11
MODELS FOR PANEL DATA

11.1 INTRODUCTION

Data sets that combine time series and cross sections are common in economics. The published statistics of the OECD contain numerous series of economic aggregates observed yearly for many countries. The Penn World Tables [CIC (2010)] is a data bank that contains national income data on 167 countries for more than 60 years. Recently constructed **longitudinal data sets** contain observations on thousands of individuals or families, each observed at several points in time. Other empirical studies have examined time-series data on sets of firms, states, countries, or industries simultaneously. These data sets provide rich sources of information about the economy. The analysis of panel data allows the model builder to learn about economic processes while accounting for both heterogeneity across individuals, firms, countries, and so on and for dynamic effects that are not visible in cross sections. Modeling in this context often calls for complex stochastic specifications. In this chapter, we will survey the most commonly used techniques for time-series—cross-section (e.g., cross-country) and panel (e.g., longitudinal)—data. The methods considered here provide extensions to most of the models we have examined in the preceding chapters. Section 11.2 describes the specific features of panel data. Most of this analysis is focused on individual data, rather than cross-country aggregates. We will examine some aspects of aggregate data modeling in Section 11.10. Sections 11.3, 11.4, and 11.5 consider in turn the three main approaches to regression analysis with panel data, pooled regression, the fixed effects model, and the random effects model. Section 11.6 considers robust estimation of covariance matrices for the panel data estimators, including a general treatment of cluster effects. Sections 11.7 through 11.10 examine some specific applications and extensions of panel data methods. Spatial autocorrelation is discussed in Section 11.7. In Section 11.8, we consider sources of endogeneity in the random effects model, including a model of the sort considered in Chapter 8 with an endogenous right-hand-side variable and then two approaches to dynamic models. Section 11.9 builds the fixed and random effects models into nonlinear regression models. Finally, Section 11.10 examines random parameter models. The random parameters approach is an extension of the fixed and random effects model in which the heterogeneity that the FE and RE models build into the constant terms is extended to other parameters as well.

Panel data methods are used throughout the remainder of this book. We will develop several extensions of the fixed and random effects models in Chapter 14 on maximum likelihood methods, and in Chapter 15 where we will continue the development of random parameter models that is begun in Section 11.10. Chapter 14 will also present methods for handling discrete distributions of random parameters under the heading of latent class models. In Chapter 21, we will return to the models of nonstationary panel data that are suggested in Section 11.8.4. The fixed and random effects approaches will be used throughout the applications of discrete and limited dependent variables models in microeconometrics in Chapters 17, 18, and 19.

11.2 PANEL DATA MODELING

Many recent studies have analyzed panel, or longitudinal, data sets. Two very famous ones are the *National Longitudinal Survey of Labor Market Experience* (NLS, www.bls.gov/nls/nlsdoc.htm) and the *Michigan Panel Study of Income Dynamics* (PSID, http://psidonline.isr.umich.edu/). In these data sets, very large cross sections, consisting of thousands of microunits, are followed through time, but the number of periods is often quite small. The PSID, for example, is a study of roughly 6,000 families and 15,000 individuals who have been interviewed periodically from 1968 to the present. In contrast, the *European Community Household Panel* (ECHP, http://ec.europa.eu/eurostat/web/microdata/european-community-household-panel) ran for a total of eight years (waves). An ongoing study in the United Kingdom is the *Understanding Society* survey (www.understandingsociety.ac.uk/about) that grew out of the *British Household Panel Survey* (BHPS). This survey that was begun in 1991 with about 5,000 households has expanded to over 40,000 participants. Many very rich data sets have recently been developed in the area of health care and health economics, including the *German Socioeconomic Panel* (GSOEP, www.eui.eu/Research/Library/ResearchGuides/Economics/Statistics/DataPortal/GSOEP.aspx), AHRQ's *Medical Expenditure Panel Survey* (MEPS, www.meps.ahrq.gov/), and the *Household Income and Labour Dynamics in Australia* (HILDA, www.melbourneinstitute.com/hilda/). Constructing long, evenly spaced time series in contexts such as these would be prohibitively expensive, but for the purposes for which these data are typically used, it is unnecessary. Time effects are often viewed as transitions or discrete changes of state. The Current Population Survey (CPS, www.census.gov/cps/), for example, is a monthly survey of about 50,000 households that interviews households monthly for four months, waits for eight months, then reinterviews. This two-wave, **rotating panel** format allows analysis of short-term changes as well as a more general analysis of the U.S. national labor market. They are typically modeled as specific to the period in which they occur and are not carried across periods within a cross-sectional unit.[1] Panel data sets are more oriented toward cross-section analyses; they are wide but typically short. Heterogeneity across units is an integral part—indeed, often the central focus—of the analysis. [See, e.g., Jones and Schurer (2011).]

The analysis of panel or longitudinal data is the subject of one of the most active and innovative bodies of literature in econometrics,[2] partly because panel data provide such a rich environment for the development of estimation techniques and theoretical results. In more practical terms, however, researchers have been able to use time-series cross-sectional data to examine issues that could not be studied in either cross-sectional

[1] Formal time-series modeling for panel data is briefly examined in Section 21.4.

[2] A compendium of the earliest literature is Maddala (1993). Book-length surveys on the econometrics of panel data include Hsiao (2003), Dielman (1989), Matyas and Sevestre (1996), Raj and Baltagi (1992), Nerlove (2002), Arellano (2003), and Baltagi (2001, 2013, 2015). There are also lengthy surveys devoted to specific topics, such as limited dependent variable models [Hsiao, Lahiri, Lee, and Pesaran (1999)], discrete choice models [Greene (2015)] and semiparametric methods [Lee (1998)].

or time-series settings alone. Recent applications have allowed researchers to study the impact of health policy changes[3] and, more generally, the dynamics of labor market behavior. In principle, the methods of Chapters 6 and 21 can be applied to longitudinal data sets. In the typical panel, however, there are a large number of cross-sectional units and only a few periods. Thus, the time-series methods discussed there may be somewhat problematic. Recent work has generally concentrated on models better suited to these short and wide data sets. The techniques are focused on cross-sectional variation, or heterogeneity. In this chapter, we shall examine in detail the most widely used models and look briefly at some extensions.

11.2.1 GENERAL MODELING FRAMEWORK FOR ANALYZING PANEL DATA

The fundamental advantage of a panel data set over a cross section is that it will allow the researcher great flexibility in modeling differences in behavior across individuals. The basic framework for this discussion is a regression model of the form

$$\begin{aligned} y_{it} &= \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\alpha} + \varepsilon_{it} \\ &= \mathbf{x}_{it}'\boldsymbol{\beta} + c_i + \varepsilon_{it}. \end{aligned} \tag{11-1}$$

There are K regressors in $\mathbf{x}_{it}$, not including a constant term. The **heterogeneity**, or **individual effect**, is $\mathbf{z}_i'\boldsymbol{\alpha}$ where $\mathbf{z}_i$ contains a constant term and a set of individual or group-specific variables, which may be observed, such as race, sex, location, and so on; or unobserved, such as family specific characteristics, individual heterogeneity in skill or preferences, and so on, all of which are taken to be constant over time t. As it stands, this model is a classical regression model. If $\mathbf{z}_i$ is observed for all individuals, then the entire model can be treated as an ordinary linear model and fit by least squares. The complications arise when c_i is unobserved, which will be the case in most applications. Consider, for example, analyses of the effect of education and experience on earnings from which "ability" will always be a missing and unobservable variable. In health care studies, for example, of usage of the health care system, "health" and "health care" will be unobservable factors in the analysis.

The main objective of the analysis will be consistent and efficient estimation of the **partial effects**,

$$\boldsymbol{\beta} = \partial E[y_{it}|\mathbf{x}_{it}]/\partial \mathbf{x}_{it}.$$

Whether this is possible depends on the assumptions about the unobserved effects. We begin with a **strict exogeneity** assumption for the independent variables,

$$E[\varepsilon_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, c_i] = E[\varepsilon_{it}|\mathbf{X}_i, c_i] = 0.$$

This implies the current disturbance is uncorrelated with the independent variables in every period, past, present, and future. A looser assumption of contemporaneous exogeneity is sometimes useful. If

$$E[y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, c_i] = E[y_{it}|\mathbf{x}_{it}, c_i] = \mathbf{x}_{it}'\boldsymbol{\beta} + c_i,$$

then

$$E[\varepsilon_{it}|\mathbf{x}_{it}, c_i] = 0.$$

[3] For example, Riphahn et al.'s (2003) analysis of reforms in German public health insurance regulations.

The regression model with this assumption restricts influences of $\mathbf{x}$ on $E[y|\mathbf{x}, c]$ to the current period. In this form, we can see that we have ruled out dynamic models such as

$$y_{it} = \mathbf{w}_{it}'\boldsymbol{\beta} + \gamma y_{i,t-1} + c_i + \varepsilon_{it}$$

because as long as γ is nonzero, covariation between ε_{it} and $\mathbf{x}_{it} = (\mathbf{w}_{it}, y_{i,t-1})$ is transmitted through c_i in $y_{i,t-1}$. We will return to dynamic specifications in Section 11.8.3. In some settings (such as the static fixed effects model in Section 11.4), strict exogeneity is stronger than necessary. It is, however, a natural assumption. It will prove convenient to start there, and loosen the assumption in specific cases where it would be useful.

The crucial aspect of the model concerns the heterogeneity. A convenient assumption is mean independence,

$$E[c_i|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots] = \alpha.$$

If the unobserved variable(s) are uncorrelated with the included variables, then, as we shall see, they may be included in the disturbance of the model. This is the assumption that underlies the random effects model, as we will explore later. It is, however, a particularly strong assumption—it would be unlikely in the labor market and health care examples mentioned previously. The alternative would be

$$E[c_i|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots,] = h(\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots) = h(\mathbf{X}_i)$$

for some unspecified, but nonconstant function of $\mathbf{X}_i$. This formulation is more general, but at the same time, considerably more complicated, the more so because estimation may require yet further assumptions about the nature of the regression function.

11.2.2 MODEL STRUCTURES

We will examine a variety of different models for panel data. Broadly, they can be arranged as follows:

1. **Pooled Regression:** If $\mathbf{z}_i$ contains only a constant term, then ordinary least squares provides consistent and efficient estimates of the common α and the slope vector $\boldsymbol{\beta}$.
2. **Fixed Effects:** If $\mathbf{z}_i$ is unobserved, but correlated with $\mathbf{x}_{it}$, then the least squares estimator of $\boldsymbol{\beta}$ is biased and inconsistent as a consequence of an omitted variable. However, in this instance, the model

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \varepsilon_{it},$$

where $\alpha_i = \mathbf{z}_i'\boldsymbol{\alpha}$, embodies all the observable effects and specifies an estimable conditional mean. This **fixed effects** approach takes α_i to be a group-specific constant term in the regression model. It should be noted that the term "fixed" as used here signifies the correlation of c_i and $\mathbf{x}_{it}$, not that c_i is nonstochastic.
3. **Random Effects:** If the unobserved individual heterogeneity, however formulated, is uncorrelated with $\mathbf{x}_{it}$, then the model may be formulated as

$$\begin{aligned} y_{it} &= \mathbf{x}_{it}'\boldsymbol{\beta} + E[\mathbf{z}_i'\boldsymbol{\alpha}] + \{\mathbf{z}_i'\boldsymbol{\alpha} - E[\mathbf{z}_i'\boldsymbol{\alpha}]\} + \varepsilon_{it} \\ &= \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha + u_i + \varepsilon_{it}, \end{aligned}$$

that is, as a linear regression model with a compound disturbance that may be consistently, albeit inefficiently, estimated by least squares. This random effects approach specifies that u_i is a group-specific random element, similar to ε_{it} except that for each group, there is but a single draw that enters the regression identically in each period. Again, the crucial distinction between fixed and random effects is whether the unobserved individual effect embodies elements that are correlated with the regressors in the model, not whether these effects are stochastic or not. We will examine this basic formulation, then consider an extension to a dynamic model.

4. **Random Parameters:** The random effects model can be viewed as a regression model with a random constant term. With a sufficiently rich data set, we may extend this idea to a model in which the other coefficients vary randomly across individuals as well. The extension of the model might appear as

$$y_{it} = \mathbf{x}_{it}'(\boldsymbol{\beta} + \mathbf{u}_i) + (\alpha + u_i) + \varepsilon_{it},$$

where $\mathbf{u}_i$ is a random vector that induces the variation of the parameters across individuals. This random parameters model has recently enjoyed widespread attention in several fields. It represents a natural extension in which researchers broaden the amount of heterogeneity across individuals while retaining some commonalities—the parameter vectors still share a common mean. Some recent applications have extended this yet another step by allowing the mean value of the parameter distribution to be person specific, as in

$$y_{it} = \mathbf{x}_{it}'(\boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \mathbf{u}_i) + (\alpha + u_i) + \varepsilon_{it},$$

where $\mathbf{z}_i$ is a set of observable, person-specific variables, and $\boldsymbol{\Delta}$ is a matrix of parameters to be estimated. As we will examine in Chapter 17, this **hierarchical model** is extremely versatile.

11.2.3 EXTENSIONS

The short list of model types provided earlier only begins to suggest the variety of applications of panel data methods in econometrics. We will begin in this chapter to study some of the formulations and uses of linear models. The random and fixed effects models and random parameters models have also been widely used in models of censoring, binary, and other discrete choices, and models for event counts. We will examine all of these in the chapters to follow. In some cases, such as the models for count data in Chapter 18, the extension of random and fixed effects models is straightforward, if somewhat more complicated computationally. In others, such as in binary choice models in Chapter 17 and censoring models in Chapter 19, these panel data models have been used, but not before overcoming some significant methodological and computational obstacles.

11.2.4 BALANCED AND UNBALANCED PANELS

By way of preface to the analysis to follow, we note an important aspect of panel data analysis. As suggested by the preceding discussion, a panel data set will consist of n sets of observations on individuals to be denoted $i = 1, \ldots, n$. If each individual in the data set is observed the same number of times, usually denoted T, the data set is a **balanced panel**. An **unbalanced panel** data set is one in which individuals may be observed different numbers of times. We will denote this T_i. A **fixed panel** is one in which the same set of individuals is observed for the duration of the study. The data sets we will examine in this chapter, while not all balanced, are fixed.

A rotating panel is one in which the cast of individuals changes from one period to the next. For example, Gonzalez and Maloney (1999) examined self-employment decisions in Mexico using the National Urban Employment Survey. This is a quarterly data set drawn from 1987 to 1993 in which individuals are interviewed five times. Each quarter, one-fifth of the individuals is rotated out of the data set. The U.S. Census Bureau's SIPP data (Survey of Income and Program Participation, www.census.gov/programs-surveys/sipp/data.html) is another rotating panel. Some discussion and numerous references may be found in Baltagi (2013)

Example 11.1 A Rotating Panel: The Survey of Income and Program Participation (SIPP) Data

From the Census Bureau's home site for this data set:

The SIPP survey design is a continuous series of national panels, with sample size ranging from approximately 14,000 to 52,000 interviewed households. The duration of each panel ranges from $2\frac{1}{2}$ years to 4 years. The SIPP sample is a multistage-stratified sample of the U.S. civilian non-institutionalized population. From 1984 to 1993, a new panel of households was introduced each year in February. A 4-year panel was implemented in April 1996; however, a 3-year panel that was started in February 2000 was canceled after 8 months due to budget restrictions. Consequently, a 3-year panel was introduced in February 2001. The $2\frac{1}{2}$ year 2004 SIPP Panel was started in February 2004 and was the first SIPP panel to use the 2000 decennial-based redesign of the sample. The 2014 panel, starting in February 2014, is the first SIPP panel to use the 2010 decennial as the basis for its sample.

11.2.5 ATTRITION AND UNBALANCED PANELS

Unbalanced panels arise in part because of nonrandom attrition from the sample. Individuals may appear for only a subset of the waves. In general, if the attrition is systematically related to the outcome variable in the model being studied, then it may induce conditions of nonrandom sampling bias—sometimes called *sample selection*. The nature of the bias is unclear, but sample selection bias as a general aspect of econometric analysis is well documented. [An example would be attrition of subjects from a medical clinical trial for reasons related to the efficacy (or lack of) of the drug under study.] Verbeek and Nijman (1992) proposed a nonconstructive test for attrition in panel data models—the test results detect the condition but do not imply a strategy if the hypothesis of no nonrandom attrition is rejected. Wooldridge (2002 and 2010, pp. 837–844) describes an *inverse probability weighting* (IPW) approach for correcting for nonrandom attrition.

Example 11.2 Attrition and Inverse Probability Weighting in a Model for Health

Contoyannis, Jones, and Rice (2004) employed an ordered probit model to study self-assessed health in the first eight waves of the BHPS.[4] The sample exhibited some attrition as shown in Table 11.1 (from their Table V). (Although the sample size does decline after each wave, the remainder at each wave is not necessarily a subset of the previous wave. Some individuals returned to the sample. A subsample of observations for which attrition at each wave was an *absorbing state*—they did not return—was analyzed separately. This group is used for IPW-2 in the results below.) To examine the issue of nonrandom attrition, the authors first employed Nijman and Verbeek's tests. This entails adding three variables to the model:

[4] See Chapter 18 and Greene and Hensher (2010).

TABLE 11.1 Attrition from BHPS

Wave	*Individuals*	*Survival*	*Exited*	*Attrition*
1	10,256	—	—	—
2	8,957	87.33%	1299	12.67%
3	8,162	79.58%	795	8.88%
4	7,825	76.30%	337	4.13%
5	7,430	72.45%	395	5.05%
6	7,238	70.57%	192	2.58%
7	7,102	69.25%	136	1.88%
8	6,839	66.68%	263	3.70%

NEXT WAVE$_{it}$ = 1 if individual i is in the sample in wave $t + 1$,
ALL WAVE$_{it}$ = 1 if individual i is in the sample for all waves,
NUMBER OF WAVES = the number of waves for which the individual is present.

The results at this step included those in Table 11.2 (extracted from their Table IX). Curiously, at this step, the authors found strong evidence of nonrandom attrition in the subsample of men in the sample, but not in that for women. The authors then employed an inverse probability weighting approach to "correct" for the possibility of nonrandom attrition. They employed two procedures. First, for each individual in the sample, construct $\mathbf{d}_i = (d_{i1},\ . . . ., d_{iT})$ where $d_{it} = 1$ if individual i is present in wave t. By construction, $d_{i1} = 1$ for everyone. A vector of covariates observed at the baseline that is thought to be relevant to attrition in each period is designated $\mathbf{z}_{i1}$. This includes ln Income, marital status, age, race, education, household size, and health status, and some indicators of morbidity. For each period, a probit model is fit for $\text{Prob}(d_{it} = 1 | \mathbf{z}_{i1})$ and fitted probabilities, $\hat{p}_{it}$ are computed. (Note: $\hat{p}_{i1} = 1$.) With these fitted probabilities in hand, the model is estimated by maximizing the criterion function, in their case, the log-likelihood function, $\ln L = \Sigma_i \Sigma_t (d_{it}/\hat{p}_{it}) \ln L_{it}$. (For the models examined in this chapter, the log-likelihood term would be the negative of a squared residuals to maximize the negative of the sum or squares.) These results are labeled IPW-1 in Table 11.3. For the second method, the sample is restricted to the subset for which attrition was permanent. For each period, the list of variables is expanded to include z_{i1} and $z_{i,t-1.}$ The predicted probabilities at each, computed using the probit model, are denoted $\hat{\pi}_{is}$. Finally, to account for the fact that the sample at each wave is based on selection from the previous wave (so that $d_{it} = \Pi_{s \le t} d_{is}$) the probabilities are likewise adjusted: $\hat{p}_{it} = \Pi_{s=1}^{t} \hat{\pi}_{is}$. The results below show the influence of the sample treatment on one of the estimated coefficients in the full model.

TABLE 11.2 Tests for Attrition Bias

	Men		*Women*	
	β	*t Ratio*	β	*t Ratio*
NEXT WAVE	0.199	5.67	0.060	1.77
ALL WAVES	0.139	4.46	0.071	2.45
NUMBER OF WAVES	0.031	3.54	0.016	1.88

TABLE 11.3 Estimated Coefficients* on ln Income in Ordered Probit Models (Standard errors in Parentheses)

	Balanced Sample	*Unbalanced*	*IPW-1*	*IPW-2*
	NT = 19,460	NT = 24,371	NT = 24,370	NT = 23,211
Men	0.036 (0.022)	0.035 (0.019)	0.035 (0.020)	0.043 (0.021)
Women	0.029 (0.021)	0.033 (0.018)	0.021 (0.019)	0.018 (0.020)

*Coefficient on ln Income in Dynamic Ordered Probit Model. (Extracted from Table X and Table XI.)

Example 11.3 Attrition and Sample Selection in an Earnings Model for Physicians

Cheng and Trivedi (2015) approached the attrition question from a nonrandom sample selection perspective in their panel data study of Australian physicians' earnings. The starting point is a "missing at random" (MAR) interpretation of attrition. If individuals exit the sample for reasons that are unrelated to the variable under study—specifically, unrelated to the unobservables in the equation being used to model that variable—then attrition has no direct implications for the estimation of the parameters of the model.

Table 11.4 (derived from Table I in the article) shows that about one-third of the initial sample in their four-wave panel ultimately exited the sample. (Some individuals did return. The table shows the net effect.)

The model is a structural system,

$$\begin{aligned} &\textit{Attrition}: \; A_{it}^* = \mathbf{z}_{it}'\boldsymbol{\gamma} + u_{it}; \qquad A_{it} = 1 \text{ if } A_{it}^* > 0,\\ &\ln \textit{Wages}: \; y_{it}^* = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{f}_i'\boldsymbol{\delta} + \alpha_i + \varepsilon_{it};\; y_{it} = y_{it}^* \text{ if } A_{it} = 0, \text{ unobserved otherwise,}\end{aligned}$$

where $\mathbf{x}_{it}$ and $\mathbf{z}_{it}$ are time-varying exogenous variables, $\mathbf{f}_i$ is time-invariant, possibly endogenous variables, and α_i is a fixed effect. This setup is an application of Heckman's (1979) sample selection framework. (See Section 19.5.) The implication of the observation mechanism for the observed data is

$$\begin{aligned} E[y_{it} \mid \mathbf{x}_{it}, \mathbf{f}_i, \alpha_i, A_{it} = 0] &= \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{f}_i'\boldsymbol{\delta} + \alpha_i + E[\varepsilon_{it} \mid u_{it} \leq -\mathbf{z}_{it}'\boldsymbol{\gamma}]\\ &= \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{f}_i'\boldsymbol{\delta} + \alpha_i + \theta\, \lambda(\mathbf{z}_{it}'\boldsymbol{\gamma}).\end{aligned}$$

[In this reduced form of the model, θ is not (yet) a structural parameter. A nonzero value of this coefficient implies the presence of the attrition (selection) effect. The effect is generic until some structure is placed on the joint observation and attrition mechanism.] If ε_{it} and u_{it} are correlated, then $[\theta\lambda(\mathbf{z}_{it}'\boldsymbol{\gamma})]$ will be nonzero. Regression of y_{it} on $\mathbf{x}_{it}$, $\mathbf{f}_i$, and whatever device is used to control for the fixed effects will be affected by the missing *selection effect*, $\lambda_{it} = \lambda(\mathbf{z}_{it}'\boldsymbol{\gamma})$. If this omitted variable is correlated with $(\mathbf{x}_{it}, \mathbf{f}_i, \alpha_i)$, then the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are likely to be distorted. A partial solution is obtained by using first differences in the

TABLE 11.4 Attrition from the Medicine in Australia Balancing Employment and Life Data

	General Practitioners			*Specialists*		
Year	*N*	*Attrition**	*Survival*	*N*	*Attrition*	*Survival*
1	3906	840	100.0%	4596	926	100.0%
2	3066	242	78.5%	3670	303	79.9%
3	2824	270	72.3%	3367	299	73.3%
4	2554	—	65.4%	3068	—	66.8%

* Net attrition takes place after the indicated year.

regression. First differences will eliminate the time-invariant components of the regression, ($\mathbf{f}_i$, α_i), but will not solve the selection problem unless the attrition mechanism is also time invariant, which is not assumed. This nonzero correlation will be the *attrition effect*.

If there is attrition bias (in the estimator that ignores attrition), then the sample should become progressively less random as the observation period progresses. This suggests a possible indirect test for attrition bias. The full *unbalanced* sample contains a *balanced* subsample of individuals who are present for all waves of the panel. (Individuals who left and rejoined the panel would be bypassed for purposes of this exercise.) Under the MAR assumption, estimation of $\boldsymbol{\beta}$ based on the unbalanced full sample and the balanced subsample should produce the same results (aside from some sampling variability). This suggests one might employ a Hausman style test. (See Section 11.5.6.) The authors employed a more direct strategy. A narrow assumption that (ε_{it}, u_{it}) are bivariate normally distributed with zero means, variances, σ^2 and 1, and correlation ρ (a variance for u_{it} is not identified) produces

$$\theta_t\,\lambda(\mathbf{z}_{it}'\gamma_t) = \theta_t\frac{\phi(-\mathbf{z}_{it}'\gamma_t)}{\Phi(-\mathbf{z}_{it}'\gamma_t)}.$$

Estimates of the coefficients in this "control function" regression are computed for each of waves 2–4 and added to the first difference regression,

$$y_{it} - y_{i,t-1} = (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta} + \textstyle\sum_{t=2}^{4}\theta_t\hat{\lambda}_{it} + w_{it},$$

which is then estimated using least squares. Standard errors are computed using bootstrapping. Under the joint normality assumption, this control function estimator is robust, in that if there is an attrition effect (nonzero ρ), the effect is accounted for while if $\rho = 0$, the original estimator (within or first differences) will be consistent on its own. A second approach that loosens the bivariate normality assumption is based on a copula model (Section 12.2.2) that is estimated by maximum likelihood.

Table 11.5 below (derived from Tables III and IV in the paper) summarizes the results. The bivariate normal model strongly suggests the presence of the attrition effect, though the impact on the main estimation result is relatively modest. But the results for the copula are quite different. The effect is found to be significant only for the specialists. The impact on the hours coefficient is quite large for this group as well.

TABLE 11.5 Earnings Models and Tests for Attrition Bias

	General Practitioners	*Specialists*
Fixed Effects Hours Coefficient		
Unbalanced*	0.460 (0.027) [7776]	0.287 (0.022) [8904]
Balanced	0.407 (0.038) [3464]	0.356 (0.029) [4204]
First Differences Hours Coefficient		
Unbalanced	0.428 (0.042) [4106]	0.174 (0.038) [4291]
Balanced	0.387 (0.055) [2598]	0.244 (0.053) [3153]
Bivariate Normal Hazards Attrition Model		
Hours coefficient	0.422 (0.041) [4043]	0.180 (0.035) [4875]
Wald Statistic (3 df)	42.47	38.65
p Value	0.000	0.000
Frank Copula Attrition Model		
Marginals	Probit, Student's *t*	Logit, logistic
Hours coefficient	0.315 (0.043) [5166]	0.104 (0.026) [6109]
Wald Statistic (1 df)	1.862	7535.119
p Value	0.172	0.000

* Standard errors in parentheses. Sample size in brackets.

Unbalanced panels may arise for systematic reasons that induce problems that look like sample selection issues. But the attrition from a panel data set may also be completely ignorable, that is, due to issues that are out of the view of the analyst. In such cases, it is reasonable simply to treat the unbalanced nature of the data as a characteristic of the random sampling. Almost none of the useful theory that we will examine here relies on an assumption that the panel is balanced. The development to follow is structured so that the distinction between balanced and unbalanced panels, beyond the attrition issue, will entail little more than a trivial change in notation—where for convenience we write T suggesting a balanced panel, merely changing T to T_i generalizes the results. We will note specifically when this is not the case, such as in Breusch and Pagan's (1980) LM statistic.

11.2.6 WELL-BEHAVED PANEL DATA

The asymptotic properties of the estimators in the classical regression model were established in Section 4.4 under the following assumptions:

A.1. ***Linearity:*** $y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K + \varepsilon_i$.
A.2. ***Full rank:*** The $n \times K$ sample data matrix, $\mathbf{X}$ has full column rank for every $n > K$.
A.3. ***Strict exogeneity of the independent variables:*** $E[\varepsilon_i | x_{j1}, x_{j2}, \ldots, x_{jK}] = 0, i, j = 1, \ldots, n$.
A.4. ***Homoscedasticity and nonautocorrelation:*** $E[\varepsilon_i\varepsilon_j | \mathbf{X}] = \sigma_\varepsilon^2$ if $i = j$ and 0 otherwise.

The following are the crucial results needed: For consistency of **b**, we need

$$\begin{aligned}\text{plim}(1/n)\mathbf{X}'\mathbf{X} &= \text{plim}\,\overline{\mathbf{Q}}_n = \mathbf{Q}, \text{ a positive definite matrix,}\\ \text{plim}(1/n)\mathbf{X}'\boldsymbol{\varepsilon} &= \text{plim}\,\overline{\mathbf{w}}_n = E[\overline{\mathbf{w}}_n] = \mathbf{0}.\end{aligned}$$

(For consistency of s^2, we added a fairly weak assumption about the moments of the disturbances.) To establish asymptotic normality, we required consistency and

$$\sqrt{n}\,\overline{\mathbf{w}}_n \xrightarrow{d} N[0, \sigma^2\mathbf{Q}].$$

With these in place, the desired characteristics are then established by the methods of Sections 4.4.1 and 4.4.2.

Exceptions to the assumptions are likely to arise in a **panel data** set. The sample will consist of multiple observations on each of many observational units. For example, a study might consist of a set of observations made at different points in time on a large number of families. In this case, the **x**'s will surely be correlated across observations, at least within observational units. They might even be the same for all the observations on a single family.

The panel data set could be treated as follows. Assume for the moment that the data consist of a fixed number of observations, say T, on a set of n families, so that the total number of rows in $\mathbf{X}$ is $N = nT$. The matrix $\overline{\mathbf{Q}}_n$, in which n is all the observations in the sample, is

$$\overline{\mathbf{Q}}_n = \frac{1}{n}\sum_i \frac{1}{T}\mathbf{X}_i'\mathbf{X}_i = \frac{1}{n}\sum_{i=1}^{n}\mathbf{Q}_i.$$

We then view the set of observations on the *i*th unit as if they were a single observation and apply our convergence arguments to the number of units increasing without bound. The point is that the conditions that are needed to establish convergence will apply with respect to the number of observational units. The number of observations taken for each observation unit might be fixed and could be quite small.

This chapter will contain relatively little development of the properties of estimators as was done in Chapter 4. We will rely on earlier results in Chapters 4, 8, and 9 and focus instead on a variety of models and specifications.

11.3 THE POOLED REGRESSION MODEL

We begin the analysis by assuming the simplest version of the model, the **pooled model**,

$$y_{it} = \alpha + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, i = 1, \ldots, n, t = 1, \ldots, T_i, \qquad \textbf{(11-2)}$$

$$E[\varepsilon_{it}, |\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT_i}] = 0,$$
$$E[\varepsilon_{it}\varepsilon_{js}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT_i}] = \sigma_\varepsilon^2 \text{ if } i = j \text{ and } t = s \text{ and } = 0 \text{ if } i \neq j \text{ or } t \neq s.$$

In this form, if the remaining assumptions of the classical model are met (zero conditional mean of ε_{it}, homoscedasticity, uncorrelatedness across observations, *i* and strict exogeneity of $\mathbf{x}_{it}$), then no further analysis beyond the results of Chapter 4 is needed. Ordinary least squares is the efficient estimator and inference can reliably proceed along the lines developed in Chapter 5.

11.3.1 LEAST SQUARES ESTIMATION OF THE POOLED MODEL

The crux of the panel data analysis in this chapter is that the assumptions underlying ordinary least squares estimation of the pooled model are unlikely to be met. The question, then, is what can be expected of the estimator when the heterogeneity does differ across individuals? The fixed effects case is obvious. As we will examine later, omitting (or ignoring) the heterogeneity when the fixed effects model is appropriate renders the least squares estimator inconsistent—sometimes wildly so. In the random effects case, in which the true model is

$$y_{it} = c_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it},$$

where $E[c_i|\mathbf{X}_i] = \alpha$, we can write the model

$$\begin{aligned} y_{it} &= \alpha + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} + (c_i - E[c_i|\mathbf{X}_i]) \\ &= \alpha + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} + u_i \\ &= \alpha + \mathbf{x}_{it}'\boldsymbol{\beta} + w_{it}. \end{aligned}$$

In this form, we can see that the unobserved heterogeneity induces autocorrelation; $E[w_{it}w_{is}] = \sigma_u^2$ when $t \neq s$. As we explored in Chapter 9—we will revisit it in Chapter 20—the ordinary least squares estimator in the generalized regression model may be consistent, but the conventional estimator of its asymptotic variance is likely to underestimate the true variance of the estimator.

11.3.2 ROBUST COVARIANCE MATRIX ESTIMATION AND BOOTSTRAPPING

Suppose we consider the model more generally. Stack the T_i observations for individual i in a single equation,

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

where $\boldsymbol{\beta}$ now includes the constant term. In this setting, there may be heteroscedasticity across individuals. However, in a panel data set, the more substantive effect is cross-observation correlation, or autocorrelation. In a longitudinal data set, the group of observations may all pertain to the same individual, so any latent effects left out of the model will carry across all periods. Suppose, then, we assume that the disturbance vector consists of ε_{it} plus these omitted components. Then,

$$\begin{aligned}\text{Var}[\mathbf{w}_i \mid \mathbf{X}_i] &= \sigma_\varepsilon^2 \mathbf{I}_{T_i} + \boldsymbol{\Sigma}_i \\ &= \boldsymbol{\Omega}_i.\end{aligned}$$

(The subscript i on $\boldsymbol{\Omega}_i$ does not necessarily indicate a different variance for each i. The designation is necessary because the matrix is $T_i \times T_i$.) The ordinary least squares estimator of $\boldsymbol{\beta}$ is

$$\begin{aligned}\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} &= \left[\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i\right]^{-1}\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{y}_i \\ &= \left[\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i\right]^{-1}\sum_{i=1}^{n}\mathbf{X}_i'(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{w}_i) \\ &= \boldsymbol{\beta} + \left[\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i\right]^{-1}\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{w}_i.\end{aligned}$$

Consistency can be established along the lines developed in Chapter 4. The true asymptotic covariance matrix would take the form we saw for the generalized regression model in (9-8),

$$\begin{aligned}\text{Asy.Var}[\mathbf{b}] &= \frac{1}{n}\text{plim}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i\right]^{-1}\text{plim}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{w}_i\mathbf{w}_i'\mathbf{X}_i\right]\text{plim}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i\right]^{-1} \\ &= \frac{1}{n}\text{plim}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i\right]^{-1}\text{plim}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i'\boldsymbol{\Omega}_i\mathbf{X}_i\right]\text{plim}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i\right]^{-1}.\end{aligned}$$

This result provides the counterpart to (9-12). As before, the center matrix must be estimated. In the same fashion as the White estimator, we can estimate this matrix with

$$\text{Est.Asy.Var}[\mathbf{b}] = \frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i'\hat{\mathbf{w}}_i\hat{\mathbf{w}}_i'\mathbf{X}_i\right]\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i\right]^{-1}, \qquad \textbf{(11-3)}$$

where $\hat{\mathbf{w}}_i'$ is the vector of T_i residuals for individual i. In fact, the logic of the White estimator *does* carry over to this estimator. Note, however, this is not quite the same as (9-5). It is quite likely that the more important issue for appropriate estimation of the asymptotic covariance matrix is the correlation across observations, not heteroscedasticity. As such, it is likely that the White estimator in (9-5) is not the solution to the inference problem here. Example 11.4 shows this effect at work. This is the "cluster" robust estimator developed in Section 4.5.3.

Bootstrapping offers another approach to estimating an appropriate covariance matrix for the estimator. We used this approach earlier in a cross-section setting in Example 4.6 where we devised an estimator for the LAD estimator. Here, we will take the group or cluster as the unit of observation. For example, in the data in Example 11.4, there are 595 groups of 7 observations, so the block of 7 observations is the unit of observation. To compute the **block bootstrap** estimator, we use the following procedure. For each of R repetitions, draw random samples of $N = 595$ blocks with replacement. (Each time, some blocks are drawn more than once and others are not drawn.) After the R repetitions, compute the empirical variance of the R replicates. The estimator is

$$\text{Est.Asy.Var}[\mathbf{b}] = \frac{1}{R}\sum_{r=1}^{R}(\mathbf{b}_r - \overline{\mathbf{b}})(\mathbf{b}_r - \overline{\mathbf{b}})'.$$

Example 11.4 Wage Equation

Cornwell and Rupert (1988) analyzed the returns to schooling in a balanced panel of 595 observations on heads of households. The sample data are drawn from years 1976–1982 from the "Non-Survey of Economic Opportunity" from the Panel Study of Income Dynamics. Our estimating equation is a modified version of the one in the paper (without the time fixed effects);

$$\begin{aligned}\ln Wage_{it} = {} & \beta_1 + \beta_2\, Exp_{it} + \beta_3\, Exp_{it}^2 + \beta_4\, Wks_{it} + \beta_5\, Occ_{it} \\ & + \beta_6\, Ind_{it} + \beta_7\, South_{it} + \beta_8\, SMSA_{it} + \beta_9\, MS_{it} \\ & + \beta_{10}\, Union_{it} + \beta_{11}\, Ed_i + \beta_{12}\, Fem_i + \beta_{13}\, Blk_i + \varepsilon_{it}\end{aligned}$$

where the variables in the model are

Exp	= years of full-time work experience,
Wks	= weeks worked,
Occ	= 1 if the individual has a blue-collar occupation, 0 if not,
Ind	= 1 if the individual works in a manufacturing industry, 0 if not,
South	= 1 if the individual resides in the south, 0 if not,
SMSA	= 1 if the individual resides in an SMSA, 0 if not,
MS	= 1 if the individual is married, 0 if not
Union	= 1 if the individual's wage is set by a union contract, 0 if not
Ed	= years of education
Fem	= 1 if the individual is female, 0 if not,
Blk	= 1 if the individual is black, 0 if not.

See Appendix Table F8.1 for the data source. Note that *Ed*, *Fem*, and *Blk* are **time invariant**. The main interest of the study, beyond comparing various estimation methods, is β_{11}, the return to education. Table 11.6 reports the least squares estimates based on the full sample of 4,165 observations. [The authors do not report OLS estimates. However, they do report linear least squares estimates of the fixed effects model, which are simple least squares using deviations from individual means. (See Section 11.4.)] The conventional OLS standard errors are given in the second column of results. The third column gives the robust standard errors computed using (11-3). For these data, the computation is

$$\text{Est.Asy.Var}[b] = \left[\sum_{i=1}^{595}\mathbf{X}_i'\mathbf{X}_i\right]^{-1}\left[\sum_{i=1}^{595}\left(\sum_{t=1}^{7}\mathbf{x}_{it}e_{it}\right)\left(\sum_{t=1}^{7}\mathbf{x}_{it}e_{it}\right)'\right]\left[\sum_{i=1}^{595}\mathbf{X}_i'\mathbf{X}_i\right]^{-1}.$$

TABLE 11.6 Wage Equation Estimated by OLS

Variable	*Least Squares Estimate*	*Standard Error*	*Clustered Std.Error*	*Bootstrapped Std.Error*	*White Hetero. Robust Std. Error*
Constant	5.25112	0.07129	0.12355	0.11171	0.07435
Exp	0.00401	0.00216	0.00408	0.00434	0.00216
ExpSq	−0.00067	0.00005	0.00009	0.00010	0.00005
Wks	0.00422	0.00108	0.00154	0.00164	0.00114
Occ	−0.14001	0.01466	0.02724	0.02555	0.01494
Ind	0.04679	0.01179	0.02366	0.02153	0.01199
South	−0.05564	0.01253	0.02616	0.02414	0.01274
SMSA	0.15167	0.01207	0.02410	0.02323	0.01208
MS	0.04845	0.02057	0.04094	0.03749	0.02049
Union	0.09263	0.01280	0.02367	0.02553	0.01233
Ed	0.05670	0.00261	0.00556	0.00483	0.00273
Fem	−0.36779	0.02510	0.04557	0.04460	0.02310
Blk	−0.16694	0.02204	0.04433	0.05221	0.02075

The robust standard errors are generally about twice the uncorrected ones. In contrast, the White robust standard errors are almost the same as the uncorrected ones. This suggests that for this model, ignoring the within-group correlations does, indeed, substantially affect the inferences one would draw. The block bootstrap standard errors based on 100 replications are shown in the last column. As expected, the block bootstrap results are quite similar to the two-step residual-based method.

11.3.3 CLUSTERING AND STRATIFICATION

Many recent studies have analyzed survey data sets, such as the Current Population Survey (CPS). Survey data are often drawn in clusters, partly to reduce costs. For example, interviewers might visit all the families in a particular block. In other cases, effects that resemble the common random effects in panel data treatments might arise naturally in the sampling setting. Consider, for example, a study of student test scores across several states. Common effects could arise at many levels in such a data set. Education curriculum or funding policies in a state could cause a "state effect"; there could be school district effects, school effects within districts, and even teacher effects within a particular school. Each of these is likely to induce correlation across observations that resembles the random (or fixed) effects we have identified. One might be reluctant to assume that a tightly structured model such as the simple random effects specification is at work. But, as we saw in Example 11.1, ignoring common effects can lead to serious inference errors.

Moulton (1986, 1990) examined the bias of the conventional least squares estimator of Asy.Var[$\mathbf{b}$], $s^2(\mathbf{X}'\mathbf{X})^{-1}$. The calculation is complicated because the comparison ultimately depends on the group sizes, the data themselves, and the within-group cross-observation correlation of the common effects. For a simple case,

$$y_{i,g} = \beta_1 + x_{i,g}\beta_2 + u_{i,g} + w_g,$$

a broad, approximate result is the Moulton factor,

$$\frac{\text{Cluster Corrected Variance}}{\text{OLS Uncorrected Variance}} \approx [1 + (n_g - 1)r_x r_u],$$

where n_g is the group size, r_x is the cross-observation correlation (within a group) of $x_{i,g}$ and r_u is the "intraclass correlation," $\sigma_w^2/(\sigma_w^2 + \sigma_u^2)$. The Moulton bias factor suggests that the conventional standard error is biased downward, potentially quite substantially if n_g is large. It is worth noting the Moulton bias might create the impression that the correction of the standard errors *always* increases the standard errors. Algebraically, this is not true—a counterexample appears in Example 4.5. The Moulton result suggests a correction to the OLS standard errors. However, using it would require several approximations of unknown size (based on there being more than one regressor, variable cluster sizes, and needing an estimator for r_u). The robust estimator suggested in Section 11.3.2 will be a preferable approach.

A refinement to (11-3) is sometimes employed to account for small-sample effects when the number of clusters is likely to be a significant proportion of a finite total, such as the number of school districts in a state. A degrees of freedom correction as shown in (11-4) is often employed for this purpose. The robust covariance matrix estimator would be

$$\begin{aligned}\text{Est.Asy.Var}[\mathbf{b}] &= \left[\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{X}_g\right]^{-1}\left[\frac{G}{G-1}\sum_{g=1}^{G}\left(\sum_{i=1}^{n_g}\mathbf{x}_{ig}\hat{w}_{ig}\right)\left(\sum_{i=1}^{n_g}\mathbf{x}_{ig}\hat{w}_{ig}\right)'\right]\left[\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{X}_g\right]^{-1} \\ &= \left[\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{X}_g\right]^{-1}\left[\frac{G}{G-1}\sum_{g=1}^{G}(\mathbf{X}_g'\hat{\mathbf{w}}_g)(\hat{\mathbf{w}}_g'\mathbf{X}_g)\right]\left[\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{X}_g\right]^{-1}, \qquad \textbf{(11-4)}\end{aligned}$$

where G is the number of clusters in the sample and each cluster consists of n_g, $g = 1, \ldots, G$ observations. [Note that this matrix is simply $G/(G - 1)$ times the matrix in (11-3).] A further correction (without obvious formal motivation) sometimes employed is a degrees of freedom correction, $[(\Sigma_g n_g) - 1]/[(\Sigma_g n_g) - K]$.

Many further refinements for more complex samples—consider the test scores example—have been suggested. For a detailed analysis, see Cameron and Trivedi (2005, Chapter 24) and Cameron and Miller (2015). Several aspects of the computation are discussed in Wooldridge (2010, Chapter 20) as well. An important question arises concerning the use of asymptotic distributional results in cases in which the number of clusters might be relatively small. Angrist and Lavy (2002) find that the clustering correction after pooled OLS, as we have done in Example 11.3, is not as helpful as might be hoped for (though our correction with 595 clusters each of size 7 would be "safe" by these standards). But, the difficulty might arise, at least in part, from the use of OLS in the presence of the common effects. Kezde (2001) and Bertrand, Dufflo, and Mullainathan (2002) find more encouraging results when the correction is applied after estimation of the fixed effects regression. Yet another complication arises when the groups are very large and the number of groups is relatively small, for example, when the panel consists of many large samples from a subset (or even all) of the U.S. states. Since the asymptotic theory we have used to this point assumes the opposite, the results will be less reliable in this case. Donald and Lang (2007) find that this case gravitates toward analysis of group means rather than the individual data. Wooldridge (2003) provides results that help explain this finding. Finally, there is a natural question as to whether the correction is even called for if one has used a random effects, generalized least squares procedure (see Section 11.5) to do the estimation at the first step. If the data-generating mechanism were strictly consistent with the random effects model, the answer would clearly be negative. Under the view that the random effects specification is only an approximation to the correlation across observations in a cluster, then there would remain residual correlation that would be accommodated by the correction in (11-4) (or some GLS counterpart). (This would call the specific random effects correction in Section 11.5 into question, however.) A similar argument would motivate the correction after fitting the fixed effects model as well. We will pursue these possibilities in Section 11.6.4 after we develop the fixed and random effects estimator in detail.

11.3.4 ROBUST ESTIMATION USING GROUP MEANS

The pooled regression model can also be estimated using the sample means of the data. The implied regression model is obtained by premultiplying each group by $(1/T)\mathbf{i}'$ where $\mathbf{i}'$ is a row vector of ones,

$$(1/T)\mathbf{i}'\mathbf{y}_i = (1/T)\mathbf{i}'\mathbf{X}_i\boldsymbol{\beta} + (1/T)\mathbf{i}'w_i$$

or

$$\overline{y}_{i.} = \overline{\mathbf{x}}_i'\boldsymbol{\beta} + \overline{w}_i.$$

In the transformed linear regression, the disturbances continue to have zero conditional means but heteroscedastic variances $\sigma_i^2 = (1/T^2)\mathbf{i}'\boldsymbol{\Omega}_i\mathbf{i}$. With $\boldsymbol{\Omega}_i$ unspecified, this is a heteroscedastic regression for which we would use the White estimator for appropriate inference. Why might one want to use this estimator when the full data set is available? If the classical assumptions are met, then it is straightforward to show that the asymptotic covariance matrix for the group means estimator is unambiguously larger, and the answer would be that there is no benefit. But failure of the classical assumptions is what brought us to this point, and then the issue is less clear-cut. In the presence of unstructured cluster effects the efficiency of least squares can be considerably diminished, as we saw in the preceding example. The loss of information that occurs through the averaging might be relatively small, though in principle the disaggregated data should still be better.

We emphasize that using **group means** does not solve the problem that is addressed by the fixed effects estimator. Consider the general model,

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + c_i\boldsymbol{i} + \mathbf{w}_i,$$

where as before, c_i is the latent effect. If the mean independence assumption, $E[c_i|\mathbf{X}_i] = \alpha$, is not met, then the effect will be transmitted to the group means as well. In this case, $E[c_i|\mathbf{X}_i] = h(\mathbf{X}_i)$. A common specification is Mundlak's (1978), where we employ the projection of c_i on the group means (see Section 4.4.5),

$$c_i|\mathbf{X}_i = \overline{\mathbf{x}}_i'.\boldsymbol{\gamma} + v_i.$$

Then,

$$\begin{aligned} y_{it} &= \mathbf{x}_{it}'\boldsymbol{\beta} + c_i + \varepsilon_{it} \\ &= \mathbf{x}_{it}'\boldsymbol{\beta} + \overline{\mathbf{x}}_{i.}'\boldsymbol{\gamma} + [\varepsilon_{it} + v_i] \\ &= \mathbf{x}_{it}'\boldsymbol{\beta} + \overline{\mathbf{x}}_{i.}'\boldsymbol{\gamma} + u_{it}, \end{aligned}$$

where, by construction, $Cov[u_{it}, \bar{\mathbf{x}}_i] = 0$. Taking means as before,

$$\begin{aligned}\bar{y}_{i.} &= \bar{\mathbf{x}}_{i.}'\boldsymbol{\beta} + \bar{\mathbf{x}}_{i.}'\boldsymbol{\gamma} + \bar{u}_{i.}\\ &= \bar{\mathbf{x}}_{i.}'(\boldsymbol{\beta} + \boldsymbol{\gamma}) + \bar{u}_{i.}.\end{aligned}$$

The implication is that the group means estimator estimates not $\boldsymbol{\beta}$, but $\boldsymbol{\beta} + \boldsymbol{\gamma}$. Averaging the observations in the group collects the entire set of effects, observed and latent, in the group means.

One consideration that remains, which, unfortunately, we cannot resolve analytically, is the possibility of measurement error. If the regressors are measured with error, then, as we examined in Section 8.7, the least squares estimator is inconsistent and, as a consequence, efficiency is a moot point. In the panel data setting, if the measurement error is random, then using group means would work in the direction of averaging it out—indeed, in this instance, assuming the benchmark case $\mathbf{x}_{itk} = \mathbf{x}_{itk}^* + u_{itk}$, one could show that the group means estimator would be consistent as $T \rightarrow \infty$ while the OLS estimator would not.

Example 11.5 Robust Estimators of the Wage Equation

Table 11.7 shows the group means estimates of the wage equation shown in Example 11.4 with the original least squares estimates. In both cases, a robust estimator is used for the covariance matrix of the estimator. It appears that similar results are obtained with the means.

11.3.5 ESTIMATION WITH FIRST DIFFERENCES

First differencing is another approach to estimation. Here, the intent would explicitly be to transform latent heterogeneity out of the model. The base case would be

$$y_{it} = c_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it},$$

TABLE 11.7 Wage Equation Estimated by OLS

Coefficient	*OLS Estimated Coefficient*	*Cluster Robust Standard Error*	*Group Means Estimates*	*White Robust Standard Error*
Constant	5.25112	0.12330	5.12143	0.20425
Exp	0.04010	0.00408	0.03190	0.00478
Exp^2	−0.00067	0.00009	−0.00057	0.00010
Wks	0.00422	0.00154	0.00919	0.00360
Occ	−0.14001	0.02724	−0.16762	0.03382
Ind	0.04679	0.02366	0.05792	0.02554
South	−0.05564	0.02616	−0.05705	0.02597
SMSA	0.15167	0.02410	0.17578	0.02576
MS	0.04845	0.04094	0.11478	0.04770
Union	0.09263	0.02367	0.10907	0.02923
Ed	0.05670	0.00556	0.05144	0.00555
Fem	−0.36779	0.04557	−0.31706	0.05473
Blk	−0.16694	0.04433	−0.15780	0.04501

which implies the first differences equation,

$$\Delta y_{it} = \Delta c_i + (\Delta \mathbf{x}_{it})'\boldsymbol{\beta} + \Delta \varepsilon_{it},$$

or

$$\begin{aligned}\Delta y_{it} &= (\Delta \mathbf{x}_{it})'\boldsymbol{\beta} + \varepsilon_{it} - \varepsilon_{i,t-1}\\ &= (\Delta \mathbf{x}_{it})'\boldsymbol{\beta} + u_{it}.\end{aligned}$$

The advantage of the **first difference** approach is that it removes the latent heterogeneity from the model whether the fixed or random effects model is appropriate. The disadvantage is that the differencing also removes any time-invariant variables from the model. In our example, we had three, *Ed*, *Fem*, and *Blk*. If the time-invariant variables in the model are of no interest, then this is a robust approach that can estimate the parameters of the time-varying variables consistently. Of course, this is not helpful for the application in the example because the impact of *Ed* on ln *Wage* was the primary object of the analysis. Note, as well, that the differencing procedure trades the cross-observation correlation in c_i for a moving average (MA) disturbance, $u_{i,t} = \varepsilon_{i,t} - \varepsilon_{i,t-\mathbf{1}}$.[5] The new disturbance, $u_{i,t}$, is autocorrelated, though across only one period. Nonetheless, in order to proceed, it would have to be true that $\Delta \mathbf{x}_t$ is uncorrelated with $\Delta \varepsilon_t$. Strict exogeneity of $\mathbf{x}_{it}$ is sufficient, but in the absence of that assumption, such as if only $\text{Cov}(\varepsilon_{it}, \mathbf{x}_{it}) = 0$ has been assumed, then it is conceivable that $\Delta \mathbf{x}_t$ and $\Delta \varepsilon_t$ could be correlated. The presence of a lagged value of y_{it} in the original equation would be such a case. Procedures are available for using two-step feasible GLS for an MA disturbance (see Chapter 20). Alternatively, this model is a natural candidate for OLS with the Newey–West robust covariance estimator because the right number of lags (one) is known. (See Section 20.5.2.)

As a general observation, with a variety of approaches available, the first difference estimator does not have much to recommend it, save for one very important application. Many studies involve two period panels, a before and an after treatment. In these cases, as often as not, the phenomenon of interest may well specifically be the change in the outcome variable—the "treatment effect." Consider the model

$$y_{it} = c_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \theta S_{it} + \varepsilon_{it},$$

where $t = 1, 2$ and $S_{it} = 0$ in period 1 and 1 in period 2; S_{it} indicates a treatment that takes place between the two observations. The treatment effect would be

$$E[\Delta y_i | (\Delta \mathbf{x}_i = 0)] = \theta,$$

which is precisely the constant term in the first difference regression,

$$\Delta y_i = \theta + (\Delta \mathbf{x}_i)'\boldsymbol{\beta} + u_i.$$

We examined cases like these in detail in Section 6.3.

11.3.6 THE WITHIN- AND BETWEEN-GROUPS ESTIMATORS

The pooled regression model is

$$y_{it} = \alpha + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}. \tag{11-5a}$$

[5] If the original disturbance, ε_{it} were a random walk, $\varepsilon_{i,t} = \varepsilon_{i,t-1} + u_{it}$, then the disturbance in the first differenced equation would be homoscedastic and nonautocorrelated. This would be a narrow assumption that might apply in a particular situation. This would not seem to be a natural specification for the model in Example 11.4, for example.

In terms of the group means,

$$\bar{y}_{i.} = \alpha + \bar{\mathbf{x}}_{i.}'\boldsymbol{\beta} + \bar{\varepsilon}_{i.}, \tag{11-5b}$$

while in terms of deviations from the group means,

$$y_{it} - \bar{y}_{i.} = (\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})'\boldsymbol{\beta} + \varepsilon_{it} - \bar{\varepsilon}_{i.}.$$

For convenience later, write this as

$$\ddot{y} = \ddot{\mathbf{x}}_{it}'\boldsymbol{\beta} + \ddot{\varepsilon}_{it}. \tag{11-5c}$$

[We are assuming there are no time-invariant variables in $\mathbf{x}_{it}$, such as *Ed* in Example 11.4. These would become all zeros in (11-5c).] All three are classical regression models, and in principle, all three could be estimated, at least consistently if not efficiently, by ordinary least squares. [Note that (11-5b) defines only n observations, the group means.] Consider then the matrices of sums of squares and cross products that would be used in each case, where we focus only on estimation of $\boldsymbol{\beta}$. In (11-5a), the moments would accumulate variation about the overall means, $\bar{\bar{y}}$ and $\bar{\bar{\mathbf{x}}}$, and we would use the total sums of squares and cross products,

$$\mathbf{S}_{xx}^{total} = \sum_{i=\mathbf{1}}^{n}\sum_{t=\mathbf{1}}^{T}(\mathbf{x}_{it} - \bar{\bar{\mathbf{x}}})(\mathbf{x}_{it} - \bar{\bar{\mathbf{x}}})' \quad \text{and} \quad \mathbf{S}_{xy}^{total} = \sum_{i=\mathbf{1}}^{n}\sum_{t=\mathbf{1}}^{T}(\mathbf{x}_{it} - \bar{\bar{\mathbf{x}}})(y_{it} - \bar{\bar{y}}). \tag{11-6}$$

For (11-5c), because the data are in deviations already, the means of $(y_{it} - \bar{y}_{i.})$ and $(\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})$ are zero. The moment matrices are **within-groups** (i.e., variation around group means) sums of squares and cross products,

$$\mathbf{S}_{xx}^{within} = \sum_{i=1}^{n}\sum_{t=1}^{T}(\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})(\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})' \quad \text{and} \quad \mathbf{S}_{xy}^{within} = \sum_{i=1}^{n}\sum_{t=1}^{T}(\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})(y_{it} - \bar{y}_{i.}).$$

Finally, for (11-5b), the mean of group means is the overall mean. The moment matrices are the **between-groups** sums of squares and cross products—that is, the variation of the group means around the overall means,

$$\mathbf{S}_{xx}^{between} = \sum_{i=1}^{n}T(\bar{\mathbf{x}}_{i.} - \bar{\bar{\mathbf{x}}})(\bar{\mathbf{x}}_{i.} - \bar{\bar{\mathbf{x}}})' \quad \text{and} \quad \mathbf{S}_{xy}^{between} = \sum_{i=1}^{n}T(\bar{\mathbf{x}}_{i.} - \bar{\bar{\mathbf{x}}})(\bar{y}_{i.} - \bar{\bar{y}}).$$

It is easy to verify that

$$\mathbf{S}_{xx}^{total} = \mathbf{S}_{xx}^{within} + \mathbf{S}_{xx}^{between} \quad \text{and} \quad \mathbf{S}_{xy}^{total} = \mathbf{S}_{xy}^{within} + \mathbf{S}_{xy}^{between}.$$

Therefore, there are three possible least squares estimators of $\boldsymbol{\beta}$ corresponding to the decomposition. The least squares estimator is

$$\mathbf{b}^{total} = \left[\mathbf{S}_{xx}^{total}\right]^{-1}\mathbf{S}_{xy}^{total} = \left[\mathbf{S}_{xx}^{within} + \mathbf{S}_{xx}^{between}\right]^{-1}\left[\mathbf{S}_{xy}^{within} + \mathbf{S}_{xy}^{between}\right]. \tag{11-7}$$

The within-groups estimator is

$$\mathbf{b}^{within} = \left[\mathbf{S}_{xx}^{within}\right]^{-1}\mathbf{S}_{xy}^{within}. \tag{11-8}$$

This is the dummy variable estimator developed in Section 11.4. An alternative estimator would be the between-groups estimator,

$$\mathbf{b}^{between} = \left[\mathbf{S}_{xx}^{between}\right]^{\mathbf{-1}} \mathbf{S}_{xy}^{between}. \qquad \textbf{(11-9)}$$

This is the **group means estimator**. This least squares estimator of (11-5b) is based on the n sets of groups means. (Note that we are assuming that n is at least as large as K.) From the preceding expressions (and familiar previous results),

$$\mathbf{S}_{xy}^{within} = \mathbf{S}_{xx}^{within}\mathbf{b}^{within} \quad \text{and} \quad \mathbf{S}_{xy}^{between} = \mathbf{S}_{xx}^{between}\mathbf{b}^{between}.$$

Inserting these in (11-7), we see that the least squares estimator is a **matrix weighted average** of the within- and between-groups estimators:

$$\mathbf{b}^{total} = \mathbf{F}^{within}\mathbf{b}^{within} + \mathbf{F}^{between}\mathbf{b}^{between}, \qquad \textbf{(11-10)}$$

where

$$\mathbf{F}^{within} = \left[\mathbf{S}_{xx}^{within} + \mathbf{S}_{xx}^{between}\right]^{-1} \mathbf{S}_{xx}^{within} = \mathbf{I} - \mathbf{F}^{between}.$$

The form of this result resembles the Bayesian estimator in the classical model discussed in Chapter 16. The resemblance is more than passing; it can be shown[6] that

$$\mathbf{F}^{within} = \{[\text{Asy.Var}(\mathbf{b}^{within})]^{-1} + [\text{Asy.Var}(\mathbf{b}^{between})]^{-1}\}^{-1}[\text{Asy.Var}(\mathbf{b}^{within})]^{-1},$$

which is essentially the same mixing result we have for the Bayesian estimator. In the weighted average, the estimator with the smaller variance receives the greater weight.

Example 11.6 Analysis of Covariance and the World Health Organization (WHO) Data

The decomposition of the total variation in Section 11.3.6 extends to the linear regression model the familiar *analysis of variance*, or ANOVA, that is often used to decompose the variation in a variable in a clustered or stratified sample, or in a panel data set. One of the useful features of panel data analysis as we are doing here is the ability to analyze the between-groups variation (heterogeneity) to learn about the main regression relationships and the within-groups variation to learn about dynamic effects.

The WHO data used in Example 6.22 is an unbalanced panel data set—we used only one year of the data in Example 6.22. Of the 191 countries in the sample, 140 are observed in the full five years, one is observed four times, and 50 are observed only once. The original WHO studies (2000a, 2000b) analyzed these data using the fixed effects model developed in the next section. The estimator is that in (11-8). It is easy to see that groups with one observation will fall out of the computation, because if $T_i = 1$, then the observation equals the group mean. These data have been used by many researchers in similar panel data analyses.[7] Gravelle et al. (2002a) have strongly criticized these analyses, arguing that the WHO data are much more like a cross section than a panel data set.

From Example 6.22, the model used by the researchers at WHO was

$$\ln DALE_{it} = \alpha_i + \beta_1 \ln Health\ Expenditure_{it} + \beta_2 \ln Education_{it} + \beta_3 \ln^2 Education_{it} + \varepsilon_{it}.$$

Additional models were estimated using WHO's composite measure of health care attainment, *COMP*. The analysis of variance for a variable x_{it} is based on the decomposition

$$\sum_{i=1}^{n}\sum_{t=1}^{T_i}(x_{it} - \bar{\bar{x}})^2 = \sum_{i=1}^{n}\sum_{t=1}^{T_i}(x_{it} - \bar{x}_{i.})^2 + \sum_{t=1}^{n} T_i(\bar{x}_{i.} - \bar{\bar{x}})^2.$$

[6] See, for example, Judge et al. (1985).

[7] See, e.g., Greene (2004c) and several references.

TABLE 11.8 Analysis of Variance for WHO Data on Health Care Attainment

Variable	*Within-Groups Variation (%)*	*Between-Groups Variation (%)*
COMP	5.635	94.635
DALE	0.150	99.850
Expenditure	0.635	99.365
Education	0.177	99.823

Dividing both sides of the equation by the left-hand side produces the decomposition

$$1 = \textit{Within-groups proportion} + \textit{Between-groups proportion}.$$

The first term on the right-hand side is the within-group variation that differentiates a panel data set from a cross section (or simply multiple observations on the same variable). Table 11.8 lists the decomposition of the variation in the variables used in the WHO studies.

The results suggest the reasons for the authors' concern about the data. For all but COMP, virtually all the variation in the data is between groups—that is cross-sectional variation. As the authors argue, these data are only slightly different from a cross section.

11.4 THE FIXED EFFECTS MODEL

The fixed effects model arises from the assumption that the omitted effects, c_i, in the regression model of (11-1),

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + c_i + \varepsilon_{it}, i = 1, \ldots, n, t = 1, \ldots, T_i,$$
$$E[\varepsilon_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT_i}] = 0, \quad \textbf{(11-11)}$$
$$E[\varepsilon_{it}\varepsilon_{js}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT_i}] = \sigma_\varepsilon^2 \text{ if } i = j \text{ and } t = s \text{ and } = 0 \text{ if } i \neq j \text{ or } t \neq s,$$

can be arbitrarily correlated with the included variables. In a generic form,

$$E[c_i|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}] = E[c_i|\mathbf{X}_i] = h(\mathbf{X}_i). \quad \textbf{(11-12)}$$

We also assume that $\text{Var}[c_i|\mathbf{X}_i]$ is constant and all observations c_i and c_j are independent. We emphasize it is (11-12) that signifies the fixed effects model, not that any variable is fixed in this context and random elsewhere. The formulation implies that the heterogeneity across groups is captured in the constant term.[8] In (11-1), $\mathbf{z}_i = (1)$ and

$$y_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}.$$

Each α_i can be treated as an unknown parameter to be estimated.

11.4.1 LEAST SQUARES ESTIMATION

Let $\mathbf{y}_i$ and $\mathbf{X}_i$ be the T observations for the ith unit, let $\mathbf{i}$ be a $T \times 1$ column of ones, and let $\boldsymbol{\varepsilon}_i$ be the associated $T \times 1$ vector of disturbances.[9] Then,

$$\mathbf{y}_i = \mathbf{i}\alpha_i + \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i.$$

[8] It is also possible to allow the slopes to vary across i. A study on the topic is Cornwell and Schmidt (1984). We will examine this case in Section 11.4.6.

[9] The assumption of a fixed group size, T, at this point is purely for convenience. As noted in Section 11.2.4, the unbalanced case is a minor variation.

Collecting these terms gives

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_n \end{bmatrix} = \begin{bmatrix} \mathbf{i} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{i} & \cdots & \mathbf{0} \\ & & \vdots & \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{i} \end{bmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix} + \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{bmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{bmatrix}$$

or

$$\mathbf{y} = [\mathbf{d}_1 \quad \mathbf{d}_2, \ \ldots, \mathbf{d}_n \quad \mathbf{X}] \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} + \boldsymbol{\varepsilon},$$

where $\mathbf{d}_i$ is a dummy variable indicating the ith unit. Let the $nT \times n$ matrix $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \ \ldots, \mathbf{d}_n]$. Then, assembling all nT rows gives

$$\mathbf{y} = \mathbf{D}\,\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \qquad \textbf{(11-13)}$$

This model is occasionally referred to as the **least squares dummy variable (LSDV) model** (although the "least squares" part of the name refers to the technique usually used to estimate it, not to the model itself).

This model is a classical regression model, so no new results are needed to analyze it. If n is small enough, then the model can be estimated by ordinary least squares with K regressors in $\mathbf{X}$ and n columns in $\mathbf{D}$, as a multiple regression with $K + n$ parameters. Of course, if n is thousands, as is typical, then treating (11-13) as an ordinary regression will be extremely cumbersome. But, by using familiar results for a partitioned regression, we can reduce the size of the computation.[10] We write the least squares estimator of $\boldsymbol{\beta}$ as

$$\mathbf{b}_{\text{LSDV}} = [\mathbf{X}'\mathbf{M}_\mathbf{D}\mathbf{X}]^{-1}[\mathbf{X}'\mathbf{M}_\mathbf{D}\mathbf{y}] = \mathbf{b}^{within}, \qquad \textbf{(11-14)}$$

where

$$\mathbf{M}_\mathbf{D} = \mathbf{I}_{nT} - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'.$$

Because $\mathbf{M}_\mathbf{D}$ is symmetric and idempotent, $\mathbf{b}_{\text{LSDV}} = [(\mathbf{X}'\mathbf{M}_\mathbf{D})(\mathbf{M}_\mathbf{D}\mathbf{X})]^{-1}[(\mathbf{X}'\mathbf{M}_\mathbf{D})(\mathbf{M}_\mathbf{D}\mathbf{y})]$. This amounts to a least squares regression using the transformed data $\mathbf{M}_\mathbf{D}\mathbf{X} = \ddot{\mathbf{X}}$ and $\mathbf{M}_\mathbf{D}\mathbf{y} = \ddot{\mathbf{y}}$. The structure of $\mathbf{D}$ is particularly convenient; its columns are orthogonal, so

$$\mathbf{M}_\mathbf{D} = \begin{bmatrix} \mathbf{M}^0 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{M}^0 & \mathbf{0} & \cdots & \mathbf{0} \\ & & \cdots & & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{M}^0 \end{bmatrix}.$$

Each matrix on the diagonal is

$$\mathbf{M}^0 = \mathbf{I}_T - \frac{1}{T}\mathbf{i}\mathbf{i}'. \qquad \textbf{(11-15)}$$

Premultiplying any $T \times 1$ *vector* $\mathbf{z}_i$ by $\mathbf{M}^0$ creates $\mathbf{M}^0\mathbf{z}_i = \mathbf{z}_i - \bar{z}\mathbf{i}$. (Note that the mean is taken over only the T observations for unit i.) Therefore, the least squares regression of $\mathbf{M}_\mathbf{D}\mathbf{y} = \ddot{\mathbf{y}}$ on $\mathbf{M}_\mathbf{D}\mathbf{X} = \ddot{\mathbf{X}}$ is equivalent to a regression of $[y_{it} - \bar{y}_{i.}] = \ddot{y}_{it}$ on

[10] See Theorem 3.2.

$[\mathbf{x}_{it} - \overline{\mathbf{x}}_{i.}] = \ddot{\mathbf{x}}_{it}$, where $\overline{y}_{i.}$ and $\overline{\mathbf{x}}_{i.}$ are the scalar and $K \times 1$ vector of means of y_{it} and $\mathbf{x}_{it}$ over the T observations for group i.[11]

In terms of the within transformed data, then,

$$\begin{aligned}\mathbf{b}_{LSDV} &= \left[\sum_{i=1}^{n}(\mathbf{M}^0\mathbf{X}_i)'(\mathbf{M}^0\mathbf{X}_i)\right]^{-1}\left[\sum_{i=1}^{n}(\mathbf{M}^0\mathbf{X}_i)'(\mathbf{M}^0\mathbf{y}_i)\right]\\ &= \left[\sum_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right]^{-1}\left[\sum_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{y}}_i\right] \qquad \textbf{(11-16a)}\\ &= (\ddot{\mathbf{X}}'\ddot{\mathbf{X}})^{-1}\ddot{\mathbf{X}}'\ddot{\mathbf{y}}.\end{aligned}$$

The dummy variable coefficients can be recovered from the other normal equation in the partitioned regression,

$$\mathbf{D}'\mathbf{D}\mathbf{a} + \mathbf{D}'\mathbf{X}\mathbf{b}_{\text{LSDV}} = \mathbf{D}'\mathbf{y}$$

or

$$\mathbf{a} = [\mathbf{D}'\mathbf{D}]^{-1}\mathbf{D}'(\mathbf{y} - \mathbf{X}\mathbf{b}_{\text{LSDV}}).$$

This implies that for each i,

$$a_i = \overline{y}_{i.} - \mathbf{x}_{i.}'\mathbf{b}_{\text{LSDV}}. \qquad \textbf{(11-16b)}$$

The appropriate estimator of the asymptotic covariance matrix for $\mathbf{b}$ is

$$\text{Est.Asy.Var.}[\mathbf{b}_{\text{LSDV}}] = s^2\left[\sum_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right]^{-1}. \qquad \textbf{(11-17)}$$

Based on (11-14) and (11-16), the disturbance variance estimator is

$$\begin{aligned}s^2 = \frac{(\ddot{\mathbf{y}} - \ddot{\mathbf{X}}\mathbf{b})'(\ddot{\mathbf{y}} - \ddot{\mathbf{X}}\mathbf{b}_{\text{LSDV}})}{nT - n - K} &= \frac{\sum_{i=1}^{n}(\ddot{\mathbf{y}}_i - \ddot{\mathbf{X}}_i\mathbf{b}_{\text{LSDV}})'(\ddot{\mathbf{y}}_i - \ddot{\mathbf{X}}_i\mathbf{b}_{\text{LSDV}})}{nT - n - K} \qquad \textbf{(11-18)}\\ &= \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it} - \mathbf{x}_{it}'\mathbf{b}_{\text{LSDV}} - a_i)^2}{nT - n - K}.\end{aligned}$$

The itth residual used in this computation is

$$\begin{aligned}e_{it} = y_{it} - \mathbf{x}_{it}'\mathbf{b}_{\text{LSDV}} - a_i &= y_{it} - \mathbf{x}_{it}'\mathbf{b}_{\text{LSDV}} - (\overline{y}_{i.} - \overline{\mathbf{x}}_{i.}'\mathbf{b}_{\text{LSDV}})\\ &= (y_{it} - \overline{y}_{i.}) - (\mathbf{x}_{it} - \overline{\mathbf{x}}_{i.})'\,\mathbf{b}_{\text{LSDV}}.\end{aligned}$$

Thus, the numerator in s^2 is exactly the sum of squared residuals using the least squares slopes and the data in group mean deviation form. But, done in this fashion, one might then use $nT - K$ instead of $nT - n - K$ for the denominator in computing s^2, so a correction would be necessary.[12] For the individual effects,

[11] An interesting special case arises if $T = 2$. In the two-period case, you can show—we leave it as an exercise—that this least squares regression is done with nT first difference observations, by regressing observation $(y_{i2} - y_{i1})$ (and its negative) on $(\mathbf{x}_{i2} - \mathbf{x}_{i1})$ (and its negative).

[12] The maximum likelihood estimator of σ^2 for the fixed effects model with normally distributed disturbances is $\Sigma_i\Sigma_t e_{it}^2/nT$, with no degrees of freedom correction. This is a case in which the MLE is biased, given (11-18) which gives the unbiased estimator. This bias in the MLE for a fixed effects model is an example (actually, the first example) of the incidental parameters problem. [See Neyman and Scott (1948) and Lancaster (2000).] With a bit of manipulation it is clear that although the estimator is biased, if T increases asymptotically, then the bias eventually diminishes to zero. This is the signature feature of estimators that are affected by the incidental parameters problem.

$$\text{Asy.Var}[a_i] = \frac{\sigma_\varepsilon^2}{T} + \bar{\mathbf{x}}_{i.}{}'\{\text{Asy.Var}[\mathbf{b}]\}\bar{\mathbf{x}}_{i.}, \tag{11-19}$$

so a simple estimator based on s^2 can be computed.

With increasing n, the asymptotic variance of a_i declines to a lower bound of σ_ε^2/T *which does not converge to zero*. The constant term estimators in the fixed effects model are not consistent estimators of α_i. They are not inconsistent because they gravitate toward the wrong parameter. They are so because their asymptotic variances do not converge to zero, even as the sample size grows. It is easy to see why this is the case. We see that each a_i is estimated using only T observations—assume n were infinite, so that $\boldsymbol{\beta}$ were known. Because T is not assumed to be increasing, we have the surprising result. The constant terms are inconsistent unless $T \rightarrow \infty$, which is not part of the model.

We note a major shortcoming of the fixed effects approach. Any **time-invariant** variables in $\mathbf{x}_{it}$ will mimic the individual specific constant term. Consider the application of Example 11.3. We could write the fixed effects formulation as

$$\ln Wage_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + [\beta_{10}Ed_i + \beta_{11}Fem_i + \beta_{12}Blk_i + c_i] + \varepsilon_{it}.$$

The fixed effects formulation of the model will absorb the last four terms in the regression in α_i. The coefficients on the time-invariant variables cannot be estimated. For any $\mathbf{x}_k$ that is time invariant, every observation is the group mean, so $\mathbf{M}_\mathbf{D}\mathbf{x}_k = \ddot{\mathbf{x}}_k = 0$ so the corresponding column of $\ddot{\mathbf{X}}$ becomes a column of zeros and $(\ddot{\mathbf{X}}'\ddot{\mathbf{X}})^{-1}$ will not exist.

11.4.2 A ROBUST COVARIANCE MATRIX FOR $\mathbf{b}_{\text{LSDV}}$

The LSDV estimator is computed as

$$\mathbf{b}_{\text{LSDV}} = \left[\sum\nolimits_{\mathrm{i}=1}^{\mathrm{n}}\ddot{\mathbf{X}}_{\mathrm{i}}'\ddot{\mathbf{X}}_{\mathrm{i}}\right]^{-1}\left[\sum\nolimits_{\mathrm{i}=1}^{\mathrm{n}}\ddot{\mathbf{X}}_{\mathrm{i}}'\ddot{\mathbf{y}}_{\mathrm{i}}\right] = \boldsymbol{\beta} + \left[\sum\nolimits_{\mathrm{i}=1}^{\mathrm{n}}\ddot{\mathbf{X}}_{\mathrm{i}}'\ddot{\mathbf{X}}_{\mathrm{i}}\right]^{-1}\left[\sum\nolimits_{\mathrm{i}=1}^{\mathrm{n}}\ddot{\mathbf{X}}_{\mathrm{i}}'\ddot{\boldsymbol{\varepsilon}}_{\mathrm{i}}\right]. \tag{11-20}$$

The asymptotic covariance matrix for the estimator derives from

$$\text{Var}[(\mathbf{b}_{\text{LSDV}} - \boldsymbol{\beta})|\mathbf{X}] = \left[\sum\nolimits_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right]^{-1}E\left\{\left[\sum\nolimits_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\boldsymbol{\varepsilon}}_i\right]\left[\sum\nolimits_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\boldsymbol{\varepsilon}}_i\right]'|\mathbf{X}\right\}\left[\sum\nolimits_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right]^{-1}.$$

The center matrix is a double sum over $i,j = 1, \ldots, n$, but terms with $i \neq j$ are independent and have expectation zero, so the matrix is

$$E\left\{\left[\sum\nolimits_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\boldsymbol{\varepsilon}}_i\right]\left[\sum\nolimits_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\boldsymbol{\varepsilon}}_i\right]'|\mathbf{X}\right\} = E\left\{\left[\sum\nolimits_{i=1}^{n}(\ddot{\mathbf{X}}_i\ddot{\boldsymbol{\varepsilon}}_i)(\ddot{\boldsymbol{\varepsilon}}_i'\ddot{\mathbf{X}}_i)\right]|\mathbf{X}\right\}.$$

Each term in the sum is $(\ddot{\mathbf{X}}_i'\ddot{\boldsymbol{\varepsilon}}_i)(\ddot{\boldsymbol{\varepsilon}}_i'\ddot{\mathbf{X}}_i) = (\mathbf{X}_i'\mathbf{M}^0\mathbf{M}^0\boldsymbol{\varepsilon}_i)(\boldsymbol{\varepsilon}_i'\mathbf{M}^0\mathbf{M}^0\mathbf{X}_i')$. But $\mathbf{M}^0$ is idempotent, so $\ddot{\mathbf{X}}_i'\ddot{\boldsymbol{\varepsilon}}_i = \ddot{\mathbf{X}}_i\boldsymbol{\varepsilon}_i$, and we have assumed that $E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'|\mathbf{X}] = \sigma_\varepsilon{}^2\mathbf{I}$. Collecting the terms,

$$\begin{aligned}
\text{Var}[(\mathbf{b}_{\text{LSDV}} - \boldsymbol{\beta})|\mathbf{X}] &= \left[\sum\nolimits_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right]^{-1}E\left\{\left[\sum\nolimits_{i=1}^{n}\ddot{\mathbf{X}}_i'\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'\ddot{\mathbf{X}}_i\right]|\mathbf{X}\right\}\left[\sum\nolimits_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right]^{-1} \\
&= \left[\sum\nolimits_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right]^{-1}\left\{\left[\sum\nolimits_{i=1}^{n}\ddot{\mathbf{X}}_i'(\sigma_\varepsilon^2\mathbf{I})\ddot{\mathbf{X}}_i\right]\right\}\left[\sum\nolimits_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right]^{-1} \\
&= \sigma_\varepsilon^2\left[\sum\nolimits_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right]^{-1},
\end{aligned}$$

which produces the estimator in (11-17). If the disturbances in (11-11) are heteroscedastic and/or autocorrelated, then $E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'|\mathbf{X}] \neq \sigma_\varepsilon{}^2\mathbf{I}$. A robust counterpart to (11-4) would be

$$\text{Est.Asy.Var}[\mathbf{b}_{\text{LSDV}}] = \left[\sum\nolimits_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right]^{-1}\left\{\left[\sum\nolimits_{i=1}^{n}(\ddot{\mathbf{X}}_i'\mathbf{e}_i)(\mathbf{e}_i'\ddot{\mathbf{X}}_i)\right]\right\}\left[\sum\nolimits_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right]^{-1}, \quad \textbf{(11-21)}$$

where e_{it} is the residual shown after (11-18). Note that using $\ddot{e}_{it}$ in this calculation gives exactly the same result because $\bar{e}_{i.} = 0.$ [13]

11.4.3 TESTING THE SIGNIFICANCE OF THE GROUP EFFECTS

The t ratio for a_i can be used for a test of the hypothesis that α_i equals zero. This hypothesis about one specific group, however, is typically not useful for testing in this regression context. If we are interested in differences across groups, then we can test the hypothesis that the constant terms are all equal with an F test. Under the null hypothesis of equality, the efficient estimator is pooled least squares. The F ratio used for this test is

$$F(n-1, nT-n-K) = \frac{(R^2_{LSDV} - R^2_{Pooled})/(n-1)}{(1-R^2_{LSDV})/(nT-n-K)},$$

where *LSDV* indicates the dummy variable model and *Pooled* indicates the pooled or restricted model with only a single overall constant term. Alternatively, the model may have been estimated with an overall constant and $n - 1$ dummy variables instead. All other results (i.e., the least squares slopes, s^2, R^2) will be unchanged, but rather than estimate α_i, each dummy variable coefficient will now be an estimate of $\alpha_i - \alpha_1$ where group "1" is the omitted group. The F test that the coefficients on these $n - 1$ dummy variables are zero is identical to the one above. It is important to keep in mind, however, that although the statistical results are the same, the interpretation of the dummy variable coefficients in the two formulations is different.[14]

Example 11.7 Fixed Effects Estimates of a Wage Equation

We continue Example 11.4 by computing the fixed effects estimates of the wage equation, now

$$\begin{aligned}\ln Wage_{it} = {} & \alpha_i + \beta_2\, Exp_{it} + \beta_3\, Exp^2_{it} + \beta_4\, Wks_{it} + \beta_5\, Occ_{it} \\ & + \beta_6\, Ind_{it} + \beta_7\, South_{it} + \beta_8\, SMSA_{it} + \beta_9\, MS_{it} \\ & + \beta_{10}\, Union_{it} + 0 \times Ed_i + 0 \times Fem_i + 0 \times Blk_i + \varepsilon_{it}.\end{aligned}$$

Because *Ed, Fem*, and *Blk* are time invariant, their coefficients will not be estimable, and will be set to zero. The OLS and fixed effects estimates are presented in Table 11.9. Each is accompanied by the conventional standard errors and the robust standard errors. We note, first, the rather large change in the parameters that occurs when the fixed effects specification is used. Even some statistically significant coefficients in the least squares results change sign in the fixed effects results. Likewise, the robust standard errors are characteristically much larger than the conventional counterparts. The fixed effects standard errors increased

[13] See Arellano (1987) and Arellano and Bover (1995).

[14] The F statistic can also be based on the sum of squared residuals rather than the R^{2}s, [See (5-29) and (5-30).] In this connection, we note that the software package Stata contains two estimators for the fixed effects linear regression, `areg` and `xtreg`. In computing the former, Stata uses $\Sigma_i\Sigma_t(y_{tt} - \bar{\bar{y}})^2$ as the denominator, as it would in computing the counterpart for the constrained regression. But `xtreg` (which is the procedure typically used) uses $\Sigma_i\Sigma_t(y_{it} - \bar{y}_i)^2$, which is smaller. The R^2 produced by `xtreg` will be smaller, as will be the F statistic, possibly substantially so.

TABLE 11.9 Wage Equation Estimated by OLS and LSDV

	Pooled OLS			*Fixed Effects LSDV*		
Variable	*Least Squares Estimate*	*Standard Error*	*Clustered Std. Error*	*Fixed Effects Estimates*	*Standard Error*	*Robust Std. Error*
R^2	0.42861			0.90724		
Constant	5.25112	0.07129	0.12355	—	—	—
Exp	0.00401	0.00216	0.00408	0.11321	0.00247	0.00438
ExpSq	−0.00067	0.00005	0.00009	−0.00042	0.00006	0.00009
Wks	0.00422	0.00108	0.00154	0.00084	0.00060	0.00094
Occ	−0.14001	0.01466	0.02724	−0.02148	0.01379	0.02053
Ind	0.04679	0.01179	0.02366	0.01921	0.01545	0.02451
South	−0.05564	0.01253	0.02616	−0.00186	0.03431	0.09650
SMSA	0.15167	0.01207	0.02410	−0.04247	0.01944	0.03186
MS	0.04845	0.02057	0.04094	−0.02973	0.01899	0.02904
Union	0.09263	0.01280	0.02367	0.03278	0.01493	0.02709
Ed	0.05670	0.00261	0.00556	—	—	—
Fem	−0.36779	0.02510	0.04557	—	—	—
Blk	−0.16694	0.02204	0.04433	—	—	—

more than might have been expected, given that heteroscedasticity is not a major issue, but a source of autocorrelation is in the equation (as the fixed effects). The large changes suggest that there may yet be some additional, unstructured correlation remaining in ε_{it}. The test for the presence of the fixed effects is based on

$$F = [(0.90724 - 0.42861)/594]/[(1 - 0.90724)/(4165 - 595 - 9)] = 30.933.$$

The critical value from the F table would be less than 1.3, so the hypothesis of homogeneity is rejected.

11.4.4 FIXED TIME AND GROUP EFFECTS

The least squares dummy variable approach can be extended to include a time-specific effect as well. One way to formulate the extended model is simply to add the time effect, as in

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \delta_t + \varepsilon_{it}. \quad \textbf{(11-22)}$$

This model is obtained from the preceding one by the inclusion of an additional $T - 1$ dummy variables. (One of the time effects must be dropped to avoid perfect collinearity—the group effects and time effects both sum to one.) If the number of variables is too large to handle by ordinary regression, then this model can also be estimated by using the partitioned regression. There is an asymmetry in this formulation, however, because each of the group effects is a group-specific intercept, whereas the time effects are **contrasts**—that is, comparisons to a base period (the one that is excluded). A symmetric form of the model is

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \mu + \alpha_i + \delta_t + \varepsilon_{it}, \quad \textbf{(11-23)}$$

where a full n and T effects are included, but the restrictions

$$\sum_i \alpha_i = \sum_t \delta_t = 0$$

are imposed. Least squares estimates of the slopes in this model are obtained by regression of

$$\begin{aligned} & y_{*it} = y_{it} - \overline{y}_{i.} - \overline{y}_{.t} + \overline{\overline{y}} \\ \text{on} \quad & \mathbf{x}_{*it} = \mathbf{x}_{it} - \overline{\mathbf{x}}_{i.} - \overline{\mathbf{x}}_{.t} + \overline{\overline{\mathbf{x}}}, \end{aligned} \qquad \textbf{(11-24)}$$

where the period-specific and overall means are

$$\overline{y}_{.t} = \frac{1}{n}\sum_{i=1}^{n} y_{it} \qquad \text{and} \qquad \overline{\overline{y}} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} y_{it},$$

and likewise for $\overline{\mathbf{x}}_{.t}$ and $\overline{\overline{\mathbf{x}}}$. The overall constant and the dummy variable coefficients can then be recovered from the normal equations as

$$\begin{aligned} \hat{\mu} &= m = \overline{\overline{y}} - \overline{\overline{\mathbf{x}}}'\mathbf{b}, \\ \hat{\alpha}_i &= a_i = (\overline{y}_{i.} - \overline{\overline{y}}) - (\overline{\mathbf{x}}_{i.} - \overline{\overline{\mathbf{x}}})'\mathbf{b}, \\ \hat{\delta}_t &= d_t = (\overline{y}_{.t} - \overline{\overline{y}}) - (\overline{\mathbf{x}}_{.t} - \overline{\overline{\mathbf{x}}})'\mathbf{b}. \end{aligned} \qquad \textbf{(11-25)}$$

The estimator of the asymptotic covariance matrix for $\mathbf{b}$ is computed using the sums of squares and cross products of $\mathbf{x}_{*it}$ computed in (11-24) and

$$s^2 = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it} - \mathbf{x}_{it}'\mathbf{b} - m - a_i - d_t)^2}{nT - (n-1) - (T-1) - K - 1}. \qquad \textbf{(11-26)}$$

The algebra of the two-way fixed effects estimator is rather complex—see, for example, Baltagi (2014). It is not obvious from the presentation so far, but the template result in (11-24) is incorrect if the panel is unbalanced. Unfortunately, for the unwary, the result does not fail in a way that would make the mistake obvious; if the panel is unbalanced, (11-24) simply leads to the wrong answer, but one that could look right. A numerical example is shown in Example 11.8. The conclusion for the practitioner is that (11-24) should only be used with balanced panels, but the augmented one-way estimator can be used in all cases.

Example 11.8 Two-Way Fixed Effects with Unbalanced Panel Data

The following experiment is done with the Cornwell and Rupert data used in Examples 11.4, 11.5, and 11.7. There are 595 individuals and 7 periods. Each group is 7 observations. Based on the balanced panel using all 595 individuals, in the fixed effects regression of ln *Wage* on just *Wks*, both methods give the answer $b = 0.00095$. If the first 300 groups are shortened by dropping the last 3 years of data, the unbalanced panel now has 300 groups with $T = 4$ and 295 with $T = 7$. For the same regression, the one-way estimate with time dummy variables is 0.00050 but the template result in (11-24) (which is incorrect) gives 0.00283.

11.4.5 REINTERPRETING THE WITHIN ESTIMATOR: INSTRUMENTAL VARIABLES AND CONTROL FUNCTIONS

The fixed effects model, in basic form, is

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + (c_i + \varepsilon_{it}).$$

We once again first consider least squares estimation. As we have already noted, for this case, $\mathbf{b}_{\text{OLS}}$ is inconsistent because of the correlation between $\mathbf{x}_{it}$ and c_i. Therefore, in the absence of the dummy variables, $\mathbf{x}_{it}$ is endogenous in this model. We used the within estimator in Section 11.4.1 instead of least squares to remedy the problem. The LSDV estimator is

$$\mathbf{b}_{\text{LSDV}} = (\ddot{\mathbf{X}}'\ddot{\mathbf{X}})^{-1}\ddot{\mathbf{X}}'\ddot{\mathbf{y}}.$$

The LSDV estimator is computed by regressing $\mathbf{y}$ transformed to deviations from group means on the same transformation of $\mathbf{X}$; that is, $\mathbf{M_D y}$ on $\mathbf{M_D X}$. But, because $\mathbf{M_D}$ is idempotent, we may also write $\mathbf{b}_{LSDV} = (\ddot{\mathbf{X}}'\mathbf{X})^{-1}\ddot{\mathbf{X}}'\mathbf{y}$. In this form, $\ddot{\mathbf{X}}$ appears to be a set of instrumental variables, precisely in the form of (8-6). We have already demonstrated the consistency of the estimator, though it remains to verify the exogeneity and relevance conditions. These are both straightforward to verify. For the exogeneity condition, let $\mathbf{c}$ denote the full set of common effects. By construction, $(1/nT)\ddot{\mathbf{X}}'\mathbf{c} = \mathbf{0}$. We have assumed at the outset that $\text{plim}(1/nT)\mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{0}$. We need $\text{plim}(1/nT)\mathbf{X}'\mathbf{M_D}\boldsymbol{\varepsilon} = \text{plim}(1/nT)\mathbf{X}'(\mathbf{M_D}\boldsymbol{\varepsilon})$. If $\mathbf{X}$ is uncorrelated with ε, it will be uncorrelated with ε in deviations from its group means. For the relevance condition, all that will be needed is full rank of $(1/nT)\ddot{\mathbf{X}}'\mathbf{X}$, which is equivalent to $(1/nT)(\ddot{\mathbf{X}}'\ddot{\mathbf{X}})$. This matrix will have full rank so long as no variables in $\mathbf{X}$ are time invariant—note that $(\ddot{\mathbf{X}}'\ddot{\mathbf{X}})^{-1}$ is used to compute $\mathbf{b}_{\text{LSDV}}$. The conclusion is that the data in group mean deviations form, that is, $\ddot{\mathbf{X}}$, are valid instrumental variables for estimation of the fixed effects model. This useful result will reappear when we examine Hausman and Taylor's model in Section 11.8.2.

We continue to assume that there are no time-invariant variables in $\mathbf{X}$. The matrix of group means is obtained as $\mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'\mathbf{X} = \mathbf{P_D X} = (\mathbf{I} - \mathbf{M_D})\mathbf{X}$. [See (11-14)–(11-17).] Consider, then, least squares regression of $\mathbf{y}$ on $\mathbf{X}$ and $\mathbf{P_D X}$, that is, on $\mathbf{X}$ and the group means, $\overline{\mathbf{X}}$. Using the partitioned regression formulation [Theorem 3.2 and (3-19)], we find this estimator of $\boldsymbol{\beta}$ is

$$\begin{aligned}\mathbf{b}_{\text{Mundlak}} &= (\mathbf{X}'\mathbf{M_{PX}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M_{PX}}\mathbf{y}\\ &= \{\mathbf{X}'[\mathbf{I} - \overline{\mathbf{X}}(\overline{\mathbf{X}}'\overline{\mathbf{X}})^{-1}\overline{\mathbf{X}}']\mathbf{X}\}^{-1} \times \{\mathbf{X}'[\mathbf{I} - \overline{\mathbf{X}}(\overline{\mathbf{X}}'\overline{\mathbf{X}})^{-1}\overline{\mathbf{X}}']\mathbf{y}\}.\end{aligned}$$

This simplifies considerably. Recall $\overline{\mathbf{X}} = \mathbf{P_D X}$ and $\mathbf{P_D}$ is idempotent. We expand the first matrix in braces.

$$\begin{aligned}\{\mathbf{X}'[\mathbf{I} - (\mathbf{P_D X})[(\mathbf{P_D X})'(\mathbf{P_D X})]^{-1}(\mathbf{P_D X})']\mathbf{X}\} &= \mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{P_D X}[\mathbf{X}'\mathbf{P}_\mathbf{D}'\mathbf{P_D X}]^{-1}\mathbf{X}'\mathbf{P}_\mathbf{D}'\mathbf{X}\\ &= \mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{P_D X}\\ &= \mathbf{X}'[\mathbf{I} - \mathbf{P_D}]\mathbf{X}\\ &= \mathbf{X}'\mathbf{M_D X}.\end{aligned}$$

The same result will emerge for the second term, which implies that the coefficients on $\mathbf{X}$ in the regression of $\mathbf{y}$ on $(\mathbf{X}, \overline{\mathbf{X}})$ is the within estimator, $\mathbf{b}_{\text{LSDV}}$. So, the group means qualify as a control function, as defined in Section 8.4.2. This useful insight makes the Mundlak approach a very useful method of dealing with fixed effects in regression, and by extension, in many other settings that appear in the literature.

11.4.6 PARAMETER HETEROGENEITY

With a small change in notation, the common effects model in (11-1) becomes

$$\begin{aligned} y_{it} &= c_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} \\ &= (\alpha + u_i) + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} \\ &= \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \end{aligned}$$

where $E[u_i] = 0$ and $E[\alpha_i] = \alpha$. The heterogeneity affects the constant term. We can extend the model to allow other parameters to be heterogeneous as well. In the labor market model examined in Example 11.3, an extension in which the partial effect of weeks worked depends on both market and individual characteristics, might appear as

$$\begin{aligned} \ln Wage_{it} = \alpha_{i1} &+ \alpha_{i2}\, Wks_{it} + \beta_2\, Exp_{it} + \beta_3\, Exp_{it}^2 + \beta_5\, Occ_{it} \\ &+ \beta_6\, Ind_{it} + \beta_7\, South_{it} + \beta_8\, SMSA_{it} + \beta_9\, MS_{it} \\ &+ \beta_{10}\, Union_{it} + \beta_{11}\, Ed_i + \beta_{12}\, Fem_i + \beta_{13}\, Blk_i + \varepsilon_{it} \end{aligned}$$

$$\boldsymbol{\alpha}_i = \begin{pmatrix} \alpha_{i1} \\ \alpha_{i2} \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} u_{i1} \\ u_{i2} \end{pmatrix} = \boldsymbol{\alpha} + \mathbf{u}_i.$$

Another interesting case is a random trend model, $y_{it} = \alpha_{i1} + \alpha_{i2}t + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}$. As before, the difference between the random and fixed effects models is whether $E[\mathbf{u}_i | \mathbf{X}_i]$ is zero or not. For the present, we will allow this to be nonzero—a fixed effects form of the model.

The preceding developments have been concerned with a strategy for estimation and inference about $\boldsymbol{\beta}$ in the presence of $\mathbf{u}_i$. In this fixed effects setting, the dummy variable approach of Section 11.4.1 can be extended essentially with only a small change in notation. First, let's generalize the model slightly,

$$y_{it} = \mathbf{z}_{it}'\boldsymbol{\alpha}_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}.$$

In the basic common effects model, $\mathbf{z}_{it}' = (1)$; in the random trend model, $\mathbf{z}_{it}' = (1,t)$; in the suggested extension of the labor market model, $\mathbf{z}_{it}' = (1, Wks_{it})$, with $E[\mathbf{u}_i | \mathbf{X}_i, \mathbf{Z}_i] \neq \mathbf{0}$ (fixed effects) and $E[\mathbf{u}_i\mathbf{u}_i' | \mathbf{X}_i, \mathbf{Z}_i] = \boldsymbol{\Sigma}$, a constant, positive definite matrix. The strict exogeneity assumption now is $E[\varepsilon_{it} | \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, \mathbf{z}_{i1}, \ldots, \mathbf{z}_{iT}, \mathbf{u}_i] = 0$. For the present, we assume ε_{it} is homoscedastic and nonautocorrelated, so $E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i' | \mathbf{X}_i, \mathbf{Z}_i, \mathbf{u}_i] = \sigma_\varepsilon{}^2\mathbf{I}$. We can approach estimation of $\boldsymbol{\beta}$ the same way we did in Section 11.4.1. Recall the LSDV estimator is based on

$$\mathbf{y} = \mathbf{D}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{11-27}$$

where $\mathbf{D}$ is the $nT \times n$ matrix of individual specific dummy variables. The estimator of $\boldsymbol{\beta}$ is

$$\begin{aligned} \mathbf{b}_{\text{LSDV}} &= (\mathbf{X}'\mathbf{M}_\mathbf{D}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_\mathbf{D}\mathbf{y} = \left[\textstyle\sum_{i=1}^{n} \ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i \right]^{-1} \left[\textstyle\sum_{i=1}^{n} \ddot{\mathbf{X}}_i'\ddot{\mathbf{y}}_i \right], \\ \mathbf{a}_{\text{LSDV}} &= (\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'(\mathbf{y} - \mathbf{X}\text{b}_{\text{LSDV}}) = \overline{\mathbf{y}} - \overline{\mathbf{X}}\mathbf{b}_{\text{LSDV}}, \\ \mathbf{M}_\mathbf{D} &= \mathbf{I} - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'. \end{aligned}$$

The special structure of $\mathbf{D}$—the columns are orthogonal—allows the calculations to be done with two convenient steps: (1) compute $\mathbf{b}_{\text{LSDV}}$ by regression of $(y_{it} - \bar{y}_{i.})$ on $(\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})$; (2) compute a_i as $(1/T)\Sigma_t(y_{it} - \mathbf{x}_{it}'\text{b}_{\text{LSDV}})$.

No new results are needed to develop the fixed effects estimator in the extended model. In this specification, we have simply respecified $\mathbf{D}$ to contain two or more sets of N columns. For the time trend case, for example, define an $nT \times 1$ column vector of time trends, $\mathbf{t}^{*\prime} = (1, 2, \ldots, T, 1, 2, \ldots, T, \ldots, 1, 2, \ldots, T)$. Then, $\mathbf{D}$ has $2N$ columns, $\{[\mathbf{d}_1, \ldots, \mathbf{d}_n], [\mathbf{d}_1 \circ \mathbf{t}^*, \mathbf{d}_2 \circ \mathbf{t}^* \ldots, d_{\text{n}} \circ \mathbf{t}^*]\}$. This is an $nT \times 2n$ matrix of dummy variables and interactions of the dummy variables with the time trend. (The operation $\mathbf{d}_i \circ \mathbf{t}^*$ is the Hadamard product—element by element multiplication—of $\mathbf{d}_i$ and $\mathbf{t}^*$.) With $\mathbf{D}$ redefined this way, the results in Section 11.4.1 can be applied as before. For example, for the random trends model, $\ddot{\mathbf{X}}_i$ is obtained by "detrending" the columns of $\mathbf{X}_i$. Define $\mathbf{Z}_i$ to be the $T \times 2$ matrix $(\mathbf{1},\mathbf{t})$. Then, for individual i, the block of data in $\ddot{\mathbf{X}}_i$ is $[\mathbf{I} - \mathbf{Z}_i(\mathbf{Z}_i'\mathbf{Z}_i)^{-1}\mathbf{Z}_i']\mathbf{X}_i$ and $\mathbf{b}_{\text{LSDV}}$ is computed using (11-20). (Note that this requires that T be at least $J + 1$ where J is the number of variables in $\mathbf{Z}$. In the simpler fixed effects case, we require at least two observations in group i. Here, in the random trend model, that would be three observations.) In computing s^2, the appropriate degrees of freedom will be $(n(T - J) - K)$. The asymptotic covariance matrices in (11-17) and (11-21) are computed as before.[15] For each group, $\mathbf{a}_i = (\mathbf{Z}_i'\mathbf{Z}_i)^{-1}\mathbf{Z}_i'(\mathbf{y}_i - \mathbf{X}_i\mathbf{b}_{\text{LSDV}})$. The natural estimator of $\alpha = E[\alpha_i]$ would be $\bar{\mathbf{a}} = \frac{1}{n}\Sigma_{i=1}^{n}\mathbf{a}_i$. The asymptotic variance matrix for $\bar{\mathbf{a}}$ can be estimated with

$$\text{Est.Asy.Var}[\bar{\mathbf{a}}] = (1/n^2)\Sigma_i\mathbf{f}_i\mathbf{f}_i' \text{ where } \mathbf{f}_i = [(\mathbf{a}_i - \bar{\mathbf{a}}) - \mathbf{C}\mathbf{A}^{-1}\ddot{\mathbf{X}}_i'\mathbf{e}_i],\ \mathbf{A} = \tfrac{1}{n}\Sigma_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i \text{ and}$$
$$\mathbf{C} = \tfrac{1}{n}\Sigma_{i=1}^{n}(\mathbf{Z}_i'\mathbf{Z}_i)^{-1}\mathbf{Z}_i'\mathbf{X}_i.$$

[See Wooldridge (2010, p. 381).]

Example 11.9 Heterogeneity in Time Trends in an Aggregate Production Function

We extend Munnell's (1990) proposed model of productivity of public capital at the state level that was estimated in Example 10.1. The central equation of the analysis that we will extend here is a Cobb–Douglas production function,

$$\begin{aligned}\ln gsp_{it} = \alpha_{i1} + \alpha_{i2}t + \beta_1 \ln pc_{it} + \beta_2 \ln hwy_{it} + \beta_3 \ln water_{it}\\ + \beta_4 \ln util_{it} + \beta_5 \ln emp_{it} + \beta_6\, unemp_{it} + \varepsilon_{it},\end{aligned}$$

where	*gsp*	= gross state product,
	pc	= private capital,
	hwy	= highway capital,
	water	= water utility capital,
	util	= utility capital,
	emp	= employment (labor),
	unemp	= unemployment rate.

The data, measured for the lower 48 U.S. states (excluding Alaska and Hawaii) and years 1970–1986, are given in Appendix Table F10.1. Table 11.10 reports estimates of the several

[15] The random trends model is a special case that can be handled by differences rather than the partitioned regression method used here. In $y_{it} = \alpha_{i1} + \alpha_{i2}t + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}$, $(y_{it} - y_{i,t-1}) = \Delta y_{it} = \alpha_{i2} + (\Delta\mathbf{x}_{it})'\boldsymbol{\beta} + \Delta\varepsilon_{it}$. The time trend becomes the common effect. This can be treated as a fixed effects model. Or, taking a second difference, $\Delta y_{it} - \Delta y_{i,t-1} = \Delta^2 y_{it}$ removes α_{i2} and leaves a linear regression, $\Delta^2 y_{it} = \Delta^2\mathbf{x}_{it}'\boldsymbol{\beta} + \Delta^2\varepsilon_{it}$. Details are given in Wooldridge (2010, pp. 375–377).

TABLE 11.10 Estimates of Fixed Effects Statewide Production Functions

	Constant	*Trend*	ln *PC*	ln *Hwy*	ln *Water*	ln *Util*	ln *Emp*	*Unemp*	R^2
Pooled Model	1.91618	0.00108	0.30669	0.06708	0.11607	0.01054	0.54838	−0.00812	0.99307
Std.Error	0.05287	0.00072	0.01163	0.01633	0.01246	0.01241	0.01555	0.00149	
Robust S.E.	0.21420	0.00162	0.05136	0.05881	0.03481	0.04138	0.06825	0.00341	
Fixed Effects		0.00625	0.13751	0.08529	0.02966	−0.09807	0.75870	−0.00732	0.99888
Std.Error		0.00080	0.02814	0.03010	0.01574	0.01760	0.02915	0.00098	
Robust S.E.		0.00179	0.08248	0.08878	0.04147	0.05672	0.08744	0.00226	
Random Trend			0.11228	−0.01120	−0.03181	−0.08828	0.71207	−0.00793	0.99953
Std.Error	5.41942	0.01108	0.02645	0.03900	0.0166	0.02339	0.03105	0.00083	
Robust S.E	0.54657	0.00207	0.04189	0.07554	0.03110	0.04655	0.04803	0.00123	
Difference			−0.09104	−0.20767	−0.00094	0.12013	0.94832	−0.00374	
Std.Error			0.02324	0.10785	0.03582	0.05647	0.05312	0.00080	
Newey–West(2)			0.03746	0.14590	0.04343	0.07181	0.06576	0.00095	
Robust S.E.			0.04129	0.14358	0.03635	0.07552	0.06920	0.00112	

fixed effects models. The pooled estimator is computed using simple least squares for all 816 observations. The standard errors use $s^2(\mathbf{X}'\mathbf{X})^{-1}$. The robust standard errors are based on (11-4). For the two fixed effects models, the standard errors are based on (11–17) and (11–21). Finally, for the difference estimator, two sets of robust standard errors are computed. The Newey–West estimator assumes that ε_{it} in the model is homoscedastic so that $\Delta^2\varepsilon_{it} = \varepsilon_{it} - 2\varepsilon_{i,t-1} + \varepsilon_{i,t-2}$. The robust standard errors are based, once again, on (11-4). Note that two observations have been dropped from each state with the second difference estimator. The patterns of the standard errors are predictable. They all rise substantially with the correction for clustering, in spite of the presence of the fixed effects. The effect is quite substantial, with most of the standard errors rising by a factor of 2 to 4. The Newey–West correction (see Section 20.5.2) of the difference estimators seems mostly to cover the effect of the autocorrelation. The *F* test for the hypothesis that neither the constant nor the trend are heterogeneous is F[94,816-96-6] = [(0.99953 − 0.99307)/94]/[(1 − 0.99953)/(816 − 96 − 6)] = 104.40. The critical value from the *F* table is 1.273, so the hypothesis of homogeneity is rejected. The differences in the estimated parameters across the specifications are also quite pronounced. The difference between the random trend and difference estimators is striking, given that these are two different estimation approaches to the same model.

11.5 RANDOM EFFECTS

The fixed effects model allows the unobserved individual effects to be correlated with the included variables. We then modeled the differences between units as parametric shifts of the regression function. This model might be viewed as applying only to the cross-sectional units in the study, not to additional ones outside the sample. For example, an intercountry comparison may well include the full set of countries for which it is reasonable to assume that the model is constant. Example 6.5 is based on a panel consisting of data on 31 baseball teams. Save for rare discrete changes in the league, these 31 units will always be the entire population. If the individual effects are strictly uncorrelated with the regressors, then it might be appropriate to model the individual specific constant terms as randomly distributed across cross-sectional units. This view would be appropriate if we believed that sampled cross-sectional units were drawn from a large population. It would certainly be the case for the longitudinal data sets listed in the introduction to this chapter and for the labor market data we have used in several examples in this chapter.[16]

The payoff to this form is that it greatly reduces the number of parameters to be estimated. The cost is the possibility of inconsistent estimators, if the assumption is inappropriate.

Consider, then, a reformulation of the model,

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + (\alpha + u_i) + \varepsilon_{it}, \tag{11-28}$$

where there are K regressors including a constant and now the single constant term is the mean of the unobserved heterogeneity, $E[\mathbf{z}_i'\boldsymbol{\alpha}]$. The component u_i is the random heterogeneity specific to the ith observation and is constant through time; recall from Section 11.2.1, $u_i = \{\mathbf{z}_i'\boldsymbol{\alpha} - E[\mathbf{z}_i'\boldsymbol{\alpha}]\}$. For example, in an analysis of families, we can view

[16] This distinction is not hard and fast; it is purely heuristic. We shall return to this issue later. See Mundlak (1978) for a methodological discussion of the distinction between fixed and random effects.

u_i as the collection of factors, $\mathbf{z}_i'\boldsymbol{\alpha}$, not in the regression that are specific to that family. We continue to assume strict exogeneity:

$$\begin{aligned} E[\varepsilon_{it}|\mathbf{X}_i] &= E[u_i|\mathbf{X}_i] = 0, \\ E[\varepsilon_{it}^2|\mathbf{X}_i] &= \sigma_\varepsilon^2, \\ E[u_i^2|\mathbf{X}_i] &= \sigma_u^2, \\ E[\varepsilon_{it}u_j|\mathbf{X}_i] &= 0 \quad \text{for all } i, t, \text{ and } j, \\ E[\varepsilon_{it}\varepsilon_{js}|\mathbf{X}_i] &= 0 \quad \text{if } t \neq s \text{ or } i \neq j, \\ E[u_iu_j|\mathbf{X}_i, \mathbf{X}_j] &= 0 \quad \text{if } i \neq j. \end{aligned} \tag{11-29}$$

As before, it is useful to view the formulation of the model in blocks of T observations for group i, $\mathbf{y}_i$, $\mathbf{X}_i$, $u_i\mathbf{i}$, and $\boldsymbol{\varepsilon}_i$. For these T observations, let

$$\eta_{it} = \varepsilon_{it} + u_i$$

and

$$\boldsymbol{\eta}_i = [\eta_{i1}, \eta_{i2}, \ldots, \eta_{iT}]'.$$

In view of this form of $\boldsymbol{\eta}_{it}$, we have what is often called an **error components model**. For this model,

$$\begin{aligned} E[\eta_{it}^2|\mathbf{X}_i] &= \sigma_\varepsilon^2 + \sigma_u^2, \\ E[\eta_{it}\eta_{is}|\mathbf{X}_i] &= \sigma_u^2, \quad t \neq s, \\ E[\eta_{it}\eta_{js}|\mathbf{X}_i] &= 0 \qquad \text{for all } t \text{ and } s, \text{ if } i \neq j. \end{aligned} \tag{11-30}$$

For the T observations for unit i, let $\boldsymbol{\Sigma} = E[\boldsymbol{\eta}_i, \boldsymbol{\eta}_i'|\mathbf{X}]$. Then

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_\varepsilon^2 + \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ \sigma_u^2 & \sigma_\varepsilon^2 + \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ & \cdots & & & \\ \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_\varepsilon^2 + \sigma_u^2 \end{bmatrix} = \sigma_\varepsilon^2\,\mathbf{I}_T + \sigma_u^2\mathbf{i}_T\mathbf{i}_T', \tag{11-31}$$

where $\mathbf{i}_T$ is a $T \times 1$ column vector of 1s. Because observations i and j are independent, the disturbance covariance matrix for the full nT observations is

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} & \mathbf{0} & \cdots & \mathbf{0} \\ & & & \vdots & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma} \end{bmatrix} = \mathbf{I}_n \otimes \boldsymbol{\Sigma}. \tag{11-32}$$

11.5.1 LEAST SQUARES ESTIMATION

The model defined by (11-28),

$$y_{it} = \alpha + \mathbf{x}_{it}'\boldsymbol{\beta} + u_i + \varepsilon_{it},$$

with the strict exogeneity assumptions in (11-29) and the covariance matrix detailed in (11-31) and (11-32), is a generalized regression model that fits into the framework we developed in

Chapter 9. The disturbances are autocorrelated in that observations are correlated across time within a group, though not across groups. All the implications of Section 9.2 would apply here. In particular, the parameters of the random effects model can be estimated consistently, though not efficiently, by ordinary least squares (OLS). An appropriate robust asymptotic covariance matrix for the OLS estimator would be given by (11-3).

There are other consistent estimators available as well. By taking deviations from group means, we obtain

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\boldsymbol{\beta} + \varepsilon_{it} - \bar{\varepsilon}_i.$$

This implies that (assuming there are no time-invariant regressors in $\mathbf{x}_{it}$), the LSDV estimator of (11-14) is a consistent estimator of $\boldsymbol{\beta}$. An estimator based on first differences,

$$y_{it} - y_{i,t-1} = (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta} + \varepsilon_{it} - \varepsilon_{i,t-1}.$$

(The LSDV and first differences estimators are robust to whether the correct specification is actually a random or a fixed effects model.) As is OLS, LSDV is inefficient because, as we will show in Section 11.5.2, there is an efficient GLS estimator that is not equal to $\mathbf{b}_{\text{LSDV}}$. The group means (between groups) regression model,

$$\bar{y}_i = \alpha + \bar{\mathbf{x}}_i'\boldsymbol{\beta} + u_i + \bar{\varepsilon}_i, i = 1, \ldots, n,$$

provides a fourth method of consistently estimating the coefficients $\boldsymbol{\beta}$. None of these is the preferred estimator in this setting because the GLS estimator will be more efficient than any of them. However, as we saw in Chapters 9 and 10, many generalized regression models are estimated in two steps, with the first step being a robust least squares regression that is used to produce a first round estimate of the variance parameters of the model. That would be the case here as well. To suggest where this logic will lead in Section 11.5.3, note that for the four cases noted, the sum of squared residuals can produce the following consistent estimators of functions of the variances:

$$\begin{aligned}
&(\text{Pooled}) && \text{plim}\,[\mathbf{e}_{\text{pooled}}{}'\mathbf{e}_{\text{pooled}}/(nT)] = \sigma_u^2 + \sigma_\varepsilon^2,\\
&(\text{LSDV}) && \text{plim}\,[\mathbf{e}_{\text{LSDV}}{}'\mathbf{e}_{\text{LSDV}}/(n(T-1) - K)] = \sigma_\varepsilon^2,\\
&(\text{Differences}) && \text{plim}\,[\mathbf{e}_{\text{FD}}{}'\mathbf{e}_{FD}/(n(T-1))] = 2\sigma_\varepsilon^2,\\
&(\text{Means}) && \text{plim}\,[\mathbf{e}_{\text{means}}{}'\mathbf{e}_{\text{means}}/(nT)] = \sigma_u^2 + \sigma_\varepsilon^2/T.
\end{aligned}$$

Baltagi (2001) suggests yet another method of moments estimator that could be based on the pooled OLS results. Based on (11-31), Cov$(\varepsilon_{it},\varepsilon_{is}) = \sigma_{\text{u}}{}^2$ within group i for $t \neq s$. There are $T(T-1)/2$ pairs of residuals that can be used, so for each group, we could use $(1/(T(T-1)/2))\Sigma_s\Sigma_t\, e_{it}\, e_{is}$ to estimate $\sigma_u{}^2$. Because we have n groups that each provide an estimator, we can average the n implied estimators, to obtain

$$(\text{OLS}) \quad \text{plim}\,\frac{1}{n}\sum\nolimits_{i=1}^{n}\frac{\Sigma_{t=2}^{T}\Sigma_{s=1}^{t-1}e_{it}e_{is}}{T(T-1)/2} = \sigma_u^2.$$

Different pairs of these estimators (and other candidates not shown here) could provide a two-equation method of moments estimator of $(\sigma_u^2, \sigma_\varepsilon^2)$. (Note that the last of these is using a covariance to estimate a variance. Unfortunately, unlike the others, this could be negative in a finite sample.) With these in mind, we will now develop an efficient generalized least squares estimator.

11.5.2 GENERALIZED LEAST SQUARES

The generalized least squares estimator of the slope parameters is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y} = \left(\sum_{i=1}^{n}\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{y}_i\right).$$

To compute this estimator as we did in Chapter 9 by transforming the data and using ordinary least squares with the transformed data, we will require $\boldsymbol{\Omega}^{-1/2} = [\mathbf{I}_n \otimes \boldsymbol{\Sigma}]^{-1/2} = \mathbf{I}_n \otimes \boldsymbol{\Sigma}^{-1/2}$. We need only find $\boldsymbol{\Sigma}^{-1/2}$, which is

$$\boldsymbol{\Sigma}^{-1/2} = \left[\mathbf{I}_T - \frac{\theta_i}{T}\mathbf{i}_T\mathbf{i}_T'\right], \qquad \textbf{(11-33)}$$

where

$$\theta = 1 - \frac{\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2 + T\sigma_u^2}}.$$

The transformation of $\mathbf{y}_i$ and $\mathbf{X}_i$ for GLS is therefore

$$\boldsymbol{\Sigma}^{-1/2}\mathbf{y}_i = \frac{1}{\sigma_\varepsilon}\begin{bmatrix} y_{i1} - \theta\bar{y}_{i.} \\ y_{i2} - \theta\bar{y}_{i.} \\ \vdots \\ y_{i\mathrm{T}} - \theta\bar{y}_{i.} \end{bmatrix}, \qquad \textbf{(11-34)}$$

and likewise for the rows of $\mathbf{X}_i$. For the data set as a whole, then, generalized least squares is computed by the regression of these partial deviations of y_{it} on the same transformations of $\mathbf{x}_{it}$. Note the similarity of this procedure to the computation in the LSDV model, which uses $\theta = 1$ in (11-15).

It can be shown that the GLS estimator is, like the pooled OLS estimator, a matrix weighted average of the within- and between-units estimators,

$$\hat{\boldsymbol{\beta}} = \hat{\mathbf{F}}^{within}\mathbf{b}^{within} + (\mathbf{I} - \hat{\mathbf{F}}^{within})\mathbf{b}^{between},$$

where now,

$$\hat{\mathbf{F}}^{within} = [\mathbf{S}_{xx}^{within} + \lambda\mathbf{S}_{xx}^{between}]^{-1}\mathbf{S}_{xx}^{within},$$

$$\lambda = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_u^2} = (1 - \theta)^2.$$

To the extent that λ differs from one, we see that the inefficiency of ordinary least squares will follow from an inefficient weighting of the two estimators. Compared with generalized least squares, ordinary least squares places too much weight on the between-units variation. It includes all of it in the variation in $\mathbf{X}$, rather than apportioning some of it to random variation across groups attributable to the variation in u_i across units.

Unbalanced panels complicate the random effects model a bit. The matrix $\boldsymbol{\Omega}$ in (11-32) is no longer $\mathbf{I}_n \otimes \boldsymbol{\Sigma}$ because the diagonal blocks in $\boldsymbol{\Omega}$ are of different sizes. In (11-33), the ith diagonal block in $\boldsymbol{\Omega}^{-1/2}$ is

$$\boldsymbol{\Sigma}_i^{-1/2} = \frac{1}{\sigma_\varepsilon}\left[\mathbf{I}_{T_i} - \frac{\theta_i}{T_i}\mathbf{i}_{T_i}\mathbf{i}_{T_i}'\right], \; \theta_i = 1 - \frac{\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2 + T_i\sigma_u^2}}.$$

In principle, estimation is still straightforward, because the source of the groupwise heteroscedasticity is only the unequal group sizes. Thus, for GLS, or FGLS with estimated variance components, it is necessary only to use the group-specific θ_i in the transformation in (11-34).

11.5.3 FEASIBLE GENERALIZED LEAST SQUARES ESTIMATION OF THE RANDOM EFFECTS MODEL WHEN $\boldsymbol{\Sigma}$ IS UNKNOWN

If the variance components are known, generalized least squares can be computed as shown earlier. Of course, this is unlikely, so as usual, we must first estimate the disturbance variances and then use an FGLS procedure. A heuristic approach to estimation of the variance components is as follows:

$$\begin{aligned} y_{it} &= \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha + \varepsilon_{it} + u_i \\ \text{and} \qquad \bar{y}_{i.} &= \bar{\mathbf{x}}_{i.}'\boldsymbol{\beta} + \alpha + \bar{\varepsilon}_{i.} + u_i. \end{aligned} \qquad \textbf{(11-35)}$$

Therefore, taking deviations from the group means removes the heterogeneity,

$$y_{it} - \bar{y}_{i.} = [\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.}]'\boldsymbol{\beta} + [\varepsilon_{it} - \bar{\varepsilon}_{i.}]. \qquad \textbf{(11-36)}$$

Because

$$E\left[\sum_{t=1}^{T}(\varepsilon_{it} - \bar{\varepsilon}_{i.})^2\right] = (T-1)\sigma_\varepsilon^2,$$

if $\boldsymbol{\beta}$ were observed, then an unbiased estimator of σ_ε^2 based on T observations in group i would be

$$\hat{\sigma}_\varepsilon^2(i) = \frac{\sum_{t=1}^{T}(\varepsilon_{it} - \bar{\varepsilon}_{i.})^2}{T-1}. \qquad \textbf{(11-37)}$$

Because $\boldsymbol{\beta}$ must be estimated—the LSDV estimator is consistent, indeed, unbiased in general—we make the degrees of freedom correction and use the LSDV residuals in

$$s_e^2(i) = \frac{\sum_{t=1}^{T}(e_{it} - \bar{e}_{i.})^2}{T-K-1} \qquad \textbf{(11-38)}$$

(Note that based on the LSDV estimates, $\bar{e}_{i.}$ is actually zero. We will carry it through nonetheless to maintain the analogy to (11-35) where $\bar{\varepsilon}_{i.}$ is not zero but is an estimator of $E[\varepsilon_{it}] = 0$.) We have n such estimators, so we average them to obtain

$$\bar{s}_e^2 = \frac{1}{n}\sum_{i=1}^{n} s_e^2(i) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\sum_{t=1}^{T}(e_{it} - \bar{e}_{i.})^2}{T-K-1}\right] = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(e_{it} - \bar{e}_{i.})^2}{nT-nK-n} \qquad \textbf{(11-39a)}$$

The degrees of freedom correction in $\bar{s}_e^2$ is excessive because it assumes that α and $\boldsymbol{\beta}$ are reestimated for each i. The estimated parameters are the n means $\bar{y}_{i\cdot}$ and the K slopes. Therefore, we propose the unbiased estimator[17]

$$\hat{\sigma}_\varepsilon^2 = s_{LSDV}^2 = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(e_{it} - \bar{e}_{i.})^2}{nT-n-K}.$$

[17] A formal proof of this proposition may be found in Maddala (1971) or in Judge et al. (1985, p. 551).

This is the variance estimator in the fixed effects model in (11-18), appropriately corrected for degrees of freedom. It remains to estimate σ_u^2. Return to the original model specification in (11-35). In spite of the correlation across observations, this is a classical regression model in which the ordinary least squares slopes and variance estimators are both consistent and, in most cases, unbiased. Therefore, using the ordinary least squares residuals from the model with only a single overall constant, we have

$$\operatorname{plim} s_{Pooled}^2 = \operatorname{plim} \frac{\mathbf{e}'\mathbf{e}}{nT - K - 1} = \sigma_\varepsilon^2 + \sigma_u^2. \tag{11-39b}$$

This provides the two estimators needed for the variance components; the second would be $\hat{\sigma}_u^2 = s_{Pooled}^2 - s_{\text{LSDV}}^2$. As noted in Section 11.5.1, there are a variety of pairs of variance estimators that can be used to obtain estimates of $\sigma_\varepsilon{}^2$ and $\sigma_u{}^2$.[18] The estimators based on s_{LSDV}^2 and s_{Pooled}^2 are common choices. Alternatively, let $[\mathbf{b}, a]$ be any consistent estimator of $[\boldsymbol{\beta}, \alpha]$ in (11-35), such as the ordinary least squares estimator. Then, s_{Pooled}^2 provides a consistent estimator of $m_{ee} = \sigma_\varepsilon^2 + \sigma_u^2$. The mean squared residuals using a regression based only on the n group means in (11-35) provides a consistent estimator of $m_{**} = \sigma_u^2 + (\sigma_\varepsilon^2/T)$, so we can use

$$\hat{\sigma}_\varepsilon^2 = \frac{T}{T-1}(m_{ee} - m_{**}),$$
$$\hat{\sigma}_u^2 = \frac{T}{T-1}m_{**} - \frac{1}{T-1}m_{ee} = \omega m_{**} + (1-\omega)m_{ee},$$

where $\omega > 1$. A possible complication is that the estimator of $\sigma_u{}^2$ can be negative in any of these cases. This happens fairly frequently in practice, and various ad hoc solutions are typically tried. (The first approach is often to try out different pairs of moments. Unfortunately, typically, one failure is followed by another. It would seem that this failure of the estimation strategy should suggest to the analyst that there is a problem with the specification to begin with. A last solution in the face of a persistently negative estimator is to set $\sigma_u{}^2$ to the value the data are suggesting, zero, and revert to least squares.)

11.5.4 ROBUST INFERENCE AND FEASIBLE GENERALIZED LEAST SQUARES

The feasible GLS estimator based on (11-28) and (11-31) is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}X)^{-1}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}) = \left(\sum\nolimits_{i=1}^n \mathbf{X}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\right)^{-1}\left(\sum\nolimits_{i=1}^n \mathbf{X}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{y}_i\right).$$

There is a subscript i on $\boldsymbol{\Sigma}_i$ because of the consideration of unbalanced panels discussed at the end of Section 11.5.2. If the panel is unbalanced, a minor adjustment is needed because $\boldsymbol{\Sigma}_i$ is $T_i \times T_i$ and because of the specific computation of θ_i. The feasible GLS estimator is then

$$\hat{\hat{\boldsymbol{\beta}}} = \boldsymbol{\beta} + \left(\sum\nolimits_{i=1}^n \mathbf{X}_i'\hat{\boldsymbol{\Sigma}}_i^{-1}\mathbf{X}_i\right)^{-1}\left(\sum\nolimits_{i=1}^n \mathbf{X}_i'\hat{\boldsymbol{\Sigma}}_i^{-1}\boldsymbol{\varepsilon}_i\right). \tag{11-40}$$

[18] See, for example, Wallace and Hussain (1969), Maddala (1971), Fuller and Battese (1974), and Amemiya (1971). This is a point on which modern software varies. Generally, programs begin with (11-39a) and (11-39b) to estimate the variance components. Others resort to different strategies based on, for example, the group means estimator. The unfortunate implication for the unwary is that different programs can systematically produce different results using the same model and the same data. The practitioner is strongly advised to consult the program documentation for resolution.

This form suggests a way to accommodate failure of the random effects assumption in (11-28). Following the approach used in the earlier applications, the estimator would be

$$\text{Est.Asy.Var}[\hat{\hat{\boldsymbol{\beta}}}] = \left(\sum\nolimits_{i=1}^{n}\mathbf{X}_i'\hat{\boldsymbol{\Sigma}}_i^{-1}\mathbf{X}_i\right)^{-1}\left(\sum\nolimits_{i=1}^{n}(\mathbf{X}_i'\hat{\boldsymbol{\Sigma}}_i^{-1}\mathbf{e}_i)(\mathbf{X}_i'\hat{\boldsymbol{\Sigma}}_i^{-1}\mathbf{e}_i)'\right)\left(\sum\nolimits_{i=1}^{n}\mathbf{X}_i'\hat{\boldsymbol{\Sigma}}_i^{-1}\mathbf{X}_i\right)^{-1}. \tag{11-41}$$

With this estimator in hand, inference would be based on Wald statistics rather than F statistics.

There is a loose end in the proposal just made. If assumption (11-28) fails, then what are the properties of the generalized least squares estimator based on $\boldsymbol{\Sigma}$ in (11-31)? The FGLS estimator remains consistent and asymptotically normally distributed—consider that OLS is also a consistent estimator that uses the wrong covariance matrix. And (11-41) would provide an appropriate estimator to use for statistical inference about $\boldsymbol{\beta}$. However, in this case, (11-31) is the wrong starting point for FGLS estimation.

If the random effects assumption is not appropriate, then a more general starting point is

$$\mathbf{y}_i = \alpha\mathbf{i} + \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i,\ E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i' \mid \mathbf{X}_i] = \boldsymbol{\Sigma},$$

which returns us to the pooled regression model in Section 11.3.1. An appealing approach based on that would base feasible GLS on (11-32) and, assuming n is reasonably large and T is relatively small, would use $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{e}_{OLS,i}\mathbf{e}_{OLS,i}'$. Then, feasible GLS would be based on (11-40). One serious complication is how to accommodate an unbalanced panel. With the random effects formulation, the covariances in $\boldsymbol{\Sigma}$ are identical, so positioning of the observations in the matrix is arbitrary. This is not so with an unbalanced panel. We will see in the example below, in this more general case, a distinct pattern in the locations of the cells in the matrix emerges. It is unclear what should be done with the unfilled cells in $\boldsymbol{\Sigma}$.

11.5.5 TESTING FOR RANDOM EFFECTS

Breusch and Pagan (1980) have devised a **Lagrange multiplier test** for the random effects model based on the OLS residuals.[19] For

$$H_0{:}\ \sigma_u^2 = 0,$$
$$H_1{:}\ \sigma_u^2 > 0,$$

the test statistic is

$$LM = \frac{nT}{2(T-1)}\left[\frac{\sum_{i=1}^{n}\left[\sum_{t=1}^{T}e_{it}\right]^2}{\sum_{i=1}^{n}\sum_{t=1}^{T}e_{it}^2} - 1\right]^2 = \frac{nT}{2(T-1)}\left[\frac{\sum_{i=1}^{n}(T\bar{e}_{i.})^2}{\sum_{i=1}^{n}\sum_{t=1}^{T}e_{it}^2} - 1\right]^2. \tag{11-42}$$

Under the null hypothesis, the limiting distribution of LM is chi-squared with one degree of freedom. (The computation for an unbalanced panel replaces the multiple by $[(\Sigma_{i=1}^{n}T_i)^2]/[2\Sigma_{i=1}^{n}T_i(T_i - 1)]$ and replaces T with T_i in the summations.) The LM

[19] Thus far, we have focused strictly on generalized least squares and moments-based consistent estimation of the variance components. The LM test is based on maximum likelihood estimation, instead. See Maddala (1971) and Baltagi (2013) for this approach to estimation.

statistic is based on normally distributed disturbances. Wooldridge (2010) proposed a statistic that is more robust to the distribution, $z = \dfrac{\Sigma_{i=1}^{n}(\Sigma_{t=2}^{T_i}\Sigma_{s=1}^{t-1}e_{it}e_{is})}{\sqrt{\Sigma_{i=1}^{n}(\Sigma_{t=2}^{T_i}\Sigma_{s=1}^{t-1}e_{it}e_{is})^2}}$, which converges to $N[0,1]$ in all cases, or z^2 which has a limiting chi-squared distribution with one degree of freedom. The inner double sums in the statistic sum the below diagonal terms in $\mathbf{e}_i\mathbf{e}_i'$ which is one-half the sum of all the terms minus the diagonals, $\mathbf{e}_i'\mathbf{e}_i$. The ith term in the sum is $\frac{1}{2}[(\Sigma_{t=1}^{T}e_{it})^2 - (\Sigma_{t=1}^{T}e_{it}^2)] = f_i$. By manipulating this result, we find that $z^2 = (n\bar{f}^2/s_f^2)$ (where s_f^2 is computed around the assumed $E[f_i] = 0$), which would be the standard test statistic for the hypothesis that $E[f_i] = 0$. This makes sense, because f_i is essentially composed of the difference between two estimators of σ_ε^2.[20] With some tedious manipulation, we can show that the LM statistic is also a multiple of $n\bar{f}^2$.

Example 11.10 Test for Random Effects

We are interested in comparing the random and fixed effects estimators in the Cornwell and Rupert wage interested equation,

$$\begin{aligned} \ln Wage_{it} = {} & \beta_1 + \beta_2\, Exp_{it} + \beta_3\, Exp_{it}^2 + \beta_4\, Wks_{it} + \beta_5\, Occ_{it} \\ & + \beta_6\, Ind_{it} + \beta_7\, South_{it} + \beta_8\, SMSA_{it} + \beta_9\, MS_{it} \\ & + \beta_{10}\, Union_{it} + \beta_{11}\, Ed_i + \beta_{12}\, Fem_i + \beta_{13}\, Blk_i + c_i + \varepsilon_{it}. \end{aligned}$$

The least squares estimates appear in Table 11.6 in Example 11.4. We will test for the presence of random effects. The computations in the two statistics are simpler than it might appear at first. The LM statistic is

$$LM = \frac{nT}{2(T-1)}\left[\frac{\mathbf{e}'\mathbf{DD}'\mathbf{e}}{\mathbf{e}'\mathbf{e}} - 1\right]^2,$$

where $\mathbf{D}$ is the matrix of individual dummy variables in (11-13). To compute z^2, we compute

$$\mathbf{f} = \frac{1}{2}(\mathbf{D}'\mathbf{e} \circ \mathbf{D}'\mathbf{e} - \mathbf{D}'(\mathbf{e} \circ \mathbf{e})),$$

($\circ$ is the Hadamard product—element by element multiplication) then $z^2 = \mathbf{i}'\mathbf{f}/\mathbf{f}'\mathbf{f}$. The results for the two statistics are $LM = 3497.02$ and $z^2 = 179.66$. These far exceed the 95% critical value for the chi-squared distribution with one degree of freedom, 3.84. At this point, we conclude that the classical regression model without the heterogeneity term is inappropriate for these data. The result of the test is to reject the null hypothesis in favor of the random effects model. But it is best to reserve judgment on that because there is another competing specification that might induce these same results, the fixed effects model. We will examine this possibility in the subsequent examples.

With the variance estimators in hand, FGLS can be used to estimate the parameters of the model. All of our earlier results for FGLS estimators apply here. In particular, all that is needed for efficient estimation of the model parameters are consistent estimators of the variance components, and there are several.[21]

[20] Wooldridge notes that z can be negative, suggesting a negative estimate of $\sigma_u{}^2$. This counterintuitive result arises, once again (see Section 11.5.1), from using a covariance estimator to estimate a variance. However, with some additional manipulation, we find that the numerator of z is actually $(nT/2)[\hat{\sigma}_\varepsilon^2(\text{based on } \bar{e}_i) - \hat{\sigma}_\varepsilon^2(\text{based on } e_{it})]$ so the outcome is not so contradictory as it might appear—since the statistic has a standard normal distribution, the negative result should occur half of the time. The test is not actually based on the covariance; it is based on the difference of two estimators of the same variance (under the null hypothesis). The numerator of the LM statistic, $\mathbf{e}'\mathbf{DD}'\mathbf{e} - \mathbf{e}'\mathbf{e}$, is the same as that of z, though it is squared to produce the test statistic.

[21] See Hsiao (2003), Baltagi (2005), Nerlove (2002), Berzeg (1979), and Maddala and Mount (1973).

Example 11.11 Estimates of the Random Effects Model

In the previous example, we found the total sum of squares for the least squares estimator was 506.766. The fixed effects (LSDV) estimates for this model appear in Table 11.10. The sum of squares is 82.26732. Therefore, the moment estimators of the variance parameters are

$$\hat{\sigma}_\varepsilon^2 + \hat{\sigma}_\mathrm{u}^2 = \frac{506.766}{4165 - 13} = 0.122053$$

and

$$\hat{\sigma}_\varepsilon^2 = \frac{82.26732}{4165 - 595 - 9} = 0.0231023.$$

The implied estimator of σ_u^2 is 0.098951. (No problem of negative variance components has emerged. Note that the three time-invariant variables have not been used in computing the fixed effects estimator to estimate $\sigma_\varepsilon{}^2$.) The estimate of θ for FGLS is

$$\hat{\theta} = 1 - \sqrt{\frac{0.0231023}{0.0231023 + 7(0.098951)}} = 0.820343.$$

FGLS estimates are computed by regressing the partial differences of ln $Wage_{\mathrm{it}}$ on the partial differences of the constant and the 12 regressors, using this estimate of θ in (11-33). The full GLS estimates are obtained by estimating $\mathbf{\Sigma}$ using the OLS residuals. The estimate of $\mathbf{\Sigma}$ is listed below with the other estimates. Thus, $\hat{\mathbf{\Sigma}} = \frac{1}{595}\sum_{i=1}^{595}\mathbf{e}_i\mathbf{e}_i'$. The estimate of $\mathbf{\Omega} = \sigma_\varepsilon^2\mathbf{I} + \sigma_u^2\mathbf{ii}'$**.** Estimates of the parameters using the OLS and random effects estimators appear in Table 11.11. The similarity of the estimates is to be expected given that, under the hypothesis of the model, all three estimators are consistent.

The random effects specification is a substantive restriction on the stochastic part of the regression. The assumption that the disturbances are equally correlated across periods regardless of how far apart the periods are may be a particularly strong assumption, particularly if the time dimension of the panel is relatively long. The force of the restrictions can be seen in the covariance matrices shown below. In the random effects model, the cross period correlation is $\sigma_u^2/(\sigma_\varepsilon^2 + \sigma_u^2)$ which we have estimated as 0.9004 for all periods. But, the first column of the estimate of $\mathbf{\Sigma}$ suggests quite a different pattern; the cross period covariances diminish substantially with the separation in time. If an AR(1) pattern is assumed, $\varepsilon_{i,t} = \rho\varepsilon_{i,\,t-1} + v_{i,t}$ then the implied estimate of ρ would be $r = 0.1108/0.1418 = 0.7818$. The next two periods appear consistent with the pattern, ρ^2 then ρ^3. The first-order autoregression might be a reasonable candidate for the model. At the same time, the diagonal elements of $\hat{\mathbf{\Sigma}}$ do not strongly suggest much heteroscedasticity across periods.

None of the desirable properties of the estimators in the random effects model rely on T going to infinity.[22] Indeed, T is likely to be quite small. The estimator of σ_ε^2 is equal to an average of n estimators, each based on the T observations for unit i. [See (11-39a).] Each component in this average is, in principle, consistent. That is, its variance is of order $1/T$ or smaller. Because T is small, this variance may be relatively large. But each term provides some information about the parameter. The average over the n cross-sectional units has a variance of order $1/(nT)$, which will go to zero if n increases, even if we regard T as fixed. The conclusion to draw is that nothing in this treatment relies on T growing large. Although it can be shown that some consistency results will follow for T increasing, the typical panel data set is based on data sets for which it does not make sense to

[22] See Nickell (1981).

TABLE 11.11 Wage Equation Estimated by GLS

Variable	*Least Squares Estimate*	*Clustered Std. Error*	*Random Effects Ests.*	*Standard Error*	*Generalized Least Squares*	*Standard Error*
Constant	5.25112	0.12355	4.04144	0.08330	5.31019	0.07948
Exp	0.04010	0.00408	0.08748	0.00225	0.04478	0.00388
ExpSq	−0.00067	0.00009	−0.00076	0.00005	−0.00071	0.00009
Wks	0.00422	0.00154	0.00096	0.00059	0.00071	0.00055
Occ	−0.14001	0.02724	−0.04322	0.01299	−0.03842	0.01265
Ind	0.04679	0.02366	0.00378	0.01373	0.02671	0.01340
South	−0.05564	0.02616	−0.00825	0.02246	−0.06089	0.02129
SMSA	0.15167	0.02410	−0.02840	0.01616	0.06737	0.01669
MS	0.04845	0.04094	−0.07090	0.01793	−0.02610	0.02020
Union	0.09263	0.02367	0.05835	0.01350	0.03544	0.01316
Ed	0.05670	0.00556	0.10707	0.00511	0.06507	0.00429
Fem	−0.36779	0.04557	−0.30938	0.04554	−0.39606	0.03889
Blk	−0.16694	0.04433	−0.21950	0.05252	−0.15154	0.04262

GLS Estimated Covariance Matrix of ε_i

1	*2*	*3*	*4*	*5*	*6*	*7*
0.1418						
0.1108	0.1036					
0.0821	0.0748	0.1135				
0.0583	0.0579	0.0845	0.1046			
0.0368	0.0418	0.0714	0.0817	0.1008		
0.0152	0.0250	0.0627	0.0799	0.0957	0.1246	
−0.0056	0.0099	0.0585	0.0822	0.1024	0.1259	0.1629

Estimated Covariance Matrix for ε_i Based on Random Effects Model

1	*2*	*3*	*4*	*5*	*6*	*7*
0.1221						
0.0989	0.1221					
0.0989	0.0989	0.1221				
0.0989	0.0989	0.0989	0.1221			
0.0989	0.0989	0.0989	0.0989	0.1221		
0.0989	0.0989	0.0989	0.0989	0.0989	0.1221	
0.0989	0.0989	0.0989	0.0989	0.0989	0.0989	0.1221

assume that T increases without bound or, in some cases, at all.[23] As a general proposition, it is necessary to take some care in devising estimators whose properties hinge on whether T is large or not. The widely used conventional ones we have discussed here do not, but we have not exhausted the possibilities.

[23] In this connection, Chamberlain (1984) provided some innovative treatments of panel data that, in fact, take T as given in the model and that base consistency results solely on n increasing. Some additional results for dynamic models are given by Bhargava and Sargan (1983). Recent research on "bias reduction" in nonlinear panel models, such as Fernandez-Val (2010), do make use of large T approximations in explicitly small T settings.

11.5.6 HAUSMAN'S SPECIFICATION TEST FOR THE RANDOM EFFECTS MODEL

At various points, we have made the distinction between fixed and random effects models. An inevitable question is, which should be used? From a purely practical standpoint, the dummy variable approach is costly in terms of degrees of freedom lost. On the other hand, the fixed effects approach has one considerable virtue. There is little justification for treating the individual effects as uncorrelated with the other regressors, as is assumed in the random effects model. The random effects treatment, therefore, may suffer from the inconsistency due to this correlation between the included variables and the random effect.[24]

The **specification test** devised by Hausman (1978)[25] is used to test for orthogonality of the common effects and the regressors. The test is based on the idea that under the hypothesis of no correlation, both LSDV and FGLS estimators are consistent, but LSDV is inefficient,[26] whereas under the alternative, LSDV is consistent, but FGLS is not. Therefore, under the null hypothesis, the two estimates should not differ systematically, and a test can be based on the difference. The other essential ingredient for the test is the covariance matrix of the difference vector, $[\mathbf{b}_{\text{FE}} - \hat{\boldsymbol{\beta}}_{\text{RE}}]$,

$$\text{Var}[\mathbf{b}_{\text{FE}} - \hat{\boldsymbol{\beta}}_{\text{RE}}] = \text{Var}[\mathbf{b}_{\text{FE}}] + \text{Var}[\hat{\boldsymbol{\beta}}_{\text{RE}}] - \text{Cov}[\mathbf{b}_{\text{FE}}, \hat{\boldsymbol{\beta}}_{\text{RE}}] - \text{Cov}[\hat{\boldsymbol{\beta}}_{\text{RE}}, \mathbf{b}_{\text{FE}}]. \quad \textbf{(11-43)}$$

Hausman's essential result is that *the covariance of an efficient estimator with its difference from an inefficient estimator is zero*, which implies that

$$\text{Cov}[(\mathbf{b}_{\text{FE}} - \hat{\boldsymbol{\beta}}_{\text{RE}}), \hat{\boldsymbol{\beta}}_{\text{RE}}] = \text{Cov}[\mathbf{b}_{\text{FE}}, \hat{\boldsymbol{\beta}}_{\text{RE}}] - \text{Var}[\hat{\boldsymbol{\beta}}_{\text{RE}}] = \mathbf{0},$$

or that

$$\text{Cov}[\mathbf{b}_{\text{FE}}, \hat{\boldsymbol{\beta}}_{\text{RE}}] = \text{Var}[\hat{\boldsymbol{\beta}}_{\text{RE}}].$$

Inserting this result in (11-43) produces the required covariance matrix for the test,

$$\text{Var}[\mathbf{b}_{\text{FE}} - \hat{\boldsymbol{\beta}}_{\text{RE}}] = \text{Var}[\mathbf{b}_{\text{FE}}] - \text{Var}[\hat{\boldsymbol{\beta}}_{\text{RE}}] = \boldsymbol{\Psi}.$$

The chi-squared test is based on the Wald criterion,

$$W = \chi^2[K-1] = [\mathbf{b}_{\text{FE}} - \hat{\boldsymbol{\beta}}_{\text{RE}}]'\hat{\boldsymbol{\Psi}}^{-1}[\mathbf{b}_{\text{FE}} - \hat{\boldsymbol{\beta}}_{\text{RE}}]. \quad \textbf{(11-44)}$$

For $\hat{\boldsymbol{\Psi}}$, we use the estimated covariance matrices of the slope estimator in the LSDV model and the estimated covariance matrix in the random effects model, excluding the constant term. Under the null hypothesis, W has a limiting chi-squared distribution with $K - 1$ degrees of freedom.

The *Hausman test* is a useful device for determining the preferred specification of the common effects model. As developed here, it has one practical shortcoming. The construction in (11-43) conforms to the theory of the test. However, it does not guarantee that the difference of the two covariance matrices will be positive definite in a finite sample. The implication is that nothing prevents the statistic from being negative when it is computed according to (11-44). One might, in that event, conclude that the random effects model is not rejected, because the similarity of the covariance matrices is what

[24] See Hausman and Taylor (1981) and Chamberlain (1978).

[25] Related results are given by Baltagi (1986).

[26] Referring to the FGLS matrix weighted average given earlier, we see that the efficient weight uses θ, whereas LSDV sets $\theta = 1$.

is causing the problem, and under the alternative (fixed effects) hypothesis, they should be significantly different. There are, however, several alternative methods of computing the statistic for the Hausman test, some asymptotically equivalent and others actually numerically identical. Baltagi (2005, pp. 65–73) provides an extensive analysis. One particularly convenient form of the test finesses the practical problem noted here. An asymptotically equivalent test statistic is given by

$$H' = (\mathbf{b}_{\text{FE}} - \mathbf{b}_{\text{MEANS}})'[\text{Asy.Var}[\mathbf{b}_{\text{FE}}] + \text{Asy.Var}[\mathbf{b}_{\text{MEANS}}]]^{-1}(\mathbf{b}_{\text{FE}} - \mathbf{b}_{\text{MEANS}}) \quad \textbf{(11-45)}$$

where $\mathbf{b}_{\text{MEANS}}$ is the group means estimator discussed in Section 11.3.4. As noted, this is one of several equivalent forms of the test. The advantage of this form is that the covariance matrix will always be nonnegative definite.

Imbens and Wooldridge (2007) have argued that in spite of the practical considerations about the Hausman test in (11-44) and (11-45), the test should be based on robust covariance matrices that do not depend on the assumption of the null hypothesis (the random effects model).[27] Their suggested approach amounts to the variable addition test described in the next section, with a robust covariance matrix.

11.5.7 EXTENDING THE UNOBSERVED EFFECTS MODEL: MUNDLAK'S APPROACH

Even with the Hausman test available, choosing between the fixed and random effects specifications presents a bit of a dilemma. Both specifications have unattractive shortcomings. The fixed effects approach is robust to correlation between the omitted heterogeneity and the regressors, but it proliferates parameters and cannot accommodate time-invariant regressors. The random effects model hinges on an unlikely assumption, that the omitted heterogeneity is uncorrelated with the regressors. Several authors have suggested modifications of the random effects model that would at least partly overcome its deficit. The failure of the random effects approach is that the mean independence assumption, $E[c_i | \mathbf{X}_i] = 0$, is untenable. **Mundlak's approach** (1978) suggests the specification

$$E[c_i | \mathbf{X}_i] = \overline{\mathbf{x}}_{i.}'\boldsymbol{\gamma}.^{28}$$

Substituting this in the random effects model, we obtain

$$\begin{aligned} y_{it} &= \mathbf{z}_i'\boldsymbol{\alpha} + \mathbf{x}_{it}'\boldsymbol{\beta} + c_i + \varepsilon_{it} \\ &= \mathbf{z}_i'\boldsymbol{\alpha} + \mathbf{x}_{it}'\boldsymbol{\beta} + \overline{\mathbf{x}}_{i.}'\boldsymbol{\gamma} + \varepsilon_{it} + (c_i - E[c_i | \mathbf{X}_i]) \\ &= \mathbf{z}_i'\boldsymbol{\alpha} + \mathbf{x}_{it}'\boldsymbol{\beta} + \overline{\mathbf{x}}_{i.}'\boldsymbol{\gamma} + \varepsilon_{it} + u_i. \end{aligned} \quad \textbf{(11-46)}$$

This preserves the specification of the random effects model, but (one hopes) deals directly with the problem of correlation of the effects and the regressors. Note that the

[27] That is, "It makes no sense to report a fully robust variance matrix for FE and RE but then to compute a Hausman test that maintains the full set of RE assumptions."

[28] Other analyses, for example, Chamberlain (1982) and Wooldridge (2010), interpret the linear function as the *projection* of c_i on the group means, rather than the conditional mean. The difference is that we need not make any particular assumptions about the conditional mean function while there always exists a linear projection. The conditional mean interpretation does impose an additional assumption on the model but brings considerable simplification. Several authors have analyzed the extension of the model to projection on the full set of individual observations rather than the means. The additional generality provides the bases of several other estimators including minimum distance [Chamberlain (1982)], GMM [Arellano and Bover (1995)], and constrained seemingly unrelated regressions and three-stage least squares [Wooldridge (2010)].

additional terms in $\overline{\mathbf{x}}_{i.}'\,\boldsymbol{\gamma}$ will only include the time-varying variables—the time-invariant variables are already group means.

Mundlak's approach is frequently used as a compromise between the fixed and random effects models. One side benefit of the specification is that it provides another convenient approach to the Hausman test. As the model is formulated above, the difference between the fixed effects model and the random effects model is the nonzero $\boldsymbol{\gamma}$. As such, a statistical test of the null hypothesis that $\boldsymbol{\gamma}$ equals zero should provide an alternative approach to the two methods suggested earlier. Estimation of (11-46) can be based on either pooled OLS (with a robust covariance matrix) or random effects FGLS. It turns out the coefficient vectors for the two estimators are identical, though the asymptotic covariance matrices will not be. The pooled OLS estimator is fully robust and seems preferable. The test of the null hypothesis that the common effects are uncorrelated with the regressors is then based on a Wald test.

Example 11.12 Hausman and Variable Addition Tests for Fixed versus Random Effects

Using the results in Examples 11.7 (fixed effects) and 11.11 (random effects), we retrieved the coefficient vector and estimated robust asymptotic covariance matrix, $\mathbf{b}_{FE}$ and $\mathbf{V}_{FE}$, from the fixed effects results and the nine elements of $\hat{\boldsymbol{\beta}}_{RE}$ and $\mathbf{V}_{RE}$ (excluding the constant term and the time-invariant variables) from the random effects results. The test statistic is

$$H = (\mathbf{b}_{FE} - \hat{\boldsymbol{\beta}}_{RE})'[\mathbf{V}_{FE} - \mathbf{V}_{RE}]^{-1}(\mathbf{b}_{FE} - \hat{\boldsymbol{\beta}}_{RE}),$$

The value of the test statistic is 739.374. The critical value from the chi-squared table is 16.919 so the null hypothesis of the random effects model is rejected. There is an additional subtle point to be checked. The difference of the covariance matrices, $\mathbf{V}_{FE} - \mathbf{V}_{RE}$, may not be positive definite. That might not prevent calculation of H if the analyst uses an ordinary inverse in the computation. In that case, a positive statistic might be obtained anyway. The statistic should not be used in this instance. However, that outcome should not lead one to conclude that the correct value for H is zero. The better response is to use the variable addition test we consider next. (For the example here, the smallest characteristic root of the difference matrix was, indeed positive.)

We conclude that the fixed effects model is the preferred specification for these data. This is an unfortunate turn of events, as the main object of the study is the impact of education, which is a time-invariant variable in this sample. We then used the variable addition test instead, based on the regression results in Table 11.12. We recovered the subvector of the estimates at the right in Table 11.12 corresponding to $\boldsymbol{\gamma}$**,** and the corresponding submatrix of the full covariance matrix. The test statistic is

$$H' = \hat{\boldsymbol{\gamma}}'[\text{Est.Asy.Var}(\hat{\boldsymbol{\gamma}})]^{-1}\,\hat{\boldsymbol{\gamma}}.$$

We obtained a value of 2267.32. This does not change the conclusion, so the null hypothesis of the random effects model is rejected. We conclude as before that the fixed effects estimator is the preferred specification for this model.

11.5.8 EXTENDING THE RANDOM AND FIXED EFFECTS MODELS: CHAMBERLAIN'S APPROACH

The linear unobserved effects model is

$$y_{it} = c_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}. \qquad \textbf{(11-47)}$$

The *random effects* model assumes that $E[c_i\,|\,\mathbf{X}_i] = \alpha$, where the T rows of $\mathbf{X}_i$ are $\mathbf{x}_{it}'$. As we saw in Section 11.5.1, this model can be estimated consistently by ordinary

TABLE 11.12 Wage Equation Estimated by OLS and LSDV

	Pooled OLS		*Augmented Regression*		*Group Means*	
Variable	*Least Squares Estimate*	*Clustered Std. Error*	*Least Squares Estimates*	*Robust Std. Error*	*Least Squares Estimates*	*Robust Std. Error*
R^2	*0.42861*		*0.57518*			
Constant	5.25112	0.12355	5.12143	0.20847	—	—
Exp	0.00401	0.00408	0.11321	0.00406	−0.08131	0.00614
ExpSq	−0.00067	0.00009	−0.00042	0.00008	−0.00015	0.00013
Wks	0.00422	0.00154	0.00084	0.00087	0.00835	0.00361
Occ	−0.14001	0.02724	−0.02148	0.01902	−0.14614	0.03821
Ind	0.04679	0.02366	0.01921	0.02271	0.03871	0.03509
South	−0.05564	0.02616	−0.00186	0.08943	−0.05519	0.09371
SMSA	0.15167	0.02410	−0.04247	0.02953	0.21824	0.03859
MS	0.04845	0.04094	−0.02973	0.02691	0.14451	0.05569
Union	0.09263	0.02367	0.03278	0.02510	0.07628	0.03828
Ed	0.05670	0.00556	0.05144	0.00588	—	—
Fem	−0.36779	0.04557	−0.31706	0.05122	—	—
Blk	−0.16694	0.04433	−0.15780	0.04367	—	—

least squares. Regardless of how ε_{it} is modeled, there is autocorrelation induced by the common, unobserved c_i, so the generalized regression model applies. The random effects formulation is based on the assumption $E[\mathbf{w}_i\mathbf{w}_i'|\mathbf{X}_i] = \sigma_\varepsilon^2\mathbf{I}_T + \sigma_u^2\mathbf{ii}'$, where $w_{it} = (\varepsilon_{it} + u_i)$. We developed the GLS and FGLS estimators for this formulation as well as a strategy for robust estimation of the OLS and LSDV covariance matrices. Among the implications of the development of Section 11.5 is that this formulation of the disturbance covariance matrix is more restrictive than necessary, given the information contained in the data. The assumption that $E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'|\mathbf{X}_i] = \sigma_\varepsilon^2\mathbf{I}_T$ assumes that the correlation across periods is equal for all pairs of observations, and arises solely through the persistent c_i. We found some contradictory empirical evidence in Example 11.11—the OLS covariances across periods in the Cornwell and Rupert model do not appear to conform to this specification. In Example 11.11, we estimated the equivalent model with an unrestricted covariance matrix, $E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'|\mathbf{X}_i] = \boldsymbol{\Sigma}$. The implication is that the random effects treatment includes two restrictive assumptions, mean independence, $E[c_i|\mathbf{X}_i] = \alpha$, and homoscedasticity, $E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'|\mathbf{X}_i] = \sigma_\varepsilon^2\mathbf{I}_T$. [We do note that dropping the second assumption will cost us the identification of σ_u^2 as an estimable parameter. This makes sense—if the correlation across periods t and s can arise from either their common u_{i} or from correlation of $(\varepsilon_{it}, \varepsilon_{is})$ then there is no way for us separately to estimate a variance for u_i apart from the covariances of ε_{it} and ε_{is}.] It is useful to note, however, that the panel data model can be viewed and formulated as a seemingly unrelated regressions model with common coefficients in which each period constitutes an equation, Indeed, it is possible, albeit unnecessary, to impose the restriction $E[\mathbf{w}_i\mathbf{w}_i'|\mathbf{X}_i] = \sigma_\varepsilon^2\mathbf{I}_T + \sigma_u^2\mathbf{ii}'$.

The mean independence assumption is the major shortcoming of the random effects model. The central feature of the fixed effects model in Section 11.4 is the possibility that

$E[c_i|\mathbf{X}_i]$ is a nonconstant $h(\mathbf{X}_i)$. As such, least squares regression of y_{it} on $\mathbf{x}_{it}$ produces an inconsistent estimator of $\boldsymbol{\beta}$. The dummy variable model considered in Section 11.4 is the natural alternative. The **fixed effects** approach has the advantage of dispensing with the unlikely assumption that c_i and $\mathbf{x}_{it}$ are uncorrelated. However, it has the shortcoming of requiring estimation of the n parameters, α_i.

Chamberlain (1982, 1984) and Mundlak (1978) suggested alternative approaches that lie between these two. Their modifications of the fixed effects model augment it with the **projections** of c_i on all the rows of $\mathbf{X}_i$ (Chamberlain) or the group means (Mundlak). (See Section 11.5.7.) Consider the first of these, and assume (as it requires) a balanced panel of T observations per group. For purposes of this development, we will assume $T = 3$. The generalization will be obvious at the conclusion. Then, the projection suggested by Chamberlain is

$$c_i = \alpha + \mathbf{x}_{i1}'\boldsymbol{\gamma}_1 + \mathbf{x}_{i2}'\boldsymbol{\gamma}_2 + \mathbf{x}_{i3}'\boldsymbol{\gamma}_3 + r_i, \qquad \textbf{(11-48)}$$

where now, by construction, r_i is orthogonal to $\mathbf{x}_{it}$.[29] Insert (11-48) into (11-47) to obtain

$$y_{it} = \alpha + \mathbf{x}_{i1}'\boldsymbol{\gamma}_1 + \mathbf{x}_{i2}'\boldsymbol{\gamma}_2 + \mathbf{x}_{i3}'\boldsymbol{\gamma}_3 + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} + r_i.$$

Estimation of the $1 + 3K + K$ parameters of this model presents a number of complications. [We do note that this approach has the potential to (wildly) proliferate parameters. For our quite small regional productivity model in Example 11.22. the original model with six main coefficients plus the treatment of the constants becomes a model with $1 + 6 + 17(6) = 109$ parameters to be estimated.]

If only the n observations for period 1 are used, then the parameter vector,

$$\boldsymbol{\theta}_1 = \alpha, (\boldsymbol{\beta} + \boldsymbol{\gamma}_1), \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3 = \alpha, \boldsymbol{\pi}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3, \qquad \textbf{(11-49)}$$

can be estimated consistently, albeit inefficiently, by ordinary least squares. The model is

$$y_{i1} = \mathbf{z}_{i1}'\boldsymbol{\theta}_1 + w_{i1}, i = 1, \ldots, n.$$

Collecting the n observations, we have

$$\mathbf{y}_1 = \mathbf{Z}_1\boldsymbol{\theta}_1 + \mathbf{w}_1.$$

If, instead, only the n observations from period 2 or period 3 are used, then OLS estimates, in turn,

$$\boldsymbol{\theta}_2 = \boldsymbol{\alpha}, \boldsymbol{\gamma}_1, (\boldsymbol{\beta} + \boldsymbol{\gamma}_2), \boldsymbol{\gamma}_3 = \alpha, \boldsymbol{\gamma}_1, \boldsymbol{\pi}_2, \boldsymbol{\gamma}_3,$$

or

$$\boldsymbol{\theta}_3 = \boldsymbol{\alpha}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, (\boldsymbol{\beta} + \boldsymbol{\gamma}_3) = \alpha, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\pi}_3.$$

[29] There are some fine points here that can only be resolved theoretically. If the projection in (11-48) is not the conditional mean, then we have $E[r_i \times \mathbf{x}_{it}] = 0, t = 1, \ldots, T$ but not $E[r_i|\mathbf{X}_i] = 0$. This does not affect the asymptotic properties of the FGLS estimator to be developed here, although it does have implications, for example, for unbiasedness. Consistency will hold regardless. The assumptions behind (11-48) do not include that $\text{Var}[r_i|\mathbf{X}_i]$ is homoscedastic. It might not be. This *could* be investigated empirically. The implication here concerns efficiency, not consistency. The FGLS estimator to be developed here would remain consistent, but a GMM estimator would be more efficient—see Chapter 13. Moreover, without homoscedasticity, it is not certain that the FGLS estimator suggested here is more efficient than OLS (with a robust covariance matrix estimator). Our intent is to begin the investigation here. Further details can be found in Chamberlain (1984) and, for example, Im, Ahn, Schmidt, and Wooldridge (1999).

It remains to reconcile the multiple estimates of the same parameter vectors. In terms of the preceding layouts above, we have the following:

$$\begin{array}{llll}
\text{OLS Estimates:} & a_1, \mathbf{p}_1, \mathbf{c}_{2,1}, \mathbf{c}_{3,1}, & a_2\ \mathbf{c}_{1,2}, \mathbf{p}_2, \mathbf{c}_{3,2}, & a_3, \mathbf{c}_{1,3}, \mathbf{c}_{2,3}, \mathbf{p}_3; \\
\text{Estimated Parameters:} & \alpha, (\boldsymbol{\beta} + \boldsymbol{\gamma}_1), \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3, & \alpha, \boldsymbol{\gamma}_1, (\boldsymbol{\beta} + \boldsymbol{\gamma}_2), \boldsymbol{\gamma}_3, & \alpha, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, (\boldsymbol{\beta} + \boldsymbol{\gamma}_3); \\
\text{Structural Parameters:} & \alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3. & &
\end{array} \quad \textbf{(11-50)}$$

Chamberlain suggested a minimum distance estimator (MDE). For this problem, the MDE is essentially a weighted average of the several estimators of each part of the parameter vector. We will examine the MDE for this application in more detail in Chapter 13. (For another simpler application of minimum distance estimation that shows the weighting procedure at work, see the reconciliation of four competing estimators of a single parameter at the end of Example 11.23.) There is an alternative way to formulate the estimator that is a bit more transparent. For the first period,

$$\mathbf{y}_1 = \begin{pmatrix} y_{1,1} \\ y_{2,1} \\ \vdots \\ y_{n,1} \end{pmatrix} = \begin{bmatrix} 1 & \mathbf{x}_{1,1} & \mathbf{x}_{1,1} & \mathbf{x}_{1,2} & \mathbf{x}_{1,3} \\ 1 & \mathbf{x}_{2,1} & \mathbf{x}_{2,1} & \mathbf{x}_{2,2} & \mathbf{x}_{2,3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \mathbf{x}_{n,1} & \mathbf{x}_{n,1} & \mathbf{x}_{n,2} & \mathbf{x}_{n,3} \end{bmatrix} \begin{pmatrix} \alpha \\ \boldsymbol{\beta} \\ \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \\ \boldsymbol{\gamma}_3 \end{pmatrix} + \begin{pmatrix} r_{1,1} \\ r_{2,1} \\ \vdots \\ r_{n,1} \end{pmatrix} = \widetilde{\mathbf{X}}_1 \boldsymbol{\theta} + \mathbf{r}_1. \quad \textbf{(11-51)}$$

We treat this as the first equation in a T equation seemingly unrelated regressions model. The second equation, for period 2, is the same (same coefficients), with the data from the second period appearing in the blocks, then likewise for period 3 (and periods 4, . . ., T in the general case). Stacking the data for the T equations (periods), we have

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_T \end{pmatrix} = \begin{pmatrix} \widetilde{\mathbf{X}}_1 \\ \widetilde{\mathbf{X}}_2 \\ \vdots \\ \widetilde{\mathbf{X}}_T \end{pmatrix} \begin{pmatrix} \alpha \\ \boldsymbol{\beta} \\ \boldsymbol{\gamma}_1 \\ \vdots \\ \boldsymbol{\gamma}_T \end{pmatrix} + \begin{pmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \vdots \\ \mathbf{r}_T \end{pmatrix} = \widetilde{\mathbf{X}} \boldsymbol{\theta} + \mathbf{r}, \quad \textbf{(11-52)}$$

where $E[\widetilde{\mathbf{X}}'\mathbf{r}] = \mathbf{0}$ and (by assumption), $E[\mathbf{rr}' \,|\, \widetilde{\mathbf{X}}] = \boldsymbol{\Sigma} \otimes \mathbf{I}_n$. With the homoscedasticity assumption for $r_{i,t}$, this is precisely the application in Section 10.2.5. The parameters can be estimated by FGLS as shown in Section 10.2.5.

Example 11.13 Hospital Costs

Carey (1997) examined hospital costs for a sample of 1,733 hospitals observed in five years, 1987–1991. The model estimated is

$$\begin{aligned} \ln (TC/P)_{it} = {} & \alpha_i + \beta_D\, DIS_{it} + \beta_O\, OPV_{it} + \beta_3\, ALS_{it} + \beta_4\, CM_{it} \\ & + \beta_5\, DIS_{it}^2 + \beta_6\, DIS_{it}^3 + \beta_7\, OPV_{it}^2 + \beta_8\, OPV_{it}^3 \\ & + \beta_9\, ALS_{it}^2 + \beta_{10}\, ALS_{it}^3 + \beta_{11} DIS_{it} \times OPV_{it} \\ & + \beta_{12} FA_{it} + \beta_{13} HI_{it} + \beta_{14} HT_i + \beta_{15} LT_i + \beta_{16}\, Large_i \\ & + \beta_{17}\, Small_i + \beta_{18}\, NonProfit_i + \beta_{19}\, Profit_i \\ & + \varepsilon_{it}, \end{aligned}$$

where

TC = total cost,
P = input price index,
DIS = discharges
OPV = outpatient visits,
ALS = average length of stay,
CM = case mix index,
FA = fixed assets,
HI = Hirfindahl index of market concentration at county level,
HT = dummy variable for high teaching load hospital,
LT = dummy variable for low teaching load hospital
$Large$ = dummy variable for large urban area
$Small$ = dummy variable for small urban area,
$Nonprofit$ = dummy variable for nonprofit hospital,
$Profit$ = dummy variable for for-profit hospital.

We have used subscripts “D” and “O” for the coefficients on DIS and OPV as these will be isolated in the following discussion. The model employed in the study is that in (11-47) and (11-48). Initial OLS estimates are obtained for the full cost function in each year. SUR estimates are then obtained using a restricted version of the Chamberlain system. This second step involved a hybrid model that modified (11-49) so that in each period the coefficient vector was

$$\boldsymbol{\theta}_t = [\alpha_t, \beta_{Dt}(\boldsymbol{\gamma}), \beta_{Ot}(\boldsymbol{\gamma}), \beta_{3t}(\boldsymbol{\gamma}), \beta_{4t}(\boldsymbol{\gamma}), \beta_{5t}, \ \ldots, \beta_{19t}],$$

where $\beta_{Dt}(\boldsymbol{\gamma})$ indicates that all five years of the variable (DIS_{it}) are included in the equation, and likewise, for $\beta_{Ot}(\boldsymbol{\gamma})(OPV)$, $\beta_{3t}(\boldsymbol{\gamma})(ALS)$, and $\beta_{4t}(\boldsymbol{\gamma})(CM)$. This is equivalent to using

$$c_i = \alpha + \boldsymbol{\Sigma}_{t=1987}^{1991} (DIS, OPV, ALS, CM)'_{it}\boldsymbol{\gamma}_t + r_i$$

in (11-48).

The unrestricted SUR system estimated at the second step provides multiple estimates of the various model parameters. For example, each of the five equations provides an estimate of $(\beta_5, \ \ldots, \beta_{19})$. The author added one more layer to the model in allowing the coefficients on DIS_{it} and OPV_{it} to vary over time. Therefore, the structural parameters of interest are $(\beta_{D1}, \ \ldots, \beta_{D5})$, $(\gamma_{D1} \ldots, \gamma_{D5})$ (the coefficients on DIS) and $(\beta_{O1}, \ \ldots, \beta_{O5})$, $(\gamma_{O1} \ldots, \gamma_{O5})$ (the coefficients on OPV). There are, altogether, 20 parameters of interest. The SUR estimates produce, in each year (equation), parameters on DIS for the five years and on OPV for the five years, so there is a total of 50 estimates. Reconciling all of them means imposing a total of 30 restrictions. Table 11.13 shows the relationships for the time-varying parameter on DIS_{it} in the five-equation model. The numerical values reported by the author are shown following the theoretical results. A similar table would apply for the coefficients on OPV, ALS, and CM. (In the latter two, the β coefficient was not assumed to be time varying.) It can be seen in the table, for example, that there are directly four different estimates of $\gamma_{D,87}$ in the second to fifth equations, and likewise for each of the other parameters. Combining the entries in Table 11.13 with the counterparts for the coefficients on OPV, we see 50 SUR/FGLS estimates to be used to estimate 20 underlying parameters. The author used a minimum distance approach to reconcile the different estimates. We will return to this example in Example 13.6, where we will develop the MDE in more detail.

TABLE 11.13 Coefficient Estimates in SUR Model for Hospital Costs

	Coefficient on Variable in the Equation				
Equation	***DIS87***	***DIS88***	***DIS89***	***DIS90***	***DIS91***
SUR87	$\beta_{D,87} + \gamma_{D,87}$ 1.76	$\gamma_{D,88}$ 0.116	$\gamma_{D,89}$ -0.0881	$\gamma_{D,90}$ 0.0570	$\gamma_{D,91}$ -0.0617
SUR88	$\gamma_{D,87}$ 0.254	$\beta_{D,88} + \gamma_{D,88}$ 1.61	$\gamma_{D,89}$ -0.0934	$\gamma_{D,90}$ 0.0610	$\gamma_{D,91}$ -0.0514
SUR89	$\gamma_{D,87}$ 0.217	$\gamma_{D,88}$ 0.0846	$\beta_{D,89} + \gamma_{D,89}$ 1.51	$\gamma_{D,90}$ 0.0454	$\gamma_{D,91}$ -0.0253
SUR90	$\gamma_{D,87}$ 0.179	$\gamma_{D,88}$ 0.0822[a]	$\gamma_{D,89}$ 0.0295	$\beta_{D,90} + \gamma_{D,90}$ 1.57	$\gamma_{D,91}$ 0.0244
SUR91	$\gamma_{D,87}$ 0.153	$\gamma_{D,88}$ 0.0363	$\gamma_{D,89}$ -0.0422	$\gamma_{D,90}$ 0.0813	$\beta_{D,91} + \gamma_{D,91}$ 1.70

[a] The value reported in the published paper is 8.22. The correct value is 0.0822. (Personal communication with the author.)

11.6 NONSPHERICAL DISTURBANCES AND ROBUST COVARIANCE MATRIX ESTIMATION

Because the models considered here are extensions of the classical regression model, we can treat heteroscedasticity in the same way that we did in Chapter 9. That is, we can compute the ordinary or feasible generalized least squares estimators and obtain an appropriate robust covariance matrix estimator, or we can impose some structure on the disturbance variances and use generalized least squares. In the panel data settings, there is greater flexibility for the second of these without making strong assumptions about the nature of the heteroscedasticity.

11.6.1 HETEROSCEDASTICITY IN THE RANDOM EFFECTS MODEL

Because the random effects model is a generalized regression model with a known structure, OLS with a robust estimator of the asymptotic covariance matrix is not the best use of the data. The GLS estimator is efficient whereas the OLS estimator is not. If a perfectly general covariance structure is assumed, then one might simply use Arellano's estimator, described in Section 11.4.3, with a single overall constant term rather than a set of fixed effects. But, within the setting of the random effects model, $\eta_{it} = \varepsilon_{it} + u_i$, allowing the disturbance variance to vary across groups would seem to be a useful extension.

The calculation in (11-33) has a type of heteroscedasticity due to the varying group sizes. The estimator there (and its feasible counterpart) would be the same if, instead of $\theta_i = 1 - \sigma_\varepsilon/(T_i\sigma_u^2 + \sigma_\varepsilon^2)^{1/2}$, the disturbances were specifically heteroscedastic with $E[\varepsilon_{it}{}^2|\mathbf{X}_i] = \sigma_{\varepsilon i}{}^2$ and

$$\theta_i = 1 - \frac{\sigma_{\varepsilon i}}{\sqrt{\sigma_{\varepsilon i}^2 + T_i\sigma_u^2}}.$$

Therefore, for computing the appropriate feasible generalized least squares estimator, once again we need only devise consistent estimators for the variance components and then apply the GLS transformation shown earlier. One possible way to proceed is as follows: Because pooled OLS is still consistent, OLS provides a usable set of residuals. Using the OLS residuals for the specific groups, we would have, for each group,

$$\sigma_{\varepsilon i}^2 + u_i^2 = \frac{\mathbf{e}_i'\mathbf{e}_i}{T}.$$

The residuals from the dummy variable model are purged of the individual specific effect, u_i, so $\sigma_{\varepsilon i}^2$ may be consistently (in T) estimated with

$$\sigma_{\varepsilon i}^2 = \frac{\mathbf{e}_i^{lsdv\prime}\mathbf{e}_i^{lsdv}}{T},$$

where $e_{it}^{lsdv} = y_{it} - \mathbf{x}_{it}'\mathbf{b}^{lsdv} - a_i$. Combining terms, then,

$$\hat{\sigma}_u^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{\mathbf{e}_i^{ols\prime}\mathbf{e}_i^{ols}}{T}\right) - \left(\frac{\mathbf{e}_i^{lsdv\prime}\mathbf{e}_i^{lsdv}}{T}\right)\right] = \frac{1}{n}\sum_{i=1}^{n}(u_i^2).$$

We can now compute the FGLS estimator as before.

11.6.2 AUTOCORRELATION IN PANEL DATA MODELS

As we saw in Section 11.3.2 and Example 11.4, autocorrelation—that is, correlation across the observations in the groups in a panel—is likely to be a substantive feature of the model. Our treatment of the effect there, however, was meant to accommodate autocorrelation in its broadest sense, that is, nonzero covariances across observations in a group. The results there would apply equally to clustered observations, as observed in Section 11.3.3. An important element of that specification was that with clustered data, there might be no obvious structure to the autocorrelation. When the panel data set consists explicitly of groups of time series, and especially if the time series are relatively long as in Example 11.9, one might want to begin to invoke the more detailed, structured time-series models which are discussed in Chapter 20.

11.7 SPATIAL AUTOCORRELATION

The clustering effects suggested in Section 11.3.3 are motivated by an expectation that effects of neighboring locations would spill over into each other, creating a sort of correlation across space, rather than across time as we have focused on thus far. The effect should be common in cross-region studies, such as in agriculture, urban economics, and regional science. Studies of the phenomenon include Case's (1991) study of expenditure patterns, Bell and Bockstael's (2000) study of real estate prices, Baltagi and Li's (2001) analysis of R&D spillovers, Fowler, Cover and Kleit's (2014) study of fringe banking, Klier and McMillen's (2012) analysis of clustering of auto parts suppliers, and Flores-Lagunes and Schnier's (2012) model of cod fishing performance. Models of spatial regression and **spatial autocorrelation** are constructed to formalize this notion.[30]

[30] See Anselin (1988, 2001) for the canonical reference and Le Sage and Pace (2009) for a recent survey.

A model with spatial autocorrelation can be formulated as follows: The regression model takes the familiar panel structure,

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} + u_{i,}i = 1, \ldots, n; t = 1, \ldots, T.$$

The common u_i is the usual unit (e.g., country) effect. The correlation across space is implied by the spatial autocorrelation structure,

$$\varepsilon_{it} = \lambda \sum_{j=1}^{n} W_{ij}\varepsilon_{jt} + v_t.$$

The scalar λ is the **spatial autocorrelation coefficient**. The elements W_{ij} are spatial (or **contiguity**) weights that are assumed known. The elements that appear in the sum above are a row of the spatial weight or **contiguity matrix**, **W**, so that for the n units, we have

$$\boldsymbol{\varepsilon}_t = \lambda \mathbf{W}\boldsymbol{\varepsilon}_t + \mathbf{v}_t, \mathbf{v}_t = v_t\mathbf{i}.$$

The structure of the model is embodied in the symmetric weight matrix, **W**. Consider for an example counties or states arranged geographically on a grid or some linear scale such as a line from one coast of the country to another. Typically W_{ij} will equal one for *i,j* pairs that are neighbors and zero otherwise. Alternatively, W_{ij} may reflect distances across space, so that W_{ij} decreases with increases in $|i - j|$. In Flores-Lagunes and Schnier's (2012) study, the spatial weights were inversely proportional to the Euclidean distances between points in a grid. This would be similar to a temporal autocorrelation matrix. Assuming that $|\lambda|$ is less than one, and that the elements of **W** are such that $(\mathbf{I} - \lambda\mathbf{W})$ is nonsingular, we may write

$$\boldsymbol{\varepsilon}_t = (\mathbf{I}_n - \lambda\mathbf{W})^{-1}\mathbf{v}_t,$$

so for the n observations at time t,

$$\mathbf{y}_t = \mathbf{X}_t\boldsymbol{\beta} + (\mathbf{I}_n - \lambda\mathbf{W})^{-1}\mathbf{v}_t + \mathbf{u}.$$

We further assume that u_i and v_i have zero means, variances σ_u^2 and σ_v^2, and are independent across countries and of each other. It follows that a generalized regression model applies to the n observations at time t,

$$E[\mathbf{y}_t|\mathbf{X}_t] = \mathbf{X}_t\,\boldsymbol{\beta},$$
$$\text{Var}[\mathbf{y}_t|\mathbf{X}_t] = (\mathbf{I}_n - \lambda\mathbf{W})^{-1}[\sigma_v^2\mathbf{ii}'](\mathbf{I}_n - \lambda\mathbf{W})^{-1} + \sigma_u^2\mathbf{I}_n.$$

At this point, estimation could proceed along the lines of Chapter 9, save for the need to estimate λ. There is no natural residual-based estimator of λ. Recent treatments of this model have added a normality assumption and employed maximum likelihood methods.[31]

A natural first step in the analysis is a test for spatial effects. The standard procedure for a cross section is Moran's (1950) I statistic, which would be computed for each set of residuals, $\mathbf{e}_t$, using

$$I_t = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n}W_{ij}(e_{it} - \bar{e}_t)(e_{jt} - \bar{e}_t)}{\left(\sum_{i=1}^{n}\sum_{j=1}^{n}W_{i,j}\right)\sum_{i=1}^{n}(e_{it} - \bar{e}_t)^2}. \tag{11-53}$$

[31] The log-likelihood function for this model and numerous references appear in Baltagi (2005, p. 196). Extensive analysis of the estimation problem is given in Bell and Bockstael (2000).

For a panel of T independent sets of observations, $\bar{I} = \frac{1}{T}\sum_{t=1}^{T} I_t$ would use the full set of information. A large sample approximation to the variance of the statistic under the null hypothesis of no spatial autocorrelation is

$$V^2 = \frac{1}{T}\frac{n^2\sum_{i=1}^{n}\sum_{j=1}^{n}W_{ij}^2 + 3\left(\sum_{i=1}^{n}\sum_{j=1}^{n}W_{ij}\right)^2 - n\sum_{i=1}^{n}\left(\sum_{j=1}^{n}W_{ij}\right)^2}{(n^2-1)\left(\sum_{i=1}^{n}\sum_{j=1}^{n}W_{ij}\right)^2}. \quad \textbf{(11-54)}$$

The statistic $\bar{I}/V$ will converge to standard normality under the null hypothesis and can form the basis of the test. (The assumption of independence across time is likely to be dubious at best, however.) Baltagi, Song, and Koh (2003) identify a variety of LM tests based on the assumption of normality. Two that apply to cross-section analysis are[32]

$$LM(1) = \frac{(\mathbf{e}'\mathbf{W}\mathbf{e}/s^2)^2}{tr(\mathbf{W}'\mathbf{W} + \mathbf{W}^2)}$$

for spatial autocorrelation and

$$LM(2) = \frac{(\mathbf{e}'\mathbf{W}\mathbf{y}/s^2)^2}{\mathbf{b}'\mathbf{X}'\mathbf{W}\mathbf{M}\mathbf{W}\mathbf{X}\mathbf{b}/s^2 + tr(\mathbf{W}'\mathbf{W} + \mathbf{W}^2)}$$

for spatially lagged dependent variables, where $\mathbf{e}$ is the vector of OLS residuals, $s^2 = \mathbf{e}'\mathbf{e}/n$, and $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. [33]

Anselin (1988) identifies several possible extensions of the spatial model to dynamic regressions. A “pure space-recursive model” specifies that the autocorrelation pertains to neighbors in the previous period,

$$y_{it} = \gamma[\mathbf{W}\mathbf{y}_{t-1}]_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}.$$

A “time-space recursive model” specifies dependence that is purely autoregressive with respect to neighbors in the previous period,

$$y_{it} = \rho y_{i,t-1} + \gamma[\mathbf{W}\mathbf{y}_{t-1}]_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}.$$

A “time-space simultaneous” model specifies that the spatial dependence is with respect to neighbors in the current period,

$$y_{it} = \rho y_{i,t-1} + \lambda[\mathbf{W}\mathbf{y}_t]_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}.$$

Finally, a “time-space dynamic model” specifies that autoregression depends on neighbors in both the current and last period,

$$y_{it} = \rho y_{i,t-1} + \lambda[\mathbf{W}\mathbf{y}_t]_i + \gamma[\mathbf{W}\mathbf{y}_{t-1}]_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}.$$

Example 11.14 Spatial Autocorrelation in Real Estate Sales

Bell and Bockstael analyzed the problem of modeling spatial autocorrelation in large samples. This is a common problem with GIS (geographic information system) data sets. The central problem is maximization of a likelihood function that involves a sparse matrix, ($\mathbf{I} - \lambda\mathbf{W}$). Direct approaches to the problem can encounter severe inaccuracies in evaluation of the inverse

[32] See Bell and Bockstael (2000, p. 78).

[33] See Anselin and Hudak (1992).

and determinant. Kelejian and Prucha (1999) have developed a moment-based estimator for λ that helps alleviate the problem. Once the estimate of λ is in hand, estimation of the spatial autocorrelation model is done by FGLS. The authors applied the method to analysis of a cross section of 1,000 residential sales in Anne Arundel County, Maryland, from 1993 to 1996. The parcels sold all involved houses built within one year prior to the sale. GIS software was used to measure attributes of interest.

The model is

$$
\begin{aligned}
&+ \beta_2 \ln \textit{Lot size } (\textit{LLT}) \\
&+ \beta_3 \ln \textit{Distance in km to Washington, DC } (\textit{LDC}) \\
&+ \beta_4 \ln \textit{Distance in km to Baltimore } (\textit{LBA}) \\
&+ \beta_5 \textit{\% land surrounding parcel in publicly owned space } (\textit{POPN}) \\
&+ \beta_6 \textit{\% land surrounding parcel in natural privately owned space } (\textit{PNAT}) \\
&+ \beta_7 \textit{\% land surrounding parcel in intensively developed use } (\textit{PDEV}) \\
&+ \beta_8 \textit{\% land surrounding parcel in low density residential use } (\textit{PLOW}) \\
&+ \beta_9 \textit{ Public sewer service } (1 \textit{ if existing or planned}, 0 \textit{ if not})(\textit{PSEW}) \\
&+ \varepsilon.
\end{aligned}
$$

(Land surrounding the parcel is all parcels in the GIS data whose centroids are within 500 meters of the transacted parcel.) For the full model, the specification is

$$
\begin{aligned}
\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \\
\boldsymbol{\varepsilon} &= \lambda \mathbf{W}\boldsymbol{\varepsilon} + \mathbf{v}.
\end{aligned}
$$

The authors defined four contiguity matrices:

W1: $W_{ij} = 1/\text{distance between } i \text{ and } j$ if distance < 600 meters, 0 otherwise,
W2: $W_{ij} = 1$ if distance between i and $j < 200$ meters, 0 otherwise,
W3: $W_{ij} = 1$ if distance between i and $j < 400$ meters, 0 otherwise,
W4: $W_{ij} = 1$ if distance between i and $j < 600$ meters, 0 otherwise.

All contiguity matrices were row-standardized. That is, elements in each row are scaled so that the row sums to one. One of the objectives of the study was to examine the impact of row standardization on the estimation. It is done to improve the numerical stability of the optimization process. Because the estimates depend numerically on the normalization, it is not completely innocent.

Test statistics for spatial autocorrelation based on the OLS residuals are shown in Table 11.14. (These are taken from the authors' Table 3.) The Moran statistics are distributed as standard normal while the LM statistics are distributed as chi-squared with one degree of freedom. All but the LM(2) statistic for W3 are larger than the 99 percent critical value from the respective table, so we would conclude that there is evidence of spatial autocorrelation. Estimates from some of the regressions are shown in Table 11.15. In the remaining results in the study, the authors find that the outcomes are somewhat sensitive to the specification of the spatial weight matrix, but not particularly so to the method of estimating λ.

TABLE 11.14 Test Statistics for Spatial Autocorrelation

	W1	*W2*	*W3*	*W4*
Moran's *I*	7.89	9.67	13.66	6.88
LM(1)	49.95	84.93	156.48	36.46
LM(2)	7.40	17.22	2.33	7.42

TABLE 11.15 Estimated Spatial Regression Models

	OLS		*FGLS*[a]		*Spatial Based on W1 ML*		*Spatial Based on W1 Gen. Moments*	
Parameter	*Estimate*	*Std.Err.*	*Estimate*	*Std.Err.*	*Estimate*	*Std.Err.*	*Estimate*	*Std.Err.*
α	4.7332	0.2047	4.7380	0.2048	5.1277	0.2204	5.0648	0.2169
β_1	0.6926	0.0124	0.6924	0.0214	0.6537	0.0135	0.6638	0.0132
β_2	0.0079	0.0052	0.0078	0.0052	0.0002	0.0052	0.0020	0.0053
β_3	−0.1494	0.0195	−0.1501	0.0195	−0.1774	0.0245	−0.1691	0.0230
β_4	−0.0453	0.0114	−0.0455	0.0114	−0.0169	0.0156	−0.0278	0.0143
β_5	−0.0493	0.0408	−0.0484	0.0408	−0.0149	0.0414	−0.0269	0.0413
β_6	0.0799	0.0177	0.0800	0.0177	0.0586	0.0213	0.0644	0.0204
β_7	0.0677	0.0180	0.0680	0.0180	0.0253	0.0221	0.0394	0.0211
β_8	−0.0166	0.0194	−0.0168	0.0194	−0.0374	0.0224	−0.0313	0.0215
β_9	−0.1187	0.0173	−0.1192	0.0174	−0.0828	0.0180	−0.0939	0.0179
λ	—	—	—	—	0.4582	0.0454	0.3517	—

[a] The authors report using a heteroscedasticity model $\sigma_i^2 \times f(LIV_i, LIV_i^2)$. The function $f(.)$ is not identified.

Example 11.15 Spatial Lags in Health Expenditures

Moscone, Knapp, and Tosetti (2007) investigated the determinants of mental health expenditure over six years in 148 British local authorities using two forms of the spatial correlation model to incorporate possible interaction among authorities as well as unobserved spatial heterogeneity. The models estimated, in addition to pooled regression and a random effects model, were as follows. The first is a model with **spatial lags**,

$$\mathbf{y}_t = \gamma_t\mathbf{i} + \rho\mathbf{W}\mathbf{y}_t + \mathbf{X}_t\boldsymbol{\beta} + \mathbf{u} + \boldsymbol{\varepsilon}_t,$$

where **u** is a 148 × 1 vector of random effects and **i** is a 148 × 1 column of ones. For each local authority,

$$y_{it} = \gamma_t + \rho(\mathbf{w}_i'\mathbf{y}_t) + \mathbf{x}_{it}'\boldsymbol{\beta} + u_i + \varepsilon_{it},$$

where $\mathbf{w}_i'$ is the *i*th row of the contiguity matrix, **W**. Contiguities were defined in **W** as one if the locality shared a border or vertex and zero otherwise. (The authors also experimented with other contiguity matrices based on “sociodemographic” differences.) The second model estimated is of **spatial error correlation,**

$$\mathbf{y}_t = \gamma_t\mathbf{i} + \mathbf{X}_t\boldsymbol{\beta} + \mathbf{u} + \boldsymbol{\varepsilon}_t,$$
$$\boldsymbol{\varepsilon}_t = \lambda\,\mathbf{W}\boldsymbol{\varepsilon}_t + \mathbf{v}_t.$$

For each local authority, this model implies

$$y_{it} = \gamma_t + \mathbf{x}_{it}'\boldsymbol{\beta} + u_i + \lambda\Sigma_j w_{ij}\varepsilon_{jt} + v_{it}.$$

The authors use maximum likelihood to estimate the parameters of the model. To simplify the computations, they note that the maximization can be done using a two-step procedure. As we have seen in other applications, when $\boldsymbol{\Omega}$ in a generalized regression model is known, the appropriate estimator is GLS. For both of these models, with known spatial autocorrelation parameter, a GLS transformation of the data produces a classical regression model. [See (9-11).] The method used is to iterate back and forth between simple OLS estimation of γ_t, $\boldsymbol{\beta}$, and σ_ε^2 and maximization of the concentrated log-likelihood function which, given the other estimates, is a function of the spatial autocorrelation parameter, ρ or λ, and the variance of the heterogeneity, σ_u^2.

The dependent variable in the models is the log of per capita mental health expenditures. The covariates are the percentage of males and of people under 20 in the area, average mortgage rates, numbers of unemployment claims, employment, average house price, median weekly wage, percent of single parent households, dummy variables for Labour party or Liberal Democrat party authorities, and the density of population ("to control for supply-side factors"). The estimated spatial autocorrelation coefficients for the two models are 0.1579 and 0.1220, both more than twice as large as the estimated standard error. Based on the simple Wald tests, the hypothesis of no spatial correlation would be rejected. The log-likelihood values for the two spatial models were +206.3 and +202.8, compared to −211.1 for the model with no spatial effects or region effects, so the results seem to favor the spatial models based on a chi-squared test statistic (with one degree of freedom) of twice the difference. However, there is an ambiguity in this result as the improved "fit" could be due to the region effects rather than the spatial effects. A simple random effects model shows a log-likelihood value of +202.3, which bears this out. Measured against this value, the spatial lag model seems the preferred specification, whereas the spatial autocorrelation model does not add significantly to the log-likelihood function compared to the basic random effects model.

11.8 ENDOGENEITY

Recent **panel data** applications have relied heavily on the methods of instrumental variables. We will develop some of this methodology in detail in Chapter 13 where we consider generalized method of moments (GMM) estimation. At this point, we can examine three major building blocks in this set of methods, a panel data counterpart to two-stage least squares developed in Chapter 8, Hausman and Taylor's (1981) estimator for the random effects model and Bhargava and Sargan's (1983) proposals for estimating a dynamic panel data model. These tools play a significant role in the GMM estimators of dynamic panel models in Chapter 13.

11.8.1 INSTRUMENTAL VARIABLE ESTIMATION

The exogeneity assumption, $E[\mathbf{x}_{it}\varepsilon_{it}] = \mathbf{0}$, has been essential to the estimation strategies suggested thus far. For the generalized regression model (random effects), it was necessary to strengthen this to strict exogeneity, $E[\mathbf{x}_{it}\varepsilon_{is}] = \mathbf{0}$ for all t,s for given i. If these assumptions are not met, then $\mathbf{x}_{it}$ is endogenous in the model, and typically an instrumental variable approach to consistent estimation would be called for.

The fixed effects case is simpler, and can be based entirely on results we have already obtained. The model is $y_{it} = c_i + \mathbf{x}_{it}{}'\boldsymbol{\beta} + \varepsilon_{it}$. We assume there is a set of $L \geq K$ instrumental variables, $\mathbf{z}_{it}$. The set of instrumental variables must be exogenous, that is, orthogonal to ε_{it}; the minimal assumption is $E[\mathbf{z}_{it}\varepsilon_{it}] = \mathbf{0}$. (It will turn out, at least initially, to be immaterial to estimation of $\boldsymbol{\beta}$ whether $E[\mathbf{z}_{it}c_i] = \mathbf{0}$, though one would expect it would be.) Then, the model in deviation form,

$$\begin{aligned} y_{it} - \overline{y}_{i.} &= (\mathbf{x}_{it} - \overline{\mathbf{x}}_{i.})'\boldsymbol{\beta} + (\varepsilon_{it} - \overline{\varepsilon}_{i.}) \\ \ddot{y} &= \ddot{\mathbf{x}}_{it}'\boldsymbol{\beta} + \ddot{\varepsilon}_{it}, \end{aligned}$$

is amenable to 2SLS. The IV estimator can be written

$$\mathbf{b}_{IV,FE} = (\ddot{\mathbf{X}}'\ddot{\mathbf{Z}}(\ddot{\mathbf{Z}}'\ddot{\mathbf{Z}})^{-1}\ddot{\mathbf{Z}}'\ddot{\mathbf{X}})^{-1}(\ddot{\mathbf{X}}'\ddot{\mathbf{Z}}(\ddot{\mathbf{Z}}'\ddot{\mathbf{Z}})^{-1}\ddot{\mathbf{Z}}'\ddot{\mathbf{y}}).$$

We can see from this expression that this computation will break down if $\mathbf{Z}$ contains any time-invariant variables. Clearly if there are, then the corresponding columns in $\ddot{\mathbf{Z}}$ will be zero. But, even if $\mathbf{Z}$ is not transformed, columns of $\ddot{\mathbf{X}}'\mathbf{Z}$ will still turn to zeros because $\ddot{\mathbf{X}}'\mathbf{Z} = \mathbf{X}'\ddot{\mathbf{Z}}$ ($\mathbf{M}^0$ is idempotent). Assuming, then, that $\mathbf{Z}$ is also transformed to deviations from group means, the 2SLS estimator is

$$\begin{aligned}\mathbf{b}_{IV,FE} &= \left[\sum_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{Z}}_i(\ddot{\mathbf{Z}}_i'\ddot{\mathbf{Z}}_i)^{-1}\ddot{\mathbf{Z}}_i'\ddot{\mathbf{X}}_i\right]^{-1}\left[\sum_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{Z}}_i(\ddot{\mathbf{Z}}_i'\ddot{\mathbf{Z}}_i)^{-1}\ddot{\mathbf{Z}}_i'\ddot{\mathbf{y}}_i\right] \\ &= \left[\sum_{i=1}^{n}\hat{\ddot{\mathbf{X}}}_i'\hat{\ddot{\mathbf{X}}}_i\right]^{-1}\left[\sum_{i=1}^{n}\hat{\ddot{\mathbf{X}}}_i'\ddot{\mathbf{y}}_i\right].\end{aligned} \quad \textbf{(11-55)}$$

For computing the asymptotic covariance matrix, without correction, we would use

$$\text{Est.Asy.Var}[\mathbf{b}_{IV,FE}] = \hat{\sigma}_\varepsilon^2\left[\sum_{i=1}^{n}\hat{\ddot{\mathbf{X}}}_i'\hat{\ddot{\mathbf{X}}}_i\right]^{-1}$$

where

$$\hat{\sigma}_\varepsilon^2 = \frac{\Sigma_{i=1}^{n}\Sigma_{t=1}^{t}(\ddot{y}_{it} - \ddot{\mathbf{X}}_{it}'\mathbf{b}_{IV,FE})^2}{n(T-1) - K}. \quad \textbf{(11-56)}$$

An asymptotic covariance matrix that is robust to heteroscedasticity and autocorrelation is

$$\text{Est.Asy.Var}[\mathbf{b}_{IV,FE}] = \left[\sum_{i=1}^{n}\hat{\ddot{\mathbf{X}}}_i'\hat{\ddot{\mathbf{X}}}_i\right]^{-1}\left[\sum_{i=1}^{n}\left(\hat{\ddot{\mathbf{X}}}_i'\ddot{\mathbf{e}}_i\right)\left(\ddot{\mathbf{e}}_i'\hat{\ddot{\mathbf{X}}}_i\right)\right]\left[\sum_{i=1}^{n}\hat{\ddot{\mathbf{X}}}_i'\hat{\ddot{\mathbf{X}}}_i\right]^{-1}. \quad \textbf{(11-57)}$$

The procedure would be similar for the random effects model, but would (as before) require a first step to estimate the variances of ε and u. The steps follow the earlier prescription:

1. Use pooled 2SLS to compute $\hat{\boldsymbol{\beta}}_{IV,Pooled}$ and obtain residuals $\mathbf{w}$. The estimator of $\sigma_\varepsilon^2 + \sigma_u^2$ is $\mathbf{w}'\mathbf{w}/(nT\text{-}K)$. Use FE 2SLS as described above to obtain $\mathbf{b}_{\text{IV,FE}}$, then use (11-56) to estimate $\sigma_\varepsilon{}^2$. Use these two estimators to compute the estimator of $\sigma_u{}^2$, then $\boldsymbol{\Sigma}^{-1} = (1/\sigma_\varepsilon^2)[\mathbf{I}_T - (\theta(2-\theta)/T)\mathbf{ii}']$. [The result for $\boldsymbol{\Sigma}^{-1/2}$ is given in (11-33).]
2. Use IV for the generalized regression model,

$$\hat{\boldsymbol{\beta}}_{IV,RE} = \left[\sum_{i=1}^{n}\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{Z}_i(\mathbf{Z}_i'\boldsymbol{\Sigma}^{-1}\mathbf{Z}_i)^{-1}\mathbf{Z}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_i'\right]^{-1}\left[\sum_{i=1}^{n}\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{Z}_i(\mathbf{Z}_i'\boldsymbol{\Sigma}^{-1}\mathbf{Z}_i)^{-1}\mathbf{Z}_i'\boldsymbol{\Sigma}^{-1}\mathbf{y}_i\right]. \quad \textbf{(11-58)}$$

3. The estimator for the asymptotic covariance matrix is the bracketed inverse. A robust covariance matrix is computed with

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\beta}}_{IV,RE}] =$$

$$\begin{aligned}&\mathbf{A}^{-1}\left[\sum_{i=1}^{n}(\mathbf{X}_i'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Z}_i(\mathbf{Z}_i'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Z})^{-1}\mathbf{Z}_i'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\varepsilon}}_i)(\mathbf{X}_i'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Z}_i(\mathbf{Z}_i'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Z}_i)^{-1}\mathbf{Z}_i'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\varepsilon}}_i)'\right]\mathbf{A}^{-1} \\ &\mathbf{A} = \left[\sum_{i=1}^{n}\mathbf{X}_i'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Z}_i(\mathbf{Z}_i'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Z}_1)^{-1})\mathbf{Z}_i'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}_i\right].\end{aligned} \quad \textbf{(11-59)}$$

TABLE 11.16 Estimated Health Satisfaction Equations (Robust standard errors in parentheses)

Variable	*OLS*	*2SLS*	*FE*	*RE*	*FE/2SLS*	*RE/2SLS*
Constant	9.17989	10.7061	—	9.69595	—	12.1185
	(0.36704)	(0.36931)	—	(0.28573)	—	(0.75062)
ln *Income*	0.18045	1.16373	0.13957	0.13001	0.99046	1.24378
	(0.10931)	(0.20863)	(0.10246)	(0.06970)	(0.48337)	(0.33140)
Working	0.63475	0.34196	0.12963	0.29491	−0.05739	0.00243
	(0.12705)	(0.09007)	(0.11656)	(0.07392)	(0.15171)	(0.12932)
Public	−0.78176	−0.52551	−0.20282	−0.48854	−0.15991	−0.29334
	(0.15438)	(0.10963)	(0.17409)	(0.12775)	(0.16779)	(0.13964)
Add On	0.18664	−0.06131	−0.03252	0.04340	−0.01482	−0.02720
	(0.29279)	(0.24477)	(0.17287)	(0.21060)	(0.16327)	(0.15847)
Age	−0.04606	−0.05523	−0.07178	−0.05926	−0.10419	−0.08409
	(0.00583)	(0.00369)	(0.00900)	(0.00468)	(0.01992)	(0.00882)
σ_ε	2.17305	2.21080	1.57382	2.47692	1.59032	2.57864
σ_u	—	—	—	1.49841	—	1.53728

Example 11.16 Endogenous Income in a Health Production Model

In Example 10.8, we examined a health outcome, health satisfaction, in a two-equation model,

$$\begin{aligned} \textit{Health Satisfaction} &= \alpha_1 + \gamma_1 \ln \textit{Income} + \alpha_2 \textit{Female} + \alpha_3 \textit{Working} + \alpha_4 \textit{Public} + \alpha_5 \textit{AddOn} \\ &\quad + \alpha_6 \textit{Age} + \varepsilon_{H,} \\ \ln \textit{Income} &= \beta_1 + \gamma_2 \textit{Health Satisfaction} + \beta_2 \textit{Female} + \beta_3 \textit{Education} + \beta_4 \textit{Married} \\ &\quad + \beta_5 \textit{HHKids} + \beta_6 \textit{Age} + \varepsilon_I. \end{aligned}$$

The data are an unbalanced panel of 7,293 households. For simplicity, we will focus on the balanced panel of 887 households that were present for all 7 waves. The variable ln *Income* is endogenous in the health equation. There is also a time-invariant variable, *Female*, in the equation that will have to be dropped in this application as we are going to fit a fixed effects model. The instrumental variables are the constant, *Working*, *Public*, *AddOn*, *Age*, *Education*, *Married,* and *HHKids*. Table 11.16 presents the OLS, 2SLS, FE, RE, FE2SLS, and RE2SLS estimates for the health satisfaction equation. Robust standard errors are reported for each case. There is a clear pattern in the results; the instrumental variable estimates of the coefficient on ln Income are 7 to 10 times as large as the least squares estimates, and the estimated standard errors increase comparably.

11.8.2 HAUSMAN AND TAYLOR'S INSTRUMENTAL VARIABLES ESTIMATOR

Recall the original specification of the linear model for panel data in (11-1),

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\alpha} + \varepsilon_{it}. \qquad \textbf{(11-60)}$$

The random effects model is based on the assumption that the unobserved person-specific effects, $\mathbf{z}_i$, are uncorrelated with the included variables, $\mathbf{x}_{it}$. This assumption is a major shortcoming of the model. However, the random effects treatment does allow the model to contain observed time-invariant characteristics, such as demographic characteristics, while the fixed effects model does not—if present, they are simply absorbed into the fixed effects. **Hausman and Taylor's** (1981) **estimator** for the random effects model suggests a way to overcome the first of these while accommodating the second.

Their model is of the form

$$y_{it} = \mathbf{x}_{1it}'\boldsymbol{\beta}_1 + \mathbf{x}_{2it}'\boldsymbol{\beta}_2 + \mathbf{z}_{1i}'\boldsymbol{\alpha}_1 + z_{2i}'\boldsymbol{\alpha}_2 + \varepsilon_{it} + u_i,$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1', \boldsymbol{\alpha}_2')'$. In this formulation, all individual effects denoted $\mathbf{z}_i$ are observed. As before, unobserved individual effects that are contained in $\mathbf{z}_i'\boldsymbol{\alpha}$ in (11-60) are contained in the person-specific random term, u_i. Hausman and Taylor define four sets of *observed* variables in the model:

$\mathbf{x}_{1it}$ is K_1 variables that are time varying and uncorrelated with u_i,
$\mathbf{z}_{1i}$ is L_1 variables that are time invariant and uncorrelated with u_i,
$\mathbf{x}_{2it}$ is K_2 variables that are time varying and are correlated with u_i,
$\mathbf{z}_{2i}$ is L_2 variables that are time invariant and are correlated with u_i.

The assumptions about the random terms in the model are

$$E[u_i | \mathbf{x}_{1it}, \mathbf{z}_{1i}] = 0 \text{ though } E[u_i | \mathbf{x}_{2it}, \mathbf{z}_{2i}] \neq 0,$$

$$\text{Var}[u_i | \mathbf{x}_{1it}, \mathbf{z}_{1i}, \mathbf{x}_{2it}, \mathbf{z}_{2i}] = \sigma_u^2,$$

$$\text{Cov}[\varepsilon_{it}, u_i | \mathbf{x}_{1it}, \mathbf{z}_{1i}, \mathbf{x}_{2it}, \mathbf{z}_{2i}] = 0,$$

$$\text{Var}[\varepsilon_{it}+u_i | \mathbf{x}_{1it}, \mathbf{z}_{1i}, \mathbf{x}_{2it}, \mathbf{z}_{2i}] = \sigma^2 = \sigma_\varepsilon^2 + \sigma_u^2,$$

$$\text{Corr}[\varepsilon_{it} + u_i, \varepsilon_{is} + u_i | \mathbf{x}_{1it}, \mathbf{z}_{1i}, \mathbf{x}_{2it}, \mathbf{z}_{2i}] = \rho = \sigma_u^2/\sigma^2.$$

Note the crucial assumption that one can distinguish sets of variables $\mathbf{x}_1$ and $\mathbf{z}_1$ that are uncorrelated with u_i from $\mathbf{x}_2$ and $\mathbf{z}_2$ which are not. The likely presence of $\mathbf{x}_2$ and $\mathbf{z}_2$ is what complicates specification and estimation of the random effects model in the first place.

By construction, any OLS or GLS estimators of this model are inconsistent when the model contains variables that are correlated with the random effects. Hausman and Taylor have proposed an instrumental variables estimator that uses only the information within the model (i.e., as already stated). The strategy for estimation is based on the following logic: First, by taking deviations from group means, we find that

$$y_{it} - \overline{y}_{i.} = (\mathbf{x}_{1it} - \overline{\mathbf{x}}_{1i.})'\boldsymbol{\beta}_1 + (\mathbf{x}_{2it} - \overline{\mathbf{x}}_{2i.})'\boldsymbol{\beta}_2 + \varepsilon_{it} - \overline{\varepsilon}_{i.}, \quad \textbf{(11-61)}$$

which implies that both parts of $\boldsymbol{\beta}$ can be consistently estimated by least squares, in spite of the correlation between $\mathbf{x}_2$ and u. This is the familiar, fixed effects, least squares dummy variable estimator—the transformation to deviations from group means removes from the model the part of the disturbance that is correlated with $\mathbf{x}_{2it}$. In the original model, Hausman and Taylor show that the group mean deviations can be used as $(K_1 + K_2)$ instrumental variables for estimation of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$. That is the implication of (11-61). Because $\mathbf{z}_1$ is uncorrelated with the disturbances, it can likewise serve as a set of L_1 instrumental variables. That leaves a necessity for L_2 instrumental variables. The authors show that the group means for $\mathbf{x}_1$ can serve as these remaining instruments, and the model will be identified so long as K_1 is greater than or equal to L_2. For identification purposes, then, K_1 must be at least as large as L_2. As usual, **feasible GLS** is better than OLS, and available. Likewise, FGLS is an improvement over simple instrumental variable estimation of the model, which is consistent but inefficient.

The authors propose the following set of steps for consistent and efficient estimation:

Step 1. Obtain the LSDV (fixed effects) estimator of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ based on $\mathbf{x}_1$ and $\mathbf{x}_2$. The residual variance estimator from this step is a consistent estimator of σ_ε^2.

Step 2. Form the within-groups residuals, e_{it}, from the LSDV regression at step 1. Stack the group means of these residuals in a full-sample-length data vector. Thus, $e_{it}^* = \bar{e}_{i.} = \frac{1}{T}\sum_{t=1}^{T}(y_{it} - \mathbf{x}_{it}'\mathbf{b}_w)$, $t = 1, \ldots, T$, $i = 1, \ldots, n$. (The individual constant term, a_i, is not included in e_{it}^*.) (Note, from (11-16b), $e_{it}^* = \bar{e}_{i.}$ is a_i, the ith constant term.) These group means are used as the dependent variable in an instrumental variable regression on $\mathbf{z}_1$ and $\mathbf{z}_2$ with instrumental variables $\mathbf{z}_1$ and $\mathbf{x}_1$. (Note the identification requirement that K_1, the number of variables in $\mathbf{x}_1$, be at least as large as L_2, the number of variables in $\mathbf{z}_2$.) The time-invariant variables are each repeated T times in the data matrices in this regression. This provides a consistent estimator of $\boldsymbol{\alpha}$.

Step 3. The residual variance in the regression in step 2 is a consistent estimator of $\sigma_*^2 = \sigma_u^2 + \sigma_\varepsilon^2/T$. From this estimator and the estimator of σ_ε^2 in step 1, we deduce an estimator of $\sigma_u^2 = \sigma_*^2 - \sigma_\varepsilon^2/T$. We then form the weight for feasible GLS in this model by forming the estimate of

$$\theta = 1 - \sqrt{\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_u^2}}.$$

Step 4. The final step is a weighted instrumental variable estimator. Let the full set of variables in the model be

$$\mathbf{w}_{it}' = (\mathbf{x}_{1it}', \mathbf{x}_{2it}', \mathbf{z}_{1i}', \mathbf{z}_{2i}').$$

Collect these nT observations in the rows of data matrix $\mathbf{W}$. The transformed variables for GLS are, as before when we first fit the random effects model,

$$\mathbf{w}_{it}^{*\prime} = \mathbf{w}_{it}' - \hat{\theta}\overline{\mathbf{w}}_{i.}' \quad \text{and} \quad y_{it}^* = y_{it} - \hat{\theta}\bar{y}_{i.},$$

where $\hat{\theta}$ denotes the sample estimate of θ. The transformed data are collected in the rows data matrix $\mathbf{W}^*$ and in column vector $\mathbf{y}^*$. Note in the case of the time-invariant variables in $\mathbf{w}_{it}$, the group mean is the original variable, and the transformation just multiplies the variable by $1 - \hat{\theta}$. The instrumental variables are

$$\mathbf{v}_{it}' = [(\mathbf{x}_{1it} - \overline{\mathbf{x}}_{1i.})', (\mathbf{x}_{2it} - \overline{\mathbf{x}}_{2i.})', \mathbf{z}_{1i}' \, \overline{\mathbf{x}}_{1i.}].$$

These are stacked in the rows of the $nT \times (K_1 + K_2 + L_1 + K_1)$ matrix $\mathbf{V}$. Note for the third and fourth sets of instruments, the time-invariant variables and group means are repeated for each member of the group. The instrumental variable estimator would be

$$(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\alpha}}')_{\text{IV}}' = [(\mathbf{W}^{*\prime}\mathbf{V})(\mathbf{V}'\mathbf{V})^{-1}(\mathbf{V}'\mathbf{W}^*)]^{-1}[(\mathbf{W}^{*\prime}\mathbf{V})(\mathbf{V}'\mathbf{V})^{-1}(\mathbf{V}'\mathbf{y}^*)].^{34} \qquad \textbf{(11-62)}$$

The instrumental variable estimator is consistent if the data are not weighted, that is, if $\mathbf{W}$ rather than $\mathbf{W}^*$ is used in the computation. But this is inefficient, in the same way that OLS is consistent but inefficient in estimation of the simpler random effects model.

[34] Note that the FGLS random effects estimator would be $(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\alpha}}')_{RE}' = [\mathbf{W}^{*\prime}\mathbf{W}^*]^{-1}\mathbf{W}^{*\prime}\mathbf{y}^*$.

Example 11.17 The Returns to Schooling

The economic returns to schooling have been a frequent topic of study by econometricians. The PSID and NLS data sets have provided a rich source of panel data for this effort. In wage (or log wage) equations, it is clear that the economic benefits of schooling are correlated with latent, unmeasured characteristics of the individual such as innate ability, intelligence, drive, or perseverance. As such, there is little question that simple random effects models based on panel data will suffer from the effects noted earlier. The fixed effects model is the obvious alternative, but these rich data sets contain many useful variables, such as race, union membership, and marital status, which are generally time invariant. Worse yet, the variable most of interest, years of schooling, is also time invariant. Hausman and Taylor (1981) proposed the estimator described here as a solution to these problems. The authors studied the effect of schooling on (the log of) wages using a random sample from the PSID of 750 men aged 25 to 55, observed in two years, 1968 and 1972. The two years were chosen so as to minimize the effect of serial correlation apart from the persistent unmeasured individual effects. The variables used in their model were as follows:

Experience	= age − years of schooling − 5,
Years of schooling	= continuous variable
Bad Health	= a dummy variable indicating general health,
Race	= adummy variable indicating nonwhite (70 of 750 observations),
Union	= a dummy variable indicating union membership,
Unemployed	= a dummy variable indicating previous year's unemployment.

The model also included a constant term and a period indicator.[35]

The primary focus of the study is the coefficient on schooling in the log wage equation. Because Schooling and, probably, Experience and Unemployed, are correlated with the latent effect, there is likely to be serious bias in conventional estimates of this equation. Table 11.17 reports some of their reported results. The OLS and random effects GLS results in the first two columns provide the benchmark for the rest of the study. The schooling coefficient is estimated at 0.0669, a value which the authors suspected was far too small. As we saw earlier, even in the presence of correlation between measured and latent effects, in this model, the LSDV estimator provides a consistent estimator of the coefficients on the time-varying variables. Therefore, we can use it in the **Hausman specification test** for correlation between the included variables and the latent heterogeneity. The calculations are shown in Section 11.5.5, result (11-44). Because there are three variables remaining in the LSDV equation, the chi-squared statistic has three degrees of freedom. The reported value of 20.2 is far larger than the 95% critical value of 7.81, so the results suggest that the random effects model is misspecified.

Hausman and Taylor proceeded to reestimate the log wage equation using their proposed estimator. The fourth and fifth sets of results in Table 11.17 present the instrumental variable estimates. The specification test given with the fourth set of results suggests that the procedure has produced the expected result. The hypothesis of the modified random effects model is now not rejected; the chi-squared value of 2.24 is much smaller than the critical value. The schooling variable is treated as endogenous (correlated with u_i) in both cases. The difference between the two is the treatment of Unemployed and Experience. In the preferred equation, they are included in $\mathbf{x}_2$ rather than $\mathbf{x}_1$. The end result of the exercise is, again, the coefficient on schooling, which has risen from 0.0669 in the worst specification (OLS) to 0.2169 in the last one, an increase of over 200 %. As the authors note, at the same time, the measured effect of race nearly vanishes.

[35] The coding of the latter is not given, but any two distinct values, including 0 for 1968 and 1 for 1972, would produce identical results. (Why?)

TABLE 11.17 Estimated Log Wage Equations

	Variables	*OLS*	*GLS/RE*	*LSDV*	*HT/IV-GLS*	*HT/IV-GLS*
$\mathbf{x}_1$	*Experience*	0.0132	0.0133	0.0241	0.0217	
		(0.0011)[a]	(0.0017)	(0.0042)	(0.0031)	
	Bad health	−0.0843	−0.0300	−0.0388	−0.0278	−0.0388
		(0.0412)	(0.0363)	(0.0460)	(0.0307)	(0.0348)
	Unemployed	−0.0015	−0.0402	−0.0560	−0.0559	
	Last Year	(0.0267)	(0.0207)	(0.0295)	(0.0246)	
	Time	NR[b]	NR	NR	NR	NR
$\mathbf{x}_2$	*Experience*					0.0241
						(0.0045)
	Unemployed					−0.0560
						(0.0279)
$\mathbf{z}_1$	*Race*	−0.0853	−0.0878		−0.0278	−0.0175
		(0.0328)	(0.0518)		(0.0752)	(0.0764)
	Union	0.0450	0.0374		0.1227	0.2240
		(0.0191)	(0.0296)		(0.0473)	(0.2863)
	Schooling	**0.0669**	**0.0676**			
		(0.0033)	(0.0052)			
	Constant	NR	NR	NR	NR	NR
$\mathbf{z}_2$	*Schooling*				**0.1246**	**0.2169**
					(0.0434)	(0.0979)
	σ_ε	0.321	0.192	0.160	0.190	0.629
	$\rho = \sigma_u^2/(\sigma_u^2 + \sigma_\varepsilon^2)$		0.632		0.661	0.817
	Spec. Test [3]		20.2		2.24	0.00

[a] Estimated asymptotic standard errors are given in parentheses.

[b] NR indicates that the coefficient estimate was not reported in the study.

Example 11.18 The Returns to Schooling

In Example 11.17, Hausman and Taylor find that the estimated effect of education in a wage equation increases substantially (nearly doubles from 0.0676 to 0.1246) when it is treated as endogenous in a random effects model, then increases again by 75% to 0.2169 when experience and unemployment status are also treated as endogenous. In this exercise, we will examine whether these results reappear in Cornwell and Rupert's application. (We do not have the unemployment indicator.) Three sets of least squares results, ordinary, fixed effects, and feasible GLS random effects, appear at the left of Table 11.18. The education effect in the RE model is about 11%. (Time-invariant education falls out of the fixed effects model.) The effect increases by 29% to 13.8% when education is treated as endogenous, which is similar to Hausman and Taylor's 12.5%. When experience is treated as exogenous, instead, the education effect rises again by 72%. (The second such increase in the Hausman/Taylor results resulted from treating experience as endogenous, not exogenous.)

11.8.3 CONSISTENT ESTIMATION OF DYNAMIC PANEL DATA MODELS: ANDERSON AND HSIAO'S IV ESTIMATOR

Consider a heterogeneous dynamic panel data model,

$$y_{it} = \gamma y_{i,t-1} + \mathbf{x}_{it}'\boldsymbol{\beta} + c_i + \varepsilon_{it}, \qquad \textbf{(11-63)}$$

TABLE 11.18 Hausman–Taylor Estimates of Wage Equation

	OLS	*LGLS/RE*	*FE*	*HT-RE/FGLS*		
				$x_1 =$ Exogenous Time Varying		
OCC	−0.14001	−0.04322	−0.02148	−0.02004	−0.02070	−0.01445
South	−0.05564	−0.00825	−0.00186	0.00821	0.00746	0.01512
SMSA	0.15167	−0.02840	−0.04247	−0.04227	−0.04183	−0.05219
IND	0.04679	0.00378	0.01921	0.01392	0.01359	0.01971
Exp	0.04010	0.08748	0.11321			0.10919
Expsq	−0.00067	−0.00076	−0.00042			−0.00048
				$x_2 =$ Endogenous Time Varying		
Exp				0.11313	0.11313	
ExpSq				−0.00042	−0.00042	
WKS	0.00422	0.00096	0.00084	0.00084	0.00084	0.00080
MS	0.04845	−0.07090	−0.02973	−0.02980	−0.02985	−0.03850
Union	0.09263	0.05835	0.03278	0.03293	0.03277	0.03773
				$f_1 =$ Exogenous Time Invariant		
Constant	5.25112	4.04144		2.82907	2.91273	1.74978
FEM	−0.36779	−0.30938		−0.13209	−0.13093	−0.18008
Blk	−0.16694	−0.21950		−0.27726	−0.28575	−0.13633
Education	**0.05670**	**0.10707**		**0.14440**		
				$f_2 =$ Endogenous Time Invariant		
Education					**0.13794**	**0.23726**
σ_ε	0.34936	0.15206	0.15206	0.15199	0.15199	0.15199
σ_u	—	0.31453		0.94179	0.94180	0.99443

where c_i is, as in the preceding sections of this chapter, individual unmeasured heterogeneity, that may or may not be correlated with $\mathbf{x}_{it}$. We consider methods of estimation for this model when T is fixed and relatively small, and n may be large and increasing.

Pooled OLS is obviously inconsistent. Rewrite (11-63) as

$$y_{it} = \gamma y_{i,t-1} + \mathbf{x}_{it}'\boldsymbol{\beta} + w_{it}.$$

The disturbance in this pooled regression may be correlated with $\mathbf{x}_{it}$, but either way, it is surely correlated with $y_{i,t-1}$. By substitution,

$$\text{Cov}[y_{i,t-1}, (c_i + \varepsilon_{it})] = \sigma_c^2 + \gamma\,\text{Cov}[y_{i,t-2}, (c_i + \varepsilon_{it})],$$

and so on. By repeated substitution, it can be seen that for $|\gamma| < 1$ and moderately large T,

$$\text{Cov}[y_{i,t-1}, (c_i + \varepsilon_{it})] \approx \sigma_c^2/(1 - \gamma). \qquad \textbf{(11-64)}$$

[It is useful to obtain this result from a different direction. If the stochastic process that is generating (y_{it}, c_i) is stationary, then $\text{Cov}[y_{i,t-1}, c_i] = \text{Cov}[y_{i,t-2}, c_i]$, from which we would obtain (11-64) directly. The assumption $|\gamma| < 1$ would be required for stationarity.]

Consequently, OLS and GLS are inconsistent. The fixed effects approach does not solve the problem either. Taking deviations from individual means, we have

$$y_{it} - \bar{y}_{i.} = (\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})'\boldsymbol{\beta} + \gamma(y_{i,t-1} - \bar{y}_{i.}) + (\varepsilon_{it} - \bar{\varepsilon}_{i.}).$$

Anderson and Hsiao (1981, 1982) show that

$$\begin{aligned}\text{Cov}[(y_{it} - \bar{y}_{i.}), (\varepsilon_{it} - \bar{\varepsilon}_{i.})] &\approx \frac{-\sigma_\varepsilon^2}{T(1-\gamma)^2}\left[\frac{(T-1) - T\gamma + \gamma^T}{T}\right]\\ &= \frac{-\sigma_\varepsilon^2}{T(1-\gamma)^2}\left[(1-\gamma) - \frac{1-\gamma^T}{T}\right].\end{aligned}$$

This does converge to zero as T increases, but, again, we are considering cases in which T is small or moderate, say 5 to 15, in which case the bias in the OLS estimator could be 15% to 60%. The implication is that the "within" transformation does not produce a consistent estimator.

It is easy to see that taking first differences is likewise ineffective. The first differences of the observations are

$$y_{it} - y_{i,t-1} = (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta} + \gamma(y_{i,t-1} - y_{i,t-2}) + (\varepsilon_{it} - \varepsilon_{i,t-1}). \tag{11-65}$$

As before, the correlation between the last regressor and the disturbance persists, so OLS or GLS based on first differences would also be inconsistent. There is another approach. Write the regression in differenced form as

$$\Delta y_{it} = \Delta\mathbf{x}_{it}'\boldsymbol{\beta} + \gamma\,\Delta y_{i,t-1} + \Delta\varepsilon_{it},$$

or, defining $\mathbf{x}_{it}^* = [\Delta\mathbf{x}_{it}, \Delta y_{i,t-1}]$, $\varepsilon_{it}^* = \Delta\varepsilon_{it}$ and $\boldsymbol{\theta} = [\boldsymbol{\beta}', \gamma]'$,

$$y_{it}^* = \mathbf{x}_{it}^{*\prime}\boldsymbol{\theta} + \varepsilon_{it}^*.$$

For the pooled sample, beginning with $t = 3$, write this as

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\theta} + \boldsymbol{\varepsilon}^*.$$

The least squares estimator based on the first differenced data is

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \left[\frac{1}{n(T-3)}\mathbf{X}^{*\prime}\mathbf{X}^*\right]^{-1}\left(\frac{1}{n(T-3)}\mathbf{X}^{*\prime}\mathbf{y}^*\right)\\ &= \boldsymbol{\theta} + \left[\frac{1}{n(T-3)}\mathbf{X}^{*\prime}\mathbf{X}^*\right]^{-1}\left(\frac{1}{n(T-3)}\mathbf{X}^{*\prime}\boldsymbol{\varepsilon}^*\right).\end{aligned}$$

Assuming that the inverse matrix in brackets converges to a positive definite matrix—that remains to be shown—the inconsistency in this estimator arises because the vector in parentheses does not converge to zero. The last element is $\text{plim}_{n\to\infty}[1/(n(T-3))]\Sigma_{i=1}^n\Sigma_{t=3}^T(y_{i,t-1} - y_{i,t-2})(\varepsilon_{it} - \varepsilon_{i,t-1})$, which is not zero.

Suppose there were a variable $\mathbf{z}^*$ such that plim $[1/(n(T-3))]\mathbf{z}^{*\prime}\boldsymbol{\varepsilon}^* = 0$ (exogenous) and plim$[1/(n(T-3))]\mathbf{z}^{*\prime}\mathbf{X}^* \neq \mathbf{0}$ (relevant). Let $\mathbf{Z} = [\Delta\mathbf{X}, \mathbf{z}^*]$; z_{it}^* replaces $\Delta y_{i,t-1}$ in $\mathbf{x}_{it}^*$. By this construction, it appears we have a consistent estimator. Consider

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{IV} &= (\mathbf{Z}'\mathbf{X}^*)^{-1}\mathbf{Z}'\mathbf{y}^*.\\ &= (\mathbf{Z}'\mathbf{X}^*)^{-1}\mathbf{Z}'(\mathbf{X}^*\boldsymbol{\theta} + \boldsymbol{\varepsilon}^*)\\ &= \boldsymbol{\theta} + (\mathbf{Z}'\mathbf{X}^*)^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}^*.\end{aligned}$$

Then, after multiplying throughout by $1/(n(T-3))$ as before, we find

$$\text{Plim } \hat{\boldsymbol{\theta}}_{\text{IV}} = \boldsymbol{\theta} + \text{plim}\{[1/(n(T-3))](\mathbf{Z}'\mathbf{X}^*)\}^{-1} \times \mathbf{0},$$

which seems to solve the problem of consistent estimation.

The variable z^* is an **instrumental variable**, and the estimator is an **instrumental variable estimator** (hence the subscript on the preceding estimator). Finding suitable, valid instruments, that is, variables that satisfy the necessary assumptions, for models in which the right-hand variables are correlated with omitted factors is often challenging. In this setting, there is a natural candidate—in fact, there are several. From (11-65), we have at period $t = 3$,

$$y_{i3} - y_{i2} = (\mathbf{x}_{i3} - \mathbf{x}_{i2})'\boldsymbol{\beta} + \gamma(y_{i2} - y_{i1}) + (\varepsilon_{i3} - \varepsilon_{i2}).$$

We could use y_{i1} as the needed variable because it is not correlated $\varepsilon_{i3} - \varepsilon_{i2}$. Continuing in this fashion, we see that for $t = 3, 4, \ldots, T$, $y_{i,t-2}$ satisfies our requirements. Alternatively, beginning from period $t = 4$, we can see that $z_{it} = (y_{i,t-2} - y_{i,t-3})$ once again satisfies our requirements. This is Anderson and Hsiao's (1981) result for instrumental variable estimation of the dynamic panel data model. It now becomes a question of which approach, levels ($y_{i,t-2}$, $t = 3, \ldots, T$), or differences ($y_{i,t-2} - y_{i,t-3}$, $t = 4, \ldots, T$) is a preferable approach. Arellano (1989) and Kiviet (1995) obtain results that suggest that the estimator based on levels is more efficient.

11.8.4 EFFICIENT ESTIMATION OF DYNAMIC PANEL DATA MODELS: THE ARELLANO/BOND ESTIMATORS

A leading application of the methods of this chapter is the **dynamic panel data model**, which we now write as

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \delta y_{i,t-1} + c_i + \varepsilon_{it}.$$

Several applications are described in Example 11.21. The basic assumptions of the model are

1. Strict exogeneity: $E[\varepsilon_{it} | \mathbf{X}_i, c_i] = 0$,
2. Homoscedasticity and Nonautocorrelation:

$$E[\varepsilon_{it}\varepsilon_{is} | \mathbf{X}_i, c_i] = \sigma_\varepsilon^2 \text{ if } i = j \text{ and } t = s \text{ and } = 0 \text{ if } i \neq j \text{ or } t \neq s,$$

3. Common effects: The rows of the $T \times K$ data matrix $\mathbf{X}_i$ are $\mathbf{x}_{it}'$. We will not assume mean independence. The "effects" may be fixed or random, so we allow

$$E[c_i | \mathbf{X}_i] = h(\mathbf{X}_i).$$

(See Section 11.2.1.) We will also assume a fixed number of periods, T, for convenience. The treatment here (and in the literature) can be modified to accommodate unbalanced panels, but it is a bit inconvenient. (It involves the placement of zeros at various places in the data matrices defined below and changing the terminal indexes in summations from 1 to T.)

The presence of the lagged dependent variable in this model presents a considerable obstacle to estimation. Consider, first, the straightforward application of Assumption A.I3 in Section 8.2. The compound disturbance in the model is

$(c_i + \varepsilon_{it})$. The correlation between $y_{i,t-1}$ and $(c_i + \varepsilon_{i,t})$ is obviously nonzero because $y_{i,t-1} = \mathbf{x}_{i,t-1}'\boldsymbol{\beta} + \delta y_{i,t-2} + c_i + \varepsilon_{i,t-1}$,

$$\text{Cov}[y_{i,t-1}, (c_i + \varepsilon_{it})] = \sigma_c^2 + \delta\,\text{Cov}[y_{i,t-2}, (c_i + \varepsilon_{it})].$$

If T is large and $0 < \delta < 1$, then this covariance will be approximately $\sigma_c^2/(1 - \delta)$. The large T assumption is not going to be met in most cases. But because δ will generally be positive, we can expect that this covariance will be at least larger than σ_c^2. The implication is that both (pooled) OLS and GLS in this model will be inconsistent. Unlike the case for the static model ($\delta = 0$), the fixed effects treatment does not solve the problem. Taking group mean differences, we obtain

$$y_{i,t} - \bar{y}_{i.} = (\mathbf{x}_{i,t} - \overline{\mathbf{x}}_{i.})'\boldsymbol{\beta} + \delta(y_{i,t-1} - \bar{y}_{i.}) + (\varepsilon_{i,t} - \bar{\varepsilon}_{i.}).$$

As shown in Anderson and Hsiao (1981, 1982),

$$\text{Cov}[(y_{i,t-1} - \bar{y}_{i.}), (\varepsilon_{i,t} - \bar{\varepsilon}_{i.})] \approx \frac{-\sigma_\varepsilon^2}{T^2}\frac{(T-1) - T\delta + \delta^T}{(1-\delta)^2}.$$

This result is O($1/T$), which would generally be no problem if the asymptotics in the model were with respect to increasing T. But, in this panel data model, T is assumed to be fixed and relatively small. For conventional values of T, say 5 to 15, the proportional bias in estimation of δ could be on the order of, say, 15 to 60 percent.

Neither OLS nor GLS are useful as estimators. There are, however, instrumental variables available within the structure of the model. Anderson and Hsiao (1981, 1982) proposed an approach based on first differences rather than differences from group means,

$$y_{it} - y_{i,t-1} = (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta} + \delta(y_{i,t-1} - y_{i,t-2}) + \varepsilon_{it} - \varepsilon_{i,t-1}.$$

For the first full observation,

$$y_{i3} - y_{i2} = (\mathbf{x}_{i3} - \mathbf{x}_{i2})'\boldsymbol{\beta} + \delta(y_{i2} - y_{i1}) + \varepsilon_{i3} - \varepsilon_{i2}, \tag{11-66}$$

the variable y_{i1} (assuming initial point $t = 0$ is where our data-generating process begins) satisfies the requirements, because ε_{i1} is predetermined with respect to $(\varepsilon_{i3} - \varepsilon_{i2})$. [That is, if we used only the data from periods 1 to 3 constructed as in (11-66), then the instrumental variables for $(y_{i2} - y_{i1})$ would be $\mathbf{z}_{i(3)}$ where $\mathbf{z}_{i(3)} = (y_{1,1}, y_{2,1}, \ldots, y_{n,1})$ for the n observations.] For the next observation,

$$y_{i4} - y_{i3} = (\mathbf{x}_{i4} - \mathbf{x}_{i3})'\boldsymbol{\beta} + \delta(y_{i3} - y_{i2}) + \varepsilon_{i4} - \varepsilon_{i3},$$

variables y_{i2} and $(y_{i2} - y_{i1})$ are both available.

Based on the preceding paragraph, one might begin to suspect that there is, in fact, rather than a paucity of instruments, a large surplus. In this limited development, we have a choice between differences and levels. Indeed, we could use both and, moreover, in any period after the fourth, not only is y_{i2} available as an instrument, but so also is y_{i1}, and so on. This is the essential observation behind the Arellano, Bover, and Bond (1991, 1995) estimators, which are based on the very large number of candidates for instrumental variables in this panel data model. To begin, with the model in first differences form, for $y_{i3} - y_{i2}$, variable y_{i1} is available. For $y_{i4} - y_{i3}$, y_{i1} and y_{i2} are both available; for $y_{i5} - y_{i4}$, we have y_{i1}, y_{i2}, and y_{i3}, and so on. Consider, as well, that we have not used the exogenous variables. With strictly exogenous regressors, not only are all lagged values of y_{is} for s previous to $t - 1$, but all values of $\mathbf{x}_{it}$ are also available as instruments. For example, for $y_{i4} - y_{i3}$, the candidates are y_{i1}, y_{i2} and $(\mathbf{x}'_{i1}, \mathbf{x}'_{i2}, \ldots, \mathbf{x}'_{iT})$ for all T periods. The number of candidates for instruments is, in fact, potentially huge.[36] If the exogenous variables are only predetermined, rather than strictly exogenous, then only $E[\varepsilon_{it} | \mathbf{x}_{i,t}, \mathbf{x}_{i,t-1}, \ldots, \mathbf{x}_{i1}] = 0$, and only vectors $\mathbf{x}_{is}$ from 1 to $t - 1$ will be valid instruments in the differenced equation that contains $\varepsilon_{it} - \varepsilon_{i,t-1}$.[37] This is hardly a limitation, given that in the end, for a moderate sized model, we may be considering potentially hundreds or thousands of instrumental variables for estimation of what is usually a small handful of parameters.

We now formulate the model in a more familiar form, so we can apply the instrumental variable estimator. In terms of the differenced data, the basic equation is

$$y_{it} - y_{i,t-1} = (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta} + \delta(y_{i,t-1} - y_{i,t-2}) + \varepsilon_{it} - \varepsilon_{i,t-1},$$

or

$$\Delta y_{it} = (\Delta \mathbf{x}_{it})'\boldsymbol{\beta} + \delta(\Delta y_{i,t-1}) + \Delta\varepsilon_{it}, \tag{11-67}$$

where Δ is the first difference operator, $\Delta a_t = a_t - a_{t-1}$ for any time-series variable (or vector) a_t. (It should be noted that a constant term and any time-invariant variables in $\mathbf{x}_{it}$ will fall out of the first differences. We will recover these below after we develop the estimator for $\boldsymbol{\beta}$.) The parameters of the model to be estimated are $\boldsymbol{\theta} = (\boldsymbol{\beta}', \delta)'$ and σ_ε^2. For convenience, write the model as

$$\widetilde{y}_{it} = \widetilde{\mathbf{x}}'_{it}\boldsymbol{\theta} + \widetilde{\varepsilon}_{it}.$$

We are going to define an instrumental variable estimator along the lines of (8-9) and (8-10). Because our data set is a panel, the counterpart to

$$\mathbf{Z}'\widetilde{\mathbf{X}} = \sum_{i=1}^{n} \mathbf{z}_i \widetilde{\mathbf{x}}'_i \tag{11-68}$$

in the cross-section case would seem to be

$$\mathbf{Z}'\widetilde{\mathbf{X}} = \sum_{i=1}^{n}\sum_{t=3}^{T} \mathbf{z}_{it}\widetilde{\mathbf{x}}'_{it} = \sum_{i=1}^{n} \mathbf{Z}'_i \widetilde{\mathbf{X}}'_i, \tag{11-69}$$

$$\widetilde{\mathbf{y}}_i = \begin{bmatrix} \Delta y_{i3} \\ \Delta y_{i4} \\ \vdots \\ \Delta y_{iT_i} \end{bmatrix}, \widetilde{\mathbf{X}}_i = \begin{bmatrix} \Delta \mathbf{x}'_{i3} & \Delta y_{i2} \\ \Delta \mathbf{x}'_{i4} & \Delta y_{i3} \\ \ldots & \\ \Delta \mathbf{x}'_{iT} & \Delta y_{i,T-1} \end{bmatrix},$$

where there are $(T - 2)$ observations (rows) and $K + 1$ columns in $\widetilde{\mathbf{X}}_i$. There is a complication, however, in that the number of instruments we have defined may vary by period, so the matrix computation in (11-69) appears to sum matrices of different sizes.

[36] See Ahn and Schmidt (1995) for a very detailed analysis.

[37] See Baltagi and Levin (1986) for an application.

Consider an alternative approach. If we used only the first full observations defined in (11-67), then the cross-section version would apply, and the set of instruments **Z** in (11-68) with strictly exogenous variables would be the $n \times (1 + KT)$ matrix,

$$\mathbf{Z}_{(3)} = \begin{bmatrix} y_{1,1}, \mathbf{x}_{1,1}', \mathbf{x}_{1,2}', \cdots \mathbf{x}_{1,T}' \\ y_{2,1}, \mathbf{x}_{2,1}', \mathbf{x}_{2,2}', \cdots \mathbf{x}_{2,T}' \\ \vdots \\ y_{n,1}, \mathbf{x}_{n,1}', \mathbf{x}_{n,2}', \cdots \mathbf{x}_{n,T}' \end{bmatrix},$$

and the instrumental variable estimator of (8-9) would be based on

$$\widetilde{\mathbf{X}}_{(3)} = \begin{bmatrix} \mathbf{x}_{1,3}' - \mathbf{x}_{1,2}' & y_{1,4} - y_{1,3} \\ \mathbf{x}_{2,3}' - \mathbf{x}_{2,2}' & y_{2,4} - y_{2,3} \\ \vdots & \vdots \\ \mathbf{x}_{n,3}' - \mathbf{x}_{n,2}' & y_{n,4} - y_{n,3} \end{bmatrix} \text{ and } \widetilde{\mathbf{y}}_{(3)} = \begin{bmatrix} y_{1,3} - y_{1,2} \\ y_{2,3} - y_{2,2} \\ \vdots \\ y_{n,3} - y_{n,2} \end{bmatrix}.$$

The subscript "(3)" indicates the first observation used for the left-hand side of the equation. Neglecting the other observations, then, we could use these data to form the IV estimator in (8-9), which we label for the moment $\hat{\boldsymbol{\theta}}_{\text{IV}(3)}$. Now, repeat the construction using the next (fourth) observation as the first, and, again, using only a single year of the panel. The data matrices are now

$$\widetilde{\mathbf{X}}_{(4)} = \begin{bmatrix} \mathbf{x}_{1,4}' - \mathbf{x}_{1,3}' & y_{1,3} - y_{1,2} \\ \mathbf{x}_{2,4}' - \mathbf{x}_{2,3}' & y_{2,3} - y_{2,2} \\ \vdots & \vdots \\ \mathbf{x}_{n,4}' - \mathbf{x}_{n,3}' & y_{n,3} - y_{n,2} \end{bmatrix}, \widetilde{\mathbf{y}}_{(4)} = \begin{bmatrix} y_{1,4} - y_{1,3} \\ y_{2,4} - y_{2,3} \\ \vdots \\ y_{n,4} - y_{n,3} \end{bmatrix}, \text{ and}$$
$$\mathbf{Z}_{(4)} = \begin{bmatrix} y_{1,1}, y_{1,2}, \mathbf{x}_{1,1}', \mathbf{x}_{1,2}', \cdots \mathbf{x}_{1,T}' \\ y_{2,1}, y_{2,2}, \mathbf{x}_{2,1}', \mathbf{x}_{2,2}', \cdots \mathbf{x}_{2,T}' \\ \vdots \\ y_{n,1}, y_{n,2}, \mathbf{x}_{n,1}', \mathbf{x}_{n,2}', \cdots \mathbf{x}_{n,T} \end{bmatrix}, \tag{11-70}$$

and we have a second IV estimator, $\hat{\boldsymbol{\theta}}_{\text{IV}(4)}$, also based on n observations, but, now, $2 + KT$ instruments. And so on.

We now need to reconcile the $T - 2$ estimators of $\boldsymbol{\theta}$ that we have constructed, $\hat{\boldsymbol{\theta}}_{\text{IV}(3)}, \hat{\boldsymbol{\theta}}_{\text{IV}(4)}, \ldots, \hat{\boldsymbol{\theta}}_{\text{IV}(T)}$. We faced this problem in Section 11.5.8 where we examined Chamberlain's formulation of the fixed effects model. The minimum distance estimator suggested there and used in Carey's (1997) study of hospital costs in Example 11.13 provides a means of efficiently "averaging" the multiple estimators of the parameter vector. We will return to the MDE in Chapter 13. For the present, we consider, instead, **Arellano and Bond's approach** (1991)[38] to this problem. We will collect the full set of estimators in a counterpart to (11-56) and (11-57). First, combine the sets of instruments in a single matrix, **Z**, where for each individual, we obtain the $(T - 2) \times L$ matrix $\mathbf{Z}_i$. The definition of the rows of $\mathbf{Z}_i$ depend on whether the regressors are assumed to be strictly exogenous or predetermined. For strictly exogenous variables,

[38] And Arellano and Bover's (1995).

$$\mathbf{Z}_i = \begin{bmatrix} y_{i,1}, \mathbf{x}'_{i,1}, \mathbf{x}'_{i,2}, \ldots \mathbf{x}'_{i,T} & 0 & \ldots & 0 \\ 0 & y_{i,1}, y_{i,2}, \mathbf{x}'_{i,1}, \mathbf{x}'_{i,2}, \ldots \mathbf{x}'_{i,T} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & y_{i,1}, y_{i,2}, \ldots, y_{i,T-2}, \mathbf{x}'_{i,1}, \mathbf{x}'_{i,2}, \ldots \mathbf{x}'_{i,T} \end{bmatrix}, \tag{11.71a}$$

and $L = \sum_{i=1}^{T-2}(i + TK) = (T - 2)(T - 1)/2 + (T - 2)TK$. For only predetermined variables, the matrix of instrumental variables is

$$\mathbf{Z}_i = \begin{bmatrix} y_{i,1}, \mathbf{x}'_{i,1}, \mathbf{x}'_{i,2} & 0 & \ldots & 0 \\ 0 & y_{i,1}, y_{i,2}, \mathbf{x}'_{i,1}, \mathbf{x}'_{i,2}, \mathbf{x}'_{i,3} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & y_{i,1}, y_{i,2}, \ \ldots, y_{i,T-2}, \mathbf{x}'_{i,1}, \mathbf{x}'_{i,2}, \ \ldots \mathbf{x}'_{i,T-1} \end{bmatrix}, \tag{11.71b}$$

and $L = \Sigma_{i=1}^{T-2}\,(i(K + 1) + K) = [(T - 2)(T - 1)/2](1 + K) + (T - 2)K$. This construction does proliferate instruments (moment conditions, as we will see in Chapter 13). In the application in Example 11.18, we have a small panel with only $T = 7$ periods, and we fit a model with only $K = 4$ regressors in $\mathbf{x}_{it}$, plus the lagged dependent variable. The strict exogeneity assumption produces a $\mathbf{Z}_i$ matrix that is (5×135) for this case. With only the assumption of predetermined $\mathbf{x}_{it}$, $\mathbf{Z}_i$ collapses slightly to (5×95). For purposes of the illustration, we have used only the two previous observations on $\mathbf{x}_{it}$. This further reduces the matrix to

$$\mathbf{Z}_i = \begin{bmatrix} y_{i,1}, \mathbf{x}'_{i,1}, \mathbf{x}'_{i,2} & 0 & \ldots & 0 \\ 0 & y_{i,1}, y_{i,2}, \mathbf{x}_{i,2}, \mathbf{x}'_{i,3} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & y_{i,1}, y_{i,2}, \ \ldots, y_{i,T-2}, \mathbf{x}'_{i,T-2}, \mathbf{x}'_{i,T-1} \end{bmatrix}, \tag{11.71c}$$

which, with $T = 7$ and $K = 4$, will be (5×55).[39]

Now, we can compute the two-stage least squares estimator in (11-55) using our definitions of the data matrices $\mathbf{Z}_i$, $\widetilde{\mathbf{X}}_i$, and $\widetilde{\mathbf{y}}_i$ and (11-69). This will be

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{\text{IV}} &= \left[\left(\sum_{i=1}^{n}\widetilde{\mathbf{X}}_i'\mathbf{Z}_i\right)\left(\sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{Z}_i'\widetilde{\mathbf{X}}_i\right)\right]^{-1} \\ &\quad \times \left[\left(\sum_{i=1}^{n}\widetilde{\mathbf{X}}_i'\mathbf{Z}_i\right)\left(\sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{Z}_i'\widetilde{\mathbf{y}}_i\right)\right]. \end{aligned} \tag{11-72}$$

The natural estimator of the asymptotic covariance matrix for the estimator would be

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\theta}}_{\text{IV}}] = \hat{\sigma}^2_{\Delta\varepsilon}\left[\left(\sum_{i=1}^{n}\widetilde{\mathbf{X}}_i'\mathbf{Z}_i\right)\left(\sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{Z}_i'\widetilde{\mathbf{X}}_i\right)\right]^{-1}, \tag{11-73}$$

[39] Baltagi (2005, Chapter 8) presents some alternative configurations of $\mathbf{Z}_i$ that allow for mixtures of strictly exogenous and predetermined variables.

where

$$\hat{\sigma}^2_{\Delta\varepsilon} = \frac{\sum_{i=1}^{n}\sum_{t=3}^{T}[(y_{it} - y_{i,t-1}) - (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\hat{\boldsymbol{\beta}} - \hat{\delta}(y_{i,t-1} - y_{i,t-2})]^2}{n(T-2)}. \tag{11-74}$$

However, this variance estimator is likely to understate the true asymptotic variance because the observations are autocorrelated for one period. Because $(y_{it} - y_{i,t-1}) = \widetilde{\mathbf{x}}_{it}'\boldsymbol{\theta} + (\varepsilon_{it} - \varepsilon_{i,t-1}) = \widetilde{\mathbf{x}}_{it}'\boldsymbol{\theta} + v_{it}$, $\text{Cov}[v_{it}, v_{i,t-1}] = \text{Cov}[v_{it}, v_{i,t+1}] = -\sigma_{\varepsilon}^2$. Covariances at longer lags or leads are zero. In the differenced model, though the disturbance covariance matrix is not $\sigma_v^2\mathbf{I}$, it does take a particularly simple form,

$$\text{Cov}\begin{pmatrix} \varepsilon_{i,3} - \varepsilon_{i,2} \\ \varepsilon_{i,4} - \varepsilon_{i,3} \\ \varepsilon_{i,5} - \varepsilon_{i,4} \\ \cdots \\ \varepsilon_{i,T} - \varepsilon_{i,T-1} \end{pmatrix} = \sigma_{\varepsilon}^2 \begin{bmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \dots & 0 \\ 0 & -1 & 2 & \dots & 0 \\ \dots & \dots & -1 & \dots & -1 \\ 0 & 0 & \dots & -1 & 2 \end{bmatrix} = \sigma_{\varepsilon}^2\boldsymbol{\Omega}_i. \tag{11-75}$$

The implication is that the estimator in (11-74) estimates not σ_{ε}^2 but $2\sigma_{\varepsilon}^2$. However, simply dividing the estimator by two does not produce the correct asymptotic covariance matrix because the observations themselves are autocorrelated. As such, the matrix in (11-73) is inappropriate. A robust correction can be based on the counterpart to the White estimator that we developed in (11-3). For simplicity, let

$$\hat{\mathbf{A}} = \left[\left(\sum_{i=1}^{n}\widetilde{\mathbf{X}}_i'\mathbf{Z}_i\right)\left(\sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{Z}_i'\widetilde{\mathbf{X}}_i\right)\right]^{-1}.$$

Then, a robust covariance matrix that accounts for the autocorrelation would be

$$\hat{\mathbf{A}}\left[\left(\sum_{i=1}^{n}\widetilde{\mathbf{X}}_i'\mathbf{Z}_i\right)\left(\sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{Z}_i'\hat{\mathbf{v}}_i\hat{\mathbf{v}}_i'\mathbf{Z}_i\right)\left(\sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{Z}_i'\widetilde{\mathbf{X}}_i\right)\right]\hat{\mathbf{A}}. \tag{11-76}$$

[One could also replace the $\hat{\mathbf{v}}_i\hat{\mathbf{v}}_i'$ in (11-76) with $\hat{\sigma}_{\varepsilon}^2\boldsymbol{\Omega}_i$ in (11-75) because this is the known expectation.]

It will be useful to digress briefly and examine the estimator in (11-72). The computations are less formidable than it might appear. Note that the rows of $\mathbf{Z}_i$ in (11-71a,b,c) are orthogonal. It follows that the matrix $\mathbf{F} = \sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{Z}_i$ in (11-72) is block-diagonal with $T - 2$ blocks. The specific blocks in $\mathbf{F}$ are $\mathbf{F}_t = \sum_{i=1}^{n}\mathbf{z}_{it}\mathbf{z}_{it}' = \mathbf{Z}_{(t)}'\mathbf{Z}_{(t)}$, for $t = 3, \dots, T$. Because the number of instruments is different in each period—see (11-71)—these blocks are of different sizes, say, $(L_t \times L_t)$. The same construction shows that the matrix $\sum_{i=1}^{n}\widetilde{\mathbf{X}}_i'\mathbf{Z}_i$ is actually a partitioned matrix of the form

$$\sum_{i=1}^{n}\widetilde{\mathbf{X}}_i'\mathbf{Z}_i = \left[\widetilde{\mathbf{X}}_{(3)}'\mathbf{Z}_{(3)} \quad \widetilde{\mathbf{X}}_{(4)}'\mathbf{Z}_{(4)} \quad \dots \quad \widetilde{\mathbf{X}}_{(T)}'\mathbf{Z}_{(T)}\right],$$

where, again, the matrices are of different sizes; there are $T - 2$ rows in each but the number of columns differs. It follows that the inverse matrix, $\left(\sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{Z}_i\right)^{-1}$, is also block-diagonal, and that the matrix quadratic form in (11-72) can be written

$$\left(\sum_{i=1}^{n}\widetilde{\mathbf{X}}_i'\mathbf{Z}_i\right)\left(\sum_{i=1}^{n}\widetilde{\mathbf{Z}}_i'\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{Z}_i'\widetilde{\mathbf{X}}_i\right) = \sum_{t=3}^{T}(\widetilde{\mathbf{X}}_{(t)}'\mathbf{Z}_{(t)})(\mathbf{Z}_{(t)}'\mathbf{Z}_{(t)})^{-1}(\mathbf{Z}_{(t)}'\widetilde{\mathbf{X}}_{(t)})$$
$$= \sum_{t=3}^{T}\left(\hat{\widetilde{\mathbf{X}}}_{(t)}'\hat{\widetilde{\mathbf{X}}}_{(t)}\right)$$
$$= \sum_{t=3}^{T}\mathbf{W}_{(t)},$$

[see (8-9) and the preceding result]. Continuing in this fashion, we find

$$\left(\sum_{i=1}^{n}\widetilde{\mathbf{X}}_i'\mathbf{Z}_i\right)\left(\sum_{i=1}^{n}\widetilde{\mathbf{Z}}_i'\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{Z}_i'\widetilde{\mathbf{y}}_i\right) = \sum_{t=3}^{T}\hat{\widetilde{\mathbf{X}}}_{(t)}'\mathbf{y}_{(t)}.$$

From (8-10), we can see that

$$\hat{\widetilde{\mathbf{X}}}_{(t)}'\mathbf{y}_{(t)} = \left(\hat{\widetilde{\mathbf{X}}}_{(t)}'\hat{\widetilde{\mathbf{X}}}_{(t)}\right)\hat{\boldsymbol{\theta}}_{\text{IV}}(t)$$
$$= \mathbf{W}_{(t)}\hat{\boldsymbol{\theta}}_{\text{IV}}(t).$$

Combining the terms constructed thus far, we find that the estimator in (11-72) can be written in the form

$$\hat{\boldsymbol{\theta}}_{\text{IV}} = \left(\sum_{t=3}^{T}\mathbf{W}_{(t)}\right)^{-1}\left(\sum_{t=3}^{T}\mathbf{W}_{(t)}\hat{\boldsymbol{\theta}}_{\text{IV}}(t)\right)$$
$$= \sum_{t=3}^{T}\mathbf{R}_{(t)}\hat{\boldsymbol{\theta}}_{\text{IV}}(t),$$

where

$$\mathbf{R}_{(t)} = \left(\sum_{t=3}^{T}\mathbf{W}_{(t)}\right)^{-1}\mathbf{W}_{(t)} \text{ and } \sum_{t=3}^{T}\mathbf{R}_{(t)} = \mathbf{I}.$$

In words, we find that, as might be expected, the Arellano and Bond estimator of the parameter vector is a matrix weighted average of the $T - 2$ period-specific two-stage least squares estimators, where the instruments used in each period may differ. Because the estimator is an average of estimators, a question arises, is it an efficient average—are the weights chosen to produce an efficient estimator? Perhaps not surprisingly, the answer for this $\hat{\boldsymbol{\theta}}$ is no; there is a more efficient set of weights that can be constructed for this model. We will assemble them when we examine the generalized method of moments estimator in Chapter 13.

There remains a loose end in the preceding. After (11-67), it was noted that this treatment discards a constant term and any time-invariant variables that appear in the model. The Hausman and Taylor (1981) approach developed in the preceding section suggests a means by which the model could be completed to accommodate this possibility. Expand the basic formulation to include the time-invariant effects, as

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \delta y_{i,t-1} + \boldsymbol{\alpha} + \mathbf{f}_i'\boldsymbol{\gamma} + c_i + \varepsilon_{it},$$

where $\mathbf{f}_i$ is the set of time-invariant variables and $\boldsymbol{\gamma}$ is the parameter vector yet to be estimated. This model is consistent with the entire preceding development, as the component $\alpha + \mathbf{f}_i'\boldsymbol{\gamma}$ would have fallen out of the differenced equation along with c_i at the first step at (11-63). Having developed a consistent estimator for $\boldsymbol{\theta} = (\boldsymbol{\beta}', \delta)'$, we now turn to estimation of $(\alpha, \boldsymbol{\gamma}')'$. The residuals from the IV regression (11-72),

$$w_{it} = \mathbf{x}_{it}'\hat{\boldsymbol{\beta}}_{IV} - \hat{\delta}_{IV} y_{i,t-1},$$

are pointwise consistent estimators of

$$\omega_{it} = \alpha + \mathbf{f}_i'\boldsymbol{\gamma} + c_i + \varepsilon_{it}.$$

Thus, the group means of the residuals can form the basis of a second-step regression,

$$\overline{w}_i = \alpha + \mathbf{f}_i'\boldsymbol{\gamma} + c_i + \bar{\varepsilon}_i + \eta_i, \tag{11-77}$$

where $\eta_i = (\overline{w}_{i\cdot} - \overline{\omega}_{i\cdot})$ is the estimation error that converges to zero as $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}$. The implication would seem to be that we can now linearly regress these group mean residuals on a constant and the time-invariant variables $\mathbf{f}_i$ to estimate α and $\boldsymbol{\gamma}$. The flaw in the strategy, however, is that the initial assumptions of the model do not state that c_i is uncorrelated with the other variables in the model, including the implicit time-invariant terms, $\mathbf{f}_i$. Therefore, least squares is not a usable estimator here unless the random effects model is assumed, which we specifically sought to avoid at the outset. As in Hausman and Taylor's treatment, there is a workable strategy if it can be assumed that there are some variables in the model, including possibly some among the $\mathbf{f}_i$ as well as others among $\mathbf{x}_{it}$ that are uncorrelated with c_i and ε_{it}. These are the $\mathbf{z}_1$ and $\mathbf{x}_1$ in the Hausman and Taylor estimator (see step 2 in the development of the preceding section). Assuming that these variables are available—this is an identification assumption that must be added to the model—then we do have a usable instrumental variable estimator, using as instruments the constant term (1), any variables in $\mathbf{f}_i$ that are uncorrelated with the latent effects or the disturbances (call this $\mathbf{f}_{i1}$), and the group means of any variables in $\mathbf{x}_{it}$ that are also exogenous. There must be enough of these to provide a sufficiently large set of instruments to fit all the parameters in (11-77). This is, once again, the same identification we saw in step 2 of the Hausman and Taylor estimator, K_1, the number of exogenous variables in $\mathbf{x}_{it}$ must be at least as large as L_2, which is the number of endogenous variables in $\mathbf{f}_i$. With all this in place, we then have the instrumental variable estimator in which the dependent variable is $\overline{w}_{i.}$, the right-hand-side variables are $(1, \mathbf{f}_i)$, and the instrumental variables are $(1, \mathbf{f}_{i1}, \overline{\mathbf{x}}_{i1}.)$.

There is yet another direction that we might extend this estimation method. In (11-76), we have implicitly allowed a more general covariance matrix to govern the generation of the disturbances ε_{it} and computed a robust covariance matrix for the simple IV estimator. We could take this a step further and look for a more efficient estimator. As a library of recent studies has shown, panel data sets are rich in information that allows the analyst to specify highly general models and to exploit the implied relationships among the variables to construct much more efficient generalized method of moments (GMM) estimators.[40] We will return to this development in Chapter 13.

Example 11.19 Dynamic Labor Supply Equation

In Example 8.5, we used instrumental variables to fit a labor supply equation,

$$Wks_{it} = \gamma_1 + \gamma_2 \ln Wage_{it} + \gamma_3\, Ed_i + \gamma_4\, Union_{it} + \gamma_5\, Fem_i + u_{it}.$$

[40] See, in particular, Arellano and Bover (1995) and Blundell and Bond (1998).

To illustrate the computations of this section, we will extend this model as follows,

$$Wks_{it} = \beta_1 \ln Wage_{it} + \beta_2\, Union_{it} + \beta_3\, Occ_{it} + \beta_4\, Exp_{it} + \delta\, Wks_{i,t-1}$$
$$+ \alpha + \gamma_1\, Ed_i + \gamma_2\, Fem_i + c_i + \varepsilon_{it}.$$

(We have rearranged the variables and parameter names to conform to the notation in this section.) We note, in theoretical terms, as suggested in the earlier example, it may not be appropriate to treat $\ln Wage_{it}$ as uncorrelated with ε_{it} or c_i. However, we will be analyzing the model in first differences. It may well be appropriate to treat changes in wages as exogenous. That would depend on the theoretical underpinnings of the model. We will treat the variable as predetermined here, and proceed. There are two time-invariant variables in the model, Fem_i, which is clearly exogenous, and Ed_i, which might be endogenous. The identification requirement for estimation of $(\alpha, \gamma_1, \gamma_2)$ is met by the presence of three exogenous variables, $Union_{it}$, Occ_{it}, and Exp_{it} ($K_1 = 3$ and $L_2 = 1$).

The differenced equation analyzed at the first step is

$$\Delta Wks_{it} = \beta_1 \Delta \ln Wage_{it} + \beta_2 \Delta Union_{it} + \beta_3 \Delta Occ_{it} + \beta_4 \Delta Exp_{it} + \delta \Delta Wks_{i,t-1} + \Delta\varepsilon_{it}.$$

We estimated the parameters and the asymptotic covariance matrix according to (11-73) and (11-76). For specification of the instrumental variables, we used the one previous observation on $\mathbf{x}_{it}$, as shown in the text. Table 11.19 presents the computations with several other inconsistent estimators.

The various estimates are quite far apart. In the absence of the common effects (and autocorrelation of the disturbances), all five estimators shown would be consistent. Given the very wide disparities, one might suspect that common effects are an important feature

TABLE 11.19 Estimated Dynamic Panel Data Model Using Arellano and Bond Estimator

(Estimated standard errors in parentheses)

Variable	*OLS Full Equation*	*OLS Differenced*	*IV Differenced*	*Random Effects*	*Fixed Effects*
ln *Wage*	0.2966	−0.1100	−1.1402	0.2281	0.5886
	(0.2052)	(0.4565)	(0.2639) [0.8768]	(0.2405)	(0.4790)
Union	−1.2945	1.1640	2.7089	−1.4104	0.1444
	(0.1713)	(0.4222)	(0.3684) [0.8676]	(0.2199)	(0.4369)
Occ	0.4163	0.8142	2.2808	0.5191	1.0064
	(0.2005)	(0.3924)	(1.3105) [0.7220]	(2.2484)	(0.4030)
Exp	−0.0295	−0.0742	−0.0208	−0.0353	−0.1683
	(0.0073)	(0.0975)	(0.1126) [0.1104]	(0.0102)	(0.0595)
Wks_{t-1}	0.3804	−0.3527	0.1304	0.2100	0.0148
	(0.0148)	(0.0161)	(0.0476) [0.0213]	(0.0151)	(0.0171)
Constant	28.918	—	−0.4110	37.4610	—
	(1.4490)	—	(0.3364)	(1.6778)	—
Ed	−0.0690	—	0.0321	−0.0657	—
	(0.0370)	—	(0.0259)	(0.0499)	—
Fem	−0.8607	—	−0.0122	−1.1463	—
	(0.2544)	—	(0.1554)	(0.3513)	—
Sample	$t = 2$ to 7	$t = 3$ to 7	$t = 3$ to 7	$t = 2$ to 7	$t = 2$ to 7
Observations	595	595	595, Means used $t = 7$	595	595

of the data. The second standard errors given in brackets with the IV estimates are based on the uncorrected matrix in (11-73) with $\hat{\sigma}^2_{\Delta\varepsilon}$ in (11-74) divided by two. We found the estimator to be quite volatile, as can be seen in the table. The estimator is also very sensitive to the choice of instruments that comprise $\mathbf{Z}_i$. Using (11-71a) instead of (11-71b) produces wild swings in the estimates and, in fact, produces implausible results. One possible explanation in this particular example is that the instrumental variables we are using are dummy variables that have relatively little variation over time.

11.8.5 NONSTATIONARY DATA AND PANEL DATA MODELS

Some of the discussion thus far (and to follow) focuses on "small T" statistical results. Panels are taken to contain a fixed and small T observations on a large n individual units. Recent research using cross-country data sets such as the Penn World Tables (http://cid.econ.ucdavis.edu/pwt.html), which now include data on over 150 countries for well over 50 years, have begun to analyze panels with T sufficiently large that the time-series properties of the data become an important consideration. In particular, the recognition and accommodation of nonstationarity that is now a standard part of single time-series analyses (as in Chapter 21) are now seen to be appropriate for large-scale cross-country studies, such as income growth studies based on the Penn World Tables, cross-country studies of health care expenditure, and analyses of purchasing power parity.

The analysis of long panels, such as in the growth and convergence literature, typically involves dynamic models, such as

$$y_{it} = \alpha_i + \gamma_i y_{i,t-1} + \mathbf{x}_{it}'\boldsymbol{\beta}_i + \varepsilon_{it}. \tag{11-78}$$

In single time-series analysis involving low-frequency macroeconomic flow data such as income, consumption, investment, the current account deficit, and so on, it has long been recognized that estimated regression relations can be distorted by nonstationarity in the data. What appear to be persistent and strong regression relationships can be entirely spurious and due to underlying characteristics of the time-series processes rather than actual connections among the variables. Hypothesis tests about long-run effects will be considerably distorted by unit roots in the data. It has become evident that the same influences, with the same deletarious effects, will be found in long panel data sets. The panel data application is further complicated by the possible heterogeneity of the parameters. The coefficients of interest in many cross-country studies are the lagged effects, such as γ_i in (11-78), and it is precisely here that the received results on nonstationary data have revealed the problems of estimation and inference. Valid tests for unit roots in panel data have been proposed in many studies. Three that are frequently cited are Levin and Lin (1992), Im, Pesaran, and Shin (2003), and Maddala and Wu (1999).

There have been numerous empirical applications of time-series methods for nonstationary data in panel data settings, including Frankel and Rose's (1996) and Pedroni's (2001) studies of purchasing power parity, Fleissig and Strauss (1997) on real wage stationarity, Culver and Papell (1997) on inflation, Wu (2000) on the current account balance, McCoskey and Selden (1998) on health care expenditure, Sala-i-Martin (1996) on growth and convergence, McCoskey and Kao (1999) on urbanization and production, and Coakely et al. (1996) on savings and investment. An extensive enumeration appears in Baltagi (2005, Chapter 12).

A subtle problem arises in obtaining results useful for characterizing the properties of estimators of the model in (11-78). The asymptotic results based on large n and large

T are not necessarily obtainable simultaneously, and great care is needed in deriving the asymptotic behavior of useful statistics. Phillips and Moon (1999, 2000) are standard references on the subject.

We will return to the topic of nonstationary data in Chapter 21. This is an emerging literature, most of which is beyond the level of this text. We will rely on the several detailed received surveys, such as Bannerjee (1999), Smith (2000), and Baltagi and Kao (2000), to fill in the details.

11.9 NONLINEAR REGRESSION WITH PANEL DATA

The extension of the panel data models to the nonlinear regression case is, perhaps surprisingly, not at all straightforward. Thus far, to accommodate the nonlinear model, we have generally applied familiar results to the linearized regression. This approach will carry forward to the case of clustered data. (See Section 11.3.3.) Unfortunately, this will not work with the standard panel data methods. The nonlinear regression will be the first of numerous panel data applications that we will consider in which the wisdom of the linear regression model cannot be extended to the more general framework.

11.9.1 A ROBUST COVARIANCE MATRIX FOR NONLINEAR LEAST SQUARES

The counterpart to (11-3) or (11-4) would simply replace $\mathbf{X}_i$ with $\hat{\mathbf{X}}_i^0$ where the rows are the pseudo regressors for cluster i as defined in (7-12) and "$^\wedge$" indicates that it is computed using the nonlinear least squares estimates of the parameters.

Example 11.20 Health Care Utilization

The recent literature in health economics includes many studies of health care utilization. A common measure of the dependent variable of interest is a count of the number of encounters with the health care system, either through visits to a physician or to a hospital. These counts of occurrences are usually studied with the Poisson regression model described in Section 18.4. The nonlinear regression model is

$$E[y_i \mid \mathbf{x}_i] = \exp(\mathbf{x}_i'\boldsymbol{\beta}).$$

A recent study in this genre is "Incentive Effects in the Demand for Health Care: A Bivariate Panel Count Data Estimation" by Riphahn, Wambach, and Million (2003). The authors were interested in counts of physician visits and hospital visits. In this application, they were particularly interested in the impact of the presence of private insurance on the utilization counts of interest, that is, whether the data contain evidence of moral hazard.

The raw data are published on the *Journal of Applied Econometrics* data archive Web site, The URL for the data file is http://qed.econ.queensu.ca/jae/2003-v18.4/riphahn-wambach-million/. The variables in the data file are listed in Appendix Table F7.1. The sample is an unbalanced panel of 7,293 households, the German Socioeconomic Panel data set. The number of observations varies from one to seven (1,525; 1,079; 825; 926; 1,311; 1,000; 887), with a total number of observations of 27,326. We will use these data in several examples here and later in the book.

The following model uses a simple specification for the count of number of visits to the physican in the observation year,

$$\mathbf{x}_{it} = (1, age_{it}, educ_{it}, income_{it}, kids_{it}).$$

Table 11.20 details the nonlinear least squares iterations and the results. The convergence criterion for the iterations is $\mathbf{e}^{0\prime}\,\mathbf{X}^0\,(\mathbf{X}^{0\prime}\,\mathbf{X}^0)^{-1}\,\mathbf{X}^{0\prime}\mathbf{e}^0 < 10^{-10}$. Although this requires 11 iterations,

TABLE 11.20 Nonlinear Least Squares Estimates of a Health Care Utilization Equation

```
Begin NLSQ iterations. Linearized regression.
Iteration = 1; Sum of squares = 1014865.00; Gradient = 156281.794
Iteration = 2; Sum of squares = 8995221.17; Gradient = 8131951.67
Iteration = 3; Sum of squares = 1757006.18; Gradient = 897066.012
Iteration = 4; Sum of squares = 930876.806; Gradient = 73036.2457
Iteration = 5; Sum of squares = 860068.332; Gradient = 2430.80472
Iteration = 6; Sum of squares = 857614.333; Gradient = 12.8270683
Iteration = 7; Sum of squares = 857600.927; Gradient = 0.411851239E-01
Iteration = 8; Sum of squares = 857600.883; Gradient = 0.190628165E-03
Iteration = 9; Sum of squares = 857600.883; Gradient = 0.904650588E-06
Iteration = 10; Sum of squares = 857600.883; Gradient = 0.430441193E-08
Iteration = 11; Sum of squares = 857600.883; Gradient = 0.204875467E-10
Convergence achieved
```

Variable	*Estimate*	*Std. Error*	*Robust Std. Error*
Constant	0.9801	0.08927	0.12522
Age	0.0187	0.00105	0.00142
Education	−0.0361	0.00573	0.00780
Income	−0.5911	0.07173	0.09702
Kids	−0.1692	0.02642	0.03330

the function actually reaches the minimum in 7. The estimates of the asymptotic standard errors are computed using the conventional method, $s^2(\hat{\mathbf{X}}^{0\prime}\hat{\mathbf{X}}^0)^{-1}$, and then by the cluster correction in (11-4). The corrected standard errors are considerably larger, as might be expected given that these are a panel data set.

11.9.2 FIXED EFFECTS IN NONLINEAR REGRESSION MODELS

The nonlinear panel data regression model would appear as

$$y_{it} = h(\mathbf{x}_{it}, \boldsymbol{\beta}) + \varepsilon_{it}, t = 1, \ldots, T_i, i = 1, \ldots, n.$$

Consider a model with latent heterogeneity, c_i. An ambiguity immediately emerges; how should heterogeneity enter the model? Building on the linear model, an additive term might seem natural, as in

$$y_{it} = h(\mathbf{x}_{it}, \boldsymbol{\beta}) + c_i + \varepsilon_{it}, t = 1, \ldots, T_i, i = 1, \ldots, n. \quad \textbf{(11-79)}$$

But we can see in the previous application that this is likely to be inappropriate. The loglinear model of the previous section is constrained to ensure that $E[y_{it}|\mathbf{x}_{it}]$ is positive. But an additive random term c_i as in (11-79) could subvert this; unless the range of c_i is restricted, the conditional mean could be negative. The most common application of nonlinear models is the **index function model**,

$$y_{it} = h(\mathbf{x}_{it}'\boldsymbol{\beta} + c_i) + \varepsilon_{it}.$$

This is the natural extension of the linear model, but only in the appearance of the conditional mean. Neither the fixed effects nor the random effects model can be estimated as they were in the linear case.

Consider the fixed effects model first. We would write this as

$$y_{it} = h(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i) + \varepsilon_{it}, \tag{11-80}$$

where the parameters to be estimated are $\boldsymbol{\beta}$ and α_i, $i = 1, \ldots, n$. Transforming the data to deviations from group means does not remove the fixed effects from the model. For example,

$$y_{it} - \bar{y}_{i.} = h(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i) - \frac{1}{T_i}\sum_{s=1}^{T_i} h(\mathbf{x}_{is}'\boldsymbol{\beta} + \alpha_i),$$

which does not simplify things at all. Transforming the regressors to deviations is likewise pointless. To estimate the parameters, it is necessary to minimize the sum of squares with respect to all $n + K$ parameters simultaneously. Because the number of dummy variable coefficients can be huge—the preceding example is based on a data set with 7,293 groups—this can be a difficult or impractical computation. A method of maximizing a function (such as the negative of the sum of squares) that contains an unlimited number of dummy variable coefficients is shown in Chapter 17. As we will examine later in the book, the difficulty with nonlinear models that contain large numbers of dummy variable coefficients is not necessarily the practical one of computing the estimates. That is generally a solvable problem. The difficulty with such models is an intriguing phenomenon known as the **incidental parameters problem**. (See footnote 12.) In most (not all, as we shall find) nonlinear panel data models that contain n dummy variable coefficients, such as the one in (11-80), as a consequence of the fact that the number of parameters increases with the number of individuals in the sample, the estimator of $\boldsymbol{\beta}$ is biased and inconsistent, to a degree that is $O(1/T)$. Because T is only 7 or less in our application, this would seem to be a case in point.

Example 11.21 Exponential Model with Fixed Effects

The exponential model of the preceding example is actually one of a small handful of known special cases in which it is possible to "condition" out the dummy variables. Consider the sum of squared residuals,

$$S_n = \frac{1}{2}\sum_{i=1}^{n}\sum_{t=1}^{T_i}[y_{it} - \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i)]^2.$$

The first-order condition for minimizing S_n with respect to α_i is

$$\frac{\partial S_n}{\partial \alpha_i} = \sum_{t=1}^{T_i} - [y_{it} - \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i)]\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i) = 0. \tag{11-81}$$

Let $\gamma_i = \exp(\alpha_i)$. Then, an equivalent necessary condition would be

$$\frac{\partial S_n}{\partial \gamma_i} = \sum_{t=1}^{T_i} - [y_{it} - \gamma_i \exp(\mathbf{x}_{it}'\boldsymbol{\beta})][\gamma_i \exp(\mathbf{x}_{it}'\boldsymbol{\beta})] = 0,$$

or

$$\gamma_i \sum_{t=1}^{T_i}[y_{it}\exp(\mathbf{x}_{it}'\boldsymbol{\beta})] = \gamma_i^2 \sum_{t=1}^{T_i}[\exp(\mathbf{x}_{it}'\boldsymbol{\beta})]^2.$$

Obviously, if we can solve the equation for γ_i, we can obtain $\alpha_i = \ln \gamma_i$. The preceding equation can, indeed, be solved for γ_i, at least conditionally. At the minimum of the sum of squares, it will be true that

$$\hat{\gamma}_i = \frac{\sum_{t=1}^{T_i} y_{it}\exp(\mathbf{x}_{it}'\hat{\boldsymbol{\beta}})}{\sum_{t=1}^{T_i}[\exp(\mathbf{x}_{it}'\hat{\boldsymbol{\beta}})]^2}. \tag{11-82}$$

We can now insert (11-82) into (11-81) to eliminate α_i. (This is a counterpart to taking deviations from means in the linear case. As noted, this is possible only for a very few special models—this happens to be one of them. The process is also known as "concentrating out" the parameters γ_i. Note that at the solution, $\hat{\gamma}_i$ is obtained as the slope in a regression without a constant term of y_{it} on $\hat{\mathbf{z}}_{it} = \exp(\mathbf{x}_{it}'\hat{\boldsymbol{\beta}})$ using T_i observations.) The result in (11-82) must hold at the solution. Thus, (11-82) inserted in (11-81) restricts the search for $\boldsymbol{\beta}$ to those values that satisfy the restrictions in (11-82). The resulting sum of squares function is now a function only of the data and $\boldsymbol{\beta}$**,** and can be minimized with respect to this vector of K parameters. With the estimate of $\boldsymbol{\beta}$ in hand, α_i can be estimated using the log of the result in (11-82) (which is positive by construction).

The preceding example presents a mixed picture for the fixed effects model. In nonlinear cases, two problems emerge that were not present earlier, the practical one of actually computing the dummy variable parameters and the theoretical incidental parameters problem that we have yet to investigate, but which promises to be a significant shortcoming of the fixed effects model. We also note we have focused on a particular form of the model, the single index function, in which the conditional mean is a nonlinear function of a linear function. In more general cases, it may be unclear how the unobserved heterogeneity should enter the regression function.

11.9.3 RANDOM EFFECTS

The random effects nonlinear model also presents complications both for specification and for estimation. We might begin with a general model,

$$y_{it} = h(\mathbf{x}_{it}, \boldsymbol{\beta}, u_i) + \varepsilon_{it}.$$

The "random effects" assumption would be, as usual, mean independence,

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

Unlike the linear model, the nonlinear regression cannot be consistently estimated by (nonlinear) least squares. In practical terms, we can see why in (7-28) through (7-30). In the linearized regression, the conditional mean at the expansion point $\boldsymbol{\beta}^0$ [see (7-28)] as well as the pseudoregressors are both functions of the unobserved u_i. This is true in the general case as well as the simpler case of a single index model,

$$y_{it} = h(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i) + \varepsilon_{it}. \tag{11-83}$$

Thus, it is not possible to compute the iterations for nonlinear least squares. As in the fixed effects case, neither deviations from group means nor first differences solves the problem. Ignoring the problem—that is, simply computing the nonlinear least squares estimator without accounting for heterogeneity—does not produce a consistent estimator, for the same reasons. In general, the benign effect of latent heterogeneity (random effects) that we observe in the linear model only carries over to a very few nonlinear models and, unfortunately, this is not one of them.

The problem of computing partial effects in a random effects model such as (11-83) is that when $E[y_{it} \,|\, \mathbf{x}_{it}, u_i]$ is given by (11-83), then

$$\frac{\partial E[y_{it} \,|\, \mathbf{x}_{it}'\boldsymbol{\beta} + u_i]}{\partial \mathbf{x}_{it}} = [h'(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i)]\boldsymbol{\beta}$$

is a function of the unobservable u_i. Two ways to proceed from here are the fixed effects approach of the previous section and a random effects approach. The fixed effects approach is feasible but may be hindered by the incidental parameters problem noted earlier. A random effects approach might be preferable, but comes at the price of assuming that $\mathbf{x}_{it}$ and u_i are uncorrelated, which may be unreasonable. Papke and Wooldridge (2008) examined several cases and proposed the Mundlak approach of projecting u_i on the group means of $\mathbf{x}_{it}$. The working specification of the model is then

$$E^*[y_{it}|\mathbf{x}_{it}, \overline{\mathbf{x}}_i, v_i] = h(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha + \overline{\mathbf{x}}_i'\boldsymbol{\theta} + v_i).$$

This leaves the practical problem of how to compute the estimates of the parameters and how to compute the partial effects. Papke and Wooldridge (2008) suggest a useful result if it can be assumed that v_i is normally distributed with mean zero and variance σ_v^2. In that case,

$$E[y_{it}|\mathbf{x}_{it}, \overline{\mathbf{x}}] = E_{v_i}E[y_{it}|\mathbf{x}_{it}, \overline{\mathbf{x}}, v_i] = h\left(\frac{\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha + \overline{\mathbf{x}}_i'\boldsymbol{\theta}}{\sqrt{1 + \sigma_v^2}}\right) = h(\mathbf{x}_{it}'\boldsymbol{\beta}_v + \alpha_v + \mathbf{x}_i'\boldsymbol{\theta}_v).$$

The implication is that nonlinear least squares regression will estimate the scaled coefficients, after which the average partial effect can be estimated for a particular value of the covariates, $\mathbf{x}_0$, with

$$\hat{\Delta}(\mathbf{x}_0) = \frac{1}{n}\sum_{i=1}^{n} h'(\mathbf{x}_0'\hat{\boldsymbol{\beta}}_v + \hat{\alpha}_v + \overline{\mathbf{x}}_i'\hat{\boldsymbol{\theta}}_v)\hat{\boldsymbol{\beta}}_v.$$

They applied the technique to a case of test pass rates, which are a fraction bounded by zero and one. Loudermilk (2007) is another application with an extension to a dynamic model.

11.10 PARAMETER HETEROGENEITY

The treatment so far has assumed that the slope parameters of the model are fixed constants, and the intercept varies randomly from group to group. An equivalent formulation of the pooled, fixed, and random effects models is

$$y_{it} = (\alpha + u_i) + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it},$$

where u_i is a person-specific random variable with conditional variance zero in the pooled model, positive in the others, and conditional mean dependent on $\mathbf{X}_i$ in the fixed effects model and constant in the random effects model. By any of these, the heterogeneity in the model shows up as variation in the constant terms in the regression model. There is ample evidence in many studies—we will examine two later—that suggests that the other parameters in the model also vary across individuals. In the dynamic model we consider in Section 11.10.3, cross-country variation in the slope parameter in a production function is the central focus of the analysis. This section will consider several approaches to analyzing parameter heterogeneity in panel data models.

11.10.1 A RANDOM COEFFICIENTS MODEL

Parameter heterogeneity across individuals or groups can be modeled as stochastic variation.[41] Suppose that we write

[41] The most widely cited studies are Hildreth and Houck (1968), Swamy (1970, 1971, 1974), Hsiao (1975), and Chow (1984). See also Breusch and Pagan (1979). Some recent discussions are Swamy and Tavlas (1995, 2001) and Hsiao (2003). The model bears some resemblance to the Bayesian approach of Chapter 16. But the similarity is only superficial. We are maintaining the classical approach to estimation throughout.

$$\begin{aligned}\mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta}_i + \varepsilon_i,\\ E[\boldsymbol{\varepsilon}_i\,|\,\mathbf{X}_i] &= \mathbf{0},\\ E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'\,|\,\mathbf{X}_i] &= \sigma_\varepsilon^2\mathbf{I}_T,\end{aligned} \tag{11-84}$$

where

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{u}_i \tag{11-85}$$

and

$$\begin{aligned}E[\mathbf{u}_i\,|\,\mathbf{X}_i] &= \mathbf{0},\\ E[\mathbf{u}_i\mathbf{u}_i',\,|\,\mathbf{X}_i] &= \boldsymbol{\Gamma}.\end{aligned} \tag{11-86}$$

(Note that if only the constant term in $\boldsymbol{\beta}$ is random in this fashion and the other parameters are fixed as before, then this reproduces the random effects model we studied in Section 11.5.) Assume for now that there is no autocorrelation or cross-section correlation in $\boldsymbol{\varepsilon}_i$. We also assume for now that $T > K$, so that, when desired, it is possible to compute the linear regression of $\mathbf{y}_i$ on $\mathbf{X}_i$ for each group. Thus, the $\boldsymbol{\beta}_i$ that applies to a particular cross-sectional unit is the outcome of a random process with mean vector $\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{\Gamma}$.[42] By inserting (11-85) into (11-84) and expanding the result, we obtain a generalized regression model for each block of observations,

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + (\boldsymbol{\varepsilon}_i + \mathbf{X}_i\mathbf{u}_i),$$

so

$$\boldsymbol{\Omega}_{ii} = E[(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\,|\,\mathbf{X}_i] = \sigma_\varepsilon^2\mathbf{I}_T + \mathbf{X}_i\boldsymbol{\Gamma}\mathbf{X}_i'.$$

For the system as a whole, the disturbance covariance matrix is block diagonal, with $T \times T$ diagonal block $\boldsymbol{\Omega}_{ii}$. We can write the GLS estimator as a matrix weighted average of the group-specific OLS estimators,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y} = \sum_{i=1}^{n}\mathbf{W}_i\mathbf{b}_i, \tag{11-87}$$

where

$$\mathbf{W}_i = \left[\sum_{i=1}^{n}\left(\boldsymbol{\Gamma} + \sigma_\varepsilon^2(\mathbf{X}_i'\mathbf{X}_i)^{-1}\right)^{-1}\right]^{-1}\left(\boldsymbol{\Gamma} + \sigma_\varepsilon^2(\mathbf{X}_i'\mathbf{X}_i)^{-1}\right)^{-1}.$$

Empirical implementation of this model requires an estimator of $\boldsymbol{\Gamma}$. One approach[43] is to use the empirical variance of the set of n least squares estimates, $\mathbf{b}_i$ minus the average value of $s_i^2(\mathbf{X}_i'\mathbf{X}_i)^{-1}$,

$$\mathbf{G} = [1/(n-1)][\Sigma_i\mathbf{b}_i\mathbf{b}_i' - n\bar{\mathbf{b}}\bar{\mathbf{b}}'] - (1/N)\Sigma_i\mathbf{V}_i, \tag{11-88}$$

where

$$\bar{\mathbf{b}} = (1/n)\Sigma_i\mathbf{b}_i$$

and

$$\mathbf{V}_i = s_i^2(\mathbf{X}_i'\mathbf{X}_i)^{-1}.$$

[42] Swamy and Tavlas (2001) label this the "first-generation random coefficients model" (RCM). We will examine the "second generation" (the current generation) of random coefficients models in the next section.

[43] See, for example, Swamy (1971).

This matrix may not be positive definite, however, in which case [as Baltagi (2005) suggests], one might drop the second term.

A chi-squared test of the random coefficients model against the alternative of the classical regression[44] (no randomness of the coefficients) can be based on

$$C = \Sigma_i(\mathbf{b}_i - \mathbf{b}_*)'\mathbf{V}_i^{-1}(\mathbf{b}_i - \mathbf{b}_*),$$

where

$$\mathbf{b}_* = [\Sigma_i \mathbf{V}_i^{-1}]^{-1}\Sigma_i \mathbf{V}_i^{-1}\mathbf{b}_i.$$

Under the null hypothesis of homogeneity, C has a limiting chi-squared distribution with $(n - 1)K$ degrees of freedom. The best linear unbiased individual predictors of the group-specific coefficient vectors are matrix weighted averages of the GLS estimator, $\hat{\boldsymbol{\beta}}$, and the group-specific OLS estimates, $\mathbf{b}_i$,[45]

$$\hat{\boldsymbol{\beta}}_i = \mathbf{Q}_i\hat{\boldsymbol{\beta}} + [\mathbf{I} - \mathbf{Q}_i]\mathbf{b}_i, \qquad \textbf{(11-89)}$$

where

$$\mathbf{Q}_i = [(1/s_i^2)\mathbf{X}_i'\mathbf{X}_i + \mathbf{G}^{-1}]^{-1}\mathbf{G}^{-1}.$$

Example 11.22 Random Coefficients Model

In Examples 10.1 and 11.9, we examined Munell's production model for gross state product,

$$\ln gsp_{it} = \beta_1 + \beta_2 \ln pc_{it} + \beta_3 \ln hwy_{it} + \beta_4 \ln \text{water}_{it}$$
$$+ \beta_5 \ln util_{it} + \beta_6 \ln emp_{it} + \beta_7\, unemp_{it} + \varepsilon_{it}, \quad i = 1, \ldots, 48; t = 1, \ldots, 17.$$

The panel consists of state-level data for 17 years. The model in Example 10.1 (and Munnell's) provides no means for parameter heterogeneity save for the constant term. We have reestimated the model using the Hildreth and Houck approach. The OLS and Feasible GLS estimates are given in Table 11.21. The chi-squared statistic for testing the null hypothesis of parameter homogeneity is 25,556.26, with 7(47) = 329 degrees of freedom. The critical value from the table is 372.299, so the hypothesis would be rejected.

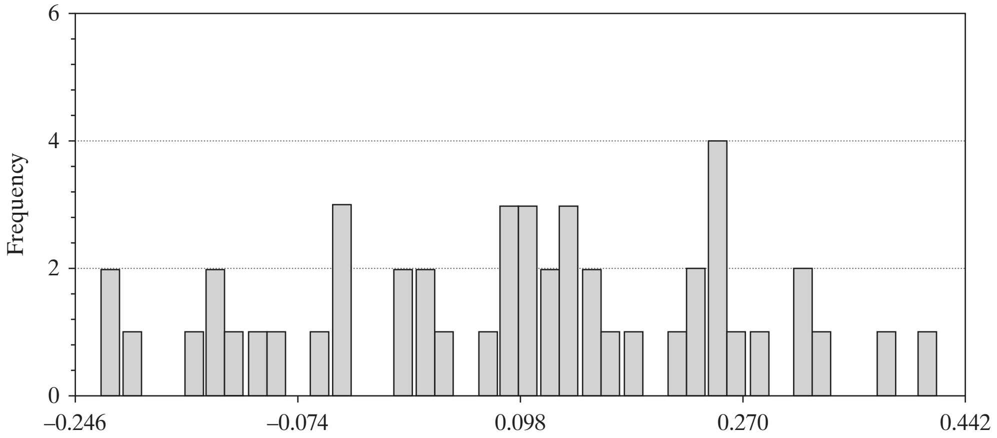
TABLE 11.21 Estimated Random Coefficients Models

	Least Squares		*Feasible GLS*		
Variable	*Estimate*	*Standard Error*	*Estimate*	*Std. Error*	*Popn. Std. Deviation*
Constant	1.9260	0.05250	1.6533	1.08331	7.0782
ln *pc*	0.3120	0.01109	0.09409	0.05152	0.3036
ln *hwy*	0.05888	0.01541	0.1050	0.1736	1.1112
ln *water*	0.1186	0.01236	0.07672	0.06743	0.4340
ln *util*	0.00856	0.01235	−0.01489	0.09886	0.6322
ln *emp*	0.5497	0.01554	0.9190	0.1044	0.6595
unemp	−0.00727	0.00138	−0.00471	0.00207	0.01266
σ_ε	0.08542			0.2129	
ln *L*	853.13720				

[44] See Swamy (1971).

[45] See Hsiao (2003, pp. 144–149).

FIGURE 11.1 Estimates of Coefficient on Private Capital.



Unlike the other cases we have examined in this chapter, the FGLS estimates are very different from OLS in these estimates, in spite of the fact that both estimators are consistent and the sample is fairly large. The underlying standard deviations are computed using **G** as the covariance matrix. [For these data, subtracting the second matrix rendered **G** not positive definite, so in the table, the standard deviations are based on the estimates using only the first term in (11-88).] The increase in the standard errors is striking. This suggests that there is considerable variation in the parameters across states. We have used (11-89) to compute the estimates of the state-specific coefficients. Figure 11.1 shows a histogram for the coefficient on private capital. As suggested, there is a wide variation in the estimates.

11.10.2 A HIERARCHICAL LINEAR MODEL

Many researchers have employed a two-step approach to estimate two-level models. In a common form of the application, a panel data set is employed to estimate the model,

$$\mathbf{y}_{it} = \mathbf{x}_{it}'\boldsymbol{\beta}_i + \varepsilon_{it}, i = 1, \ldots n, t = 1, \ldots, T,$$
$$\beta_{i,k} = \mathbf{z}_i'\boldsymbol{\alpha}_k + u_{i,k}, i = 1, \ldots, n.$$

Assuming the panel is long enough, the first equation is estimated n times, once for each individual i, and then the estimated coefficient on x_{itk} in each regression forms an observation for the second-step regression.[46] [This is the approach we took in (11-16) in Section 11.4; each a_i is computed by a linear regression of $\mathbf{y}_i - \mathbf{X}_i\mathbf{b}_{LSDV}$ on a column of ones.]

Example 11.23 Fannie Mae's Pass Through

Fannie Mae is the popular name for the Federal National Mortgage Corporation. Fannie Mae is the secondary provider for mortgage money for nearly all the small- and moderate-sized home mortgages in the United States. Loans in the study described here are termed "small" if they are for less than $100,000. A loan is termed as *conforming* in the language

[46] An extension of the model in which "u_i" is heteroscedastic is developed at length in Saxonhouse (1976) and revisited by Achen (2005).

of the literature on this market if (as of 2016), it is for no more than \$417,000. A larger than conforming loan is called a *jumbo* mortgage. Fannie Mae provides the capital for nearly all conforming loans and no nonconforming loans. (See Exercise 6.14 for another study of Fannie Mae and Freddie Mac.) The question pursued in the study described here was whether the clearly observable spread between the rates on jumbo loans and conforming loans reflects the cost of raising the capital in the market. Fannie Mae is a government sponsored enterprice (GSE). It was created by the U.S. Congress, but it is not an arm of the government; it is a private corporation. In spite of, or perhaps because of, this ambiguous relationship to the government, apparently, capital markets believe that there is some benefit to Fannie Mae in raising capital. Purchasers of the GSE's debt securities seem to believe that the debt is implicitly backed by the government—this in spite of the fact that Fannie Mae explicitly states otherwise in its publications. This emerges as a funding advantage (GFA) estimated by the authors of the study of about 16 basis points (hundredths of one percent). In a study of the residential mortgage market, Passmore (2005) and Passmore, Sherlund, and Burgess (2005) sought to determine whether this implicit subsidy to the GSE was passed on to the mortgagees or was, instead, passed on to the stockholders. Their approach utilitized a very large data set and a two-level, two-step estimation procedure. The first step equation estimated was a mortgage rate equation using a sample of roughly 1 million closed mortgages. All were conventional 30-year, fixed-rate loans closed between April 1997 and May 2003. The dependent variable of interest is the rate on the mortgage, RM_{it}. The first-level equation is

$$\begin{gathered} RM_{it} = \beta_{1i} + \beta_{2,i}\,J_{it} + \text{terms for ``loan to value ratio,'' ``new home dummy variable,''} \\ \text{``small mortgage''} \\ + \text{ terms for ``fees charged'' and whether the mortgage was originated} \\ \text{by a mortgage company} + \varepsilon_{it}. \end{gathered}$$

The main variable of interest in this model is J_{it}, which is a dummy variable for whether the loan is a jumbo mortgage. The "*i*" in this setting is a (state, time) pair for California, New Jersey, Maryland, Virginia, and all other states, and months from April 1997 to May 2003. There were 370 groups in total. The regression model was estimated for each group. At the second step, the coefficient of interest is $\beta_{2,i}$. On overall average, the spread between jumbo and conforming loans at the time was roughly 16 basis points. The second-level equation is

$$\begin{aligned} \beta_{2,i} = {} & \alpha_1 + \alpha_2\,\text{GFA}_i \\ & + \alpha_3 \text{ one-year treasury rate} \\ & + \alpha_4 \text{ 10-year treasury rate} \\ & + \alpha_5 \text{ credit risk} \\ & + \alpha_6 \text{ prepayment risk} \\ & + \text{measures of maturity mismatch risk} \\ & + \text{quarter and state fixed effects} \\ & + \text{mortgage market capacity} \\ & + \text{mortgage market development} \\ & + u_i. \end{aligned}$$

The result ultimately of interest is the coefficient on GFA, α_2, which is interpreted as the fraction of the GSE funding advantage that is passed through to the mortgage holders. Four different estimates of α_2 were obtained, based on four different measures of corporate debt liquidity; the estimated values were $(\hat{\alpha}_2^1, \hat{\alpha}_2^2, \hat{\alpha}_2^3, \hat{\alpha}_2^4) = (0.07, 0.31, 0.17, 0.10)$. The four estimates were averaged using a minimum distance estimator (MDE). Let $\hat{\boldsymbol{\Omega}}$ denote the estimated 4 × 4 asymptotic covariance matrix for the estimators. Denote the distance vector

$$\mathbf{d} = (\hat{\alpha}_2^1 - \alpha_2, \hat{\alpha}_2^2 - \alpha_2, \hat{\alpha}_2^3 - \alpha_2, \hat{\alpha}_2^4 - \alpha_2)'.$$

The minimum distance estimator is the value for α_2 that minimizes $\mathbf{d}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{d}$. For this study, $\hat{\boldsymbol{\Omega}}$ is a diagonal matrix. It is straightforward to show that in this case, the MDE is

$$\hat{\alpha}_2 = \sum_{j=1}^{4} \hat{\alpha}_2^j \left(\frac{1/\hat{\omega}_j}{\Sigma_{m=1}^4 1/\hat{\omega}_m} \right).$$

The final answer is roughly 16%. By implication, then, the authors estimated that 100 − 16 = 84 percent of the GSE funding advantage was kept within the company or passed through to stockholders.

11.10.3 PARAMETER HETEROGENEITY AND DYNAMIC PANEL DATA MODELS

The analysis in this section has involved static models and relatively straightforward estimation problems. We have seen as this section has progressed that parameter heterogeneity introduces a fair degree of complexity to the treatment. Dynamic effects in the model, with or without heterogeneity, also raise complex new issues in estimation and inference. There are numerous cases in which dynamic effects and parameter heterogeneity coincide in panel data models. This section will explore a few of the specifications and some applications. The familiar estimation techniques (OLS, FGLS, etc.) are not effective in these cases. The proposed solutions are developed in Chapter 8 where we present the technique of instrumental variables and in Chapter 13 where we present the GMM estimator and its application to dynamic panel data models.

Example 11.24 Dynamic Panel Data Models

The antecedent of much of the current research on panel data is Balestra and Nerlove's (1966) study of the natural gas market.[47] The model is a stock-flow description of the derived demand for fuel for gas using appliances. The central equation is a model for total demand,

$$G_{it} = G_{it}^* + (1 - r)G_{i,t-1},$$

where G_{it} is current total demand. Current demand consists of new demand, G_{it}^*, that is created by additions to the stock of appliances plus old demand, which is a proportion of the previous period's demand, r being the depreciation rate for gas using appliances. New demand is due to net increases in the stock of gas using appliances, which is modeled as

$$G_{it}^* = \beta_0 + \beta_1 Price_{it} + \beta_2 \Delta Pop_{it} + \beta_3 Pop_{it} + \beta_4 \Delta Income_{it} + \beta_5 Income_{it} + \varepsilon_{it},$$

where Δ is the first difference (change) operator, $\Delta X_t = X_t - X_{t-1}$. The reduced form of the model is a dynamic equation,

$$G_{it} = \beta_0 + \beta_1 Price_{it} + \beta_2 \Delta Pop_{it} + \beta_3 Pop_{it} + \beta_4 \Delta Income_{it} + \beta_5 Income_{it} + \gamma G_{i,t-1} + \varepsilon_{it}.$$

The authors analyzed a panel of 36 states over a six-year period (1957–1962). Both fixed effects and random effects approaches were considered.

An equilibrium model for steady-state growth has been used by numerous authors [e.g., Robertson and Symons (1992), Pesaran and Smith (1995), Lee, Pesaran, and Smith (1997),

[47] See, also, Nerlove (2002, Chapter 2).

Pesaran, Shin, and Smith (1999), Nerlove (2002) and Hsiao, Pesaran, and Tahmiscioglu (2002)] for cross-industry or -country comparisons. Robertson and Symons modeled real wages in 13 OECD countries over the period 1958–1986 with a wage equation

$$W_{it} = \alpha_i + \beta_{1i}k_{it} + \beta_{2i}\Delta\ wedge_{it} + \gamma_i W_{i,t-1} + \varepsilon_{it},$$

where W_{it} is the real product wage for country i in year t, k_{it} is the capital-labor ratio, and *wedge* is the "tax and import price wedge."

Lee, Pesaran, and Smith (1997) compared income growth across countries with a steady-state income growth model of the form

$$\ln y_{it} = \alpha_i + \theta_i t + \lambda_i \ln y_{i,t-1} + \varepsilon_{it},$$

where $\theta_i = (1 - \lambda_i)\delta_i$, δ_i is the technological growth rate for country i, and λ_i is the convergence parameter. The rate of convergence to a steady state is $1 - \lambda_i$.

Pesaran and Smith (1995) analyzed employment in a panel of 38 UK industries observed over 29 years, 1956–1984. The main estimating equation was

$$\ln e_{it} = \alpha_i + \beta_{1i}t + \beta_{2i}\ln y_{it} + \beta_{3i}\ln y_{i,t-1} + \beta_{4i}\ln \bar{y}_t + \beta_{5i}\ln \bar{y}_{t-1}$$
$$+ \beta_{6i}\ln w_{it} + \beta_{7i}\ln w_{i,t-1} + \gamma_{1i}\ln e_{i,t-1} + \gamma_{2i}\ln e_{i,t-2} + \varepsilon_{it},$$

where y_{it} is industry output, $\bar{y}_t$ is total (not average) output, and w_{it} is real wages.

In the growth models, a quantity of interest is the **long-run multiplier** or **long-run elasticity**. Long-run effects are derived through the following conceptual experiment. The essential feature of the models above is a dynamic equation of the form

$$y_t = \alpha + \beta x_t + \gamma y_{t-1}.$$

Suppose at time t, x_t is fixed from that point forward at $\bar{x}$. The value of y_t at that time will then be $\alpha + \beta\bar{x} + \gamma y_{t-1}$, given the previous value. If this process continues, and if $|\gamma| < 1$, then eventually y_s will reach an equilibrium at a value such that $y_s = y_{s-1} = \bar{y}$. If so, then $\bar{y} = \alpha + \beta\bar{x} + \gamma\bar{y}$, from which we can deduce that $\bar{y} = (\alpha + \bar{x})/(1 - \gamma)$. The path to this equilibrium from time t into the future is governed by the **adjustment equation**

$$y_s - \bar{y} = (y_t - \bar{y})\gamma^{s-t}, s \geq t.$$

The experiment, then, is to ask: What is the impact on the equilibrium of a change in the input, $\bar{x}$? The result is $\partial\bar{y}/\partial\bar{x} = \beta/(1 - \gamma)$. This is the long-run multiplier, or **equilibrium multiplier**, in the model. In the preceding Pesaran and Smith model, the inputs are in logarithms, so the multipliers are long-run elasticities. For example, with two lags of $\ln e_{it}$ in Pesaran and Smith's model, the long-run effects for wages are

$$\phi_i = (\beta_{6i} + \beta_{7i})/(1 - \gamma_{1i} - \gamma_{2i}).$$

In this setting, in contrast to the preceding treatments, the number of units, n, is generally taken to be fixed, though often it will be fairly large. The Penn World Tables (http://cid.econ.ucdavis.edu/pwt.html) that provide the database for many of these analyses now contain information on more than 150 countries for well more than 50 years. Asymptotic results for the estimators are with respect to increasing T, though we will consider, in general, cases in which T is small. Surprisingly, increasing T and n at the same time need not simplify the derivations.

The parameter of interest in many studies is the average long-run effect, say $\bar{\phi} = (1/n)\Sigma_i\phi_i$, in the Pesaran and Smith example. Because n is taken to be fixed, the "parameter" $\bar{\phi}$ is a definable object of estimation—that is, with n fixed, we can speak of $\overline{\phi}$ as a parameter rather than as an estimator of a parameter. There are numerous approaches one might take. For estimation purposes, pooling, fixed effects, random effects, group means, or separate regressions are all possibilities. (Unfortunately, nearly all are inconsistent.) In addition, there is a choice to be made whether to compute the average of long-run effects or to compute the long-run effect from averages of the parameters. The choice of the average of functions, $\overline{\phi}$ versus the function of averages,

$$\overline{\phi}^* = \frac{\frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_{6i} + \hat{\beta}_{7i})}{1 - \frac{1}{n}\sum_{i=1}^{n}(\hat{\gamma}_{1i} + \hat{\gamma}_{2i})},$$

turns out to be of substance. For their UK industry study, Pesaran and Smith report estimates of −0.33 for $\overline{\phi}$ and −0.45 for $\overline{\phi}^*$. (The authors do not express a preference for one over the other.)

The development to this point is implicitly based on estimation of separate models for each unit (country, industry, etc.). There are also a variety of other estimation strategies one might consider. We will assume for the moment that the data series are stationary in the dimension of T. (See Chapter 21.) This is a transparently false assumption, as revealed by a simple look at the trends in macroeconomic data, but maintaining it for the moment allows us to proceed. We will reconsider it later.

We consider the generic, dynamic panel data model,

$$y_{it} = \alpha_i + \beta_i x_{it} + \gamma_i y_{i,t-1} + \varepsilon_{it}. \qquad \textbf{(11-90)}$$

Assume that T is large enough that the individual regressions can be computed. In the absence of autocorrelation in ε_{it}, it has been shown[48] that the OLS estimator of γ_i is biased downward, but consistent in T. Thus, $E[\hat{\gamma}_i - \gamma_i] = \theta_i/T$ for some θ_i. The implication for the individual estimator of the long-run multiplier, $\phi_i = \beta_i/(1 - \gamma_i)$, is unclear in this case, however. The denominator is overestimated. But it is not clear whether the estimator of β_i is overestimated or underestimated. It is true that whatever bias there is is O(1/T). For this application, T is fixed and possibly quite small. The end result is that it is unlikely that the individual estimator of ϕ_i is unbiased, and by construction, it is inconsistent, because T cannot be assumed to be increasing. If that is the case, then $\hat{\overline{\phi}}$ is likewise inconsistent for $\overline{\phi}$. We are averaging n estimators, each of which has bias and variance that are O(1/T). The variance of the mean is, therefore, O(1/nT) which goes to zero, but the bias remains O(1/T). It follows that the average of the n means is not converging to $\overline{\phi}$; it is converging to the average of whatever these biased estimators are estimating. The problem vanishes with large T, but that is not relevant to the current context. However, in the Pesaran and Smith study, T was 29, which is large enough that these effects are probably moderate. For macroeconomic cross-country studies such as those based on the Penn World Tables, the data series may be even longer than this.

One might consider aggregating the data to improve the results. Pesaran and Smith (1995) suggest an average based on country means. Averaging the observations over T in (11-90) produces

$$\overline{y}_{i.} = \alpha_i + \beta_i \overline{x}_{i.} + \gamma_i \overline{y}_{-1,i} + \overline{\varepsilon}_{i.}. \qquad \textbf{(11-91)}$$

A linear regression using the n observations would be inconsistent for two reasons: First, $\overline{\varepsilon}_{i.}$ and $\overline{y}_{-1,i}$ must be correlated. Second, because of the parameter heterogeneity, it is not clear

[48] For example, Griliches (1961) and Maddala and Rao (1973).

without further assumptions what the OLS slopes estimate under the false assumption that all coefficients are equal. But $\bar{y}_{i.}$ and $\bar{y}_{-1,i}$ differ by only the first and last observations; $\bar{y}_{-1,i} = \bar{y}_{i.} - (y_{iT} - y_{i0})/T = \bar{y}_{i.} - [\Delta_T(y)/T]$. Inserting this in (11-91) produces

$$\begin{aligned}\bar{y}_{i.} &= \alpha_i + \beta_i\bar{x}_{i.} + \gamma_i\bar{y}_{i.} - \gamma_i[\Delta_T(y)/T] + \bar{\varepsilon}_{i.}\\ &= \frac{\alpha_i}{1-\gamma_i} + \frac{\beta_i}{1-\gamma_i}\bar{x}_{i.} - \frac{\gamma_i}{1-\gamma_i}[\Delta_T(y)/T] + \bar{\varepsilon}_{i.}\\ &= \delta_i + \phi_i\bar{x}_{i.} + \tau_i[\Delta_T(y)/T] + \bar{\varepsilon}_{i.}.\end{aligned} \tag{11-92}$$

We still seek to estimate $\bar{\phi}$. The form in (11-92) does not solve the estimation problem, because the regression suggested using the group means is still heterogeneous. If it could be assumed that the individual long-run coefficients differ randomly from the averages in the fashion of the random parameters model of Section 11.10.1, so $\delta_i = \bar{\delta} + u_{\delta,i}$ and likewise for the other parameters, then the model could be written

$$\begin{aligned}\bar{y}_{i.} &= \bar{\delta} + \bar{\phi}\bar{x}_{i.} + \bar{\tau}[\Delta_T(y)/T]_i + \bar{\varepsilon}_{i.} + \{u_{\delta,i} + u_{\phi,i}\bar{x}_i + u_{\tau,i}[\Delta_T(y)/T]_i\}\\ &= \bar{\delta} + \bar{\phi}\bar{x}_{i.} + \bar{\tau}[\Delta_T(y)/T]_i + \bar{\varepsilon}_i + w_i.\end{aligned}$$

At this point, the equation appears to be a heteroscedastic regression amenable to least squares estimation, but for one loose end. Consistency follows if the terms $[\Delta_T(y)/T]_i$ and $\bar{\varepsilon}_i$ are uncorrelated. Because the first is a rate of change and the second is in levels, this should generally be the case. Another interpretation that serves the same purpose is that the rates of change in $[\Delta_T(y)/T]_i$ should be uncorrelated with the levels in $\bar{x}_{i.}$, in which case, the regression can be partitioned, and simple linear regression of the country means of y_{it} on the country means of x_{it} and a constant produces consistent estimates of $\bar{\phi}$ and $\bar{\delta}$.

Alternatively, consider a time-series approach. We average the observation in (11-90) across countries at each time period rather than across time within countries. In this case, we have

$$\bar{y}_{.t} = \bar{\alpha} + \frac{1}{n}\sum_{i=1}^{n}\beta_i x_{it} + \frac{1}{n}\sum_{i=1}^{n}\gamma_i y_{i,t-1} + \frac{1}{n}\sum_{i=1}^{n}\varepsilon_{it}.$$

Let $\bar{\gamma} = \frac{1}{n}\sum_{i=1}^{n}\gamma_i$ so that $\gamma_i = \bar{\gamma} + (\gamma_i - \bar{\gamma})$ and $\beta_i = \bar{\beta} + (\beta_i - \bar{\beta})$. Then,

$$\begin{aligned}\bar{y}_{.t} &= \bar{\alpha} + \bar{\beta}\bar{x}_{.t} + \bar{\gamma}\bar{y}_{-1,t} + [\bar{\varepsilon}_{.t} + (\beta_i - \bar{\beta})\bar{x}_{.t} + (\gamma_i - \bar{\gamma})\bar{y}_{-1,t}]\\ &= \bar{\alpha} + \bar{\beta}\bar{x}_{.t} + \bar{\gamma}\bar{y}_{-1,t} + \bar{\varepsilon}_{.t} + w_{.t}.\end{aligned}$$

Unfortunately, the regressor, $\bar{\gamma}\bar{y}_{-1,t}$ is surely correlated with $w_{.t}$, so neither OLS or GLS will provide a consistent estimator for this model. (One might consider an instrumental variable estimator; however, there is no natural instrument available in the model as constructed.) Another possibility is to pool the entire data set, possibly with random or fixed effects for the constant terms. Because pooling, even with country-specific constant terms, imposes homogeneity on the other parameters, the same problems we have just observed persist.

Finally, returning to (11-90), one might treat it as a formal random parameters model,

$$\begin{aligned}y_{it} &= \alpha_i + \beta_i x_{it} + \gamma_i y_{i,t-1} + \varepsilon_{it},\\ \alpha_i &= \alpha + u_{\alpha,i},\\ \beta_i &= \beta + u_{\beta,i},\\ \gamma_i &= \gamma + u_{\gamma,i}.\end{aligned} \tag{11-93}$$

The assumptions needed to formulate the model in this fashion are those of the previous section. As Pesaran and Smith (1995) observe, this model can be estimated using the Swamy (1971) estimator, which is the matrix weighted average of the least squares estimators discussed in Section 11.11.1. The estimator requires that T be large enough to fit each country regression by least squares. That has been the case for the received applications. Indeed, for the applications we have examined, both n and T are relatively large. If not, then one could still use the mixed models approach developed in Chapter 15. A compromise that appears to work well for panels with moderate sized n and T is the "mixed-fixed" model suggested in Hsiao (1986, 2003) and Weinhold (1999). The dynamic model in (11-92) is formulated as a partial fixed effects model,

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

where d_{it} is a dummy variable that equals one for country i in every period and zero otherwise (i.e., the usual fixed effects approach). Note that d_{it} also appears with $y_{i,t-1}$. As stated, the model has "fixed effects," one random coefficient, and a total of $2n + 1$ coefficients to estimate, in addition to the two variance components, σ_ε^2 and σ_u^2. The model could be estimated inefficiently by using ordinary least squares—the random coefficient induces heteroscedasticity (see Section 11.10.1)—by using the Hildreth–Houck–Swamy approach, or with the mixed linear model approach developed in Chapter 15.

Example 11.25 A Mixed Fixed Growth Model for Developing Countries

Weinhold (1996) and Nair–Reichert and Weinhold (2001) analyzed growth and development in a panel of 24 developing countries observed for 25 years, 1971–1995. The model they employed was a variant of the mixed-fixed model proposed by Hsiao (1986, 2003). In their specification,

$$\begin{aligned} GGDP_{i,t} = {} & \alpha_i d_{it} + \gamma_i d_{it} GGDP_{i,t-1} \\ & + \beta_{1i} GGDI_{i,t-1} + \beta_{2i} GFDI_{i,t-1} + \beta_{3i} GEXP_{i,t-1} + \beta_4 INFL_{i,t-1} + \varepsilon_{it}, \end{aligned}$$

where

$GGDP$ = Growth rate of gross domestic product,
$GGDI$ = Growth rate of gross domestic investment,
$GFDI$ = Growthrate of foreign direct investment (inflows),
$GEXP$ = Growth rate of exports of goods and services,
$INFL$ = Inflation rate.

11.11 SUMMARY AND CONCLUSIONS

This chapter has shown a few of the extensions of the classical model that can be obtained when panel data are available. In principle, any of the models we have examined before this chapter and all those we will consider later, including the multiple equation models, can be extended in the same way. The main advantage, as we noted at the outset, is that with panel data, one can formally model dynamic effects and the heterogeneity across groups that are typical in microeconomic data.

Key Terms and Concepts

- Adjustment equation
- Arellano and Bond's estimator
- Balanced panel
- Between groups
- Contiguity
- Contiguity matrix
- Contrasts
- Dynamic panel data model
- Equilibrium multiplier
- Error components model
- Estimator
- Feasible GLS
- First difference
- Fixed effects
- Fixed panel
- Group means
- Group means estimator
- Hausman specification test
- Heterogeneity
- Hierarchical model
- Incidental parameters problem
- Index function model
- Individual effect
- Instrumental variable
- Instrumental variable estimator
- Lagrange multiplier test
- Least squares dummy variable model (LSDV)
- Long run elasticity
- Long run multiplier
- Longitudinal data set
- Matrix weighted average
- Mundlak's approach
- Panel data
- Partial effects
- Pooled model
- Projections
- Rotating panel
- Spatial autocorrelation
- Spatial autoregression coefficient
- Spatial error correlation
- Spatial lags
- Specification test
- Strict exogeneity
- Time invariant
- Unbalanced panel
- Within groups

Exercises

1. The following is a panel of data on investment (y) and profit (x) for $n = 3$ firms over $T = 10$ periods.

	$i = 1$		$i = 2$		$i = 3$	
t	y	x	y	x	y	x
1	13.32	12.85	20.30	22.93	8.85	8.65
2	26.30	25.69	17.47	17.96	19.60	16.55
3	2.62	5.48	9.31	9.16	3.87	1.47
4	14.94	13.79	18.01	18.73	24.19	24.91
5	15.80	15.41	7.63	11.31	3.99	5.01
6	12.20	12.59	19.84	21.15	5.73	8.34
7	14.93	16.64	13.76	16.13	26.68	22.70
8	29.82	26.45	10.00	11.61	11.49	8.36
9	20.32	19.64	19.51	19.55	18.49	15.44
10	4.77	5.43	18.32	17.06	20.84	17.87

a. Pool the data and compute the least squares regression coefficients of the model

$$y_{it} = \alpha + \beta x_{it} + \varepsilon_{it}.$$

b. Estimate the fixed effects model of (11-11), and then test the hypothesis that the constant term is the same for all three firms.

c. Estimate the random effects model of (11-28), and then carry out the Lagrange multiplier test of the hypothesis that the classical model without the common effect applies.

d. Carry out Hausman's specification test for the random versus the fixed effect model.

2. Suppose that the fixed effects model is formulated with an overall constant term and $n - 1$ dummy variables (dropping, say, the last one). Investigate the effect that this supposition has on the set of dummy variable coefficients and on the least squares estimates of the slopes, compared to (11-13).
3. *Unbalanced design for random effects.* Suppose that the random effects model of Section 11.5 is to be estimated with a panel in which the groups have different numbers of observations. Let T_i be the number of observations in group i.
   a. Show that the pooled least squares estimator is unbiased and consistent despite this complication.
   b. Show that the estimator in (11-40) based on the pooled least squares estimator of $\boldsymbol{\beta}$ (or, for that matter, *any* consistent estimator of $\boldsymbol{\beta}$) is a consistent estimator of σ_ε^2.
4. What are the probability limits of $(1/n)$ LM, where LM is defined in (11-42) under the null hypothesis that $\sigma_u^2 = 0$ and under the alternative that $\sigma_u^2 \neq 0$?
5. *A two-way fixed effects model.* Suppose that the fixed effects model is modified to include a time-specific dummy variable as well as an individual-specific variable. Then $y_{it} = \alpha_i + \gamma_t + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}$. At every observation, the individual- and time-specific dummy variables sum to 1, so there are some redundant coefficients. The discussion in Section 11.4.4 shows that one way to remove the redundancy is to include an overall constant and drop one of the time-specific *and* one of the time dummy variables. The model is, thus,

$$y_{it} = \mu + (\alpha_i - \alpha_1) + (\gamma_t - \gamma_1) + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}.$$

(Note that the respective time- or individual-specific variable is zero when t or i equals one.) Ordinary least squares estimates of $\boldsymbol{\beta}$ are then obtained by regression of $y_{it} - \bar{y}_{i.} - \bar{y}_{.t} + \bar{\bar{y}}$ on $\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.} - \bar{\mathbf{x}}_{.t} + \bar{\bar{\mathbf{x}}}$. Then $(\alpha_i - \alpha_1)$ and $(\gamma_t - \gamma_1)$ are estimated using the expressions in (11-25). Using the following data, estimate the full set of coefficients for the least squares dummy variable model:

	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$	$t = 7$	$t = 8$	$t = 9$	$t = 10$
					$i = 1$					
y	21.7	10.9	33.5	22.0	17.6	16.1	19.0	18.1	14.9	23.2
x_1	26.4	17.3	23.8	17.6	26.2	21.1	17.5	22.9	22.9	14.9
x_2	5.79	2.60	8.36	5.50	5.26	1.03	3.11	4.87	3.79	7.24
					$i = 2$					
y	21.8	21.0	33.8	18.0	12.2	30.0	21.7	24.9	21.9	23.6
x_1	19.6	22.8	27.8	14.0	11.4	16.0	28.8	16.8	11.8	18.6
x_2	3.36	1.59	6.19	3.75	1.59	9.87	1.31	5.42	6.32	5.35
					$i = 3$					
y	25.2	41.9	31.3	27.8	13.2	27.9	33.3	20.5	16.7	20.7
x_1	13.4	29.7	21.6	25.1	14.1	24.1	10.5	22.1	17.0	20.5
x_2	9.57	9.62	6.61	7.24	1.64	5.99	9.00	1.75	1.74	1.82
					$i = 4$					
y	15.3	25.9	21.9	15.5	16.7	26.1	34.8	22.6	29.0	37.1
x_1	14.2	18.0	29.9	14.1	18.4	20.1	27.6	27.4	28.5	28.6
x_2	4.09	9.56	2.18	5.43	6.33	8.27	9.16	5.24	7.92	9.63

Test the hypotheses that (1) the *period* effects are all zero, (2) the *group* effects are all zero, and (3) both period and group effects are zero. Use an F test in each case.

6. *Two-way random effects model.* We modify the random effects model by the addition of a time-specific disturbance. Thus,

$$y_{it} = \alpha + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} + u_i + v_t,$$

where

$$\begin{aligned} E[\varepsilon_{it}|\mathbf{X}] &= E[u_i|\mathbf{X}] = E[v_t|\mathbf{X}] = 0, \\ E[\varepsilon_{it}u_j|\mathbf{X}] &= E[\varepsilon_{it}v_s|\mathbf{X}] = E[u_iv_t|\mathbf{X}] = 0 \quad \text{for all } i, j, t, s, \\ \text{Var}[\varepsilon_{it}|\mathbf{X}] &= \sigma_\varepsilon^2, \quad \text{Cov}[\varepsilon_{it}, \varepsilon_{js}|\mathbf{X}] = 0 \quad \text{for all } i, j, t, s, \\ \text{Var}[u_i|\mathbf{X}] &= \sigma_u^2, \quad \text{Cov}[u_i, u_j|\mathbf{X}] = 0 \quad \text{for all } i, j, \\ \text{Var}[v_t|\mathbf{X}] &= \sigma_v^2, \quad \text{Cov}[v_t, v_s|\mathbf{X}] = 0 \quad \text{for all } t, s. \end{aligned}$$

Write out the full disturbance covariance matrix for a data set with $n = 2$ and $T = 2$.

7. In Section 11.4.5, we found that the group means of the time-varying variables would work as a control function in estimation of the fixed effects model. That is, although regression of $\mathbf{y}$ on $\mathbf{X}$ is inconsistent for $\boldsymbol{\beta}$, the Mundlak estimator, regression of $\mathbf{y}$ on $\mathbf{X}$ and $\overline{\mathbf{X}} = \mathbf{P_D X} = (\mathbf{I} - \mathbf{M_D})\mathbf{X}$ is a consistent estimator. Would the deviations from group means, $\ddot{\mathbf{X}} = \mathbf{M_D X} = (\mathbf{X} - \overline{\mathbf{X}})$, also be useable as a control function estimator. That is, does regression of $\mathbf{y}$ on $(\mathbf{X}, \ddot{\mathbf{X}})$ produce a consistent estimator of $\boldsymbol{\beta}$?
8. Prove plim $(1/nT)\mathbf{X}'\mathbf{M_D}\boldsymbol{\varepsilon} = \mathbf{0}$.
9. If the panel has $T = 2$ periods, the LSDV (within groups) estimator gives the same results as first differences. Prove this claim.

Applications

The following applications require econometric software.

1. Several applications in this and previous chapters have examined the returns to education in panel data sets. Specifically, we applied Hausman and Taylor's approach in Examples 11.17 and 11.18. Example 11.18 used Cornwell and Rupert's data for the analysis. Koop and Tobias's (2004) study that we used in Chapters 3 and 5 provides yet another application that we can use to continue this analysis. The data may be downloaded from the *Journal of Applied Econometrics* data archive at http://qed.econ.queensu.ca/jae/2004-vl9.7/koop-tobias/. The data file is in two parts. The first file contains the full panel of 17,919 observations on variables:

Column 1; *Person id* (ranging from 1 to 2,178),
Column 2; *Education*,
Column 3; *Log of hourly wage*,
Column 4; *Potential experience*,
Column 5; *Time trend*.

Columns 2 through 5 contain time-varying variables. The second part of the data set contains time-invariant variables for the 2,178 households. These are:

Column 1; *Ability*,
Column 2; *Mother's education*,
Column 3; *Father's education*,
Column 4; *Dummy variable for residence in a broken home*,
Column 5; *Number of siblings*.

To create the data set for this exercise, it is necessary to merge these two data files. The ith observation in the second file will be replicated T_i times for the set of T_i observations in the first file. The *person id* variable indicates which rows must contain the data from the second file. (How this preparation is carried out will vary from one computer package to another.) The panel is quite unbalanced; the number of observations by group size is:

Value of T_i

1:83, 2:104, 3:102, 4:116
5:148, 6:165, 7:201, 8:202
9:200, 10:202, 11:182, 12:148
13:136, 14:96, 15:93

a. Using these data, fit fixed and random effects models for log wage and examine the result for the return to education.

b. For a Hausman–Taylor specification, consider the following:

$$\begin{aligned}\mathbf{x}_1 &= \text{potential experience, ability}\\ \mathbf{x}_2 &= \text{education}\\ \mathbf{f}_1 &= \text{constant, number of siblings, broken home}\\ \mathbf{f}_2 &= \text{mother's education, father's education}\end{aligned}$$

Based on this specification, what is the estimated return to education? (*Note:* you may need the average value of $1/T_i$ for your calculations. This is 0.1854.)

c. It might seem natural to include ability with education in $\mathbf{x}_2$. What becomes of the Hausman and Taylor estimator if you do so?

d. Using a different specification, compute an estimate of the return to education using the instrumental variables method.

e. Compare your results in parts b and d to the results in Examples 11.17 and 11.18. The estimated return to education is surprisingly stable.

2. The data in Appendix Table F10.4 were used by Grunfeld (1958) and dozens of researchers since, including Zellner (1962, 1963) and Zellner and Huang (1962), to study different estimators for panel data and linear regression systems. [See Kleiber and Zeileis (2010).] The model is an investment equation,

$$I_{it} = \beta_1 + \beta_2 F_{it} + \beta_3 C_{it} + \varepsilon_{it}, t = 1, \ldots, 20, i = 1, \ldots, 10,$$

where

$$\begin{aligned}I_{it} &= \text{real gross investment for firm } i \text{ in year } t,\\ F_{it} &= \text{real value of the firm—shares outstanding},\\ C_{it} &= \text{real value of the capital stock.}\end{aligned}$$

For present purposes, this is a balanced panel data set.

a. Fit the pooled regression model.

b. Referring to the results in part a, is there evidence of within-groups correlation? Compute the robust standard errors for your pooled OLS estimator and compare them to the conventional ones.

c. Compute the fixed effects estimator for these data. Then, using an *F* test, test the hypothesis that the constants for the 10 firms are all the same.

d. Use a Lagrange multiplier statistic to test for the presence of common effects in the data.

e. Compute the one-way random effects estimator and report all estimation results. Explain the difference between this specification and the one in part c.

f. Use a Hausman test to determine whether a fixed or random effects specification is preferred for these data.

3. The data in Appendix Table F6.1 are an unbalanced panel on 25 U.S. airlines in the pre-deregulation days of the 1970s and 1980s. The group sizes range from 2 to 15. Data in the file are the following variables. (Variable names contained in the data file are constructed to indicate the variable contents.)

   Total cost,
   Expenditures on Capital, Labor, Fuel, Materials, Property, and Equipment,
   Price measures for the six inputs,
   Quantity measures for the six inputs,
   Output measured in revenue passenger miles, converted to an index number for the airline,
   Load factor = the average percentage capacity utilization of the airline's fleet,
   Stage = the average flight (stage) length in miles,
   Points = the number of points served by the airline,
   Year = the calendar year,
   T Year = 1969,
   TI = the number of observations for the airline, repeated for each year.

   Use these data to build a cost model for airline service. Allow for cross-airline heterogeneity in the constants in the model. Use both random and fixed effects specifications, and use available statistical tests to determine which is the preferred model. An appropriate cost model to begin the analysis with would be

$$\ln cost_{it} = \alpha_i + \sum_{k=1}^{6} \beta_k \ln Price_{k,it} + \gamma \ln Output_{it} + \varepsilon_{it}.$$

   It is necessary to impose linear homogeneity in the input prices on the cost function, which you would do by dividing five of the six prices and the total cost by the sixth price (choose any one), then using $\ln(cost/P_6)$ and $\ln(P_k/P_6)$ in the regression. You might also generalize the cost function by including a quadratic term in the log of output in the function. A translog model would include the unique squares and cross products of the input prices and products of log output with the logs of the prices. The data include three additional factors that may influence costs, stage length, load factor, and number of points served. Include them in your model, and use the appropriate test statistic to test whether they are, indeed, relevant to the determination of (log) total cost.