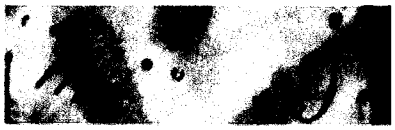
SECOND EDITION



ECONOMETRIC ANALYSIS OF CROSS SECTION AND PANEL DATA

JEFFREY M. WOOLDRIDGE

Dockets.Justia.com

© 2010, 2002, Massachusetts Institute of Technology

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

For information about special quantity discounts, please email special_sales@mitpress.mit.edu

This book was set in Times Roman by Asco Typesetters, Hong Kong. Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Wooldridge, Jeffrey M.
Econometric analysis of cross section and panel data / Jeffrey M. Wooldridge.—2nd ed.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-262-23258-6 (hardcover : alk. paper)
1. Econometrics—Asymptotic theory. I. Title.
HB139.W663 2010
330.01'5195—dc22 2010020912

10 9 8 7 6 5 4 3 2 1

Preface

It has been almost 10 years since the first edition of *Econometric Analysis of Cross Section and Panel Data* was published. The reaction to the first edition was more positive than I could have imagined when I began thinking about the project in the mid-1990s. Of course, as several of you have kindly and constructively pointed out—and as was evident to me the first time I taught out of the book—the first edition was hardly perfect. Issues of organization and gaps in coverage were shortcomings that I wanted to address in a second edition from early on. Plus, there have been some important developments in econometrics that can and should be taught to graduate students in economics.

I doubt this second edition is perfect, either. But I believe it improves the first edition in substantive ways. The structure of this edition is similar to the first edition, but I have made some changes that will contribute to the reader's understanding of several topics. For example, Chapter 11, which covers more advanced topics in linear panel data models, has been rearranged to progress more naturally through situations where instrumental variables are needed in conjuction with methods for accounting for unobserved heterogeneity. Data problems—including censoring, sample selection, attrition, and stratified sampling—are now postponed until Chapters 19 and 20, after popular nonlinear models are presented under random sampling. I think this change will further emphasize a point I tried to make in the first edition: It is critical to distinguish between specifying a population model on the one hand and the method used to sample the data on the other. As an example, consider the Tobit model. In the first edition, I presented the Tobit model as applying to two separate cases: (1) a response variable is a corner solution outcome in the population (with the corner usually at zero) and (2) the underlying variable in the population is continuously distributed but the data collection scheme involves censoring the response in some way. Many readers commented that they were happy I made this distinction, because empirical researchers often seemed to confuse a corner solution due to economic behavior and a corner that is arbitrarily created by a data censoring mechanism. Nevertheless, I still found that beginners did not always fully appreciate the difference, and poor practice in interpreting estimates lingered. Plus, combining the two types of applications of so-called "censored regression models" gave short shrift to true data censoring. In this edition, models for corner solutions in the population are treated in Chapter 17, and a variety of data censoring schemes are covered in more detail in Chapter 19.

As in the first edition, I use the approach of specifying a population model and imposing assumptions on that model. Until Chapter 19, random sampling is assumed to generate the data. Unlike traditional treatments of, say, the linear regression model, my approach forces the student to specify the population of interest, propose

a model and assumptions in the population, and then worry about data issues. The last part is easy under random sampling, and so students can focus on various models that are used for populations with different features. The students gain a clear understanding that, under random sampling, our ability to identify parameters (and other quantities of interest) is a product of our assumed model in the population. Later it becomes clear that sampling schemes that depart from random sampling can introduce complications for learning about the underlying population.

The second edition continues to omit some important topics not covered in the first edition. The leading ones are simulation methods of estimation and semiparametric/nonparametric estimation. The book by Cameron and Trivedi (2005) does an admirable job providing accessible introductions to these topics.

I have added several new problems to each of the chapters. As in the first edition, the problems are a blend of methodological questions—some of which lead to tweaking existing methods in useful directions—and empirical work. Several data sets have been added to further illustate how more advanced methods can be applied. The data sets can be accessed by visiting links at the MIT Press website for the book: http://mitpress.mit.edu/9780262232586.

New to the Second Edition

Earlier I mentioned that I have reorganized some of the material from the first edition. I have also added new material, and expanded on some of the existing topics. For example, Chapter 6 (in Part II) introduces control function methods in the context of models linear in parameters, including random coefficient models, and discusses when the method is the same as two-stage least squares and when it differs. Control function methods can be used for certain systems of equations (Chapter 9) and are used regularly for nonlinear models to deal with endogenous explanatory variables, or heterogeneity, or both (Part IV). The control function method is convenient for testing whether certain variables are endogenous, and more tests are included throughout the book. (Examples include Chapter 15 for binary response models and Chapter 18 for count data.) Chapter 6 also contains a more detailed discussion of difference-in-differences methods for independently pooled cross sections.

Chapter 7 now introduces all of the different concepts of exogeneity of the explanatory variables in the context of panel data models, without explicitly introducing unobserved heterogeneity. This chapter also contains a detailed discussion of the properties of generalized least squares when an incorrect variance-covariance structure is imposed. This general discussion is applied in Chapter 10 to models that nominally impose a random effects structure on the variance-covariance matrix.

In this edition, Chapter 8 explicitly introduces and analyzes the so-called "generalized instrumental variables" (GIV) estimator. This estimator, used implicitly in parts of the first edition, is important for discussing efficient estimation. Further, some of the instrumental variables estimators used for panel data models in Chapter 11 are GIV estimators. It is helpful for the reader to understand the general idea underlying GIV, and to see its application to classes of important models.

Chapter 10, while focusing on traditional estimation methods for unobserved effects panel data models, demonstrates more clearly the relationships among random effects, fixed effects, and "correlated random effects" (CRE) models. While the first edition used the CRE approach often—especially for nonlinear models—I never used the phrase "correlated random effects," which I got from Cameron and Trivedi (2005). Chapter 10 also provides a detailed treatment of the Hausman test for comparing the random and fixed effects estimators, including demonstrating that the traditional way of counting degrees of freedom when aggregate time effects are included is wrong and can be very misleading. The important topic of approximating the bias from fixed effects estimation and first differencing estimation, as a function of the number of available time periods, is also fleshed out.

Of the eight chapters in Part II, Chapter 11 has been changed the most. The random effects and fixed effects instrumental variables estimators are introduced and studied in some detail. These estimators form the basis for estimation of panel data models with heterogeneity and endogeneity, such as simultaneous equations models or models with measurement error, as well as models with additional orthogonality restrictions, such as Hausman and Taylor models. The method of first differencing followed by instrumental variables is also given separate treatment. This widely adopted approach can be used to estimate static models with endogeneity and dynamic models, such as those studied by Arellano and Bond (1991). The Arellano and Bond approach, along with several extensions, are now discussed in Section 11.6. Section 11.7 extends the treatment of models with individual-specific slopes, including an analysis of when traditional estimators are consistent for the population averaged effect, and new tests for individual-specific slopes.

As in the first edition, Part III of the book is the most technical, and covers general approaches to estimation. Chapter 12 contains several important additions. There is a new discussion concerning inference when the first-step estimation of a two-step procedure is ignored. Resampling schemes, such as the bootstrap, are discussed in more detail, including how one used the bootstrap in microeconometric applications with a large cross section and relatively few time periods. The most substantive additions are in Sections 12.9 and 12.10, which cover multivariate nonlinear least squares and quantile methods, respectively. An important feature of Section 12.9 is

that I make a simple link between mutivariate weighted nonlinear least squares—an estimation method familiar to economists—and the generalized estimating equations (GEE) approach. In effect, these approaches are the same, a point that hopefully allows economists to read other literature that uses the GEE nomenclature.

The section on quantile estimation covers different asymptotic variance estimators and discusses how they compare to violation of assumptions in terms of robustness. New material on estimating and inference when quantile regression is applied to panel data gives researchers simple methods for allowing unobserved effects in quantile estimation, while at the same time offering inference that is fully robust to arbitrary serial correlation.

Chapter 13, on maximum likelihood methods, also includes several additions, including a general discussion of nonlinear unobserved effects models and the different approaches to accounting for the heterogeneity (broadly, random effects, "fixed" effects, and correlated random effects) and different estimation methods (partial maximum likelihood or full maximum likelihood). Two-step maximum likelihood estimators are covered in more detail, including the case where estimating parameters in a first stage can be more efficient than simply plugging in known population values in the second stage. Section 13.11 includes new material on quasi-maximum likelihood estimation (QMLE). This section argues that, for general misspecification, only one form of asymptotic variance can be used. The QMLE perspective is attractive in that it admits that models are almost certainly wrong, thus we should conduct inference on the approximation in a valid way. Vuong's (1988) model selection tests, for nonnested models, is explicitly treated as a way to choose among competing models that are allowed to be misspecified. I show how to extend Vuong's approach to panel data applications (as usual, with a relatively small number of time periods).

Chapter 13 also includes a discussion of QMLE in the linear exponential family (LEF) of likelihoods, when the conditional mean is the object of interest. A general treatment allows me to appeal to the consistency results, and the methods for inference, at several points in Part IV. I emphasize the link between QMLE in the LEF and the so-called "generalized linear models" (GLM) framework. It turns out that GLM is just a special case of QMLE in the LEF, and this recognition should be helpful for studying research conducted from the GLM perspective. A related topic is the GEE approach to estimating panel data models. The starting point for GEE in panel data is to use (for a generic time period) a likelihood in the LEF, but to regain some efficiency that has been lost by not implementing full maximum likelihood by using a generalized least squares approach.

Chapter 14, on generalized method of moments (GMM) and minimum distance (MD) estimation, has been slightly reorganized so that the panel data applications

come at the end. These applications have also been expanded to include unobserved effects models with time-varying loads on the heterogeneity.

Perhaps for most readers the changes to Part IV will be most noticeable. The material on discrete response models has been split into two chapters (in contrast to the rather unwieldy single chapter in the first edition). Because Chapter 15 is the first applications-oriented chapter for nonlinear models, I spend more time discussing different ways of measuring the magnitudes of the effects on the response probability. The two leading choices, the partial effects evaluated at the averages and the average partial effect, are discussed in some detail. This discussion carries over for panel data models, too. A new subsection on unobserved effects panel data models with unobserved heterogeneity and a continuous endogenous explanatory variable shows how one can handle both problems in nonlinear models. This chapter contains many more empirical examples than the first edition.

Chapter 16 is new, and covers multinomial and ordered responses. These models are now treated in more detail than in the first edition. In particular, specification issues are fleshed out and the issues of endogeneity and unobserved heterogeneity (in panel data) are now covered in some detail.

Chapter 17, which was essentially Chapter 16 in the first edition, has been given a new title, Corner Solutions Responses, to reflect its focus. In reading Tobin's (1958) paper, I was struck by how he really was talking about the corner solution case—data censoring had nothing to do with his analysis. Thus, this chapter returns to the roots of the Tobit model, and covers several extensions. An important addition is a more extensive treatment of two-part models, which is now in Section 17.6. Hopefully, my unified approach in this section will help clarify the relationships among so-called "hurdle" and "selection" models, and show that the latter are not necessarily superior. Like Chapter 15, this chapter contains several more empirical applications.

Chapter 18 covers other kinds of limited dependent variables, particularly count (nonnegative integer) outcomes and fractional responses. Recent work on panel data methods for fractional responses has been incorporated into this chapter.

Chapter 19 is an amalgamation of material from several chapters in the first edition. The theme of Chapter 19 is data problems. The problem of data censoring—where a random sample of units is obtained from the population, but the response variable is censored in some way—is given a more in-depth treatment. The extreme case of binary censoring is included, along with interval censoring and top coding. Readers are shown how to allow for endogenous explanatory variables and unobserved heterogeneity in panel data.

Chapter 19 also includes the problem of not sampling at all from part of the population (truncated sampling) or not having any information about a response for a

subset of the population (incidental truncation). The material on unbalanced panel data sets and the problems of incidental truncation and attrition in panel data are studied in more detail, including the method of inverse probability weighting for correcting for missing data.

Chapter 20 continues the material on nonrandom sampling, providing a separate chapter for stratified sampling and cluster sampling. Stratification and clustering are often features of survey data sets, and it is important to know what adjustments are required to standard econometric methods. The material on cluster sampling summarizes recent work on clustering with a small number of clusters.

The material on treatment effect estimation is now in Chapter 21. While I preserved the setup from the first edition, I have added several more topics. First, I have expanded the discussion of matching estimators. Regression discontinuity designs are covered in a separate section.

The final chapter, Chapter 22, now includes the introductory material on duration analysis. I have included more empirical examples than were in the first edition.

Possible Course Outlines

At Michigan State, I teach a two-semester course to second-year, and some third-year, students that covers the material in my book—plus some additional material. I assume that the graduate students know, or will study on their own, material from Chapters 2 and 3. It helps move my courses along when students are comfortable with the basic algebra of probability (conditional expectations, conditional variances, and linear projections) as well as the basic limit theorems and manipulations. I typically spend a few lectures on Chapters 4, 5, and 6, primarily to provide a bridge between a more traditional treatment of the linear model and one that focuses on a linear population model under random sampling. Chapter 6 introduces control function methods in a simple context and so is worth spending some time on.

In the first semester (15 weeks), I cover the material (selectively) through Chapter 17. But I currently skip, in the first semester, the material in Chapter 12 on multivariate nonlinear regression and quantile estimation. Plus, I do not cover the asymptotic theory underlying M-estimation in much detail, and I pretty much skip Chapter 14 altogether. In effect, the first semester covers the popular linear and nonlinear models, for both cross section and panel data, in the context of random sampling, providing much of the background needed to justify the large-sample approximations.

In the second semester I return to Chapter 12 and cover quantile estimation. I also cover the general quasi-MLE and generalized estimating equations material in

Chapter 13. In Chapter 14, I find the minimum distance approach to estimation is important as a more advanced estimation method. I cover some of the panel data examples from this chapter. I then jump to Chapter 18, which covers count and fractional responses. I spend a fair amount of time on Chapters 19 and 20 because data problems are especially important in practice, and it is important to understand the strengths and weakness of the competing methods. After I cover the main parts of Chapter 21 (including regression discontinuity designs) and Chapter 22 (duration analysis), I sometimes have extra time. (However, if I were to cover some of the more advanced topics in Chapter 21—multivalued and multiple treatments, and dynamic treatement effects in the context of panel data—I likely would run out of time.) If I do have extra time, I like to provide an introduction to nonparametric and semi-parametric methods. Cameron and Trivedi (2005) is accessible for the basic methods, while the book by Li and Racine (2007) is comprehensive. Illustrating nonparametric methods using the treatment effects material in Chapter 21 seems particularly effective.

Supplements

A student *Solutions Manual* is available that includes answers to the odd-numbered problems (see http://mitpress.mit.edu/9780262731836). Any instructor who adopts the book for a course may have access to all solutions. In addition, I have created a set of slides for the two-semester course that I teach. They are available as Scientific Word 5.5 files—which can be edited—or as pdf files. For these teaching aids see the web page for the second edition: http://mitpress.mit.edu/9780262232586.

20 Stratified Sampling and Cluster Sampling

20.1 Introduction

In this chapter we study estimation when the data have been obtained by means of two common nonrandom sampling schemes. **Stratified sampling** occurs when units in a population are sampled with probabilities that do not reflect their frequency in the population. For example, in obtaining a data set on families, low-income families might be oversampled and high-income families undersampled. There are various mechanisms by which stratified samples are obtained, and we will cover the most common ones in this chapter.

The case of truncated sampling covered in Section 19.7 can be viewed as an extreme case of stratified sampling, where part of the population is not sampled at all. For the most part, in this chapter we focus on the case where the entire population is sampled (but where the sampling frequencies differ from the population frequencies). As we will see, in this setup simple weighting methods are available for recovering the underlying population parameters.

Cluster sampling refers to cases where clusters or groups, rather than individual units, are drawn from a population. For example, in evaluating the impact of educational policies on test performance of fourth-graders in Michigan, one might sample classrooms from the entire state (as opposed to randomly drawing fourth-graders from the population of all fourth-graders in Michigan). The classrooms constitute the clusters, and the students within the classrooms are the inividual units. The cluster sampling scheme generally implies that the outcomes of units within a cluster are correlated through unobserved "cluster effects." (In addition, some covariates, such as quality of the teacher, will be perfectly correlated because students in the same class have the same teacher. Other covariates, such as family income, are likely to have substantial correlation but would vary within a classroom.)

When a cluster sample is obtained by randomly drawing clusters from a large population of clusters, the resulting data set has features in common with the panel data sets we have studied throughout the book. Namely, we have many clusters that can be assumed to be independent of each other, but observations within a cluster are correlated. In, say, a firm-level panel data set, the firm plays the role of a cluster and time plays the role of the individual units. Because of their statistical similarity to "large N, small T" panel data sets, most of the statistical methods applied to cluster samples are familiar from our earlier analysis. We treat cluster samples separately because the nature of the within-cluster correlation is generally different from time series correlation, and cluster samples are naturally unbalanced even when there is no sample selection problem. (For example, in the population of fourth-grade classrooms in Michigan, we expect some variation in class size.)

20.2 Stratified Sampling

We begin with an analysis of **stratified samples** where, as mentioned in the introduction, different subsets of the population are sampled with different frequencies. Obtaining samples that are not representative of the underlying population is often done intentionally in obtaining survey data. Some surveys are designed primarily to learn about a particular subset of the population (perhaps based on income, education, age, or race). That group is typically overrepresented in the sample compared with its frequency in the population.

Stratification can be based on exogenous variables or endogenous variables (which are known once a model and assumptions have been specified) or some combination of these. As in the case of the sample selection problems we discussed in Chapter 19, it is important to know which is the case.

We cover the two most common types of stratified sampling in this section (and touch on a third). In Section 20.2.1 we study **standard stratified sampling**, which involves stratifying the population and then drawing random samples from the different strata. A different sampling scheme, **variable probability sampling**, is based on randomly drawing units from a population but then keeping the observations with unequal probabilities.

The section does not provide a detailed treatment of **choice-based sampling**, which occurs in discrete response models when the stratification is based entirely on the response variable. Various methods have been proposed for estimating discrete response models with choice-based samples under different assumptions. Manski and McFadden (1981) and Cosslett (1993) contain general treatments. For a class of discrete response models, Cosslett (1981) proposed an efficient estimation method with choice-based sampling, and Imbens (1992) obtained a computationally simple method-of-moments estimator that also achieves the efficiency bound. Imbens and Lancaster (1996) allow for general response variables in a maximum likelihood setting. In this section, we focus on a convenient weighted-estimation approach that applies to a variety of estimation methods. Not surprisingly, when applied in maximum likelihood contexts, weighted estimators are generally inefficient.

20.2.1 Standard Stratified Sampling and Variable Probability Sampling

The two most common stratification schemes used in obtaining data sets in the social sciences are **standard stratified sampling (SS sampling)** and **variable probability sampling (VP sampling)**. In SS sampling, the population is first partitioned into J groups, $\mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_J$, which we assume are nonoverlapping and exhaustive. We let $\mathbf{w}$ denote the random vector representing the population of interest.

STANDARD STRATIFIED SAMPLING: For $j = 1, \ldots, J$, draw a random sample of size N_j from stratum j. For each j, denote this random sample by $\{\mathbf{w}_{ij}: i = 1, 2, \ldots, N_j\}$.

The strata sample sizes N_j are nonrandom. Therefore, the total sample size, $N = N_1 + \cdots + N_J$, is also nonrandom. A randomly drawn observation from stratum j, $\mathbf{w}_{ij}$, has distribution $D(\mathbf{w} \mid \mathbf{w} \in \mathcal{W}_j)$. Therefore, while observations within a stratum are identically distributed, observations across strata are not. A scheme that is similar in nature to SS sampling is called **multinomial sampling**, where a stratum is first picked at random and then an observation is randomly drawn from the stratum. This *does* result in i.i.d. observations, but it does not correspond to how stratified samples are obtained in practice. It also leads to the same estimators as under SS sampling, so we do not discuss it further; see Cosslett (1993) or Wooldridge (1999b) for further discussion.

Variable probability samples are obtained using a different scheme. First, an observation is drawn at random from the population. If the observation falls into stratum j, it is kept with probability p_j. Thus, random draws from the population are discarded with varying frequencies depending on which stratum they fall into. This kind of sampling is appropriate when information on the variable or variables that determine the strata is relatively easy to obtain compared with the rest of the information. Survey data sets, including initial interviews to collect panel or longitudinal data, are good examples. Suppose we want to oversample individuals from, say, lower income classes. We can first ask an individual her or his income. If the response is in income class j, this person is kept in the sample with probability p_j, and then the remaining information, such as education, work history, family background, and so on can be collected; otherwise, the person is dropped without further interviewing.

A key feature of VP sampling is that observations within a stratum are discarded randomly. As discussed by Wooldridge (1999b), VP sampling is equivalent to the following:

VARIABLE PROBABILITY SAMPLING: Repeat the following steps N times:

1. Draw an observation $\mathbf{w}_i$ at random from the population.
2. If $\mathbf{w}_i$ is in stratum j, toss a (biased) coin with probability p_j of turning up heads. Let $h_{ij} = 1$ if the coin turns up heads and zero otherwise.
3. Keep observation i if $h_{ij} = 1$; otherwise, omit it from the sample.

The number of observations falling into stratum j is denoted N_j, and the number of data points we actually have for estimation is $N_0 = N_1 + N_2 + \cdots + N_J$. Notice that if N—the number of times the population is sampled—is fixed, then N_0 is a random variable: we do not know what each N_j will be prior to sampling.

The assumption that the probability of the coin turning up heads in step 2 depends only on the stratum ensures that sampling is random within each stratum. This roughly reflects how samples are obtained for certain large cross-sectional and panel data sets used in economics, including the panel study of income dynamics and the national longitudinal survey.

To see that a VP sample can be analyzed as a random sample, we construct a population that incorporates the stratification. The VP sampling scheme is equivalent to first tossing all J coins before actually observing which stratum $\mathbf{w}_i$ falls into; this gives $(h_{i1}, \ldots, h_{iJ})$. Next, $\mathbf{w}_i$ is observed to fall into one of the strata. Finally, the outcome is kept or not depending on the coin flip for that stratum. The result is that the vector $(\mathbf{w}_i, \mathbf{h}_i)$, where $\mathbf{h}_i$ is the J-vector of binary indicators h_{ij}, is a random sample from a new population with sample space $\mathcal{W} \times \mathcal{H}$, where $\mathcal{W}$ is the original sample space and $\mathcal{H}$ denotes the sample space associated with outcomes from flipping J coins. Under this alternative way of viewing the sampling scheme, $\mathbf{h}_i$ is independent of $\mathbf{w}_i$. Treating $(\mathbf{w}_i, \mathbf{h}_i)$ as a random draw from the new population is not at odds with the fact that our estimators are based on a nonrandom sample from the original population: we simply use the vector $\mathbf{h}_i$ to determine which observations are kept in the estimation procedure.

20.2.2 Weighted Estimators to Account for Stratification

With variable probability sampling, it is easy to construct weighted objective functions that produce consistent and asymptotically normal estimators of the population parameters. Initially, it is useful to define a set of binary variables that indicate whether a random draw, $\mathbf{w}_i$, is kept in the sample and, if so, which stratum it falls into. Let $z_{ij} = 1[\mathbf{w}_i \in \mathcal{W}_j]$, $j = 1, \ldots, J$ be the binary strata indicators, that is, $z_{ij} = 1$ if and only if $\mathbf{w}_i \in \mathcal{W}_j$. The vector of strata indicators is $\mathbf{z}_i = (z_{i1}, \ldots, z_{iJ})$. Then define

$$r_{ij} = h_{ij} z_{ij}, \qquad j = 1, \ldots, J. \tag{20.1}$$

By definition, $r_{ij} = 1$ for at most one j. If $h_{ij} = 1$ then $r_{ij} = z_{ij}$, so that r_{ij} is the same as the stratum indicator. If $r_{ij} = 0$ for all $j = 1, 2, \ldots, J$, then the random draw $\mathbf{w}_i$ does not appear in the sample (and we probably do not know which stratum the observation fell into).

With these definitions, we can define the **weighted M-estimator**, $\hat{\theta}_w$, as the solution to

$$\min_{\theta \in \Theta} \sum_{i=1}^{N} \sum_{j=1}^{J} p_j^{-1} r_{ij} q(\mathbf{w}_i, \theta), \tag{20.2}$$

where $q(\mathbf{w}, \theta)$ is the objective function that is chosen to identify the population parameters θ_o. Note how the outer summation is over all *potential* observations, that

is, the observations that *would* appear in a random sample. The indicators r_{ij} simply pick out the observations that actually appear in the available sample, and these indicators also attach each observed data point to its stratum. The objective function (20.2) weights each observed data point in the sample by the inverse of the sampling probability. For implementation it is useful to write the objective function as

$$\min_{\theta \in \Theta} \sum_{i=1}^{N_0} p_{j_i}^{-1} q(\mathbf{w}_i, \theta), \tag{20.3}$$

where, without loss of generality, the data points actually observed are ordered $i = 1, \ldots, N_0$. Since j_i is the stratum for observation i, $p_{j_i}^{-1}$ is the weight attached to observation i in the estimation. In practice, the $p_{j_i}^{-1}$ are the **sampling weights** reported with other variables in VP stratified samples.

The objective function $q(\mathbf{w}, \theta)$ contains all of the M-estimator examples we have covered so far in the book, including least squares (linear and nonlinear), conditional maximum likelihood, and partial maximum likelihood. In panel data applications, the probability weights are from sampling in an initial year. Weights for later years are intended to reflect both stratification (if any) and possible attrition, as discussed in Section 19.9.3.

In the case of estimating the mean from a population, the resulting weighted M-estimator has a familiar form. Let $\mu_o = \mathrm{E}(w)$ be the population mean. Then the weighted M-estimator solves

$$\min_{\mu \in \mathbb{R}} \sum_{i=1}^{N_0} p_{j_i}^{-1} (w_i - \mu)^2, \tag{20.4}$$

and the solution is easily seen to be the weighted average

$$\hat{\mu}_w = \sum_{i=1}^{N_0} v_{j_i} w_i, \tag{20.5}$$

where

$$v_{j_i} = \left(\sum_{h=1}^{N_0} p_{j_h}^{-1} \right)^{-1} p_{j_i}^{-1}. \tag{20.6}$$

In the general case, Wooldridge (1999b) shows that, under the same assumptions as Theorem 12.2 and the assumption that each sampling probability is positive, the weighted M-estimator consistently estimates θ_o, which is assumed to uniquely

minimize $\mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta})]$. Actually, as shown by Wooldridge (2007), consistency follow from our treatment of inverse-probability-weighted M-estimation in Section 19.8. A noted earlier, the vector $\mathbf{z}_i$ is the J-vector of strata indicators, $z_{ij} = 1[w_i \in W_j]$. Unde VP sampling, the sampling probability depends only on the stratum, so Assumptio 19.6 holds by design. In particular, define $s_i = h_{i1}z_{i1} + \cdots + h_{iJ}z_{iJ}$ to be the selectio indicator. Then

$$\mathrm{P}(s_i = 1 \mid \mathbf{z}_i, \mathbf{w}_i) = \mathrm{P}(s_i = 1 \mid \mathbf{z}_i) = p_1 z_{i1} + p_2 z_{i2} + \cdots + p_J z_{iJ}. \tag{20.7}$$

Therefore, we can use the consistency result for IPW M-estimation directly to estab lish the consistency of the IPW estimator for VP sampling.

Asymptotic normality also follows under the same regularity conditions as in Chapter 12. Wooldridge (1999b) shows that a valid estimator of the asymptotic variance of $\hat{\boldsymbol{\theta}}_w$ is

$$\left[\sum_{i=1}^{N_0} p_{j_i}^{-1} \nabla_{\boldsymbol{\theta}}^2 q_i(\hat{\boldsymbol{\theta}}_w)\right]^{-1} \left[\sum_{i=1}^{N_0} p_{j_i}^{-2} \nabla_{\boldsymbol{\theta}} q_i(\hat{\boldsymbol{\theta}}_w)' \nabla_{\boldsymbol{\theta}} q_i(\hat{\boldsymbol{\theta}}_w)\right] \left[\sum_{i=1}^{N_0} p_{j_i}^{-1} \nabla_{\boldsymbol{\theta}}^2 q_i(\hat{\boldsymbol{\theta}}_w)\right]^{-1}, \tag{20.8}$$

which looks like the standard formula for a robust variance matrix estimator except for the presence of the sampling probabilities p_{j_i}.

When $\mathbf{w}$ partitions as $(\mathbf{x}, \mathbf{y})$, an alternative estimator replaces the Hessian $\nabla_{\boldsymbol{\theta}}^2 q_i(\hat{\boldsymbol{\theta}}_w)$ in expression (20.8) with $\mathbf{A}(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_w)$, where $\mathbf{A}(\mathbf{x}_i, \boldsymbol{\theta}_o) \equiv \mathrm{E}[\nabla_{\boldsymbol{\theta}}^2 q(\mathbf{w}_i, \boldsymbol{\theta}_o) \mid \mathbf{x}_i]$, as in Chapter 12. Asymptotic standard errors and Wald statistics can be obtained using either estimate of the asymptotic variance.

We can also apply the "surprising" efficiency result concerning estimation of the sampling probabilities to VP stratification—at least if additional information is kept during the sampling. For a random draw i the log likelihood for the density of s_i given $\mathbf{z}_i$ can be written as

$$l_i(\mathbf{p}) = \sum_{j=1}^{J} z_{ij}[s_i \log(p_j) + (1 - s_i) \log(1 - p_j)]. \tag{20.9}$$

For each $j = 1, \ldots, J$, the maximum likelihood estimate, $\hat{p}_j$, is easily seen to be the fraction of observations retained out of all of those originally drawn from from stratum j: $\hat{p}_j = M_j/N_j$, where $M_j = \sum_{i=1}^{N} z_{ij}s_i$ and $N_j = \sum_{i=1}^{N} z_{ij}$. In other words, M_j is the number of retained data points from stratum j, and N_j is the number of times stratum j was drawn in the VP sampling scheme. If the N_j, $j = 1, \ldots, J$, are reported along with the VP sample, then we can easily obtain the $\hat{p}_j$ (because the M_j are always known). We do not need to observe the specific strata indicators for obser-

vations for which $s_i = 0$. (If the stratification is exogenous, as defined in Section 20.2.3, then it does not matter whether we use the estimated or known sampling probabilities: the asymptotic variance is unchanged in that case.)

Example 20.1 (Linear Model under Stratified Sampling): In estimating the linear model

$$y = \mathbf{x}\boldsymbol{\beta}_o + u, \qquad \mathrm{E}(\mathbf{x}'u) = \mathbf{0} \tag{20.10}$$

by IPW least squares, the asymptotic variance matrix estimator is

$$\left(\sum_{i=1}^{N_o} p_{ji}^{-1}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\left(\sum_{i=1}^{N_o} p_{ji}^{-2}\hat{u}_i^2\mathbf{x}_i'\mathbf{x}_i\right)\left(\sum_{i=1}^{N_o} p_{ji}^{-1}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}, \tag{20.11}$$

where $\hat{u}_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}_w$ is the residual after WLS estimation. Interestingly, this is simply the White (1980b) heteroskedasticity-consistent covariance matrix estimator applied to the stratified sample, where all variables for observation i are weighted by $p_{ji}^{-1/2}$ before performing the regression. This estimator has been suggested by, among others, Hausman and Wise (1981). Hausman and Wise use maximum likelihood to obtain more efficient estimators in the context of the normal linear regression model, that is, $u \mid \mathbf{x} \sim \text{Normal}(\mathbf{x}\boldsymbol{\beta}_o, \sigma_o^2)$. Because of stratification, MLE is not generally robust to failure of the homoskedastic normality assumption.

It is important to remember that the form of expression (20.11) in this example is not due to potential heteroskedasticity in the underlying population model. Even if $\mathrm{E}(u^2 \mid \mathbf{x}) = \sigma_o^2$, the estimator (20.11) is generally needed because of the stratified sampling. This estimator also works in the presence of heteroskedasticity of arbitrary and unknown form in the population, and it is routinely computed by many regression packages.

Example 20.2 (Conditional MLE under Stratified Sampling): When $f(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta})$ is a correctly specified model for the density of $\mathbf{y}_i$ given $\mathbf{x}_i$ in the population, the inverse-probability-weighted MLE is obtained with $q_i(\boldsymbol{\theta}) \equiv -\log[f(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta})]$. This estimator is consistent and asymptotically normal, with asymptotic variance estimator given by expression (20.8) [or the form that uses $\mathbf{A}(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_w)$].

A weighting scheme is also available in the standard stratified sampling case, but the weights are different from the VP sampling case. To derive them, let $Q_j = \mathrm{P}(\mathbf{w} \in \mathcal{W}_j)$ denote the population frequency for stratum j; we assume that the Q_j are *known*. By the law of iterated expectations,

$$\mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta})] = Q_1\mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta}) \mid \mathbf{w} \in \mathcal{W}_1] + \cdots + Q_J\mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta}) \mid \mathbf{w} \in \mathcal{W}_J] \tag{20.12}$$

for any θ. For each j, $\mathrm{E}[q(\mathbf{w},\theta) \mid \mathbf{w} \in \mathcal{W}_j]$ can be consistently estimated using a random sample obtained from stratum j. This scheme leads to the sample objective function

$$Q_1\left[N_1^{-1}\sum_{i=1}^{N_1} q(\mathbf{w}_{i1},\theta)\right] + \cdots + Q_J\left[N_J^{-1}\sum_{i=1}^{N_J} q(\mathbf{w}_{iJ},\theta)\right],$$

where $\mathbf{w}_{ij}$ denotes a random draw i from stratum j and N_j is the nonrandom sample size for stratum j. We can apply the uniform law of large numbers to each term, so that the sum converges uniformly to equation (20.12) under the regularity conditions in Chapter 12. By multiplying and dividing each term by the total number of observations $N = N_1 + \cdots + N_J$, we can write the sample objective function more simply as

$$N^{-1}\sum_{i=1}^{N}(Q_{j_i}/H_{j_i})q(\mathbf{w}_i,\theta), \tag{20.13}$$

where j_i denotes the stratum for observation i and $H_j \equiv N_j/N$ denotes the fraction of observations in stratum j. Because we have the stratum indicator j_i, we can drop the j subscript on $\mathbf{w}_i$. When we omit the division by N, equation (20.13) has the same form as equation (20.3), but the weights are (Q_{j_i}/H_{j_i}) rather than $p_{j_i}^{-1}$ (and the arguments for why each weighting works are very different). Also, in general, the formula for the asymptotic variance is different in the SS sampling case. In addition to the minor notational change of replacing N_0 with N, the middle matrix in equation (20.8) becomes

$$\sum_{j=1}^{J}(Q_j^2/H_j^2)\left[\sum_{i=1}^{N_j}(\nabla_\theta \hat{q}_{ij} - \overline{\nabla_\theta q_j})'(\nabla_\theta \hat{q}_{ij} - \overline{\nabla_\theta q_j})\right], \tag{20.14}$$

where $\nabla_\theta \hat{q}_{ij} \equiv \nabla_\theta q(\mathbf{w}_{ij},\hat{\theta}_w)$ and $\overline{\nabla_\theta q_j} \equiv N_j^{-1}\sum_{i=1}^{N_j}\nabla_\theta \hat{q}_{ij}$ (the within-stratum sample average). This approach requires us to explicitly partition observations into their respective strata. See Wooldridge (2001) for a detailed derivation. [If in the VP sampling case the population frequencies Q_j are known, it is better to use as weights $Q_j/(N_j/N_0)$ rather than p_j^{-1}, which makes the analysis look just like the SS sampling case. See Wooldridge (1999b) for details.]

If in Example 20.2 we have standard stratified sampling rather than VP sampling, the weighted MLE is typically called the **weighted exogenous sample MLE (WESMLE)**; this estimator was suggested by Manski and Lerman (1977) in the context of choice-based sampling in discrete response models. (Actually, Manski and Lerman (1977) use multinomial sampling where H_j is the probability of picking stratum j. But Cosslett (1981) showed that a more efficient estimator is obtained by

using N_j/N, as one always does in the case of SS sampling; see Wooldridge (1999b) for an extension of Cosslett's result to the M-estimator case.)

Provided that the sampling weights Q_{j_i}/H_{j_i} or $p_{j_i}^{-1}$ are given (along with the stratum), analysis with the weighted M-estimator under SS or VP sampling is fairly straightforward, but it is not likely to be efficient. In the conditional maximum likelihood case it is certainly possible to do better. See Imbens and Lancaster (1996) for a careful treatment.

20.2.3 Stratification Based on Exogenous Variables

When $\mathbf{w}$ partitions as $(\mathbf{x}, \mathbf{y})$, where $\mathbf{x}$ is exogenous in a sense to be made precise, and stratification is based entirely on $\mathbf{x}$, the standard unweighted estimator on the stratified sample is consistent and asymptotically normal. The sense in which $\mathbf{x}$ must be exogenous is that $\boldsymbol{\theta}_o$ solves

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta}) \mid \mathbf{x}] \tag{20.15}$$

for each possible outcome $\mathbf{x}$. This assumption holds in a variety of contexts with conditioning variables and correctly specified models. For example, as we discussed in Chapter 12, the condition holds for nonlinear regression when the conditional mean is correctly specified and $\boldsymbol{\theta}_o$ is the vector of conditional mean parameters; in Chapter 13 we showed that this holds for conditional maximum likelihood when the density of $\mathbf{y}$ given $\mathbf{x}$ is correct. It also holds in other cases, including the quasi–maximum likelihood estimators we discussed in Chapter 18 when the conditional mean is correctly specified. One interesting point—which we will rely on in our treatment of estimating average treatment effects in Chapter 21—is that, in the linear case, it will not be enough for u to be uncorrelated with $\mathbf{x}$. If we want to estimate the linear projection of y on $\mathbf{x}$, we generally need to use the weighted estimator, even if stratification is a function of $\mathbf{x}$.

In the case of VP sampling, a common form of exogenous stratification occurs when the strata are defined in terms of $\mathbf{x}$, and that is the case we treat here. (See Section 19.8 or Wooldridge (2007) for more general situations.) Then, again letting $s_i = h_{i1}z_{i1} + \cdots + h_{iJ}z_{iJ}$ be the selection indicator, where each z_{ij} is a function of $\mathbf{x}_i$, $\mathrm{P}(s_i = 1 \mid \mathbf{w}_i, \mathbf{x}_i) = \mathrm{P}(s_i = 1 \mid \mathbf{x}_i)$. We can immediately apply the results of Section 19.8 to conclude that the unweighted M-estimator is consistent.

A direct proof is also informative. The unweighted M-estimator, using the stratified sample, $\hat{\boldsymbol{\theta}}_u$, minimizes

$$\sum_{i=1}^{N} s_i q(\mathbf{w}_i, \boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{j=1}^{N} h_{ij} z_{ij} q(\mathbf{w}_i, \boldsymbol{\theta}), \tag{20.16}$$

and consistency generally follows if we can show that the population value, θ_o, uniquely minimizes

$$\sum_{j=1}^{N} \mathrm{E}[h_{ij} z_{ij} q(\mathbf{w}_i, \boldsymbol{\theta})] = \sum_{j=1}^{N} p_j \mathrm{E}[z_{ij} q(\mathbf{w}_i, \boldsymbol{\theta})], \tag{20.17}$$

where the equality follows because h_{ij} is independent of $(z_{ij}, \mathbf{w}_i)$ by the nature of VP sampling, and $p_j = \mathrm{E}(h_{ij})$. Now, because z_{ij} is a function of $\mathbf{x}_i$, it follows by iterated expectations that

$$\mathrm{E}[z_{ij} q(\mathbf{w}_i, \boldsymbol{\theta})] = \mathrm{E}\{\mathrm{E}[z_{ij} q(\mathbf{w}_i, \boldsymbol{\theta}) \mid \mathbf{x}_i]\} = \mathrm{E}\{z_{ij} \mathrm{E}[q(\mathbf{w}_i, \boldsymbol{\theta}) \mid \mathbf{x}_i]\}. \tag{20.18}$$

By assumption, $\boldsymbol{\theta}_o$ minimizes $\mathrm{E}[q(\mathbf{w}_i, \boldsymbol{\theta}) \mid \mathbf{x}_i]$, and, because z_{ij} is a zero-one variable, $\boldsymbol{\theta}_o$ is also a minimizer of $\mathrm{E}[z_{ij} q(\mathbf{w}_i, \boldsymbol{\theta})]$. Now, with $p_j \geq 0$ for all j, $\boldsymbol{\theta}_o$ is also a solution to

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{j=1}^{N} \mathrm{E}[h_{ij} z_{ij} q(\mathbf{w}_i, \boldsymbol{\theta})], \tag{20.19}$$

which is what we wanted to show.

Unlike in the case of the weighted estimator, uniqueness of $\boldsymbol{\theta}_o$ in the population is no longer sufficient for identification using the unweighted estimator. In particular, if $p_j = 0$ for some j, part of the population is not sampled at all, and this may (but need not) result in lack of identification of $\boldsymbol{\theta}_o$. As discussed in Wooldridge (2001) for the case of SS sampling, $p_j > 0$ for all j ensures identification of $\boldsymbol{\theta}_o$ when it is identified in the population.

Generally, when stratification is based on $\mathbf{x}$, one can make a case for weighting if interest lies in the solution to the population problem

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta})], \tag{20.20}$$

which we have called $\boldsymbol{\theta}_o$. The IPW estimator consistently estimates $\boldsymbol{\theta}_o$ without further assumptions, while the unweighted estimator requires the stronger assumption described surrounding equation (20.15). A special case is the linear regression model discussed in Example 20.1: to consistently estimate the linear projection, we must use weights even if selection is based on $\mathbf{x}$. Consistency of the unweighted estimator requires that we are estimating the conditional mean. In the next chapter, we will see other uses of this fact about the weighted versus unweighted estimator.

Wooldridge (1999b) shows that the usual asymptotic variance estimators (see Section 12.5) are valid when stratification is based on $\mathbf{x}$ and we ignore the stratification problem. For example, the usual conditional maximum likelihood analysis holds. In

the case of regression, we can use the usual heteroskedasticity-robust variance matrix estimator. Or, if we assume homoskedasticity in the population, the nonrobust form (see equation (12.58)) is valid with the usual estimator of the error variance.

When a generalized conditional information matrix equality holds, and stratification is based on $\mathbf{x}$, Wooldridge (1999b) shows that the unweighted estimator is more efficient than the weighted estimator. The key assumption is

$$\mathrm{E}[\nabla_\theta q(\mathbf{w}, \theta_o)' \nabla_\theta q(\mathbf{w}, \theta_o) \mid \mathbf{x}] = \sigma_o^2 \mathrm{E}[\nabla_\theta^2 q(\mathbf{w}, \theta_o) \mid \mathbf{x}] \tag{20.21}$$

for some $\sigma_o^2 > 0$. When assumption (20.21) holds and θ_o solves equation (20.19), the asymptotic variance of the unweighted M-estimator is smaller than that for the weighted M-estimator. This generalization includes conditional maximum likelihood (with $\sigma_o^2 = 1$) and nonlinear regression under homoskedasticity.

Very similar conclusions hold for standard stratified sampling. One useful fact is that, when stratification is based on $\mathbf{x}$, the estimator (20.8) is valid with $p_j = H_j/Q_j$ (and $N_0 = N$); therefore, we need not compute within-strata variation in the estimated score. The unweighted estimator is consistent when stratification is based on $\mathbf{x}$ and the usual asymptotic variance matrix estimators are valid. The unweighted estimator is also more efficient when assumption (20.21) holds. See Wooldridge (2001) for statements of assumptions and proofs of theorems.

As a practical matter, modern statistical packages that have built-in features for analyzing data from stratified samples typically ask for two pieces of information: the sampling weights and the stratum identifier. If one specifies the weights but not the stratum identifier, the middle of the "sandwich" will not be estimated as in equation (20.14). The within-stratum averages will not be subtracted off, resulting in larger estimated asymptotic variances than necessary (except in the case of exogenous sampling under exogeneity of $\mathbf{x}$). In other words, the resulting confidence intervals and inference will be (asymptotically) conservative. It is better to use the information on the strata along with the sampling weights.

20.3 Cluster Sampling

We now turn to the problem of cluster sampling, where individual units are sampled in groups or clusters. As mentioned in the introduction, the problems of cluster sampling and panel data analysis are similar in their statistical structures: each confronts the problem of correlation when observations come with a natural nesting. The similarities are strongest in the case where a large number of clusters, each relatively small, is drawn from a large population of clusters. This case is relatively easy to handle, and we treat it first in Section 20.3.1.

Many data sets have both a panel data and cluster sampling structure. Inference that is robust to serial correlation and cluster correlation is straightforward provided the number of clusters is large. We discuss this case in Section 20.3.2.

Section 20.3.3 summarizes what is known about applying the usual cluster formulas when the cluster sizes are large rather than "small."

Recently, researchers have studied data structures that can be classified as a small number of clusters and many observations per cluster. Section 20.3.4 provides two methods for analyzing such structures.

20.3.1 Inference with a Large Number of Clusters and Small Cluster Sizes

We begin with the problem of estimating linear and nonlinear models when we can sample a large number of clusters from a large population of clusters. For each group or cluster g, let $\{(y_{gm}, \mathbf{x}_g, \mathbf{z}_{gm}): m = 1, \ldots, M_g\}$ be the observable data, where M_g is the number of units in cluster g, y_{gm} is a scalar response, $\mathbf{x}_g$ is a $1 \times K$ vector containing explanatory variables that vary only at the cluster level, and $\mathbf{z}_{gm}$ is a $1 \times L$ vector of covariates that vary within (as well as across) groups. In most applications of cluster samples, at least some covariates change only at the group level; earlier we gave the example of teacher characteristics when each cluster is a classroom. In fact, it is probably a sensible rule to at least consider the data as being generated as a cluster sample whenever covariates at a level more aggregated than the individual units are included in an analysis. For example, in analyzing firm-level data, if industry-level covariates are included then we should treat the data as a cluster sample, with each industry acting as a cluster.

Throughout we assume that the sampling scheme generates observations that are independent across g. In other words, we independently draw G clusters from the population of all clusters. This assumption can be restrictive, particularly when the clusters are large geographical units. Nevertheless, in some cases we can define the clusters to allow additional "spatial correlation." For example, if originally we think of sampling fourth-grade classrooms, but then we are worried about correlation in student performance not just within class but also within school, then we can define the clusters to be the schools. What we will not cover is schemes where any two geographical units are allowed to be correlated, with correlation diminishing as the observations are farther apart in space.

The theory with $G \to \infty$ and the group sizes, M_g, fixed is well developed; see, for example, White (1984) and Arellano (1987). Of course, it is up to the researcher to decide whether the sizes of G and the M_g are suitable for this asymptotic framework. Here, we follow Wooldridge (2003a) and summarize these results and emphasize how one might have to use robust inference methods even when it is not so obvious.

Not surprisingly, linear models are easiest to analyze. The standard linear model with an additive error is

$$y_{gm} = \alpha + \mathbf{x}_g\boldsymbol{\beta} + \mathbf{z}_{gm}\boldsymbol{\gamma} + v_{gm}, \qquad m = 1, \ldots, M_g; \; g = 1, \ldots, G. \tag{20.22}$$

As with panel data, our approach to estimation and inference in equation (20.22) depends on several factors, including whether we are interested in the effects of aggregate variables ($\boldsymbol{\beta}$) or individual-specific variables ($\boldsymbol{\gamma}$). In addition, we need to make assumptions about the error terms. An important issue is whether the v_{gm} contain a common group effect that can be separated in an additive fashion, as in

$$v_{gm} = c_g + u_{gm}, \qquad m = 1, \ldots, M_g, \tag{20.23}$$

where c_g is an unobserved cluster effect and u_{gm} is the idiosyncratic error. Another important issue is whether the explanatory variables in equation (20.22) can be taken to be appropriately exogenous. If the covariates satisfy

$$\mathrm{E}(v_{gm} \mid \mathbf{x}_g, \mathbf{z}_{gm}) = 0, \qquad m = 1, \ldots, M_g; \; g = 1, \ldots, G, \tag{20.24}$$

or even a zero-correlation version, the pooled OLS estimator, where we regress y_{gm} on $1, \mathbf{x}_g, \mathbf{z}_{gm}$, $m = 1, \ldots, M_g$; $g = 1, \ldots, G$, is consistent for $\boldsymbol{\theta} \equiv (\alpha, \boldsymbol{\beta}', \boldsymbol{\gamma}')'$ as $G \to \infty$ with M_g fixed. Further, the POLS estimator is $\sqrt{G}$-asymptotically normal.

Without more assumptions, a robust variance matrix is needed to account for correlation within clusters or heteroskedasticity in $\mathrm{Var}(v_{gm} \mid \mathbf{x}_g, \mathbf{z}_{gm})$, or both. When v_{gm} has the form in equation (20.23), the amount of within-cluster correlation can be substantial, with the result that the usual OLS standard errors can be very misleading (and, in most cases, systematically too small). Write $\mathbf{W}_g$ as the $M_g \times (1 + K + L)$ matrix of all regressors for group g. Then the $(1 + K + L) \times (1 + K + L)$ variance matrix estimator is

$$\widehat{\mathrm{Avar}}(\hat{\boldsymbol{\theta}}_{POLS}) = \left(\sum_{g=1}^{G} \mathbf{W}_g'\mathbf{W}_g\right)^{-1} \left(\sum_{g=1}^{G} \mathbf{W}_g'\hat{\mathbf{v}}_g\hat{\mathbf{v}}_g'\mathbf{W}_g\right) \left(\sum_{g=1}^{G} \mathbf{W}_g'\mathbf{W}_g\right)^{-1} \tag{20.25}$$

where $\hat{\mathbf{v}}_g$ is the $M_g \times 1$ vector of pooled OLS residuals for group g. As we discussed for the panel data case, this "sandwich" variance matrix estimator is now computed routinely using "cluster" options in popular statistical packages. One simply needs to specify the cluster identifier.

Pooled OLS estimation of the parameters in equation (20.22) ignores the within-cluster correlation of the v_{gm} in estimation, so that it can be very inefficient if c_g is a part of the error v_{gm}. As we know from panel data analysis, if we strengthen the exogeneity assumption to

$$E(v_{ijm} \mid \mathbf{x}_{ij}, \mathbf{Z}_{ij}) = 0, \qquad m = 1, \ldots, M_{ij}; \; j = 1, \ldots, G, \tag{20.26}$$

where $\mathbf{Z}_{ij}$ is the $M_{ij} \times L$ matrix of unit-specific covariates, then we can exploit the presence of c_{ij} in equation (20.23) in a generalized least squares (GLS) analysis. Assumption (20.26) rules out covariates from one member of the cluster affecting the outcomes on another, holding own covariates fixed. At least nominally, this assumption appears to rule out "peer effects," but such effects can be allowed by including measures of peers in $\mathbf{z}_{ijm}$.

The standard random effects approach makes enough assumptions so that the $M_{ij} \times M_{ij}$ variance-covariance matrix of $\mathbf{v}_{ij} = (v_{ij1}, v_{ij2}, \ldots, v_{ij,M_{ij}})'$ has the "random effects" form,

$$\text{Var}(\mathbf{v}_{ij}) = \sigma_c^2 \mathbf{j}_{M_{ij}} \mathbf{j}_{M_{ij}}' + \sigma_u^2 \mathbf{I}_{M_{ij}}, \tag{20.27}$$

where $\mathbf{j}_{M_{ij}}$ is the $M_{ij} \times 1$ vector of ones and $\mathbf{I}_{M_{ij}}$ is the $M_{ij} \times M_{ij}$ identity matrix. In the standard setup, we also make the system homoskedasticity assumption, familiar from the panel data analysis in Chapter 10:

$$\text{Var}(\mathbf{v}_i \mid \mathbf{x}_i, \mathbf{Z}_i) = \text{Var}(\mathbf{v}_i). \tag{20.28}$$

As in the panel data case, it is important to understand the role of assumption (20.28): it implies that the conditional variance-covariance matrix is the same as the unconditional variance-covariance matrix, but it does not restrict $\text{Var}(\mathbf{v}_i)$; it can be any $M_i \times M_i$ matrix under assumption (20.28). The particular random effects structure on $\text{Var}(\mathbf{v}_i)$ is given by assumption (20.27). Under assumptions (20.27) and (20.28), the resulting GLS estimator is the well-known random effects (RE) estimator. The estimator has the same structure as in the unbalanced panel data case; see Section 19.9.1.

The random effects estimator $\hat{\boldsymbol{\theta}}_{RE}$ is asymptotically more efficient than pooled OLS under assumptions (20.26), (20.27), and (20.28) as $G \to \infty$ with the M_i fixed. The RE estimates and test statistics are computed routinely by popular software packages for cluster samples. Nevertheless, an important point is often overlooked in applications of RE: one can, and in many cases should, make inference completely robust to an unknown form of $\text{Var}(\mathbf{v}_i \mid \mathbf{x}_i, \mathbf{Z}_i)$.

The idea in obtaining a fully robust variance matrix for RE is straightforward, as we saw in Chapter 10 for panel data. Even if $\text{Var}(\mathbf{v}_i \mid \mathbf{x}_i, \mathbf{Z}_i)$ does not have the RE form, the RE estimator is still consistent and $\sqrt{G}$-asymptotically normal under assumption (20.26), and it is likely to be more efficient than pooled OLS even if $\text{Var}(\mathbf{v}_i \mid \mathbf{x}_i, \mathbf{Z}_i)$ does not have the RE form. The case for a fully robust variance matrix for RE is somewhat more subtle than in the panel data case, where serial correlation in the idiosyncratic errors generally invalidates assumption (20.27). Of

course, heteroskedasticity in $\text{Var}(c_g \mid \mathbf{x}_g, \mathbf{Z}_g)$ or $\text{Var}(u_{gm} \mid \mathbf{x}_g, \mathbf{Z}_g)$ is always a possibility, and either justifies robust inference. As an example, suppose that the coefficients on $\mathbf{z}_{gm}$ vary at the cluster level:

$$y_{gm} = \alpha + \mathbf{x}_g\boldsymbol{\beta} + \mathbf{z}_{gm}\boldsymbol{\gamma}_g + v_{gm}, \qquad m = 1, \ldots, M_g;\ g = 1, \ldots, G. \tag{20.29}$$

By estimating a standard random effects model that assumes common slopes $\boldsymbol{\gamma}$, we effectively include $\mathbf{z}_{gm}(\boldsymbol{\gamma}_g - \boldsymbol{\gamma})$ in the idiosyncratic error; doing so generally creates within-group correlation because $\mathbf{z}_{gm}(\boldsymbol{\gamma}_g - \boldsymbol{\gamma})$ and $\mathbf{z}_{gp}(\boldsymbol{\gamma}_g - \boldsymbol{\gamma})$ will be correlated for $m \neq p$, conditional on $\mathbf{Z}_g$. Also, the idiosyncratic error will have heteroskedasticity that is a function of $\mathbf{z}_{gm}$. Nevertheless, if we assume $\text{E}(\boldsymbol{\gamma}_g \mid \mathbf{X}_g, \mathbf{Z}_g) = \text{E}(\boldsymbol{\gamma}_g) \equiv \boldsymbol{\gamma}$ along with assumption (20.26), the random effects estimator still consistently estimates the average slopes, $\boldsymbol{\gamma}$ (and $\boldsymbol{\beta}$). Therefore, in applying random effects to panel data or cluster samples, it is sensible (with large G) to make the variance estimator of random effects robust to arbitrary heteroskedasticity and within-group correlation.

In applications, one often computes the POLS and RE estimates to see how sensitive the estimates are to choice of variance matrix. Further, one is tempted to compare estimated variance matrices—or, at least, standard errors—to see if RE is more efficient than POLS. It is fine to do so provided one uses fully robust standard errors for POLS and RE. For example, it certainly makes no sense to compare the usual POLS standard errors (which ignore the cluster sampling) with the usual RE standard errors (which account for the clustering, at least to some extent). By comparing the fully robust forms for each set of estimates, one is comparing generally reliable estimates of the sampling variation of the POLS and RE estimates.

If we are only interested in estimating $\boldsymbol{\gamma}$, the fixed effects (FE) or within estimator is attractive. The within transformation subtracts within-group averages from the dependent variable and explanatory variables:

$$y_{gm} - \bar{y}_g = (\mathbf{z}_{gm} - \bar{\mathbf{z}}_g)\boldsymbol{\gamma} + u_{gm} - \bar{u}_g, \qquad m = 1, \ldots, M_g;\ g = 1, \ldots, G, \tag{20.30}$$

and this equation is estimated by pooled OLS. (Of course, the $\mathbf{x}_g$ get swept away by the within-group demeaning.) Under a full set of FE assumptions—which, as in the panel data case, allows arbitrary correlation between c_g and the $\mathbf{z}_{gm}$—inference is straightforward using standard software. Nevertheless, analogous to the random effects case, it is prudent to allow $\text{Var}(u_g \mid \mathbf{Z}_g)$ to have an arbitrary form, including within-group correlation and heteroskedasticity. For example, if we start with model (20.29), then $(\mathbf{z}_{gm} - \bar{\mathbf{z}}_g)(\boldsymbol{\gamma}_g - \boldsymbol{\gamma})$ appears in the error term. As we discussed in Section 11.7.3, the FE estimator is still consistent if $\text{E}(\boldsymbol{\gamma}_g \mid \mathbf{z}_{gm} - \bar{\mathbf{z}}_g) = \text{E}(\boldsymbol{\gamma}_g) = \boldsymbol{\gamma}$, an assumption that allows $\boldsymbol{\gamma}_g$ to be correlated with $\bar{\mathbf{z}}_g$. Nevertheless, u_{gm} and u_{gp} will be correlated for $m \neq p$. A fully robust variance matrix estimator is

$$\widehat{\text{Avar}}(\hat{\gamma}_{FE}) = \left(\sum_{g=1}^{G} \ddot{\mathbf{Z}}_g'\ddot{\mathbf{Z}}_g\right)^{-1}\left(\sum_{g=1}^{G} \ddot{\mathbf{Z}}_g'\hat{\mathbf{u}}_g\hat{\mathbf{u}}_g'\ddot{\mathbf{Z}}_g\right)\left(\sum_{g=1}^{G} \ddot{\mathbf{Z}}_g'\ddot{\mathbf{Z}}_g\right)^{-1}, \tag{20.31}$$

where $\ddot{\mathbf{Z}}_g$ is the matrix of within-group deviations from means and $\hat{\mathbf{u}}_g$ is the $M_g \times 1$ vector of fixed effects residuals. This estimator is justified with large-G asymptotics. It has exactly the same form as the unbalanced panel data case.

One benefit of a fixed effects approach in the standard model with constant slopes but c_g in the composite error term is that no adjustments are necessary if the c_g are correlated across groups. When the groups represent different geographical units, we might expect correlation across groups close to each other. If we think such correlation is largely captured through the unobserved effect c_g, then its elimination by means of the within transformation effectively solves the problem. If we use pooled OLS or random effects, we would have to deal with spatial correlation across g, in addition to within-group correlation, a difficult statistical problem.

An alternative to FE estimation, and one that leads to a simple Hausman test for comparing FE and RE, is to add the group averages to an RE estimation. Let $\bar{\mathbf{z}}_g$ denote the vector of within-group averages, and write

$$y_{gm} = \alpha + \mathbf{x}_g\boldsymbol{\beta} + \mathbf{z}_{gm}\boldsymbol{\gamma} + \bar{\mathbf{z}}_g\boldsymbol{\xi} + a_g + u_{gm}, \qquad m = 1, \ldots, M_g;\ g = 1, \ldots, G, \tag{20.32}$$

where $c_g = \bar{\mathbf{z}}_g\boldsymbol{\xi} + a_g$ (and we absorb the intercept here into α). Estimating this equation by, say, RE allows us to easily test H_0: $\boldsymbol{\xi} = \mathbf{0}$ in a fully robust way, which tests the null that the RE estimator is consistent. It can be shown that, even though the panel is not balanced, the estimate of $\boldsymbol{\gamma}$ is the FE estimate. In addition, this approach allows us to estimate coefficients on $\mathbf{x}_g$. (Pooled OLS can also be used, and also delivers the FE estimate of $\boldsymbol{\gamma}$.)

Example 20.3 (Cluster Correlation in Teacher Compensation): The data set in BENEFITS.RAW includes average compensation, at the school level, for teachers in Michigan. Interest lies in testing for a trade-off between salary and nonsalary compensation. We view this as a cluster sample of school districts, with the schools within districts representing the individual units.

A standard approach is to estimate the equation

$$\log(avgsal_{gm}) = \alpha + \beta_1 bs_{gm} + \beta_2 \log(staff_{gm}) + \beta_3 \log(enroll_{gm}) + \beta_4 lunch_{gm} + c_g + u_{gm} \tag{20.33}$$

where $avgsal_{gm}$ is the average salary for school m in district g, $bs_{gm} = avgben_{gm}/avgsal_{gm}$, where $avgben_{gm}$ is the average benefits received by teachers, $staff_{gm}$ is the number of staff per 1,000 students, $enroll_{gm}$ is school enrollment, and $lunch_{gm}$ is the

Table 20.1
Salary-Benefits Trade-off for Michigan Teachers

Dependent Variable	log(*avgsal*)		
	(1)	(2)	(3)
Estimation Method	Pooled OLS	Random Effects	Fixed Effects
Explanatory Variable			
bs	−0.177	−0.381	−0.495
	(0.122)	(0.112)	(0.133)
	[0.260]	[0.150]	[0.194]
log(*staff*)	−0.691	−0.617	−0.622
	(0.018)	(0.015)	(0.017)
	[0.035]	[0.036]	[0.043]
log(*enroll*)	−0.0292	−0.0249	−0.0515
	(0.0085)	(0.0076)	(0.0094)
	[0.0257]	[0.0115]	[0.0131]
lunch	−0.00085	0.00030	0.00051
	(0.00016)	(0.00018)	(0.00021)
	[0.00057]	[0.00020]	[0.00021]
constant	13.724	13.367	13.618
	(0.112)	(0.098)	(0.113)
	[0.256]	[0.197]	[0.241]
Number of districts	537	537	537
Number of schools	1,848	1,848	1,848

Quantities in parentheses are the nonrobust standard errors; those in brackets are robust to arbitrary within-district correlation as well as heteroskedasticity.
The intercept reported for fixed effects is the average of the estimated district effects.
The fully robust regression based Hausman test, with four degrees-of-freedom in the chi-square distribution, yields $H = 20.70$ and p-value $= 0.0004$.

percentage of students eligible for the federal free or reduced-price lunch program. Using the approximation $\log(1 + x) \approx x$ for "small" x, it can be shown that a dollar-for-dollar trade-off in salary and benefits is the same as $\beta_1 = -1$.

We estimate the equation using three methods: pooled OLS, random effects, and fixed effects. The results are given in Table 20.1. The table contains the nonrobust standard errors for each method—that is, the standard errors computed under the "ideal" set of assumptions for the particular estimator—along with the standard errors that are robust to arbitrary within-district correlation and heteroskedasticity.

The POLS estimates provide little evidence of a trade-off between salary and benefits. The coefficient is negative, but its value, −0.177, is pretty small, and not close to −1 (the hypothesized value for a one-for-one trade-off between salary and benefits). Its fully robust t statistic is less than 0.7 in magnitude. Notice that the robust standard error, which properly accounts for the cluster nature of the data, is more than twice as large as the nonrobust one.

The magnitude of the random effects coefficient *bs* is notably larger than the pooled OLS estimate, and it is statistically different from zero, even using the fully robust standard error. The RE transformation removes a fraction of the district averages. (The fraction depends on the number of schools in a district, and it ranges from about 0.379 to 0.938, with more than 50% of the districts at 0.379.) Even though RE nominally accounts for the cluster (district) effect, the nonrobust standard errors evidently understate the actual sampling variation. The robust 95% confidence interval excludes zero, but it also excludes -1 (and the RE point estimate, -0.381, is far from -1). The robust standard error on log(*staff*), 0.036, is more than twice as large as the nonrobust one, 0.015. Again, this finding points to the importance of using robust inference even if we nominally account for the common district effect by means of random effects estimation. Incidentally, the RE robust standard errors are, except in one case, smaller than the robust pooled OLS standard errors, indicating that RE is more efficient than POLS even though RE is evidently not the most efficient estimator (because it appears there is a more complicted pattern of cluster correlation than accounted for by RE).

Column (3) in Table 20.1 contains the fixed effects estimates. The coefficient on *bs* is about -0.50, which is still pretty far from -1, and statistically different from -1 even using the fully robust standard error. Again, allowing for clustering and heteroskedasticity is important for appropriate inference: the usual FE standard errors appear to be too small. Because total compensation varies significantly by district, it is important to allow the district effects to be correlated with the explanatory variables, as FE does.

Not surprisingly, the fully robust RE standard errors are somewhat below the fully robust FE standard errors, a result which makes it tempting to use the RE estimates. But the robust Hausman test, obtained by adding the four group averages to the RE estimation and testing their joint significance, yields a low p-value, about 0.0004. It appears the district effect is systematically related to some of the variables (staff size especially), and so the safest strategy is to use the fixed effects estimates with fully robust inference.

The discussion of the previous methods extends immediately to instrumental variables versions of all estimators. With large G, one can afford to make pooled two-stage least squares (2SLS), random effects 2SLS, and fixed effects 2SLS robust to arbitrary within-cluster correlation and heteroskedasticity. Adding the group averages of the exogenous explanatory variables (including the extra instruments), estimating the resulting equation by RE 2SLS (where the group averages act as their own instruments), and jointly testing the group averages for significance leads to a simple Hausman test comparing RE 2SLS and FE 2SLS.

If the random effects variance matrix structure does not hold, more efficient estimation can be achieved by applying generalized method of moments (GMM); again, GMM is justified with large G.

As we discussed in Section 12.10.3, one might apply least absolute deviations or quantile regression directly to equation (20.32). While difficult to justify in general, adding the group averages and then applying, say, LAD, can be a useful way to approximate the effects of the variables on the median while allowing the group heterogeneity to be correlated with the individual-specific covariates. Under the kinds of symmetry assumptions discussed in Section 12.10.3, this can be a good way to account for outliers in the data.

For a the case where G is much larger than the group sizes, cluster-robust inference is available for nonlinear models, too. A general treatment based on M-estimation is possible, but most of the points can be illustrated with binary response models. Let y_{gm} be a binary response, with $\mathbf{x}_g$ and $\mathbf{z}_{gm}$, $m = 1, \ldots, M_g$, $g = 1, \ldots, G$ defined as before. Assume that

$$y_{gm} = 1[\alpha + \mathbf{x}_g\boldsymbol{\beta} + \mathbf{z}_{gm}\boldsymbol{\gamma} + c_g + u_{gm} \geq 0], \tag{20.34}$$

where c_g is the cluster effect and u_{gm} is the unit-specific error. If, say, we assume

$$u_{gm} \mid \mathbf{x}_g, \mathbf{Z}_g, c_g \sim \text{Normal}(0, 1), \tag{20.35}$$

then

$$\text{P}(y_{gm} = 1 \mid \mathbf{x}_g, \mathbf{z}_{gm}, c_g) = \text{P}(y_{gm} = 1 \mid \mathbf{x}_g, \mathbf{Z}_g, c_g) = \Phi(\alpha + \mathbf{x}_g\boldsymbol{\beta} + \mathbf{z}_{gm}\boldsymbol{\gamma} + c_g), \tag{20.36}$$

where $\Phi(\cdot)$ is the standard normal (cdf), as usual. Alternatively, if u_{gm} follows a logistic distribution, then we replace $\Phi(\cdot)$ with $\Lambda(\cdot)$. Notice that expression (20.35) assumes that, conditional on c_g, $\mathbf{x}_g$, and $\mathbf{z}_{gm}$, $\mathbf{z}_{gp}$ for $p \neq m$ does not affect the outcome. For pooled methods we could relax this restriction (as in the linear case), but, with the presence of c_g, this affords little generality in practice.

As in nonlinear panel data models, the presence of c_g in equation (20.36) raises several important issues, including how we estimate quantities of interest. As in the panel data case, we have some interest in estimating average partial or marginal effects. For example, if the first element of $\mathbf{x}_g$ is continuous,

$$\frac{\partial \text{P}(y_{gm} = 1 \mid \mathbf{x}_g, \mathbf{z}_{gm}, c_g)}{\partial x_{g1}} = \beta_1 \phi(\alpha + \mathbf{x}_g\boldsymbol{\beta} + \mathbf{z}_{gm}\boldsymbol{\gamma} + c_g), \tag{20.37}$$

where $\phi(\cdot)$ is the standard normal density function. If

$$c_g \mid \mathbf{x}_g, \mathbf{Z}_g \sim \text{Normal}(0, \sigma_c^2), \tag{20.38}$$

then the APEs are obtained from the average structural function

$$ASF(\mathbf{x}_g, \mathbf{z}_{gm}) = \Phi[(\alpha + \mathbf{x}_g\boldsymbol{\beta} + \mathbf{z}_{gm}\boldsymbol{\gamma})/(1 + \sigma_c^2)^{1/2}] = \Phi(\alpha_c + \mathbf{x}_g\boldsymbol{\beta}_c + \mathbf{z}_{gm}\boldsymbol{\gamma}_c), \tag{20.39}$$

where $\alpha_c = \alpha/(1 + \sigma_c^2)^{1/2}$, and so on. Because the right-hand side of equation (20.39) is $P(y_{gm} = 1 \mid \mathbf{x}_g, \mathbf{Z}_g)$, the scaled coefficients are conveniently estimated using pooled probit or a generalized estimating equation (GEE) approach. In either case, inference must be robust to allow general covariance structures $\text{Cov}(y_{gm}, y_{gp} \mid \mathbf{x}_g, \mathbf{Z}_g)$ for $m \neq p$. These certainly will not be zero (as would be required to ignore the clustering using pooled methods), but neither will they be constant. The same formulas used in the panel data case apply to cluster samples, with the small change that the group sizes are generally different.

The pooled and GEE approaches are attractive because they are computationally simple and do not require specification of a joint distribution within clusters. Alternatively, we can impose more assumptions — as in the panel data case — and use full maximimum likelihood (conditional on $\mathbf{x}_g$ and $\mathbf{Z}_{gm}$, of course). If we supplement assumptions (20.34), (20.35), and (20.38) with

$$\{u_{g1}, \ldots, u_{g,M_g}\} \text{ are independent conditional on } (\mathbf{x}_g, \mathbf{Z}_g, c_g), \tag{20.40}$$

then we have the random effects probit model. Details of its estimation are similar to the panel data case, with the minor exception that here we must allow for an unequal number of observations per cluster. Because we can separately identify α, $\boldsymbol{\beta}$, γ, and σ_c^2, partial effects at various values of c_g are identified in addition to the average partial effects. In one important way, a random effects approach under the conditional independence assumption (20.40) is more attractive for cluster samples than for panel data: in panel data it is often the case that time series innovations are correlated over time. With a cluster sample, independence of individual outcomes after conditioning on a common cluster effect is often more believable. (Nevertheless, we are conditioning only on a scalar heterogeneity, c_g, so such an assumption may still be too restrictive.)

As with panel data, we often want to allow the cluster heterogeneity, c_g, to be correlated with the observed covariates. When the cluster sizes are the same—that is, $M_g = M$ for all g—we can apply the same methods we used for balanced panel, including probit, logit, ordered probit, Tobit, count data, fractional responses, and so on. Calculations of average partial effects are identical, and random effects approaches under conditional independence are attractive. But pooled methods are often computationally simpler and are sufficient for identifying APEs.

A challenging task for CRE approaches using cluster samples is how to model correlation between the unobserved heterogeneity and $\{\mathbf{z}_{gm}: m = 1, \ldots, M_g\}$ when

the clusters are not balanced with respect to cluster size. One possible solution is to randomly drop observations from clusters to make them all the same size. Then we could apply the usual approach for balanced panel data. Unfortunately, this approach can be very costly in terms of lost data. For example, if the smallest group has $M_g = 3$, we would have to drop observations from all other groups until we only have three in each group.

The problem with different group sizes is that it is unclear how one should model the correlation between c_g and $(\mathbf{z}_{g1}, \ldots, \mathbf{z}_{g,M_g})$ for each g. Nevertheless, there are several possibilities. We can get some insight by assuming joint normality of $(c_g, \mathbf{z}_{g1}, \ldots, \mathbf{z}_{g,M_g})$ and then assuming that $\mathrm{E}(c_g \mid \mathbf{z}_{g1}, \ldots, \mathbf{z}_{g,M_g}) = \mathrm{E}(c_g \mid \bar{\mathbf{z}}_g) = \eta_g + \bar{\mathbf{z}}_g \xi_g$, that is, assuming $\bar{\mathbf{z}}_g$ is a sufficient statistic for the mean. Then it must follow that

$$\xi_g = [\mathrm{Var}(\bar{\mathbf{z}}_g)]^{-1} \mathrm{Cov}(\bar{\mathbf{z}}_g, c_g)$$

$$\eta_g = \mathrm{E}(c_g) - \mathrm{E}(\bar{\mathbf{z}}_g)\xi_g.$$

For the sake of argument, assume that $\{\mathbf{z}_{gm}: m = 1, \ldots, M_g\}$ has an unobserved effects structure, that is, $\mathbf{z}_{gm} = \mathbf{r}_g + \mathbf{e}_{gm}$ where $\mathbf{r}_g$ is uncorrelated with each $\mathbf{e}_{gm}$ and $\{\mathbf{e}_{gm}: m = 1, \ldots, M_g\}$ are pairwise uncorrelated with zero mean and common variance matrix $\boldsymbol{\Sigma}_\mathbf{e}$. Then $\mathrm{E}(\bar{\mathbf{z}}_g) = \boldsymbol{\mu}_\mathbf{r}$ and $\mathrm{Var}(\bar{\mathbf{z}}_g) = \boldsymbol{\Sigma}_\mathbf{r} + M_g^{-1}\boldsymbol{\Sigma}_\mathbf{e}$. Assume that c_g is uncorrelated with the $\mathbf{e}_{gm}$, and let $\boldsymbol{\sigma}_{\mathbf{r}c}$ be the vector of covariances of $\mathbf{r}_g$ with c_g. Then

$$\xi_g = (\boldsymbol{\Sigma}_\mathbf{r} + M_g^{-1}\boldsymbol{\Sigma}_\mathbf{e})^{-1}\boldsymbol{\sigma}_{\mathbf{r}c} \tag{20.41}$$

$$\eta_g = \mu_c - \boldsymbol{\mu}_\mathbf{r}\xi_g, \tag{20.42}$$

where $\mu_c = \mathrm{E}(c_g)$. Further, if we write $c_g = \eta_g + \bar{\mathbf{z}}_g\xi_g + a_g$,

$$\mathrm{Var}(a_g) = \sigma_c^2 - \boldsymbol{\sigma}_{\mathbf{r}c}'(\boldsymbol{\Sigma}_\mathbf{r} + M_g^{-1}\boldsymbol{\Sigma}_\mathbf{e})^{-1}\boldsymbol{\sigma}_{\mathbf{r}c}. \tag{20.43}$$

These calculations show that, even under fairly strong assumptions, both $\mathrm{E}(c_g \mid \bar{\mathbf{z}}_g)$ and $\mathrm{Var}(c_g \mid \bar{\mathbf{z}}_g)$ depend on the group size, M_g (and $\mathrm{E}(c_g \mid \bar{\mathbf{z}}_g)$ depends on $\bar{\mathbf{z}}_g$, too). If $\boldsymbol{\Sigma}_\mathbf{r} = \mathbf{0}$, the mean and variance depend on M_g and $M_g \cdot \bar{\mathbf{z}}_g$. If $\boldsymbol{\Sigma}_\mathbf{r}$ is "large" (in a matrix sense), or M_g is large, the mean and variance are almost independent of M_g (but the mean is a linear function of $\bar{\mathbf{z}}_g$). If $\boldsymbol{\Sigma}_\mathbf{r}$ and $\boldsymbol{\Sigma}_\mathbf{e}$ are both scalar multiples of the identity matrix, the function of M_g has the form $(\sigma_r^2 + M_g^{-1}\sigma_e^2)^{-1}$, which, for large M_g, can be approximated well by a low-order polynomial in M_g^{-1}.

How should we apply these calculations for the conditional mean and variance of c_g? First, we should recognize that they are derived under strong assumptions, so we should not use such specific forms. (Plus, they would not be very easy to handle computationally.) An approach that may be flexible enough is

$$c_{ig} = \psi_0 + \psi_1 M_g + \mathbf{z}_{ig}\psi_2 + (M_g \cdot \mathbf{z}_{ig})\psi_3 + a_{ig} \tag{20.44}$$

$$\text{Var}(a_{ig} \mid \mathbf{z}_{g1}, \ldots, \mathbf{z}_{gT}, M_g) = \text{Var}(a_{ig}) = \omega_0 + \omega_1 M_g \tag{20.45}$$

(where we expect $\omega_1 < 0$ because the conditional variance of c_{ig} shrinks as the number of explanatory variables increases). If we use these expressions in place of the usual Chamberlain-Mundlak approach (and include $\mathbf{x}_{gt}$, too), we get the following estimating equation:

$$\text{P}(y_{gm} = 1 \mid \mathbf{x}_{gt}, \mathbf{Z}_{gt}) = \Phi\left[\frac{(\alpha + \mathbf{x}_{gt}\boldsymbol{\beta} + \mathbf{z}_{gm}\boldsymbol{\gamma} + \psi_1 M_g + \mathbf{z}_{gt}\psi_2 + (M_g \cdot \mathbf{z}_{gt})\psi_3)}{\sqrt{1 + \omega_0 + \omega_1 M_g}}\right]. \tag{20.46}$$

In principle, the parameters here can be estimated by, say, a pooled heteroskedastic probit analysis. To estimate the parameters, a normalization is needed on the variance (because when $\omega_1 = 0$, only the parameters scaled by $1/\sqrt{1 + \omega_0}$ are identified). In fact, using modern software that allows for exponential forms of heteroskedasticity in probit anlaysis, an easy way to estimate the identified parameters, and then obtain average partial effects, is to specify the variance as $\exp[\delta_1 \log(M_g)]$. When $\delta_1 = 0$, the variance is one, and we estimate the scaled coefficients. Notice that specifying the composite variance as $\exp[\delta_1 \log(M_g)]$ also has the benefit of nesting the cases where $\text{Var}(a_{ig})$ is a linear function of M_g or M_g^{-1}.

A more flexible approach is to let the conditional variance of c_{ig} be as flexible as the conditional mean, but still nesting the preceding simple functional form. For example, a more flexible estimating equation is

$$\text{P}(y_{gm} = 1 \mid \mathbf{x}_g, \mathbf{Z}_g) = \Phi\left[\frac{(\alpha + \mathbf{x}_g\boldsymbol{\beta} + \mathbf{z}_{gm}\boldsymbol{\gamma} + \psi_1 M_g + \bar{\mathbf{z}}_g\psi_2 + (M_g \cdot \bar{\mathbf{z}}_g)\psi_3)}{\exp(\delta_1 \log(M_g) + \bar{\mathbf{z}}_g\boldsymbol{\delta}_2 + \log(M_g)\bar{\mathbf{z}}_g\boldsymbol{\delta}_3)}\right]. \tag{20.47}$$

We could replace M_g in the mean part with M_g^{-1}, or even use both functions. Such an equation is relatively straightforward to estimate using heteroskedastic probit software.

A very attractive alternative with large G and not much variation in the group sizes M_g is to allow a different set of parameters in $\text{D}(c_g \mid M_g, \bar{\mathbf{z}}_g)$ for each value of M_g. This is easily accomplished by including dummies for all but one group size and also interacting the dummies with $\bar{\mathbf{z}}_g$ in the mean and the variance. In equation (20.43) the variance depends only on M_g, and so one might want to simplify the estimation by including only the group-size dummies in the variance.

Regardless of the specific expression we use for $\text{P}(y_{gm} = 1 \mid \mathbf{x}_g, \mathbf{Z}_g)$, it is straightforward to estimate the average partial effects. The conditioning variables that we must average out are $(M_g, \bar{\mathbf{z}}_g)$, and we use, as usual, the discussion in Section 2.2.5.

Let $m(\mathbf{x}_g, \mathbf{z}_{gm}, M_g, \bar{\mathbf{z}}_g, \theta)$ be the response probability $P(y_{gm} = 1 \mid \mathbf{x}_g, \mathbf{Z}_g)$, where θ is the set of all parameters. Then the ASF, for fixed values of $\mathbf{x}$ and $\mathbf{z}$, is consistently estimated as

$$\widehat{ASF}(\mathbf{x}, \mathbf{z}) = G^{-1} \sum_{g=1}^{G} m(\mathbf{x}, \mathbf{z}, M_g, \bar{\mathbf{z}}_g, \hat{\theta}); \tag{20.48}$$

that is, we average out $(M_g, \bar{\mathbf{z}}_g)$. (As usual, one must use caution in interpreting the effects of the group-level variables if these are partially correlated with c_g.)

Incidentally, the methods proposed here can be applied to unbalanced panel data sets, assuming, of course, that the reason the panel is unbalanced can be ignored. With a large cross section, N (which replaces G), and a small number of time periods, T_i (which replaces M_g) for each observation i, the flexible approach of allowing different parameters for each T_i is attractive.

Rather than adopt a correlated random effects probit approach, we can apply the fixed effects logit approach, assuming that the observations within a cluster are independent conditional on the observed covariates and the cluster effect, c_g. Naturally, the cluster-level variables, $\mathbf{x}_g$, are eliminated, and one can only estimate parameters, not partial effects. The mechanics are essentially identical to the panel data case. Geronimus and Korenman (1992) use sister pairs to study the effects of teenage motherhood on subsequent economic outcomes, so $M_g = 2$ for all g. When the outcome is binary (such as an employment indicator), the authors apply fixed effects logit. CRE probit can also be used to obtain the magnitudes of the effects.

The same CRE approach can be applied to other nonlinear models, such as ordered probit and Tobit models. Generally, if we begin with a density $f(\mathbf{y}_{gm} \mid \mathbf{x}_g, \mathbf{z}_{gm}, \mathbf{c}_g; \theta)$, where both $\mathbf{y}_{gm}$ and $\mathbf{c}_g$ can be vectors, and then specify a heterogeneity density, say, $h(\mathbf{c}_g \mid M_g, \bar{\mathbf{z}}_g; \delta)$, a partial MLE analysis can be obtained by "integrating out" $\mathbf{c}_g$ to get the density

$$\int f(\mathbf{y}_{gm} \mid \mathbf{x}_g, \mathbf{z}_{gm}, \mathbf{c}; \theta) h(\mathbf{c} \mid M_g, \bar{\mathbf{z}}_g; \delta) \, d\mathbf{c}. \tag{20.49}$$

As we know from the panel data case, this density has a simple form for common models, such as Tobit, when c_g is a scalar and $h(\cdot \mid M_g, \bar{\mathbf{z}}_g; \delta)$ is chosen to have a simple form, such as normal. However, as for the probit case, one should allow heteroskedasticity in $\text{Var}(c_g \mid M_g, \bar{\mathbf{z}}_g)$, leading to a pooled estimation strategy based on the "heteroskedastic Tobit" model. We also know that using pooled (that is, partial) MLE does not always fully identify the parameters, but it does often identify scaled parameters and average partial effects. Because the observations within clusters are

almost certainly correlated, even after conditioning on $(\mathbf{x}_g, \mathbf{Z}_g)$, inference that allows within-cluster correlation is crucial.

Various quasi-MLEs can also be adapted to account for correlated random effects in the context of cluster sampling. In the exponential case, we would be led to a mean function that looks like, say,

$$\mathrm{E}(y_{gm} \mid \mathbf{Z}_g, M_g) \quad \mathrm{E}(y_{gm} \mid \mathbf{z}_{gm}, M_g, \bar{\mathbf{z}}_g) \quad \exp(\mathbf{x}_g\beta + \mathbf{z}_{gm}\gamma + \alpha_{M_g} + \bar{\mathbf{z}}_g\psi_{M_g}), \tag{20.50}$$

where α_{M_g} and ψ_{M_g} are specific to the group size, M_g (or we use linear or low-order polynomials in M_g or M_g^{-1}, or both). The parameters can be estimated by, say, pooled Poisson QMLE, or GEE using the Poisson distribution by including a full set of group-size dummies along with $\bar{\mathbf{z}}_g$ and interactions with $\bar{\mathbf{z}}_g$. The elements of (β, γ) measure semielasticities or elasticities on the mean response. The APEs on the mean are obtained by averaging out $(M_g, \bar{\mathbf{z}}_g)$. See Problem 20.8 for the case of a fractional response.

As in the panel data case, the fixed effects Poisson estimator is very convenient when we start with

$$\mathrm{E}(y_{gm} \mid \mathbf{Z}_g, c_g) = \mathrm{E}(y_{gm} \mid \mathbf{z}_{gm}, c_g) = \exp(\mathbf{z}_{gm}\gamma + c_g). \tag{20.51}$$

With arbitrarily unbalanced group sizes, the FE Poisson estimator (viewed as a quasi-MLE) consistently estimates γ (and $\mathbf{x}_g$ is eliminated). No other feature of the Poisson distribution needs to be correctly specified, and sources of within-cluster correlation other than c_g are allowed (provided, of course, we use fully robust inference).

20.3.2 Cluster Samples with Unit-Specific Panel Data

Often, cluster samples come with a time component, so that there are two potential sources of correlation across observations: across time within the same individual and across individuals within the same group. The two sources of correlation may also interact: different individuals within the same group or cluster might have unobserved shocks correlated across different time periods.

Generally, accounting for more than two data dimensions is complicated if there is not a natural nesting. Here we consider the case where each unit belongs to a cluster and the cluster identification does not change over time. In other words, we have panel data on each individual or unit, and each unit belongs to a cluster. For example, we might have annual panel data at the firm level where each firm belongs to the same industry (cluster) for all years. Or, we might have panel data for schools that each belong to a district. This is a special case of a **hierarchical linear model (HLM)**

setup. Models for data structures involving panel data and clustering are also called **mixed models** (although this latter name typically refers to the situation, which we treat later, in which some slope parameters are constant and others are unobserved heterogeneity). In the HLM/mixed models literature, more levels of nesting are allowed, but we will not consider more general structures; see, for example, Raudenbush and Bryk (2002).

Now we have three data subscripts on at least some variables that we observe. For example, the response variable is y_{gmt}, where g indexes the group or cluster, m is the unit within the group, and t is the time index. Mainly for expository purposes, assume we have a balanced panel with the time periods running from $t = 1, \ldots, T$. Within cluster g there are M_g units, and we have sampled G clusters. (In the HLM literature, g is usually called the *first level* and m the *second level*.)

As with a pure cluster sample, we assume that we have many groups, G, and relatively few members of the group. Further, our discussion of asymptotic properties of estimators assumes that T is fixed. In particular, the analysis is with the M_g and T fixed with G getting large. For example, if we can sample, say, several hundred school districts, with a few to maybe a few dozen schools per district, over a handful of years, then we have a data set that can be analyzed in the current framework.

A standard linear model with constant slopes can be written, for $t = 1, \ldots, T$, $m = 1, \ldots, M_g$, and a random draw g from the population of clusters as

$$y_{gmt} = \eta_t + \mathbf{w}_g\alpha + \mathbf{x}_{gm}\beta + \mathbf{z}_{gmt}\delta + h_g + c_{mg} + u_{gmt}, \tag{20.52}$$

where, say, h_g is the industry or district effect, c_{gm} is the firm effect or school effect (firm or school m in industry or district g), and u_{gmt} is the idiosyncratic effect. In other words, the composite error is

$$v_{gmt} = h_g + c_{gm} + u_{gmt}. \tag{20.53}$$

Generally, the model can include variables that change at any level. In equation (20.52), some elements of $\mathbf{z}_{gmt}$ might change only across g and t, and not by unit. This is an important special case for policy analysis where the policy applies at the group level and changes over time. In such cases it is crucial for obtaining correct inference to recognize the cluster correlation. In effect, if one has observables in the model measured at the group level (whether or not they change over time), it is effectively cheating to then assume there are no group-level unobservables affecting y_{gmt}. This could be the case, but one should not assume it from the outset.

A simple estimation method, assuming v_{gmt} is uncorrelated with $(\mathbf{w}_g, \mathbf{x}_{gm}, \mathbf{z}_{gmt})$, is pooled OLS, which is consistent as $G \to \infty$ for any cluster or serial correlation pattern. The most general inference for pooled OLS—maintaining independence across

clusters—is to allow any kind of serial correlation across units or time, or both, within a cluster.

Not surprisingly, one can apply a generalized least squares analysis that makes assumptions about the components of the composite error. Typically, it is assumed that the components are pairwise uncorrelated, the c_{gm} are uncorrelated within cluster (with common variance), and the u_{gmt} are uncorrelated within cluster and across time (with common variance). The resulting feasible GLS estimator is an extension of the usual random effects estimator for panel data. Because of the large-G setting, the estimator is consistent and asymptotically normal whether or not the actual variance structure we use in estimation is the proper one. To guard against heteroskedasticity in any of the errors and serial correlation in the $\{u_{gmt}\}$, one should use fully robust inference that does not rely on the form of the unconditional variance matrix (which may also differ from the conditional variance matrix).

Simple strategies are available, too. For example, one can apply random effects at the individual level, effectively ignoring the clusters in estimation. In other words, treat the data as a standard panel data set in estimation. Such an estimator might be more efficient than pooled OLS yet easier to obtain than a complete GLS analysis that also accounts for the cluster sampling. To account for the cluster sampling in inference, one computes a fully robust variance matrix estimator for the usual random effects estimator. Many statistical packages have options to allow for clustering at a higher level of aggregation than the level at which random effects is applied.

More formally, write the equation for each cluster as

$$\mathbf{y}_g = \mathbf{R}_g\theta + \mathbf{v}_g, \tag{20.54}$$

where a row of $\mathbf{R}_g$ is $(1, d2, \ldots, dT, \mathbf{w}_g, \mathbf{x}_{gm}, \mathbf{z}_{gmt})$ (which includes a full set of period dummies) and θ is the vector of all regression parameters. For cluster g, $\mathbf{y}_g$ contains M_gT elements (T periods for each unit m). In particular,

$$\mathbf{y}_g = \begin{pmatrix} \mathbf{y}_{g1} \\ \mathbf{y}_{g2} \\ \vdots \\ \mathbf{y}_{g,M_g} \end{pmatrix}, \qquad \mathbf{y}_{gm} = \begin{pmatrix} y_{gm1} \\ y_{gm2} \\ \vdots \\ y_{gmT} \end{pmatrix}, \tag{20.55}$$

so that each $\mathbf{y}_{gm}$ is $T \times 1$; $\mathbf{v}_g$ has an identical structure. Now, we can obtain $\mathbf{\Omega}_g = \text{Var}(\mathbf{v}_g)$ under various assumptions and apply feasible GLS.

Random effects estimation at the unit level is obtained by choosing $\mathbf{\Omega}_g = \mathbf{I}_{M_g} \otimes \mathbf{\Lambda}$, where $\mathbf{\Lambda}$ is the $T \times T$ matrix with the RE structure. Of course, if there is within-cluster correlation, this is not the correct form of $\text{Var}(\mathbf{v}_g)$, and that is why robust

inference generally is needed after RE estimation. Generally, to allow for an incorrect structure imposed on $\mathbf{\Omega}_g$, or to allow for system heteroskedasticity, that is, $\text{Var}(\mathbf{v}_g \mid \mathbf{R}_g) \neq \text{Var}(\mathbf{v}_g)$, we use fully robust inference. In particular, the robust asymptotic variance of $\hat{\boldsymbol{\theta}}$ is estimated as

$$\widehat{\text{Avar}}(\hat{\boldsymbol{\theta}}) = \left(\sum_{g=1}^{G} \mathbf{R}_g' \hat{\mathbf{\Omega}}_g^{-1} \mathbf{R}_g\right)^{-1} \left(\sum_{g=1}^{G} \mathbf{R}_g' \hat{\mathbf{\Omega}}_g^{-1} \hat{\mathbf{v}}_g \hat{\mathbf{v}}_g' \hat{\mathbf{\Omega}}_g^{-1} \mathbf{R}_g\right)^{-1} \left(\sum_{g=1}^{G} \mathbf{R}_g' \hat{\mathbf{\Omega}}_g^{-1} \mathbf{R}_g\right)^{-1}, \quad (20.56)$$

where $\hat{\mathbf{v}}_g = \mathbf{y}_g - \mathbf{R}_g \hat{\boldsymbol{\theta}}$. Some software packages that allow cluster-robust inference after panel data estimation compute this fully robust asymptotic variance. Unfortunately, routines intended for estimating HLMs (or mixed models) often assume that the structure imposed on $\mathbf{\Omega}_g$ is correct, and that $\text{Var}(\mathbf{v}_g \mid \mathbf{R}_g) = \text{Var}(\mathbf{v}_g)$. The resulting inference could be misleading, especially if serial correlation in $\{u_{gmt}\}$ is not allowed.

Because of the nested data structure, we have available different versions of fixed effects estimators. Subtracting cluster averages from all observations within a cluster eliminates h_g; when $\mathbf{w}_{gt} = \mathbf{w}_g$ for all t, $\mathbf{w}_g$ is also eliminated. But the unit-specific effects, c_{mg}, are still part of the error term. If we are mainly interested in $\boldsymbol{\delta}$, the coefficients on the time-varying variables $\mathbf{z}_{gmt}$, then removing c_{gm} (along with h_g) is attractive. In other words, use a standard fixed effects analysis at the individual level. (If the units were allowed to change groups over time, then we would replace h_g with h_{gt}, and then subtracting off individual-specific means would not remove the time-varying cluster effects.)

Example 20.4 (Effects of Spending on School Performance): The data in MEAP94_98, which are a subset of those used in Papke (2005), contain school-level panel data on student performance and per-pupil spending. The variable to be explained, *math4*, is the percentage of students receiving a satisfactory score on a fourth-grade math test administered by the state of Michigan. The key variable, $lavgrexp = \log(avgrexp)$, is the log of average real per-pupil spending for the current and previous year. The data set is for the years 1994 through 1998; it is unbalanced, with schools having either three, four, or five years of data (in various patterns). The other school-level controls are enrollment, in logarithmic form (*lenroll*), and the percentage of students eligible for the free lunch program (*lunch*). A full set of year dummies is also included.

We can view this as a cluster sample because schools are nested within districts. Certainly much of the variability in spending is across districts, and so it may be important to allow for district-level effects.

The results of fixed effects estimation, at the school level, are given in Table 20.2. Because schools are in the same district in every year, eliminating a school effect also

Table 20.2
Fixed Effects Estimation of Spending on Test Pass Rates

Dependent Variable	*math4*			
	FE Coefficient	Usual FE Standard Error	S.E. Clustered by School	S.E. Clustered by District
Explanatory Variable				
log(*avgrexp*)	6.29	2.10	2.43	3.13
lunch	−0.022	0.031	0.039	0.040
log(*enrol*)	−2.04	1.79	1.79	2.10
y95	11.62	0.55	0.54	0.72
y96	13.06	0.66	0.69	0.93
y97	10.15	0.70	0.73	0.96
y98	23.41	0.72	0.77	1.03
Number of districts			467	
Number of schools			1,683	

removes any additive district effect. But within-district correlation can be present if some of the slopes change by district. Along with the FE estimates, three standard errors are provided: the usual FE standard errors that ignore serial correlation and within-district correlation; the standard errors that are robust to arbitrary serial correlation within school but assume no correlation across schools within a district; and the most robust standard errors that allow within-district correlation across schools and time periods. (Remember that we are assuming independence across districts; without this assumption, proper inference becomes much more difficult.)

The FE estimate of $\beta_{lavgrexp}$ is about 6.29, which means that a 10% increase in average real spending is estimated to increase the pass rate by about 0.63 percentage points. Using the usual FE standard error, about 2.10, the t statistic for $\hat{\beta}_{lavgrexp}$ is about 3.0. Therefore, the effect of spending is statistically significant at a low significance level (about 0.3%) using the usual FE inference. The standard error that allows arbitrary serial correlation within schools (and heteroskedasticity, too) is higher, about 2.43. Naturally, this reduces the statistical significance of $\hat{\beta}_{lavgrexp}$. The standard error in column (4) is subtantially higher, about 3.13. Allowing for within-district correlation and serial correlation has practically important effects on the uncertainty associated with the estimate. Using the fully robust standard error, the 95% confidence interval for $\beta_{lavgrexp}$ excludes zero, but only barely.

If the model is given by equations (20.52) and (20.53), the unit-specific time demeaning eliminates all cluster correlation, and the inference need only be made robust to neglected serial correlation in $\{u_{gmt}\}$. But we might want to use cluster-robust inference anyway to allow for more general situations. Suppose the model is

$$y_{gmt} = \eta_t + \mathbf{w}_g\boldsymbol{\alpha} + \mathbf{x}_{gm}\boldsymbol{\beta} + \mathbf{z}_{gmt}\mathbf{d}_{mg} + h_{gt} + c_{mg} + u_{gmt}$$

$$= \eta_t + \mathbf{w}_{gt}\boldsymbol{\alpha} + \mathbf{x}_{gm}\boldsymbol{\beta} + \mathbf{z}_{gmt}\boldsymbol{\delta} + h_{gt} + c_{mg} + u_{gmt} + \mathbf{z}_{gmt}\mathbf{e}_{gm}, \tag{20.57}$$

where $\mathbf{d}_{gm} = \boldsymbol{\delta} + \mathbf{e}_{gm}$ is a set of unit-specific intercepts on the individual, time-varying covariates $\mathbf{z}_{gmt}$. The time-demeaned equation within individual m in cluster g is

$$y_{gmt} - \bar{y}_{gm} = \zeta_t + (\mathbf{z}_{gmt} - \bar{\mathbf{z}}_{gm})\boldsymbol{\delta} + (u_{gmt} - \bar{u}_{gm}) + (\mathbf{z}_{gmt} - \bar{\mathbf{z}}_{gm})\mathbf{e}_{gm}. \tag{20.58}$$

Because the $\mathbf{e}_{gm}$ are generally correlated across units within cluster g, the last term generally induces cluster correlation of a heteroskedastic nature within cluster g. From our discussion in Section 11.7.3, we know that FE is still consistent if $\mathrm{E}(\mathbf{d}_{mg} \mid \mathbf{z}_{gmt} - \bar{\mathbf{z}}_{gm}) = \mathrm{E}(\mathbf{d}_{mg})$, $m = 1, \ldots, M_g$, $t = 1, \ldots, T$, and all g, and so cluster-robust inference, which is automatically robust to serial correlation and heteroskedasticity, makes perfectly good sense.

An important feature of the HLM approach is the possibility of allowing the slopes to depend on observed covariates. Often one begins with a model at the unit–time-period level that contains heterogeneity, and then allows the intercept and slopes to depend on higher-level covariates. Write a model for unit m at time t in cluster g as

$$y_{gmt} = \mathbf{z}_{gmt}\mathbf{d}_{gm} + v_{gmt}, \tag{20.59}$$

and then decompose the idiosyncratic error, v_{gmt}, as

$$v_{gmt} = \eta_t + c_{gm} + u_{gmt}, \tag{20.60}$$

where the η_t are aggregate time effects. For notational simplicity, we absorb the group effect, h_{gt}, into u_{gmt}, and allow c_{gm} and u_{gmt} to be correlated within group. For each (g, m) define

$$\bar{\mathbf{r}}_{gm} = (\mathbf{w}_g, \bar{\mathbf{x}}_g, \mathbf{x}_{gm}, \bar{\mathbf{z}}_{gm}),$$

where $\bar{\mathbf{x}}_g = M_g^{-1}\sum_{p=1}^{M_g}\mathbf{x}_{gp}$ and $\bar{\mathbf{z}}_{gm} = T^{-1}\sum_{s=1}^{T}\mathbf{z}_{gms}$. In other words, $\bar{\mathbf{r}}_{gm}$ includes the group-level covariates along with group averages of the unit-specific covariates, the unit-specific covariates, and the time averages of the covariates that change over time. Now assume

$$c_{gm} = \alpha + \bar{\mathbf{r}}_{gm}\boldsymbol{\gamma} + a_{gm} \tag{20.61}$$

$$\mathbf{d}_{gm} = \boldsymbol{\delta} + \boldsymbol{\Pi}(\bar{\mathbf{r}}_{gm} - \boldsymbol{\mu}_{\bar{\mathbf{r}}})' + \mathbf{e}_{gm}, \tag{20.62}$$

insert these in the equation, and use basic algebra:

$$y_{gmt} = \zeta_t + \bar{\mathbf{r}}_{gm}\boldsymbol{\gamma} + \mathbf{z}_{gmt}\boldsymbol{\delta} + [(\bar{\mathbf{r}}_{gm} - \boldsymbol{\mu}_{\bar{\mathbf{r}}}) \otimes \mathbf{z}_{gmt}]\boldsymbol{\pi} + a_{gm} + \mathbf{z}_{gmt}\mathbf{e}_{gm} + u_{gmt},$$

where $\boldsymbol{\pi} = \text{vec}(\boldsymbol{\Pi})$. Importantly, centering $\bar{\mathbf{r}}_{gm}$ about its average before forming the interactions means that $\boldsymbol{\delta}$ is the average partial effect. If we instead use $\bar{\mathbf{r}}_{gm} \otimes \mathbf{z}_{gmt}$, the coefficients on the level terms, $\mathbf{z}_{gmt}$, may be of little interest because they measure the effects of the $\mathbf{z}_{gmt}$ when $\bar{\mathbf{r}}_{gm}$ is zero, which is unlikely to be an interesting segment of the population. In practice, the population mean, $\boldsymbol{\mu}_{\bar{\mathbf{r}}}$, is replaced with the sample average across m and g. The presence of $\mathbf{z}_{gmt}\mathbf{e}_{gm}$ in the error term, as well as potential serial correlation in $\{u_{gmt}\}$, makes a genuine GLS analysis difficult but possible under simple structures for the variance-covariance matrices. But we can use any of the simpler strategies mentioned earlier. For example, we can act as if $\mathbf{e}_{gm} = 0$ and as if u_{gmt} is serially uncorrelated in estimation. We can apply random effects to account for a cluster-level effect or RE at the individual level, or both. Basically, we include cluster-level variables, averages of unit-specific, time-constant variables, and time averages of the variables that change over time along with the unit-specific variables. For added flexibility, we include a full set of interactions. Regardless of the specifics, we use fully robust inference.

A very similar discussion holds in the context of instrumental variables. Suppose we start with the model

$$y_{gmt} = \eta_t + \mathbf{r}_{gmt}\boldsymbol{\theta} + v_{gmt}, \tag{20.63}$$

where $\mathbf{r}_{gmt}$ contains all covariates and v_{gmt} is the composite error. If we have exogenous variables, say $\mathbf{q}_{gmt}$, such that $\text{E}(\mathbf{q}_{gmt}'v_{gmt}) = \mathbf{0}$ and the rank condition holds, then pooled 2SLS is attractive for its simplicity. It does not matter whether elements of $\mathbf{r}_{gmt}$ or $\mathbf{q}_{gmt}$ contain elements that change only across g, across g and m, across g and t, or across g, m, and t, provided the rank condition holds. Without further assumptions, the 2SLS variance matrix estimator, as well as inference generally, should be robust to arbitrary serial correlation and cluster correlation at the most aggregated level. For example, if g indexes counties and m indexes manufacturing plants operating within a county, then we should cluster at the county level. We may have a policy and instruments that change only at the county level over time, along with exogenous explanatory variables that change at the plant level (either constant or over time). In evaluating whether the rank condition holds—say, for a single endogenous variable w_{gmt}—one can use a pooled OLS regression w_{gmt} on 1, $d2_t, \ldots, dT_t$, $\mathbf{q}_{gmt}$ (assuming that $\mathbf{q}_{gmt}$ contains all exogenous variables in equation (20.63)) to test for joint significance of the proposed instruments in $\mathbf{q}_{gmt}$. Naturally, such a test should be made robust to arbitrary cluster and serial correlation to be convincing. The test works even if w_{gmt} does not change across m (or even t for that matter), and the same with $\mathbf{q}_{gmt}$. The inference is valid with large G provided it is made fully robust.

In the previous scenario, if we apply, say, fixed effects 2SLS, where we eliminate a time-constant, plant-level effect, then we need the variables of interest to at least change over time (if not across m); the same is true of the instruments. If we have instruments that change only by g, the FE2SLS estimator—whether we remove a county-level or plant-level effect—does not identify θ.

20.3.3 Should We Apply Cluster-Robust Inference with Large Group Sizes?

Until recently, the "cluster-robust" standard errors and test statistics obtained from pooled OLS, random effects, and fixed effects were known to be valid only as $G \to \infty$ with each M_g fixed. As a practical matter, that fact means that one should have lots of small groups. Recently, because of the structure of many commonly used cluster samples, researchers have become interested in the performance of cluster-robust inference when the number of groups, G, is not substantially larger than the typical group size, M_g.

Consider the basic model without a time structure, for simplicity, and consider formula (20.25), the asymptotic variance for pooled OLS. With a large number of groups and small group sizes, we can get good estimates of the within-cluster correlations—technically, of the cluster correlations of the cross products of the regressors and errors—even if they are unrestricted, and it is for that reason that the robust variance matrix is consistent as $G \to \infty$ with M_g fixed. In fact, in this scenario, one loses nothing in terms of asymptotic local power (with local alternatives shrinking to zero at the rate $G^{-1/2}$) if c_g is not present. In other words, based on first-order asymptotic analysis, there is no cost to being fully robust to any kind of within-group correlation or heteroskedasticity. These arguments apply equally to panel data sets with a large number of cross sections and relatively few time periods, whether or not the idiosyncratic errors are serially correlated, and to the cluster sample/panel data setting considered in Section 20.3.2.

What if one applies robust inference in scenarios where the fixed M_g, $G \to \infty$ asymptotic analysis is not realistic? Hansen (2007) has recently derived properties of the cluster-robust variance matrix and related test statistics under various scenarios that help us more fully understand the properties of cluster-robust inference across different data configurations. Hansen (2007, Theorem 2) shows that, with G and M_g both getting large, the usual inference based on equation (20.25) is valid with arbitrary correlation among the errors, v_{gm}, within each group. Because we usually think of v_{gm} as including the group effect c_g, this means that, with large group sizes, we can obtain valid inference using the cluster-robust variance matrix, provided that G is also large. So, for example, if we have a sample of $G = 100$ schools and roughly $M_g = 100$ students per school, and we use pooled OLS leaving the school effects

in the error term, we should expect the inference to have roughly the correct size. Probably we leave the school effects in the error term because we are interested in a school-specific explanatory variable, perhaps indicating a policy change. Adding a short time dimension does not change these conclusions.

Unfortunately, pooled OLS with cluster effects when G is small and group sizes are large falls outside Hansen's theoretical findings: the proper asymptotic analysis would be with G fixed, $M_g \to \infty$, and persistent within-cluster correlation (because of the presence of c_g in the error) causes problems in such cases. Consequently, we should not expect good properties of the cluster-robust inference with small groups and very large group sizes when cluster effects are left in the error term. As an example, suppose that $G = 10$ hospitals have been sampled with several hundred patients per hospital. If the explanatory variable of interest is exogenous and varies only at the hospital level, it is tempting to use pooled OLS with cluster-robust inference. But we have no theoretical justification for doing so, and we have reasons to expect it will not work well—including the simulations in Hansen (2007).

If the explanatory variables of interest vary within group—say, within each hospital a subset of patients were provided with a specific kind of care—fixed effects is attractive for a couple of reasons. The first advantage is the usual one about allowing c_g to be arbitrarily correlated with the $\mathbf{z}_{gm}$. The second advantage is that, with large M_g, we can treat the c_g as parameters to estimate—because we can estimate them precisely—and then assume that the observations are independent across m (as well as g). Therefore, the usual inference is valid, perhaps with adjustments for heteroskedasticity.

In summary, for true cluster sample applications, cluster-robust inference using pooled OLS delivers statistics with proper size when G and M_g are both moderately large, but they should probably be avoided with large M_g and small G. We will discuss some approaches for handling a small number of groups in Section 20.3.4.

20.3.4 Inference When the Number of Clusters Is Small

If the explanatory variable or variables of interest do not change within cluster and the number of clusters is small, none of the previous methods can be used for reliable inference. Fixed effects eliminates the key variables, while for pooled OLS we are not justified in using cluster-robust inference. (Whether a random effects analysis produces valid inference with small G and large M_g appears to be an open, and very interesting, question.)

The problem of proper inference when M_g is large relative to G was brought to light by Moulton (1990), who was interested in studying data on individuals clustered at the state level in the United States. He proposed corrections to the usual OLS

standard errors that impose more structure than the usual cluster-robust standard errors studied by Hansen (2007). Either way, the corrections to the usual OLS inference tend to work well provided the M_g are not too much bigger than G. In this subsection we are interested in cases where a large G analysis makes no sense.

Often with small G and large M_g the sampling scheme more resembles that of standard stratified sampling, but without requiring a complete partition of the population. In other words, a small set of populations are defined, and then random samples are obtained from those populations. As an example, a random sample of adults is obtained from each of a handful of cities, some of which received federal aid for a job-training program. Labor market outcomes are recorded, possibly including changes from an early time period. In this scenario, we could analyze the data as independent outcomes across and, more importantly, within group. We will return to this point.

Recent work by Donald and Lang (2007) (hereafter, DL) treats the small G case within the context of cluster sampling. That is, presumably from a large population of clusters, only a handful or so are drawn (and then we may or may not sample every unit within each cluster). As mentioned in the previous subsection, such a scenario causes problems for cluster-robust inference. Therefore, DL propose a different approach.

Before we cover the DL approach, it is important to understand that the structure of data sets in the small G case is the same whether we think of drawing a small number of clusters from a large population or fixing a few clusters and then drawing large random samples from them. Unfortunately, how one proceeds is dependent on how we view the sampling scheme. As we will see, the DL approach is typically much more conservative than the standard approach.

To illustrate the issues considered by DL, consider the simplest case, with a single regressor that varies only by group:

$$y_{gm} = \alpha + \beta x_g + c_g + u_{gm} \tag{20.64}$$

$$= \delta_g + \beta x_g + u_{gm}, \qquad m = 1, \ldots, M_g;\ g = 1, \ldots, G. \tag{20.65}$$

Notice how equation (20.65) is written as a model with common slope, β, but intercept, δ_g, that varies across g. Donald and Lang focus on equation (20.64), where c_g is assumed to be independent of x_g with zero mean. They use this formulation to highlight the problems of applying standard inference to equation (20.64), that is, acting as if c_g is absent. We know this is a bad idea even in the large G, small M_g case, as it ignores the persistent correlation in the errors within each group. Unfortunately, while Hansen (2007) has shown that cluster-robust inference is valid with large G,

even if the M_g are also large, it is not valid when G is small. Thus other approaches are needed.

One way to see the problem in applying standard inference is to note that when $M_g = M$ for all $g = 1, \ldots, G$, the pooled OLS estimator, β, is identical to the "between" estimator obtained from the regression

$$\bar{y}_g \text{ on } 1, \bar{x}_g, \qquad g = 1, \ldots, G. \tag{20.66}$$

Conditional on the $\bar{x}_g$, the estimator β inherits its distribution from $\{\bar{v}_g: g = 1, \ldots, G\}$, the within-group averages of the composite errors $v_{gm} \equiv c_g + u_{gm}$. The presence of c_g means new observations within group do not provide additional information for estimating β beyond how they affect the group average, $\bar{y}_g$. In effect, we only have G useful pieces of information.

If we add some strong assumptions, there is a solution to the inference problem. In addition to assuming $M_g = M$ for all g, assume $c_g \mid x_g \sim \text{Normal}(0, \sigma_c^2)$ and assume $u_{gm} \mid x_g, c_g \sim \text{Normal}(0, \sigma_u^2)$. Then $\bar{v}_g$ is independent of x_g and $\bar{v}_g \sim \text{Normal}(0, \sigma_c^2 + \sigma_u^2/M)$ for all g. Because we assume independence across g, the equation

$$\bar{y}_g = \alpha + \beta \bar{x}_g + \bar{v}_g, \qquad g = 1, \ldots, G \tag{20.67}$$

satisfies the classical linear model assumptions. Therefore, we can use inference based on the t_{G-2} distribution to test hypotheses about β, provided $G > 2$. When G is very small, the requirements for a significant t statistic using the t_{G-2} distribution are much more stringent than if we use the $t_{M_1+M_2+\cdots+M_G-2}$ distribution—which is what we would be doing if we used the usual pooled OLS statistics.

When $\mathbf{x}_g$ is a $1 \times K$ vector, we need $G > K + 1$ to use the t_{G-K-1} distribution for inference. (In Moulton (1990), $G = 50$ states and $\mathbf{x}_g$ contains 17 elements.)

As pointed out by DL, performing the correct inference in the presence of c_g is *not* just a matter of correcting the pooled OLS standard errors for cluster correlation—something that does not appear to be valid for small G, anyway—or using the RE estimator. In the case of common group sizes, there is only estimator: pooled OLS. Random effects and between regression in equation (20.66) all lead to the *same* $\hat{\beta}$. The regression in equation (20.66), by using the t_{G-2} distribution, yields inference with appropriate size.

We can apply the DL method without normality of the u_{gm} if the common group size M is large: by the central limit theorem, $\bar{u}_g$ will be approximately normally distributed very generally. Then, because c_g is normally distributed, we can treat $\bar{v}_g$ as approximately normal with constant variance. Further, even if the group sizes differ across g, for very large group sizes $\bar{u}_g$ will be a negligible part of $\bar{v}_g$: $\text{Var}(\bar{v}_g) = \sigma_c^2 + \sigma_u^2/M_g$. Provided c_g is normally distributed and it dominates $\bar{v}_g$, a classical linear model analysis on equation (20.67) should be roughly valid.

The broadest applicability of DL's setup occurs when the average of the idiosyncratic errors, $\bar{u}_g$, can be ignored---either because σ_u^2 is small relative to σ_c^2, M_g is large, or both. In fact, applying DL with different group sizes or nonnormality of the u_{gm} is identical to ignoring the estimation error in the sample averages, $\bar{y}_g$. In other words, it is as if we are analyzing the simple regression $\mu_g = \alpha + \beta x_g + c_g$ using the classical linear model assumptions (where we then insert $\bar{y}_g$ in place of the unknown group mean, μ_g, and ignore the estimation error). With small G, we need to further assume that c_g is normally distributed.

If $\mathbf{z}_{gm}$ appears in the model, then we can use the averaged equation

$$\bar{y}_g = \alpha + \mathbf{x}_g\boldsymbol{\beta} + \bar{\mathbf{z}}_g\boldsymbol{\gamma} + \bar{v}_g, \qquad g = 1, \ldots, G, \tag{20.68}$$

provided $G > K + L + 1$. If c_g is independent of $(\mathbf{x}_g, \bar{\mathbf{z}}_g)$ with a homoskedastic normal distribution and the group sizes are large, inference can be carried out using the $t_{G-K-L-1}$ distribution.

The DL solution to the inference problem with small G is pretty common as a strategy to check robustness of results obtained from cluster samples, but often it is implemented with somewhat large G (say, $G = 50$). Often with cluster samples one estimates the parameters using the disaggregated data and also the averaged data. When some covariates vary within cluster, using averaged data is generally inefficient, but when estimating equation (20.68) we need not make standard errors robust to within-cluster correlation. We now know that if G is reasonably large and the group sizes not too large, the cluster-robust inference applied to the disaggregated data can be acceptable. As pointed out by DL, with small G one should use the group averages in a classical linear model analysis.

For small G and large M_g, inference obtained analyzing equation (20.67) as a classical linear model will be very conservative in the absence of a cluster effect. Thus the DL approach can be used in situations where one requires very strong statistical evidence for the effect of a policy. Nevertheless, the DL approach rules out some widely used staples of policy analysis. For example, suppose we have two populations (maybe men and women, two different cities, or a treatment and a control group) with means μ_g, $g = 1, 2$, and we would like to obtain a confidence interval for their difference. In almost all cases, it makes sense to view the data as being two random samples, one from each subgroup of the population. Under random sampling from each group, and assuming normality and equal population variances, the usual comparison-of-means statistic is distributed exactly as $t_{M_1+M_2-2}$ under the null hypothesis of equal population means. (Or, we can construct an exact 95% confidence interval of the difference in population means.) With even moderate sizes for M_1 and M_2, the $t_{M_1+M_2-2}$ distribution is close to the standard normal distribution. Also, we can relax normality to obtain approximately valid inference, and it is easy

to adjust the t statistic to allow for different population variances. With a controlled experiment, the standard difference-in-means analysis is often quite convincing. Yet we cannot even study this estimator in the DL setup because $G = 2$.

Donald and Lang criticize Card and Krueger (1994) for comparing mean wage changes of workers at a sample of fast-food restaurants across two states because Card and Krueger fail to account for the state effect (New Jersey or Pennsylvania), c_{ig}, in the composite error, v_{igm}. It is important to remember that the DL criticism of the standard difference-in-differences estimator has nothing to do with whether the increase in the minimum wage in New Jersey (in April 1992) was an exogenous event: DL's framework assumes that x_{ig}, which is an indicator for whether a fast-food restaurant is in New Jersey, is independent of the state effect, c_{ig}. Rather, DL's criticism only concerns inference. (Card and Krueger find a positive, not a negative, employment effect of increasing the minimum wage, so having a confidence interval seems to be less important in this particular case.)

To further study the $G = 2$ case with a binary policy indicator, write the difference in means as

$$\mu_2 - \mu_1 = (\delta_2 + \beta) - \delta_1 = (\alpha + c_2 + \beta) - (\alpha + c_1) = \beta + (c_2 - c_1). \tag{20.69}$$

Under the DL assumptions, $c_2 - c_1$ has mean zero, and so including it as part of the estimate, which is $\bar{y}_2 - \bar{y}_1$, does not result in bias. The authors work under the assumption that β is the parameter of interest, but, if the experiment is properly randomized—as is maintained by DL—it is harmless to include the c_g in the estimated effect, in which case the standard comparison-of-means methodology, using large M_g asymptotics, is appropriate.

Consider now a case where the DL approach to inference can be applied. Assume $G = 4$ with groups 1 and 2 the control groups ($x_1 = x_2 = 0$) and groups 3 and 4 the treatment groups ($x_3 = x_4 = 1$). The DL approach would involve computing the averages for each group, $\bar{y}_g$, and running the regression $\bar{y}_g$ on 1, x_g, $g = 1, \ldots, 4$. Inference is based on the t_2 distribution. The estimator $\hat{\beta}$ in this case can be written as

$$\hat{\beta} = (\bar{y}_3 + \bar{y}_4)/2 - (\bar{y}_1 + \bar{y}_2)/2. \tag{20.70}$$

(The pooled OLS regression using the disaggregated data results in the weighted average $(p_3\bar{y}_3 + p_4\bar{y}_4) - (p_1\bar{y}_1 + p_2\bar{y}_2)$, where $p_1 = M_1/(M_1 + M_2)$, $p_2 = M_2/(M_1 + M_2)$, $p_3 = M_3/(M_3 + M_4)$, and $p_4 = M_4/(M_3 + M_4)$ are the relative proportions within the control and treatment groups, respectively.) With $\hat{\beta}$ written as in equation (20.70), we are left to wonder why we need to use the t_2 distribution for, say, constructing a confidence interval. Each $\bar{y}_g$ is usually obtained from a large sample—$M_g = 30$ or so is usually sufficient for approximate normality of the stan-

dardized mean—and so $\hat{\beta}$, when properly standardized, has an approximate standard normal distribution quite generally.

In effect, the DL approach rejects the usual large-sample confidence interval based on group means because it may not be the case that $\mu_1 = \mu_2$ and $\mu_3 = \mu_4$. In other words, the control groups may be heterogeneous, as might be the treatment groups. This possibility in itself does not invalidate standard inference applied to equation (20.70). In fact, if we *define* the object of interest as

$$\tau = (\mu_3 + \mu_4)/2 - (\mu_1 + \mu_2)/2, \tag{20.71}$$

which is an average treatment effect of sorts, then $\hat{\beta}$ is consistent for β and (when properly scaled) asymptotically normal as the M_g get large.

The previous example suggests a different way to view the small G, large M_g setup. In this particular setup, we are estimating two parameters, α and β, given four moments that we can estimate with the data. The OLS estimates from $\bar{y}_g$ on 1, x_g, $g = 1, \ldots, G$, are minimum distance (MD) estimates that impose the restrictions $\mu_1 = \mu_2 = \alpha$ and $\mu_3 = \mu_4 = \alpha + \beta$. In particular, using the 4×4 identity matrix as the weight matrix, we get $\hat{\beta}$ as in equation (20.70) and $\hat{\alpha} = (\bar{y}_1 + \bar{y}_2)/2$. Using the MD approach, we see there are two overidentifying restrictions, which are easily tested. But even if we reject them, it simply implies that at least one pair of means within each of the control and treatment groups differs. If, say, we have four cities and random samples of workers from each city, $x_g = 1$ indicates a job-training program in two of the four cities, and y_{gm} is the change in labor market income, then it may simply be the case that the job-training program had differential effects across the two treatment cities, or that the mean change in labor market income differed across the two control cities, or both. Why should we reject the usual large M_g inference simply because the job-training program has heterogeneous effects?

With large group sizes, and whether or not G is especially large, we can put the general problem into an MD framework, as done, for example, by Loeb and Bound (1996), who had $G = 36$ cohort-division groups and many observations per group. For each group g, write

$$y_{gm} = \delta_g + \mathbf{z}_{gm}\gamma_g + u_{gm}, \qquad m = 1, \ldots, M_g, \tag{20.72}$$

where we assume random sampling within group and independent sampling across groups. We make the standard assumptions for OLS to be consistent (as $M_g \to \infty$) and $\sqrt{M_g}$-asymptotically normal, as in Chapter 4. The presence of group-level variables $\mathbf{x}_g$ in a "structural" model can be viewed as putting restrictions on the intercepts, δ_g, in the separate group models in equation (20.72). In particular,

$$\delta_g = \alpha + \mathbf{x}_g\boldsymbol{\beta}, \qquad g = 1, \ldots, G, \tag{20.73}$$

where we think of $\mathbf{x}_g$ as fixed, observed attributes of heterogeneous groups. With K attributes we must have $G \geq K + 1$ to determine α and $\boldsymbol{\beta}$. If M_g is large enough to estimate the δ_g precisely, a simple two-step estimation strategy suggests itself. First, obtain the $\hat{\delta}_g$, along with $\hat{\gamma}_g$, from an OLS regression within each group. If $G = K + 1$, then, typically, we can solve for $\hat{\theta} \equiv (\hat{\alpha}, \hat{\boldsymbol{\beta}}')'$ uniquely in terms of the $G \times 1$ vector $\hat{\delta}$: $\hat{\theta} = \mathbf{X}^{-1}\hat{\delta}$, where $\mathbf{X}$ is the $(K + 1) \times (K + 1)$ matrix with gth row $(1, \mathbf{x}_g)$. If $G > K + 1$, then, in a second step, we can use a minimum distance approach, as described in Section 14.6. If we use $\mathbf{I}_G$, the $G \times G$ identity matrix, as the weighting matrix, then the minimum distance estimator can be computed from the OLS regression

$$\hat{\delta}_g \quad \text{on} \quad 1, \mathbf{x}_g, \qquad g = 1, \ldots, G. \tag{20.74}$$

Under asymptotics such that $M_g = \rho_g M$ where $0 < \rho_g \leq 1$ and $M \to \infty$, the minimum distance estimator $\hat{\theta}$ is consistent and $\sqrt{M}$-asymptotically normal. Still, this particular MD estimator is asymptotically inefficient except under strong assumptions. Because the samples are assumed to be independent, it is not appreciably more difficult to obtain the efficient MD estimator, also called the "minimum chi-square" estimator.

First consider the case where $\mathbf{z}_{gm}$ does not appear in the first-stage estimation, so that the $\hat{\delta}_g$ is just $\bar{y}_g$, the sample mean for group g. Let $\hat{\sigma}_g^2$ denote the usual sample variance for group g. Because the $\bar{y}_g$ are independent across g, the efficient MD estimator uses a diagonal weighting matrix. As a computational device, the minimum chi-square estimator can be computed by using the weighted least squares (WLS) version of regression (20.74), where group g is weighted by $M_g/\hat{\sigma}_g^2$ (groups that have more data and smaller variance receive greater weight). Conveniently, the reported t statistics from the WLS regression are asymptotically standard normal as the group sizes M_g get large. (With fixed G, the WLS nature of the estimation is just a computational device; the standard asymptotic analysis of the WLS estimator has $G \to \infty$.) The minimum distance approach works with small G provided $G \geq K + 1$ and each M_g is large enough so that normality is a good approximation to the distribution of the (properly scaled) sample average within each group.

If $\mathbf{z}_{gm}$ is present in the first-stage estimation, we use as the minimum chi-square weights the inverses of the asymptotic variances for the g intercepts in the separate G regressions. With large M_g, we might make these fully robust to heteroskedasticity in $\mathrm{E}(u_{gm}^2 \mid \mathbf{z}_{gm})$ using the White (1980a) sandwich variance estimator. At a minimum we would want to allow different σ_g^2 even if we assume homoskedasticity within groups. Once we have the $\widehat{\mathrm{Avar}}(\hat{\delta}_g)$—which are just the squared reported standard errors for the $\hat{\delta}_g$—we use as weights $1/\widehat{\mathrm{Avar}}(\hat{\delta}_g)$ in the computationally simple WLS procedure. We are still using independence across g in obtaining a diagonal weighting matrix in the MD estimation.

An important by-product of the WLS regression is a minimum chi-square statistic that can be used to test the $G - K - 1$ overidentifying restrictions. The statistic is easily obtained as the weighted sum of squared residuals, say, SSR_w. Under the null hypothesis in equation (20.73), $SSR_w \overset{a}{\sim} \chi^2_{G-K-1}$ as the group sizes, M_g, get large. If we reject H_0 at a reasonably small significance level, the $\mathbf{x}_g$ are not sufficient for characterizing the changing intercepts across groups. If we fail to reject H_0, we can have some confidence in our specification and obtain confidence intervals for linear combinations of the population averages using the usual standard normal approximation.

We might also be interested in how one (or more) of the slopes in γ_g depends on the group features, $\mathbf{x}_g$. Then, we simple replace $\hat{\delta}_g$ with, say, $\hat{\gamma}_{g1}$, the slope on the first element of $\mathbf{z}_{gm}$. Naturally, we would use $1/\widehat{\text{Avar}}(\hat{\gamma}_{g1})$ as the weights in the MD estimation.

The minimum distance approach can also be applied if we impose $\gamma_g = \gamma$ for all g, as in the original model. Obtaining the $\hat{\delta}_g$ themselves is easy: run the pooled regression

$$y_{gm} \quad \text{on} \quad d1_g, d2_g, \ldots, dG_g, \mathbf{z}_{gm}, \qquad m = 1, \ldots, M_g; \; g = 1, \ldots, G, \tag{20.75}$$

where $d1_g, d2_g, \ldots, dG_g$ are group dummy variables. Using the $\hat{\delta}_g$ from the pooled regression (20.74) in MD estimation is complicated by the fact that the $\hat{\delta}_g$ are no longer asymptotically independent; in fact, $\hat{\delta}_g = \bar{y}_g - \bar{\mathbf{z}}_g\hat{\gamma}$, where $\hat{\gamma}$ is the vector of common slopes, and the presence of $\hat{\gamma}$ induces correlation among the intercept estimators. Let $\hat{\mathbf{V}}$ be the $G \times G$ estimated (asymptotic) variance matrix of the $G \times 1$ vector $\hat{\delta}$. Then the MD estimator is $\hat{\theta} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\hat{\delta}$, and its estimated asymptotic variance is $(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}$. If the OLS regression (20.74) is used, or even the WLS version, the resulting standard errors will be incorrect because they ignore the across-group correlation in the estimated intercepts.

Intermediate approaches are available, too. Loeb and Bound (1996) (hereafter, LB) allow different group intercepts and group-specific slopes on education, but impose common slopes on demographic and family background variables. The main group-level covariate is the student-teacher ratio. Thus LB are interested in seeing how the student-teacher ratio affects the relationship between test scores and education levels. They use both the unweighted estimator and the weighted estimator and find that the results differ in unimportant ways. Because they impose common slopes on a set of regressors, the estimated slopes on education (say $\hat{\gamma}_{g1}$) are not asymptotically independent, and perhaps using a nondiagonal estimated variance matrix $\hat{\mathbf{V}}$ (which would be 36×36 in this case) is more appropriate.

If we reject the overidentifying restrictions, we are essentially concluding that $\delta_g = \alpha + \mathbf{x}_g\boldsymbol{\beta} + c_g$, where c_g can be interpreted as the deviation from the restrictions

in equation (20.73) for group g. As G increases relative to K, the likelihood of rejecting the restrictions increases. One possibility is to apply the Donald and Lang approach, where the OLS regression (20.74) is analyzed in the context of the classical linear model (CLM) with inference based on the t_{G-K-1} distribution. Why is a CLM analysis justified? Since $\hat{\delta}_g = \delta_g + O_p(M_g^{-1/2})$, we can ingore the estimation error in $\hat{\delta}_g$ for large M_g. Then, it is as if we are estimating the equation $\delta_g = \alpha + \mathbf{x}_g\boldsymbol{\beta} + c_g$, $g = 1, \ldots, G$ by OLS. If the c_g are drawn from a normal distribution, classical analysis is applicable because c_g is assumed to be independent of $\mathbf{x}_g$. This approach is desirable when one cannot, or does not want to, find group-level observables that completely determine the δ_g. It is predicated on the assumption that the other factors in c_g are not systematically related to $\mathbf{x}_g$, a reasonable assumption if, say, $\mathbf{x}_g$ is a randomly assigned treatment at the group level, a case considered by Angrist and Lavy (2002).

Unlike in the linear case, for nonlinear models exact inference is unavailable even under the strongest set of assumptions. Nevertheless, if the group sizes M_g are reasonably large, we can extend the DL approach to nonlinear models and obtain approximate inference. In addition, the minimum distance approach carries over essentially without change.

We can apply the methods to any nonlinear model that has an index structure, which includes all the common ones, and many other models besides. Again, it is helpful to study the probit case in some detail. With small G and random sampling of $\{(y_{gm}, \mathbf{z}_{gm}): m = 1, \ldots, M_g\}$ within each g, write

$$\mathrm{P}(y_{gm} = 1 \mid \mathbf{z}_{gm}) = \Phi(\delta_g + \mathbf{z}_{gm}\gamma_g), \qquad m = 1, \ldots, M_g \tag{20.76}$$

$$\delta_g = \alpha + \mathbf{x}_g\boldsymbol{\beta}, \qquad g = 1, \ldots, G. \tag{20.77}$$

As with the linear model, we assume the intercept, δ_g in equation (20.76), is a function of the group features $\mathbf{x}_g$. With the M_g moderately large, we can get good estimates of the δ_g. The $\hat{\delta}_g$, $g = 1, \ldots, G$, are easily obtained by estimating a separate probit for each group. Or, we can impose common γ_g and just estimate different group intercepts (sometimes called "group fixed effects").

Under the restrictions in equation (20.77), we can apply the minimum distance approach just as before. Let $\widehat{\mathrm{Avar}}(\hat{\delta}_g)$ denote the estimated asymptotic variances of the $\hat{\delta}_g$ (so these shrink to zero at the rate $1/M_g$). If the $\hat{\delta}_g$ are obtained from G separate probits, they are independent, and the $\widehat{\mathrm{Avar}}(\hat{\delta}_g)$ are all we need. As in the linear case, if a pooled method is used, the $G \times G$ matrix $\widehat{\mathrm{Avar}}(\hat{\boldsymbol{\delta}})$ should be inverted and then used as the weighting matrix. For binary response, we use the usual MLE estimated variance. If we are using fractional probit for a fractional response, these

would be from a sandwich estimate of the asymptotic variance. In the case where the $\hat{\delta}_g$ are obtained from separate probits, we can obtain the minimum distance estimates as the WLS estimates from

$$\hat{\delta}_g \quad \text{on} \quad 1, \mathbf{x}_g, \qquad g = 1, \ldots, G$$

using weights $1/\widehat{\text{Avar}}(\hat{\delta}_g)$. This is the efficient minimum distance estimator and, conveniently, the proper asymptotic standard errors are reported from the WLS estimation (even though we are doing large M_g, not large G, asymptotics here). Generally, we can write the MD estimator as before: $\hat{\theta} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\hat{\boldsymbol{\delta}}$, where $\hat{\boldsymbol{\delta}}$ is the $G \times 1$ vector of $\hat{\delta}_g$ and $\hat{\mathbf{V}} = \widehat{\text{Avar}}(\hat{\boldsymbol{\delta}})$. The overidentification test is obtained exactly as in the linear case: there are $G - K - 1$ degrees of freedom in the chi-square distribution.

If we reject the overidentification restrictions, we can adapt Donald and Lang (2007) and treat

$$\hat{\delta}_g = \alpha + \mathbf{x}_g\boldsymbol{\beta} + error_g, \qquad g = 1, \ldots, G \tag{20.78}$$

as approximately satisfying the classical linear model assumptions, provided $G > K + 1$, just as before. As in the linear case, this approach is justified if $\delta_g = \alpha + \mathbf{x}_g\boldsymbol{\beta} + c_g$ with c_g independent of $\mathbf{x}_g$ and c_g drawn from a homoskedastic normal distribution. It assumes that we can ignore the estimation error in $\hat{\delta}_g$, based on $\hat{\delta}_g = \delta_g + O(1/\sqrt{M_g})$. Because the DL approach ignores the estimation error in $\hat{\delta}_g$, it is unchanged if one imposes some constant slopes across the groups.

Once we have estimated α and $\boldsymbol{\beta}$, the estimated effect on the response probability can be obtained by averaging the response probability for a given $\mathbf{x}$:

$$G^{-1}\sum_{g=1}^{G}\left(M_g^{-1}\sum_{m=1}^{M_g}\Phi(\hat{\alpha} + \mathbf{x}\hat{\boldsymbol{\beta}} + \mathbf{z}_{gm}\hat{\gamma}_g)\right), \tag{20.79}$$

where derivatives or differences with respect to the elements of $\mathbf{x}$ can be computed. Here, the minimum distance approach has an important advantage over the DL approach: the finite sample properties of estimator (20.79) are virtually impossible to obtain, whereas the large-M_g asymptotics underlying minimum distance would be straightforward using the delta method. The bootstrap should also be valid when the sampling scheme generates independent observations within each g.

With binary response problems, the two-step methods described here are problematical when the response does not vary within group. For example, suppose that x_g is a binary treatment—equal to one for receiving a voucher to attend college—and y_{gm} is an indicator of attending college. Each group is a high school class, say. If

some high schools have all students attend college, one cannot use probit (or logit) of y_{gm} on $\mathbf{z}_{gm}$, $m = 1, \ldots, M_g$. A linear regression returns zero-slope coefficients and intercept equal to unity. Of course, if randomization occurs at the group level—that is, x_g is independent of group attributes—then it is not necessary to control for the $\mathbf{z}_{gm}$. Instead, the within-group averages can be used in a simple minimum distance approach. In this case, as y_{gm} is binary, the DL approximation will not be valid, as the CLM assumptions will not even approximately hold in the model $\bar{y}_g = \alpha + \mathbf{x}_g\boldsymbol{\beta} + e_g$ (because $\bar{y}_g$ is always a fraction regardless of the size of M_g).

Naturally, there is nothing special about binary response models. It is possible to apply any nonlinear model using the invididual-specific data to obtain group-level estimates. Then, equation (20.78) can be applied.

20.4 Complex Survey Sampling

Often, survey data are characterized by clustering and variable probability sampling. For example, suppose that g represents the **primary sampling unit (PSU)** (say, city) and individuals or families (indexed by m) are **secondary sampling units**, sampled within each PSU with probability p_{gm}. Consider the problem of regression using such a data set. If $\hat{\boldsymbol{\beta}}$ is the IPW estimator pooled across PSUs and individuals, then its variance is estimated as

$$\left(\sum_{g=1}^{G}\sum_{m=1}^{M_g}\mathbf{x}_{gm}'\mathbf{x}_{gm}/p_{gm}\right)^{-1}\left[\sum_{g=1}^{G}\sum_{m=1}^{M_g}\sum_{r=1}^{M_g}\hat{u}_{gm}\hat{u}_{gr}\mathbf{x}_{gm}'\mathbf{x}_{gr}/(p_{gm}p_{gr})\right]\left(\sum_{g=1}^{G}\sum_{m=1}^{M_g}\mathbf{x}_{gm}'\mathbf{x}_{gm}/p_{gm}\right)^{-1}. \tag{20.80}$$

The middle of the sandwich accounts for cluster correlation along with unequal sampling probabilities. If the probabilities are estimated using retention frequencies, expression (20.80) is conservative, as we discussed in Section 20.2.2. A similar expression holds for general M-estimation. Typically, packages that support survey sampling require a variable defining the clusters along with a variable containing the inverse probability weights.

Multistage sampling schemes introduce more complications because standard stratification is often involved. Consider the following setup, closely related to Bhattacharya (2005). Let there be S strata (for example, states in the United States), exhaustive and mutually exclusive. Within stratum s, there are C_s clusters (for example, zip codes). In order to use large-sample approximations, we assume that in each stratum a large number of clusters is sampled. Typically, the sampling of clusters is

without replacement, but the resulting dependence across sampled clusters generated is more difficult to study. Instead, we assume sampling with replacement, which is harmless if the number of clusters sampled within each stratum, N_s, is "large." As before, we allow arbitrary correlation across units (say, households) within each cluster (say, zip code).

Within stratum s and cluster c, let there be M_{sc} total units (households or individuals). Therefore, the total number of units in the population is

$$M = \sum_{s=1}^{S} \sum_{c=1}^{C_s} M_{sc}. \tag{20.81}$$

It is convenient to start with the problem of estimating the mean of a variable that describes the population. Let z be a variable, such as family income, whose mean we want to estimate. List all population values as $\{z_{scm}^o: m = 1, \ldots, M_{sc}, c = 1, \ldots, C_s, s = 1, \ldots, S\}$, so the population mean can be written as

$$\mu = M^{-1} \sum_{s=1}^{S} \sum_{c=1}^{C_s} \sum_{m=1}^{M_{sc}} z_{scm}^o. \tag{20.82}$$

Define the total in the population as

$$\tau = \sum_{s=1}^{S} \sum_{c=1}^{C_s} \sum_{m=1}^{M_{sc}} z_{scm}^o = M\mu. \tag{20.83}$$

It is also useful to define the totals within each cluster and stratum, $\tau_{sc} = \sum_{m=1}^{M_{sc}} z_{scm}^o$ and $\tau_s = \sum_{c=1}^{C_s} \tau_{sc}$, respectively.

The specific sampling scheme is as follows: (1) for each stratum s, randomly draw N_s clusters, with replacement; (2) for each cluster c drawn in step (1), randomly sample K_{sc} households with replacement. For each pair (s, c), define the sample average

$$\hat{\mu}_{sc} = K_{sc}^{-1} \sum_{m=1}^{K_{sc}} z_{scm}. \tag{20.84}$$

Because this is an average based on a random sample within (s, c),

$$\mathrm{E}(\hat{\mu}_{sc}) = \mu_{sc} = M_{sc}^{-1} \sum_{m=1}^{M_{sc}} z_{scm}^o. \tag{20.85}$$

To continue up to the cluster level we need the total, $\tau_{sc} = M_{sc}\mu_{sc}$, for which an unbiased estimator is $\hat{\tau}_{sc} = M_{sc}\hat{\mu}_{sc}$ for all $\{(s,c)\colon c = 1,\ldots,C_s,\ s = 1,\ldots,S\}$ (even if we eventually do not use some clusters because they are not sampled). Now, for each stratum s, the estimator $N_s^{-1}\sum_{c=1}^{N_s}\hat{\tau}_{sc}$, which is the average of the cluster totals within stratum s, has expected value which is the population average (for stratum s), that is, $C_s^{-1}\sum_{c=1}^{C_s}\tau_{sc} = C_s^{-1}\sum_{c=1}^{C_s}\sum_{m=1}^{M_{sc}} z_{scm}^{o} = C_s^{-1}\tau_s$. (In general, $C_s^{-1}\tau_s \neq \mu_s = (\sum_{c=1}^{C_s} M_{sc})^{-1}\tau_s$ unless each cluster has only one observation.) It follows that an unbiased estimator of the total τ_s for stratum s is

$$C_s \cdot N_s^{-1}\sum_{c=1}^{N_s}\hat{\tau}_{sc}. \tag{20.86}$$

Finally, the total in the entire population is estimated as

$$\sum_{s=1}^{S}\left(C_s \cdot N_s^{-1}\sum_{c=1}^{N_s}\hat{\tau}_{sc}\right) = \sum_{s=1}^{S}(C_s/N_s)\sum_{c=1}^{N_s}(M_{sc}/K_{sc})\sum_{m=1}^{K_{sc}} z_{scm}$$

$$= \sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}}\left(\frac{C_s}{N_s}\cdot\frac{M_{sc}}{K_{sc}}\right)z_{scm} \equiv \sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}}\omega_{sc}z_{scm}, \tag{20.87}$$

where

$$\omega_{sc} \equiv \frac{C_s}{N_s}\cdot\frac{M_{sc}}{K_{sc}} \tag{20.88}$$

is the weight for every unit sampled in stratum-cluster pair (s,c). This weight accounts for undersampled or oversampled clusters within strata and undersampled or oversampled units within clusters. Expressions (20.87) and (20.88) appear in the literature on complex survey sampling, sometimes without M_{sc}/K_{sc} when each cluster is sampled as a complete unit, and so $M_{sc}/K_{sc} = 1$. To estimate the population mean, μ, we just divide by M, the total number of units in the population,

$$\hat{\mu} = M^{-1}\left(\sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}}\omega_{sc}z_{scm}\right). \tag{20.89}$$

In fact, we do not need to know the population size, M, to obtain an unbiased estimator of μ. We can obtain an alternative estimator that uses a modified set of weights. It falls out naturally from a regression framework, to which we now turn.

To study the asymptotic properties of regression (and many other estimation methods), it is convenient to modify the weights so that they are constant, or con-

verge to a constant. The weights ω_{sc} in expression (20.88) converge to zero at rate N_s^{-1} because C_s and M_{sc} are fixed and K_{sc} is treated as fixed. (We assume a relatively small number of households sampled per cluster.) Let $N = N_1 + N_2 + \cdots + N_S$ be the total number of clusters sampled and define

$$\upsilon_{sc} = \frac{C_s}{(N_s/N)} \cdot \frac{M_{sc}}{K_{sc}} = N\omega_{sc}. \tag{20.90}$$

As in Bhattacharya (2005), it is easiest just to assume $N_s = a_s N$ for a_s fixed, $0 < a_s < 1$, $a_1 + \cdots + a_S = 1$. But we can also just assume N_s/N converges to a_s with the same property. Therefore, by writing $\upsilon_{sc} = (C_s/a_s)(M_s/K_s)$, we see that υ_{sc} is constant. Further, any optimization problem that uses ω_{sc} as weights gives the same answer when υ_{sc} is used because the scale factor in equation (20.90) does not depend on s or c. The key in the formulas for the asymptotic variance below is that υ_{sc} is (roughly) constant.

While equation (20.90) is the most natural definition of the weights for obtaining the limiting distribution results, we can use different formulations without changing the end formulas. For example, let $C = C_1 + \cdots + C_S$ be the total number of clusters in the population, let M be the total number of units in the population, and let K be the total units samples. Then, for the final formulas, we could use the weights defined as

$$\upsilon_{sc} = \frac{(C_s/C)}{(N_s/N)} \cdot \frac{(M_{sc}/M)}{(K_{sc}/K)} = \frac{(NK)}{(CM)}\omega_{sc}. \tag{20.91}$$

Because C, M, and K are fixed, the factor $K/(CM)$ has no effect on estimation or inference. Equation (20.91) has a nice interpretation because it is expressed in terms of frequencies of the population relative to the sample frequencies. For example, if $(C_s/C) > (N_s/N)$, which means that stratum s is underrepresented in terms of number of clusters, equation (20.91) gives more weight to such strata. The same is true of the fractions involving the number of units (say, households).

While we can consider general M-estimation problems, or generalized method of moments as in Bhattacharya (2005), we consider least squares for concreteness. The weighted minimization problem is

$$\min_{\beta} N^{-1} \sum_{s=1}^{S} \sum_{c=1}^{N_s} \sum_{m=1}^{K_{sc}} \upsilon_{sc}(y_{scm} - \mathbf{x}_{scm}\beta)^2, \tag{20.92}$$

where it is helpful to divide by N to facilitate the asymptotic analysis as $N \to \infty$. The first-order condition is

$$N^{-1}\sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}} v_{sc}\mathbf{x}'_{scm}(y_{scm}-\mathbf{x}_{scm}\hat{\boldsymbol{\beta}})=\mathbf{0}. \tag{20.9}$$

Using arguments similar to the SS sampling case, but accounting for the clu tering (by, in effect, treating each cluster as its own observation), we can sho that an appropriate estimator of $\text{Avar}(\hat{\boldsymbol{\beta}})$—in the sense that it is consistent f $\text{Avar}[\sqrt{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})]$ when multiplied by N—is

$$\left(\sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}} v_{sc}\mathbf{x}'_{scm}\mathbf{x}_{scm}\right)^{-1}\hat{\mathbf{B}}\left(\sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}} v_{sc}\mathbf{x}'_{scm}\mathbf{x}_{scm}\right)^{-1}, \tag{20.94}$$

where $\hat{\mathbf{B}}$ is somewhat complicated:

$$\hat{\mathbf{B}}=\sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}} v_{sc}^2\hat{u}_{scm}^2\mathbf{x}'_{scm}\mathbf{x}_{scm}+\sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}}\sum_{r\neq m}^{K_{sc}} v_{sc}^2\hat{u}_{scm}\hat{u}_{scr}\mathbf{x}'_{scm}\mathbf{x}_{scr}$$
$$-\sum_{s=1}^{S}N_s^{-1}\left(\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}} v_{sc}\mathbf{x}'_{scm}\hat{u}_{scm}\right)\left(\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}} v_{sc}\mathbf{x}'_{scm}\hat{u}_{scm}\right)'. \tag{20.95}$$

The first part of $\hat{\mathbf{B}}$ is obtained using the White "heteroskedasticity"-robust form. Th second piece accounts for the correlation within clusters; this is typically a positiv definite matrix, and it generally increases the asymptotic standard errors. The thir piece actually reduces the variance by accounting for the nonzero means of th "score" within strata, just as in the SS sampling case.

If each cluster has just one unit, so $M_{sc}=K_{sc}=1$, then expression (20.94) reduce to

$$\left(\sum_{s=1}^{S}\sum_{c=1}^{N_s} v_{sc}\mathbf{x}'_{sc}\mathbf{x}_{sc}\right)^{-1}\left[\left(\sum_{s=1}^{S}\sum_{c=1}^{N_s} v_{sc}^2\hat{u}_{sc}^2\mathbf{x}'_{sc}\mathbf{x}_{sc}\right)\right.$$
$$\left.-\sum_{s=1}^{S}N_s^{-1}\left(\sum_{c=1}^{N_s} v_{sc}\mathbf{x}'_{sc}\hat{u}_{sc}\right)\left(\sum_{c=1}^{N_s} v_{sc}\mathbf{x}'_{sc}\hat{u}_{sc}\right)'\right]\cdot\left(\sum_{s=1}^{S}\sum_{c=1}^{N_s} v_{sc}\mathbf{x}'_{sc}\mathbf{x}_{sc}\right)^{-1}, \tag{20.96}$$

which is the formuala for standard stratified sampling with a finite number of units in each stratum.

For general M-estimation, the outer sandwich in (20.94) is replaced with the inverse of the weighted Hessian, $[\sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}} v_{sc}\mathbf{H}(\mathbf{w}_{scm},\hat{\boldsymbol{\theta}})]^{-1}$, while $\mathbf{x}'_{scm}\hat{u}_{scm}$ in equation (20.95) is replaced with the score, $\mathbf{s}(\mathbf{w}_{scm},\hat{\boldsymbol{\theta}})$. Some econometrics packages

have made implementation fairly straightforward for a variety of linear and nonlinear models. To obtain the correct asymptotic variance estimator—one that is neither too optimistic nor too conservative—one needs to specify the strata, the clusters, and the sampling weights.

Problems

20.1. Use expressions (20.4) and (20.5) to answer this question.

a. Derive the estimator in equation (20.5) from the minimization problem in expression (20.4).

b. Show directly that the estimator $\tilde{\mu}_w = N^{-1} \sum_{i=1}^{N} (s_i/p_{ji}) w_i$ is unbiased for μ.

c. What practical advantage does $\hat{\mu}_w$ have over $\tilde{\mu}_w$?

20.2. Use the log likelihood in equation (20.9) to derive $\hat{p}_j = M_j/N_j$, $j = 1, \ldots, J$, where M_j is the number of retained observations from stratum j and N_j is the number of times stratum j was drawn.

20.3. Let y be a scalar response variable and $\mathbf{x}$ a vector of explanatory variables, and let $m(\mathbf{x}, \boldsymbol{\theta})$ denote a model for $\mathrm{E}(y \mid \mathbf{x})$. The parameter space is $\boldsymbol{\Theta}$.

a. Let $\hat{\boldsymbol{\theta}}_w$ be the IPW nonlinear least squares estimator. Write down the minimization problem solved by $\hat{\boldsymbol{\theta}}_w$.

b. Assume that the model is correctly specified for $\mathrm{E}(y \mid \mathbf{x})$, and let $\boldsymbol{\theta}_o$ denote the population value; assume that $\boldsymbol{\theta}_o$ is identified in the population. Provide a set of sufficient conditions for consistency of $\hat{\boldsymbol{\theta}}_w$ for $\boldsymbol{\theta}_o$. (Hint: See Theorem 12.2.)

c. Assuming that $m(\mathbf{x}, \cdot)$ is twice continuously differentiable on the interior of $\boldsymbol{\Theta}$ and that $\boldsymbol{\theta}_o \in \mathrm{int}(\boldsymbol{\Theta})$, propose an estimator of the asymptotic variance of $\hat{\boldsymbol{\theta}}_w$ that depends only on the gradient of $m(\mathbf{x}, \cdot)$—not its Hessian.

d. If you add the homoskedasticity assumption $\mathrm{Var}(y \mid \mathbf{x}) = \sigma_o^2$, does the formula from part c simplify?

e. If $m(\mathbf{x}, \boldsymbol{\theta})$ is misspecified, how should you adjust the estimator in part c?

20.4. Consider the problem of standard stratified sampling. Assume that the sample shares, H_j, converge to $\bar{H}_j > 0$ as $N \to \infty$, $j = 1, \ldots, J$. Further, suppose that $\boldsymbol{\theta}_o$ minimizes $\mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta}) \mid \mathbf{x}]$ over $\boldsymbol{\Theta}$ for each $\mathbf{x}$ and that $\boldsymbol{\theta}_o$ uniquely minimizes $\mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta})]$ over $\boldsymbol{\Theta}$. Argue that the unweighted estimator is consistent for $\boldsymbol{\theta}_o$. (Hint: Write the unweighted objective function as

$$\sum_{j=1}^{J} H_j \left[N_j^{-1} \sum_{i=1}^{N_j} q(\mathbf{w}_{ij}, \boldsymbol{\theta}) \right]$$

and argue that this function converges uniformly to

$$\bar{H}_1 \mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta}) \mid \mathbf{x} \in \mathcal{X}_1] + \bar{H}_2 \mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta}) \mid \mathbf{x} \in \mathcal{X}_2] + \cdots + \bar{H}_J \mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta}) \mid \mathbf{x} \in \mathcal{X}_J],$$

where the strata are $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_J$. Then show that $\boldsymbol{\theta}_o$ uniquely minimizes this expression by arguing that it uniquely minimizes $\mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta}) \mid \mathbf{x} \in \mathcal{X}_j]$ for at least one j.)

20.5. Use the data in BENEFITS.RAW to answer this question.

a. To equation (20.33) add the within-district averages of *bs*, *lstaff*, *lenroll*, and *lunch*, where *lstaff* and *lenroll* denote logarithms. Estimate this equation by pooled OLS. How do the coefficients on *bs*, *lstaff*, *lenroll*, and *lunch* compare with the FE coefficients? Are the usual pooled OLS standard errors valid here?

b. Estimate the equation from part a by random effects. (That is, include the district averages along with the original variables.) How do these estimates compare with the FE estimates? How do the cluster-robust standard errors compare with the cluster-robust standard errors for FE?

c. Use the estimation in part b to obtain the value of the fully robust Wald statistic testing the RE assumption that the district effect is uncorrelated with the four district averages.

20.6. Use the data in BENEFITS.RAW to answer this question.

a. How many schools in the sample have a benefits-salary ratio of at least 0.5?

b. Estimate equation (20.33) by fixed effects omitting the observations from part a. Discuss how the estimate of β_{bs} changes, as well as its cluster-robust standard error.

c. Now add the within-district averages of all four variables and estimate the equation by least absolute deviations, using all the observations. How strong is the evidence for a trade-off using LAD?

20.7. Use the data in MEAP94_98 to answer this question.

a. How many schools have all five years of data? Are there any schools with only one year?

b. Obtain the within-school time averages of the variables *lavgrexp*, *lunch*, *lenrol*, and the four-year dummies *y95* through *y98*. Include these in a pooled OLS regression that includes the other variables in Table 20.2 (including the year dummies themselves). Verify that the coefficients on the original variables are the FE estimates.

What is the coefficient on $\overline{lunch}$ (the time average)? Is it statistically different from zero using a cluster-robust standard error at the district level?

c. Now use RE rather than pooled OLS on the equation in part b. Again verify that you obtain the FE estimates on the original variables. Is the RE coefficient on $\overline{lunch}$ identical to the POLS coefficient? Is it still statistically significant?

d. Redo part c, but do not include the time averages of the year dummies. Do you still get the FE estimates on *lavgrexp*, *lunch*, and *lenrol*? Why, with an unbalanced panel, must we include the time averages of year dummies for RE to equal FE, whereas we did not have to in the balanced case?

e. Now go back to the original FE estimation in Table 20.2, but drop the year dummies. How does the estimated spending effect change from Table 20.2? Which estimate is more reliable?

f. Return to the equation implicit in Table 20.2, but estimate the equation by pooled OLS and RE. (That is, do not include the time averages of the variables.) How do the estimates of the spending effect compare with the FE estimates? How come the *lunch* variable is much more important in the POLS and RE estimation?

g. Considering the various estimates and standard errors in Table 20.2 and obtained for this problem, which estimate of the spending variable and which standard error seem most reliable?

20.8. In the setting of Section 20.3.1, let y_{gm} be a fractional response variable, and consider the model

$$\mathrm{E}(y_{gm} \mid \mathbf{x}_g, \mathbf{Z}_g, c_g) = \Phi(\alpha + \mathbf{x}_g\boldsymbol{\beta} + \mathbf{z}_{gm}\boldsymbol{\gamma} + c_g).$$

a. Assume that $c_g = \eta_g + \bar{\mathbf{z}}_g\boldsymbol{\xi}_g + a_g$. Find $\mathrm{E}(y_{gm} \mid \mathbf{x}_g, \mathbf{Z}_g, a_g)$.

b. Add the assumption $a_g \mid \mathbf{x}_g, \mathbf{Z}_g \sim \text{Normal}(0, \tau_g^2)$ and find $\mathrm{E}(y_{gm} \mid \mathbf{x}_g, \mathbf{Z}_g)$. (Hint: It should have the probit form.)

c. Suppose that η_g, $\boldsymbol{\xi}_g$, and τ_g^2 depend only on the group size, M_g. Suggest a method for estimating all the parameters.

d. How would you perform inference on the parameters estimated in part c?