambda_p$</td>
<td>vectors of time-varying coefficients, $\lambda_i = (\lambda_{i1}, \ldots, \lambda_{ip})'$, $\lambda = (\lambda_1, \ldots, \lambda_T)'$</td>
<td>8</td>
</tr>
<tr>
<td>$z_{\alpha,iti,j}, z_{\alpha,iti}$</td>
<td>Explanatory variables associated with $\alpha_i$; they are the same as Chapter 2</td>
<td>8</td>
</tr>
<tr>
<td>$Z_{\alpha,iti}$, $Z_{\alpha}$</td>
<td>$z_{iti}, z_{iti}, Z_{\alpha}$ and $Z_i$.</td>
<td>8</td>
</tr>
<tr>
<td>$z_{\lambda,iti,j}$, $z_{\lambda,iti}$, $Z_{\lambda,i}$, $Z_{\lambda}$</td>
<td>Explanatory variables associated with $\lambda_i$; similar to $z_{\alpha,iti}, z_{\alpha,iti}, Z_{\alpha,i}$ and $Z_{\alpha}$.</td>
<td>8</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>parameter associated with lagged dependent variable model</td>
<td>8</td>
</tr>
<tr>
<td>$\Phi_1$, $\Phi_2$</td>
<td>matrix of time-varying parameters for a transition equation</td>
<td>8</td>
</tr>
<tr>
<td>$T_i$</td>
<td>matrix of time-varying parameters for a generalized transition equation</td>
<td>8</td>
</tr>
<tr>
<td>$\phi_1, \ldots, \phi_p$</td>
<td>parameters of the $AR(p)$ process, autoregressive of order $p$</td>
<td>8</td>
</tr>
<tr>
<td>$\delta_i$</td>
<td>vector of unobservables in Kalman filter algorithm, $\delta_i = (\lambda_i', \xi_i')'$</td>
<td>8</td>
</tr>
<tr>
<td>$p_{it}$</td>
<td>probability of $y_{it}$ equaling one</td>
<td>9</td>
</tr>
<tr>
<td>$\pi(z)$</td>
<td>distribution function for binary dependent variables</td>
<td>9</td>
</tr>
<tr>
<td>$u_i$</td>
<td>utility function for the $i$th subject at time $t$</td>
<td>9</td>
</tr>
<tr>
<td>$U_{ij}, V_{ij}$</td>
<td>unobserved utility and value of the $j$th choice for the $i$th subject at time $t$</td>
<td>9</td>
</tr>
<tr>
<td>logit($p$)</td>
<td>logit function, defined as $\text{logit}(p) = \ln \left( \frac{p}{1-p} \right)$</td>
<td>9</td>
</tr>
<tr>
<td>$LRT$</td>
<td>likelihood ratio test statistic</td>
<td>9</td>
</tr>
<tr>
<td>$R_{ms}^2$</td>
<td>max-scaled coefficient of determination ($R^2$)</td>
<td>9</td>
</tr>
<tr>
<td>$p(\alpha_i)$</td>
<td>conditional distribution, given the random effect $\alpha_i$</td>
<td>9</td>
</tr>
<tr>
<td>$\alpha_{i,MLE}$</td>
<td>maximum likelihood estimator of $\alpha_i$</td>
<td>9</td>
</tr>
<tr>
<td>$\mu_{it}$</td>
<td>mean of the the $i$th subject at time $t$, $\mu_{it} = \mathbb{E} y_{it}$</td>
<td>9</td>
</tr>
<tr>
<td>$\mu_i$</td>
<td>$T_i \times 1$ vector of means, $\mu_i = (\mu_{i1}, \ldots, \mu_{iT})'$</td>
<td>9</td>
</tr>
<tr>
<td>$G_{\mu}(\beta, \tau)$</td>
<td>matrix of derivatives</td>
<td>9</td>
</tr>
<tr>
<td>$b_{EE}$, $\tau_{EE}$</td>
<td>estimating equations estimators of $\beta$ and $\tau$</td>
<td>9</td>
</tr>
</tbody>
</table>
APPENDIX F. Selected Longitudinal and Panel Data Sets

In many disciplines, longitudinal and panel data sets can be readily constructed using a traditional pre- and post-event study; that is, subjects are observed prior to some condition of interest as well as after this condition, or event. It is also common to use longitudinal and panel data methods to examine data sets that follow subjects observed at an aggregate level, such as a government entity (state, province or nation, for example) or firm.

Longitudinal and panel data sets are also available that follow individuals over time. However, these data sets are generally expensive to construct and are conducted through the sponsorship of a government agency. Thus, although available, data providers are generally bound by national laws requiring some form of user agreement to protect confidential information regarding the subjects. Because of the wide interest in these data sets, most data providers make information about the data sets available on the Internet.

Despite the expense and confidentiality requirements, many countries have conducted, or are in the process of conducting, household panel studies. Socio-demographic and economic information is collected about a household as well as individuals within the household. Information may relate to income, wealth, education, health, geographic mobility, taxes, and so forth. To illustrate, one of the oldest ongoing national panels, the US Panel Study of Income Dynamics (PSID), collects 5,000 variables. Table F.1 cites some major international household panel data sets.

Education is another discipline that has a long history of interest in longitudinal methods. Table F.2 cites some major educational longitudinal data sets. Similarly, because of the dynamic nature of aging and retirement, analysts rely on longitudinal data for answers to important social science questions concerned with these issues. Table F.3 cites some major aging and retirement longitudinal data sets.

Longitudinal and panel data methods are used in many scientific disciplines. Table F.4 cites some other widely used data sets. The focus of Table F.4 is on political science through election surveys and longitudinal surveys of the firm.

<table>
<thead>
<tr>
<th>Panel Study (years available) plus web site</th>
<th>Sample Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canadian Survey of Labor Income Dynamics (1993-) <a href="http://www.statcan.ca">www.statcan.ca</a></td>
<td>Approximately 15,000 households or 31,000 individuals.</td>
</tr>
<tr>
<td>French Household Panel (1985-1900) <a href="http://www.ceps.lu/paco/pacofrpa.htm">www.ceps.lu/paco/pacofrpa.htm</a></td>
<td>There were 715 households at the baseline, increased to 2,092 in the second wave.</td>
</tr>
<tr>
<td>German Social Economic Panel (1984-) <a href="http://www.diw.de/soep">www.diw.de/soep</a></td>
<td>First wave collected in 1984, included 5,921 West German households consisting of 12,245 individuals.</td>
</tr>
</tbody>
</table>
### Table F.1 International Household Panel Studies - Continued

<table>
<thead>
<tr>
<th>Study (years available) plus web site</th>
<th>Sample Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indonesia Family Life Survey (1993-) <a href="http://www.rand.org/FLS/IFLS/">www.rand.org/FLS/IFLS/</a></td>
<td>In 1993, 7,224 households were interviewed.</td>
</tr>
<tr>
<td>Japanese Panel Survey on Consumers (1994-) <a href="http://www.kakeiken.or.jp">www.kakeiken.or.jp</a></td>
<td>National representative sample of 1,500 women age 24-34 in 1993; in 1997, 500 women were added.</td>
</tr>
<tr>
<td>Korea Labor and Income Panel Study (1998-) <a href="http://www.kli.re.kr/klips">www.kli.re.kr/klips</a></td>
<td>Sample contains 5,000 households.</td>
</tr>
<tr>
<td>Mexican Family Life Survey (2001-)</td>
<td>Will contain about 8,000 households. Plans are to collect data at two points in time, 2001 and 2004.</td>
</tr>
<tr>
<td>Polish Household Panel (1987-1990) <a href="http://www.ceps.lu/paco/pacopopa.htm">www.ceps.lu/paco/pacopopa.htm</a></td>
<td>Four waves available of a sample of persons living in private households, excluding police officers, military personnel and members of the “nomenklatura.”</td>
</tr>
<tr>
<td>Swiss Household Panel (1999-) <a href="http://www.unine.ch/psm">www.unine.ch/psm</a></td>
<td>First wave in 1999 consists of 5,074 households comprising 7,779 individuals.</td>
</tr>
<tr>
<td>U.S. Panel Study of Income Dynamics (1968-) <a href="http://www.isr.umich.edu/src/psid/index.html">www.isr.umich.edu/src/psid/index.html</a></td>
<td>Began with 4,802 families, with an oversampling of poor families. Annual interviews were conducted – over 5,000 variables were collected on roughly 31,000 individuals.</td>
</tr>
</tbody>
</table>

Sources: Institute for Social Research, University of Michigan
www.isr.umich.edu/src/psid/panelstudies.html

### Table F.2 Youth and Education

<table>
<thead>
<tr>
<th>Study (years available) plus web site</th>
<th>Sample Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early Childhood Longitudinal Study (1998-) <a href="http://www.nees.ed.gov/ecls/">www.nees.ed.gov/ecls/</a></td>
<td>Includes a Kindergarten cohort and a Birth Cohort. The Kindergarten cohort consists of a nationally representative sample of approximately 23,000 kindergartners from about 1,000 kindergarten programs. The Birth Cohort includes a nationally representative sample of approximately 15,000 children born in the calendar year 2000.</td>
</tr>
</tbody>
</table>
### Table F.2 Youth and Education - Continued

<table>
<thead>
<tr>
<th>Study (years available) plus web site</th>
<th>Sample Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>High School and Beyond (1980-1992) <a href="http://www.nces.ed.gov/surveys/hsb/">www.nces.ed.gov/surveys/hsb/</a></td>
<td>The High School and Beyond survey included two cohorts: the 1980 senior class, and the 1980 sophomore class. Both cohorts were surveyed every two years through 1986, and the 1980 sophomore class was also surveyed again in 1992.</td>
</tr>
<tr>
<td>National Longitudinal Study of the High School Class of 1972 (1972-86) <a href="http://www.nces.ed.gov/surveys/nls72/">www.nces.ed.gov/surveys/nls72/</a></td>
<td>This survey followed the 1972 cohort of high school seniors through 1986. The original sample was drawn in 1972; follow-up surveys were conducted in 1973, 1974, 1976, 1979, and 1986.</td>
</tr>
<tr>
<td>The National Longitudinal Survey of Youth 1997 (NLS) <a href="http://www.bls.gov/nls/nlsy97.htm">www.bls.gov/nls/nlsy97.htm</a></td>
<td>A nationally representative sample of approximately 9,000 youths who were 12 to 16 years old in 1996.</td>
</tr>
</tbody>
</table>

### Table F.3 Elderly and Retirement

<table>
<thead>
<tr>
<th>Study (years available) plus web site</th>
<th>Sample Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Framingham Heart Study (1948-) <a href="http://www.nhlbi.nih.gov/about/framingham/index.html">www.nhlbi.nih.gov/about/framingham/index.html</a></td>
<td>In 1948, 5,209 men and women between the ages of 30 and 62 were recruited to participate in this heart study. They are monitored every other year. In 1971, 5,124 of the original participants' adult children and their spouses were recruited to participate in similar examinations.</td>
</tr>
<tr>
<td>Health and Retirement Study (HRS) hrsonline.isr.umich.edu/</td>
<td>The original HRS cohort born 1931-1941 and first interviewed in 1992 (ages 51-61). The AHEAD cohort born before 1923 and first interviewed in 1993 (ages 70 and above). Spouses were included, regardless of age. These cohorts were merged in the 1998 wave and include over 21,000 participants. Variables collected include income, employment, wealth, health conditions, health status, health insurance coverage and so forth.</td>
</tr>
</tbody>
</table>
### Table F.3 Elderly and Retirement - Continued

<table>
<thead>
<tr>
<th>Study (years available) plus web site</th>
<th>Sample Description</th>
</tr>
</thead>
</table>
| The National Longitudinal Surveys of Labor Market Experience (NLS) www.bls.gov/nls/ | The NLS follows five distinct labor markets:  
- 5,020 older men (between 45 and 49 in 1966),  
- 5,225 young men (between 14 and 24 in 1966),  
- 5,083 mature women (between 30 and 44 in 1967),  
- 5,159 young women (between 14 and 21 in 1968), and  
- 12,686 youths (between 14 and 24 in 1979, additional cohorts were added in 1986 and 1997). The list of variables is in the thousands, with an emphasis on the supply side of the labor market. |

### Table F.4 Other Longitudinal and Panel Studies

<table>
<thead>
<tr>
<th>Study (years available) plus web site</th>
<th>Sample Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>National Election Studies 1956, 1958, 1960 American Panel Study <a href="http://www.umich.edu/~nes/studyres/nes56_60/nes56_60.htm">www.umich.edu/~nes/studyres/nes56_60/nes56_60.htm</a></td>
<td>A sample of 1,514 voters who were interviewed at most five times.</td>
</tr>
<tr>
<td>National Election Studies 1972, 1974, 1976 Series File <a href="http://www.umich.edu/~nes/studyres/nes72_76/nes72_76.htm">www.umich.edu/~nes/studyres/nes72_76/nes72_76.htm</a></td>
<td>A sample of 4,455 voters who were interviewed at most five times.</td>
</tr>
<tr>
<td>National Election Studies 1980 Panel Study <a href="http://www.umich.edu/~nes/studyres/nes80pan/nes80pan.htm">www.umich.edu/~nes/studyres/nes80pan/nes80pan.htm</a></td>
<td>Over 1,000 voters were interviewed four times over the course of the 1980 presidential election.</td>
</tr>
<tr>
<td>National Election Studies 1990-1992 Full Panel File <a href="http://www.umich.edu/~nes/studyres/nes90_92/nes90_92.htm">www.umich.edu/~nes/studyres/nes90_92/nes90_92.htm</a></td>
<td>Voter opinions are traced to follow the fortunes of the Bush presidency.</td>
</tr>
<tr>
<td>Census Bureau Longitudinal Research Database (1980-) <a href="http://www.census.gov/pub/econ/www/ma0800.html">www.census.gov/pub/econ/www/ma0800.html</a></td>
<td>Links establishment level data from several censuses and surveys of manufacturers, and can respond to diverse economic research priorities.</td>
</tr>
<tr>
<td>Medical Expenditure Panel Survey (1996-) <a href="http://www.meps.ahrq.gov">www.meps.ahrq.gov</a></td>
<td>Surveys of households, medical care providers, as well as business establishments and governments on health care use and costs.</td>
</tr>
<tr>
<td>National Association of Insurance Commissioners (NAIC) <a href="http://www.naic.org/1dbproducts/">www.naic.org/1dbproducts/</a></td>
<td>Maintains annual and quarterly data for more than 6,000 Life/Health, Property/Casualty, Fraternal, Health and Title companies.</td>
</tr>
</tbody>
</table>
Appendix G. References

Biological Sciences Longitudinal Data References


**Econometrics Panel Data References**


**Educational Science and Psychology References**


**Other Social Science References**


**Statistical Longitudinal Data References**


**General Statistics References**


Index

(page numbers correspond to double-spaced version)

aggregation bias, 191, 519
analysis of covariance model, 26, 34, 363
artificial regression, 291, 515
attrition, v, xi, xiii, 1, 6, 13, 16, 52, 270, 294, 300, 459
Bayesian, 118, 152, 183, 194, 482
canonical link, 391, 400, 412, 415, 422, 447, 475
capital asset pricing model (CAPM), 335
capital asset pricing model (CAPM), 335
categorical dependent variable, 14, 234
causal effects, 12, 228
censoring, 299, 459, 463
computational methods
  adaptive Gaussian quadrature, 366, 381
  Fisher scoring method, 116, 403, 474
  Newton-Raphson, 116, 384, 403, 411, 474
correlation, 26, 319
covariance structure analysis, 258, 521
credibility
  credibility factor, 89, 90, 143, 144, 169, 172, 173
  credibility premium, 14
  credibility ratemaking, 170, 172
credibility theory, 142, 170, 173, 175, 522
cross-sectional correlation, 47, 48, 49, 309, 320, 321
cumulative hazard function, 461
data
  balanced data, 4, 16, 47, 49, 69, 73, 167, 168, 179, 250, 279, 294, 305, 318, 321, 327
data exploration, 21, 28, 41
data snooping, 42
  unbalanced data, 4, 91, 118, 244, 275, 279, 321, 331, 343
data generating process, 14, 234
diagnostic statistic, 45, 62, 145
differenced data, 72, 241
distributions
  Bernoulli distribution, 388, 391, 407, 411, 430
  extreme-value distribution, 441, 443
  exponential family, 389, 394, 412
  generalized extreme-value distribution, 442, 443
  multinomial distribution, 416, 445
  normal distribution, 117, 146, 216, 300, 328, 355, 365, 389, 471
  Poisson distribution, vii, xiv, xv, 389, 394, 397, 408, 409, 415, 424, 430, 445
dynamic model, 246, 309
EM algorithm, 303
endogeneity, 223, 232, 234, 246, 254
  endogenous variable, 223, 231, 255, 263
  exogeneity, 222, 236, 243, 268, 281, 309
  exogenous variable, 104, 233, 254, 263
  Granger-cause, 227
  predetermined, 225, 230, 233, 236, 497
  sequentially exogenous, iv, 239, 242, 246
  strongly exogenous, 427
  weakly exogenous, 226
error components model. See random effects.
errors in variables model, 261
examples, 17
capital structure, 139
capital structure, 139
dependable contributions, 73, 137
dental, 196, 202, 267
group term life insurance, 18
hospital costs, 21, 29, 38, 43, 46, 49, 55, 87
housing price, 77, 138, 182
income tax payments, 73, 82, 92, 238, 277, 285, 292, 298, 350, 359, 366, 379, 454, 457
lottery sales, 104, 141, 155, 324
student achievement, 186, 190, 217, 520
tort filings, 75, 138, 395, 404, 410, 413
workers’ compensation, 183
yogurt, 385, 438, 446, 451
exponential correlation model, 316
factor analysis, 259, 261
factorization theorem, 415, 425, 464
failure rate, 461
feasible generalized least squares, 274, 275, 320
basic fixed effects model, 21, 25, 59, 64, 67, 68, 69, 83, 89, 162, 274
fixed effects estimator, 62, 89, 109, 128, 144, 238, 275, 280, 363, 413
fixed effects linear longitudinal data model, 56
fixed effects model, xi, 9, 20, 21, 59, 83, 105, 142, 150, 162, 182, 274, 282, 284, 338, 350, 363, 388, 517
one-way fixed effects model, 26, 75, 77, 79, 80, 182, 274, 302
two-way fixed effects model, 27, 40, 77
forecast, 16, 153, 155, 161, 162, 163, 164, 180, 181, 202, 204, 217, 221, 326, 327, 348
forward orthogonal deviations, 243
Fuller-Battese transform, 304
Gaussian correlation model, 316
Gaussian theory of errors, 24
generalized estimating equation (GEE), xiv, 375, 399
generalized linear model (GLM), 388, 421, 447
generalized method of moments (GMM), xiv, 375, 402
growth curve model, 16, 195, 267
Hausman test, 270, 275
hazard function, 460
National Association of Insurance Commissioners (NAIC), 5, 510
National Longitudinal Survey of Labor Market Experience, 5
National Longitudinal Survey of Youth (NLSY), 27, 507
Panel Study of Income Dynamics (PSID), 13, 27, 503
scanner data, 5, 386, 438
marginal distribution, 227, 418, 421, 428, 447
marginal model, xv, 350, 375, 377, 379, 388, 399, 400, 401, 404
Markov model, 447, 456
martingale difference, 226, 234
measurement equation, 259, 261, 262, 264, 267, 329, 340, 499
measurement error, 259, 266, 330
MIMIC model, 267
missing
ignorable case, 298
missing at random (MAR), 298
missing completely at random (MCAR), 296
missing data, v, 52, 268, 294, 295, 296, 297, 300, 304, 329, 531
non-ignorable case, 299
selection model, 297, 298, 301
selectivity bias, v, xiii, 294
unplanned nonresponse, 296
mixed effects
linear mixed effects model, xii, 82, 98, 141, 150, 163, 185, 187, 193, 201, 203, 206, 209, 212, 234, 282, 317, 329, 343
mixed effects models, 98, 141, 185, 212, 412
mixed linear model, xii, 82, 106, 117, 120, 122, 146, 150, 164, 165, 170, 181, 193, 210, 221, 318, 323, 343, 471, 487
nonlinear mixed effects model, 411, 423
moment generating function, 424
multilevel
cross-level effects, 190
cross-level interaction, 188
data, 185, 188
hierarchical model, 164, 188, 441, 520, 522
high order multilevel models, 208, 209
multilevel model, xii, 185, 268, 366
three-level model, 191, 193, 208, 213, 369, 496
two-level model, 187, 189, 192, 194, 201, 202, 204, 213, 369
multivariate regression model, 248, 252, 253, 261, 498
nonlinear regression model, 24, 359, 371, 388, 412
nonstationary model, 314
nuisance parameters, 25, 88, 423
observables representation, 23, 86, 99, 237. See also marginal model.
observation equation, 329, 332, 485, 487
observational data, x, 12, 229
occurrence dependence, 458
odds ratio, 356, 357, 386, 434, 438
omitted variable, xiii, 10, 40, 231, 270, 276, 281, 286, 288
model of correlated effects, 281, 283, 284
unmeasured confounder, 289
unobserved variable model, 289
unobserved variables, 24, 230, 290, 301
overdispersion, 378, 394, 398, 408, 409
partial likelihood, 450, 462, 531
path diagram, 262
plotting methods
added variable plot, 28, 29, 44, 69, 75, 77, 79, 96, 138
boxplot, 28, 95, 219
multiple time series plot, 28, 29, 31, 74, 138, 139, 158, 219, 397
partial regression plot, 44
scatter plot with symbols, 31
trellis plot, 33, 337
pooled cross-sectional model, 96, 159, 162
pooled cross-sectional time series, 5
pooling test, 42, 43, 69, 91, 92, 96, 132, 162
population parameter, 25, 26, 45, 83, 88, 106, 274
population-averaged model, 86, 100
potential outcomes, 229
prediction, xii, 14, 141, 146, 200, 222, 341, 346, 422, 485. See also forecasting.
probit regression model, vi, 302, 352, 354, 357, 385
profile log-likelihood, 115
random censoring, 459
error components model, 20, 82, 96, 98, 101, 104, 107, 109, 115, 122, 128, 129, 130, 131, 134, 137, 138, 139, 151, 159, 179, 182, 189, 198, 201, 205, 213, 219,
models with random effects, 82. See also random effects model.

one-way random effects ANOVA model, 104, 143, 148, 167, 171

random coefficients model, 102, 103, 110, 119, 122, 133, 151, 171, 179, 190, 201, 286

random effects estimators, 238, 275, 279

random intercepts model, 82, 83. See also error components model.

two-way error component, 91, 107, 255, 324

random utility interpretation, 385, 442

random walk model, 314

repeated measures design, 101

residual standard deviation, 36, 37

rotating panel, v, 294, 295

sampling based model, 14, 15, 87, 228, 230

score equations, 358, 359, 365, 370, 372

score vector, 111

seemingly unrelated regression, 251, 252, 255, 256, 257, 320, 513, 519

serial correlation, xiii, 26, 52, 98, 121, 153, 154, 195, 250, 309, 312, 320, 323, 408

autoregressive model, 54, 120, 202, 266, 312, 313, 405

moving average model, 313

serial correlation models, xiii, 312

serial covariance. See temporal variance-covariance matrix.

shrinkage estimator, 141, 143, 144, 495, 496

simultaneous equations model, 498

small area estimation, 142, 175

spatial correlations, 317, 329

standard errors (for regression coefficient estimators)

empirical, 378, 379, 404, 405

model-based, 139, 379, 404, 405

panel-corrected, 321

robust, 57, 111, 135, 245, 404

state space models, 328, 329, 528

stratified sample, 23, 56, 85, 92, 102

structural equation, 229, 231, 259, 261, 262, 264, 265, 266, 267, 268, 499

structural model, 15, 16, 87, 228, 229, 230, 231, 234

subject-specific parameters, 9, 25, 42, 45, 52, 55, 82, 166, 272, 288, 389

subject-specific intercept, 29, 101

subject-specific slopes, 55, 272

sufficiency, 358, 372, 415, 425, 464

survival model, 431, 458, 459, 461, 463

accelerated failure time model, 461

Cox proportional hazard model, 461, 462

systematic component, 390, 391, 392, 406, 412, 439, 444, 447, 449, 455, 475

temporal variance-covariance matrix, 53, 59, 195, 313

autoregressive, 53, 108, 120, 154, 202, 265, 266, 313, 405

exchangeable correlation, 378, 401

compound symmetry, 53, 54, 71, 233, 272, 313

Toeplitz matrix, 313

uniform correlation, 53

time-constant variables, 35, 88, 104, 105, 287, 288, 291

time series analysis, xiv, 2, 7, 52, 54, 313, 328, 483

time-series cross-section (TSCS) model, 319

time-series cross-section data, 319, 521

time-varying coefficients, 322, 338, 339, 340, 501

transition

equation, 329, 330, 331, 485, 487, 501

matrix, 448

model, 431, 448, 449, 455, 457, 458, 459

probability, 456

two-stage least squares, 232, 254, 255, 257, 258

two-stage sampling scheme, 84, 164, 363

unit of analysis, 1, 155, 193, 194, 200, 219


variance components estimators, 114, 122, 202, 525

minimum variance quadratic unbiased estimator (MIVQUE), 102, 120, 313

moment-based variance component estimator, 119

residual maximum likelihood estimation, 117

restricted maximum likelihood (REML), 116, 122, 143, 334, 345

testing variance components, 204

Wald test, 113, 292

working correlation, 378, 379, 404, 405