13 MODELS FOR PANEL DATA

13.1 INTRODUCTION

Data sets that combine time series and cross sections are common in economics. For example, the published statistics of the OECD contain numerous series of economic aggregates observed yearly for many countries. Recently constructed **longitudinal data sets** contain observations on thousands of individuals or families, each observed at several points in time. Other empirical studies have analyzed time-series data on sets of firms, states, countries, or industries simultaneously. These data sets provide rich sources of information about the economy. Modeling in this setting, however, calls for some complex stochastic specifications. In this chapter, we will survey the most commonly used techniques for time-series cross-section data analyses in single equation models.

13.2 PANEL DATA MODELS

Many recent studies have analyzed **panel,** or longitudinal, data sets. Two very famous ones are the National Longitudinal Survey of Labor Market Experience (NLS) and the Michigan Panel Study of Income Dynamics (PSID). 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. Another group of intensively studied panel data sets were those from the negative income tax experiments of the early 1970s in which thousands of families were followed for 8 or 13 quarters. 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. 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.

[1]Theorists have not been deterred from devising autocorrelation models applicable to panel data sets; though. See, for example, Lee (1978) or Park, Sickles, and Simar (2000). As a practical matter, however, the empirical literature in this field has focused on cross-sectional variation and less intricate time series models. Formal time-series modeling of the sort discussed in Chapter 12 is somewhat unusual in the analysis of longitudinal data.

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 or time-series settings alone. Two examples are as follows.

1. In a widely cited study of labor supply, Ben-Porath (1973) observes that at a certain point in time, in a cohort of women, 50 percent may appear to be working. It is ambiguous whether this finding implies that, in this cohort, one-half of the women on average will be working or that the same one-half will be working in every period. These have very different implications for policy and for the interpretation of any statistical results. Cross-sectional data alone will not shed any light on the question.

2. A long-standing problem in the analysis of production functions has been the inability to separate economies of scale and technological change.[3] Cross-sectional data provide information only about the former, whereas time-series data muddle the two effects, with no prospect of separation. It is common, for example, to assume constant returns to scale so as to reveal the technical change.[4] Of course, this practice assumes away the problem. A panel of data on costs or output for a number of firms each observed over several years can provide estimates of both the rate of technological change (as time progresses) and economies of scale (for the sample of different sized firms at each point in time).

In principle, the methods of Chapter 12 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.

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.

[2]The panel data literature rivals the received research on unit roots and cointegration in econometrics in its rate of growth. A compendium of the earliest literature is Maddala (1993). Book-length surveys on the econometrics of panel data include Hsiao (1986), Dielman (1989), Matyas and Sevestre (1996), Raj and Baltagi (1992), and Baltagi (1995). There are also lengthy surveys devoted to specific topics, such as limited dependent variable models [Hsiao, Lahiri, Lee, and Pesaran (1999)] and semiparametric methods [Lee (1998)]. An extensive bibliography is given in Baltagi (1995).

[3]The distinction between these two effects figured prominently in the policy question of whether it was appropriate to break up the AT&T Corporation in the 1980s and, ultimately, to allow competition in the provision of long-distance telephone service.

[4]In a classic study of this issue, Solow (1957) states: "From time series of $\Delta Q/Q$, w_K, $\Delta K/K$, w_L and $\Delta L/L$ or their discrete year-to-year analogues, we could estimate $\Delta A/A$ and thence $A(t)$ itself. Actually an amusing thing happens here. Nothing has been said so far about returns to scale. But if all factor inputs are classified either as K or L, then the available figures always show w_K and w_L adding up to one. Since we have assumed that factors are paid their marginal products, this amounts to assuming the hypothesis of Euler's theorem. The calculus being what it is, we might just as well assume the conclusion, namely, the F is homogeneous of degree one."

The basic framework for this discussion is a regression model of the form

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\alpha} + \varepsilon_{it}. \tag{13-1}$$

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 various cases we will consider are:

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 β 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 indicates that the term does not vary over time, not that it is nonstochastic, which need not be the case.

3. **Random Effects:** If the unobserved individual **heterogeneity,** however formulated, can be assumed to be uncorrelated with the included variables, then the model may be formulated as

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

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 these two cases 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

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

where $\boldsymbol{h}_i$ is a random vector which induces the variation of the parameters across individuals. This random parameters model was proposed quite early in this literature, but has only fairly 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

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

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 later, this **hierarchical model** is extremely versatile.

5. **Covariance Structures:** Lastly, we will reconsider the source of the heterogeneity in the model. In some settings, researchers have concluded that a preferable approach to modeling heterogeneity in the regression model is to layer it into the variation around the conditional mean, rather than in the placement of the mean. In a cross-country comparison of economic performance over time, Alvarez, Garrett, and Lange (1991) estimated a model of the form

$$y_{it} = f(\text{labor organization}_{it}, \text{political organization}_{it}) + \varepsilon_{it}$$

in which the regression function was fully specified by the linear part, $\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha$, but the variance of ε_{it} differed across countries. Beck et al. (1993) found evidence that the substantive conclusions of the study were dependent on the stochastic specification and on the methods used for estimation.

Example 13.1 Cost Function for Airline Production

To illustrate the computations for the various panel data models, we will revisit the airline cost data used in Example 7.2. This is a panel data study of a group of U.S. airlines. We will fit a simple model for the total cost of production:

$$\ln \text{cost}_{it} = \beta_1 + \beta_2 \ln \text{output}_{it} + \beta_3 \ln \text{fuel price}_{it} + \beta_4 \text{ load factor}_{it} + \varepsilon_{it}.$$

Output is measured in "revenue passenger miles." The load factor is a rate of capacity utilization; it is the average rate at which seats on the airline's planes are filled. More complete models of costs include other factor prices (materials, capital) and, perhaps, a quadratic term in log output to allow for variable economies of scale. We have restricted the cost function to these few variables to provide a straightforward illustration.

Ordinary least squares regression produces the following results. Estimated standard errors are given in parentheses.

$$\begin{aligned}\ln \text{cost}_{it} &= 9.5169(0.22924) + 0.88274(0.013255) \ln \text{output}_{it} \\ &\quad + 0.45398(0.020304) \ln \text{fuel price}_{it} - 1.62751(0.34540) \text{load factor}_{it} + \varepsilon_{it} \\ R^2 &= 0.9882898, s^2 = 0.015528, \mathbf{e}'\mathbf{e} = 1.335442193.\end{aligned}$$

The results so far are what one might expect. There are substantial economies of scale; $\text{e.s.}_{it} = (1/0.88274) - 1 = 0.1329$. The fuel price and load factors affect costs in the predictable fashions as well. (Fuel prices differ because of different mixes of types of planes and regional differences in supply characteristics.)

13.3 FIXED EFFECTS

This formulation of the model assumes that differences across units can be captured in differences in the constant term.[5] Each α_i is treated as an unknown parameter to be estimated. Let $\mathbf{y}_i$ and $\mathbf{X}_i$ be the T observations for the ith unit, $\mathbf{i}$ be a $T \times 1$ column of ones, and let $\boldsymbol{\varepsilon}_i$ be associated $T \times 1$ vector of disturbances. Then,

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

Collecting these terms gives

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

or

$$\mathbf{y} = [\mathbf{X} \quad \mathbf{d}_1 \quad \mathbf{d}_2 \ldots \mathbf{d}_n] \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{bmatrix} + \boldsymbol{\varepsilon}, \qquad \textbf{(13-2)}$$

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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}. \qquad \textbf{(13-3)}$$

This model is usually 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 this model is likely to exceed the storage capacity of any computer. But, by using familiar results for a partitioned regression, we can reduce the size of the computation.[6] We write the least squares estimator of $\boldsymbol{\beta}$ as

$$\mathbf{b} = [\mathbf{X}'\mathbf{M}_\mathbf{D}\mathbf{X}]^{-1}[\mathbf{X}'\mathbf{M}_\mathbf{D}\mathbf{y}], \qquad \textbf{(13-4)}$$

where

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

This amounts to a least squares regression using the transformed data $\mathbf{X}_* = \mathbf{M}_\mathbf{D}\mathbf{X}$ and

[5]It is also possible to allow the slopes to vary across i, but this method introduces some new methodological issues, as well as considerable complexity in the calculations. A study on the topic is Cornwell and Schmidt (1984). Also, the assumption of a fixed T is only for convenience. The more general case in which T_i varies across units is considered later, in the exercises, and in Greene (1995a).

[6]See Theorem 3.3.

$\mathbf{y}_* = \mathbf{M_D y}$. The structure of $\mathbf{D}$ is particularly convenient; its columns are orthogonal, so

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

Each matrix on the diagonal is

$$\mathbf{M}^0 = \mathbf{I}_T - \frac{1}{T}\mathbf{ii}'. \tag{13-5}$$

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 $\mathbf{M_D y}$ on $\mathbf{M_D X}$ is equivalent to a regression of $[y_{it} - \bar{y}_{i.}]$ on $[\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.}]$, where $\bar{y}_{i.}$ and $\bar{\mathbf{x}}_{i.}$ are the scalar and $K \times 1$ vector of means of y_{it} and $\mathbf{x}_{it}$ over the T observations for group i.[7] The dummy variable coefficients can be recovered from the other normal equation in the partitioned regression:

$$\mathbf{D'Da} + \mathbf{D'Xb} = \mathbf{D'y}$$

or

$$\mathbf{a} = [\mathbf{D'D}]^{-1}\mathbf{D'}(\mathbf{y} - \mathbf{Xb}).$$

This implies that for each i,

$$a_i = \bar{y}_{i.} - \mathbf{b}'\bar{\mathbf{x}}_{i.}. \tag{13-6}$$

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

$$\text{Est.Asy. Var}[\mathbf{b}] = s^2[\mathbf{X'M_D X}]^{-1}, \tag{13-7}$$

which uses the second moment matrix with $\mathbf{x}$'s now expressed as deviations from their respective **group means.** The disturbance variance estimator is

$$s^2 = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it} - \mathbf{x}_{it}'\mathbf{b} - a_i)^2}{nT - n - K} = \frac{(\mathbf{y} - \mathbf{M_D Xb})'(\mathbf{y} - \mathbf{M_D Xb})}{(nT - n - K)}. \tag{13-8}$$

The itth residual used in this computation is

$$e_{it} = y_{it} - \mathbf{x}_{it}'\mathbf{b} - a_i = y_{it} - \mathbf{x}_{it}'\mathbf{b} - (\bar{y}_{i.} - \bar{\mathbf{x}}_{i.}'\mathbf{b}) = (y_{it} - \bar{y}_{i.}) - (\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})'\mathbf{b}.$$

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. For the individual effects,

$$\text{Asy. Var}[a_i] = \frac{\sigma^2}{T} + \bar{\mathbf{x}}_{i.}'\{\text{Asy. Var}[\mathbf{b}]\}\bar{\mathbf{x}}_{i.},$$

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

[7] 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 $nT/2$ first difference observations, by regressing observation $(y_{i2} - y_{i1})$ (and its negative) on $(\mathbf{x}_{i2} - \mathbf{x}_{i1})$ (and its negative).

13.3.1 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

$$F(n-1, nT-n-K) = \frac{\left(R^2_{LSDV} - R^2_{Pooled}\right)/(n-1)}{\left(1 - R^2_{LSDV}\right)/(nT-n-K)}, \tag{13-9}$$

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.[8]

13.3.2 THE WITHIN- AND BETWEEN-GROUPS ESTIMATORS

We can formulate a pooled regression model in three ways. First, the original formulation is

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha + \varepsilon_{it}. \tag{13-10a}$$

In terms of deviations from the group means,

$$y_{it} - \bar{y}_{i.} = (\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})'\boldsymbol{\beta} + \varepsilon_{it} - \bar{\varepsilon}_{i.}, \tag{13-10b}$$

while in terms of the group means,

$$\bar{y}_{i.} = \bar{\mathbf{x}}_{i.}'\boldsymbol{\beta} + \alpha + \bar{\varepsilon}_{i.}. \tag{13-10c}$$

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 (13-10c) involves 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 (13-10a), 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**,

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

For (13-10b), since 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)

[8] For a discussion of the differences, see Suits (1984).

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 (13-10c), 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;

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

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

$$\mathbf{b}^{total} = \left[\mathbf{S}_{xx}^{total}\right]^{-1}\mathbf{S}_{xy}^{total} = \left[\mathbf{S}_{xx}^{within} + \mathbf{S}_{xx}^{between}\right]^{-1}\left[\mathbf{S}_{xy}^{within} + \mathbf{S}_{xy}^{between}\right]. \tag{13-11}$$

The within-groups estimator is

$$\mathbf{b}^{within} = \left[\mathbf{S}_{xx}^{within}\right]^{-1}\mathbf{S}_{xy}^{within}. \tag{13-12}$$

This is the LSDV estimator computed earlier. [See (13-4).] An alternative estimator would be the between-groups estimator,

$$\mathbf{b}^{between} = \left[\mathbf{S}_{xx}^{between}\right]^{-1}\mathbf{S}_{xy}^{between} \tag{13-13}$$

(sometimes called the **group means estimator**). This least squares estimator of (13-10c) 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 (13-11), we see that the least squares estimator is a **matrix weighted average** of the within- and between-groups estimators:

$$\mathbf{b}^{total} = \mathbf{F}^{within}\mathbf{b}^{within} + \mathbf{F}^{between}\mathbf{b}^{between}, \tag{13-14}$$

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 Section 16.2. The resemblance is more than passing; it can be shown [see, e.g., Judge (1985)] that

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

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.

13.3.3 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

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \gamma_t + \varepsilon_{it}. \tag{13-15}$$

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.[9] There is an asymmetry in this formulation, however, since 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

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \mu + \alpha_i + \gamma_t + \varepsilon_{it}, \tag{13-15$'$}$$

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

$$\sum_i \alpha_i = \sum_t \gamma_t = 0$$

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

$$y_{*_{it}} = y_{it} - \bar{y}_{i.} - \bar{y}_{.t} + \bar{\bar{y}} \tag{13-16}$$

on

$$\mathbf{x}_{*_{it}} = \mathbf{x}_{it} - \bar{\mathbf{x}}_{i.} - \bar{\mathbf{x}}_{.t} + \bar{\bar{\mathbf{x}}},$$

where the period-specific and overall means are

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

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

$$\begin{aligned}
\hat{\mu} &= m = \bar{\bar{y}} - \bar{\bar{\mathbf{x}}}'\mathbf{b}, \\
\hat{\alpha}_i &= a_i = (\bar{y}_{i.} - \bar{\bar{y}}) - (\bar{\mathbf{x}}_{i.} - \bar{\bar{\mathbf{x}}})'\mathbf{b}, \\
\hat{\gamma}_t &= c_t = (\bar{y}_{.t} - \bar{\bar{y}}) - (\bar{\mathbf{x}}_{.t} - \bar{\bar{\mathbf{x}}})'\mathbf{b}.
\end{aligned} \tag{13-17}$$

[9]The matrix algebra and the theoretical development of two-way effects in panel data models are complex. See, for example, Baltagi (1995). Fortunately, the practical application is much simpler. The number of periods analyzed in most panel data sets is rarely more than a handful. Since modern computer programs, even those written strictly for microcomputers, uniformly allow dozens (or even hundreds) of regressors, almost any application involving a second fixed effect can be handled just by literally including the second effect as a set of actual dummy variables.

The estimated asymptotic covariance matrix for **b** is computed using the sums of squares and cross products of $\mathbf{x}_{*_{it}}$ computed in (13-16) and

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

If one of n or T is small and the other is large, then it may be simpler just to treat the smaller set as an ordinary set of variables and apply the previous results to the one-way fixed effects model defined by the larger set. Although more general, this model is infrequently used in practice. There are two reasons. First, the cost in terms of degrees of freedom is often not justified. Second, in those instances in which a model of the timewise evolution of the disturbance is desired, a more general model than this simple dummy variable formulation is usually used.

Example 13.2 Fixed Effects Regressions

Table 13.1 contains the estimated cost equations with individual firm effects, specific period effects, and both firm and period effects. For comparison, the least squares and group means results are given also. The F statistic for testing the joint significance of the firm effects is

$$F[5, 81] = \frac{(0.997434 - 0.98829)/5}{(1 - 0.997434)/81} = 57.614.$$

The critical value from the F table is 2.327, so the evidence is strongly in favor of a firm specific effect in the data. The same computation for the time effects, in the absence of the firm effects produces an $F[14, 72]$ statistic of 1.170, which is considerably less than the 95 percent critical value of 1.832. Thus, on this basis, there does not appear to be a significant cost difference across the different periods that is not accounted for by the fuel price variable, output, and load factors. There is a distinctive pattern to the time effects, which we will examine more closely later. In the presence of the firm effects, the $F[14, 67]$ ratio for the joint significance of the period effects is 3.149, which is larger than the table value of 1.842.

TABLE 13.1 Cost Equations with Fixed Firm and Period Effects

	Parameter Estimates							
Specification	β_1	β_2	β_3	β_4	R^2	s^2		
No effects	9.517 (0.22924)	0.88274 (0.013255)	0.45398 (0.020304)	−1.6275 (0.34530)	0.98829	0.015528		
Group means	85.809 (56.483)	0.78246 (0.10877)	−5.5240 (4.47879)	−1.7510 (2.74319)	0.99364	0.015838		
Firm effects		0.91928 (0.029890)	0.41749 (0.015199)	−1.07040 (0.20169)	0.99743	0.003625		
$a_1 \ldots a_6$:	9.706	9.665	9.497	9.891	9.730	9.793		
Time effects		0.86773 (0.015408)	−0.48448 (0.36411)	−1.95440 (0.44238)	0.99046	0.016705		
$c_1 \ldots c_8$	20.496	20.578	20.656	20.741	21.200	21.411	21.503	21.654
$c_9 \ldots c_{15}$	21.829	22.114	22.465	22.651	22.616	22.552	22.537	
Firm and time effects	12.667 (2.0811)	0.81725 (0.031851)	0.16861 (0.16348)	−0.88281 (0.26174)	0.99845	0.002727		
$a_1 \ldots a_6$	0.12833	0.06549	−0.18947	0.13425	−0.09265	−0.04596		
$c_1 \ldots c_8$	−0.37402	−0.31932	−0.27669	−0.22304	−0.15393	−0.10809	−0.07686	−0.02073
$c_9 \ldots c_{15}$	0.04722	0.09173	0.20731	0.28547	0.30138	0.30047	0.31911	

13.3.4 UNBALANCED PANELS AND FIXED EFFECTS

Missing data are very common in panel data sets. For this reason, or perhaps just because of the way the data were recorded, panels in which the group sizes differ across groups are not unusual. These panels are called **unbalanced panels.** The preceding analysis assumed equal group sizes and relied on the assumption at several points. A modification to allow unequal group sizes is quite simple. First, the full sample size is $\sum_{i=1}^{n} T_i$ instead of nT, which calls for minor modifications in the computations of s^2, Var[$\mathbf{b}$], Var[a_i], and the F statistic. Second, group means must be based on T_i, which varies across groups. The overall means for the regressors are

$$\bar{\bar{\mathbf{x}}} = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T_i}\mathbf{x}_{it}}{\sum_{i=1}^{n}T_i} = \frac{\sum_{i=1}^{n}T_i\bar{\mathbf{x}}_{i.}}{\sum_{i=1}^{n}T_i} = \sum_{i=1}^{n} f_i\bar{\mathbf{x}}_{i.},$$

where $f_i = T_i/(\sum_{i=1}^{n} T_i)$. If the group sizes are equal, then $f_i = 1/n$. The within groups moment matrix shown in (13-4),

$$\mathbf{S}_{xx}^{within} = \mathbf{X}'\mathbf{M}_{\mathbf{D}}\mathbf{X},$$

is

$$\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{M}_i^0\mathbf{X}_i = \sum_{i=1}^{n}\left(\sum_{t=1}^{T}(\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})(\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})'\right).$$

The other moments, $\mathbf{S}_{xy}^{within}$ and $\mathbf{S}_{yy}^{within}$, are computed likewise. No other changes are necessary for the one factor LSDV estimator. The two-way model can be handled likewise, although with unequal group sizes in both directions, the algebra becomes fairly cumbersome. Once again, however, the practice is much simpler than the theory. The easiest approach for unbalanced panels is just to create the full set of T dummy variables using as T the union of the dates represented in the full data set. One (presumably the last) is dropped, so we revert back to (13-15). Then, within each group, any of the T periods represented is accounted for by using one of the dummy variables. Least squares using the LSDV approach for the group effects will then automatically take care of the messy accounting details.

13.4 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 strictly 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. 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.[10] The payoff to this form is that it greatly reduces the number of parameters to be estimated. The cost is the possibility of inconsistent estimates, should the assumption turn out to be inappropriate.

Consider, then, a reformulation of the model

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + (\alpha + u_i) + \varepsilon_{it}, \tag{13-18}$$

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 13.2, $u_i = \{\mathbf{z}_i'\boldsymbol{\alpha} - E[\mathbf{z}_i'\boldsymbol{\alpha}]\}$. For example, in an analysis of families, we can view u_i as the collection of factors, $\mathbf{z}_i'\boldsymbol{\alpha}$, not in the regression that are specific to that family. We assume further that

$$\begin{aligned}
E[\varepsilon_{it} \mid \mathbf{X}] &= E[u_i \mid \mathbf{X}] = 0,\\
E\left[\varepsilon_{it}^2 \mid \mathbf{X}\right] &= \sigma_\varepsilon^2,\\
E\left[u_i^2 \mid \mathbf{X}\right] &= \sigma_u^2,\\
E[\varepsilon_{it} u_j \mid \mathbf{X}] &= 0 \quad \text{for all } i, t, \text{ and } j,\\
E[\varepsilon_{it}\varepsilon_{js} \mid \mathbf{X}] &= 0 \quad \text{if } t \neq s \text{ or } i \neq j,\\
E[u_i u_j \mid \mathbf{X}] &= 0 \quad \text{if } i \neq j.
\end{aligned} \tag{13-19}$$

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,

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

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

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_\varepsilon^2 + \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ \sigma_u^2 & \sigma_\varepsilon^2 + \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ & \cdots & & & \\ \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_\varepsilon^2 + \sigma_u^2 \end{bmatrix} = \sigma_\varepsilon^2 \mathbf{I}_T + \sigma_u^2 \mathbf{i}_T \mathbf{i}_T', \tag{13-20}$$

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

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

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} & \mathbf{0} & \cdots & \mathbf{0} \\ & & \vdots & & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma} \end{bmatrix} = \mathbf{I}_n \otimes \boldsymbol{\Sigma}. \tag{13-21}$$

13.4.1 GENERALIZED LEAST SQUARES

The generalized least squares estimator of the slope parameters is

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

To compute this estimator as we did in Chapter 10 by transforming the data and using ordinary least squares with the transformed data, we will require $\boldsymbol{\Omega}^{-1/2} = [\mathbf{I}_n \otimes \boldsymbol{\Sigma}]^{-1/2}$. We need only find $\boldsymbol{\Sigma}^{-1/2}$, which is

$$\boldsymbol{\Sigma}^{-1/2} = \frac{1}{\sigma_\varepsilon}\left[\mathbf{I} - \frac{\theta}{T}\mathbf{i}_T\mathbf{i}_T'\right],$$

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

$$\boldsymbol{\Sigma}^{-1/2}\mathbf{y}_i = \frac{1}{\sigma_\varepsilon}\begin{bmatrix} y_{i1} - \theta\bar{y}_{i.} \\ y_{i2} - \theta\bar{y}_{i.} \\ \vdots \\ y_{iT} - \theta\bar{y}_{i.} \end{bmatrix}, \tag{13-22}$$

and likewise for the rows of $\mathbf{X}_i$.[11] 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 $\theta = 1$. (One could interpret θ as the effect that would remain if σ_ε were zero, because the only effect would then be u_i. In this case, the fixed and random effects models would be indistinguishable, so this result makes sense.)

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

$$\hat{\beta} = \hat{\mathbf{F}}^{within}\mathbf{b}^{within} + (\mathbf{I} - \hat{\mathbf{F}}^{within})\mathbf{b}^{between},\text{[12]} \tag{13-23}$$

[11] This transformation is a special case of the more general treatment in Nerlove (1971b).

[12] An alternative form of this expression, in which the weighing matrices are proportional to the covariance matrices of the two estimators, is given by Judge et al. (1985).

where now,

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

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

To the extent that λ differs from one, we see that the inefficiency of 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 it all 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.

There are some polar cases to consider. If λ equals 1, then generalized least squares is identical to ordinary least squares. This situation would occur if σ_u^2 were zero, in which case a classical regression model would apply. If λ equals zero, then the estimator is the dummy variable estimator we used in the fixed effects setting. There are two possibilities. If σ_ε^2 were zero, then all variation across units would be due to the different u_is, which, because they are constant across time, would be equivalent to the dummy variables we used in the fixed-effects model. The question of whether they were fixed or random would then become moot. They are the only source of variation across units once the regression is accounted for. The other case is $T \to \infty$. We can view it this way: If $T \to \infty$, then the unobserved u_i becomes observable. Take the T observations for the ith unit. Our estimator of $[\alpha, \boldsymbol{\beta}]$ is consistent in the dimensions T or n. Therefore,

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

becomes observable. The individual means will provide

$$\bar{y}_{i.} - \bar{\mathbf{x}}_{i.}'\boldsymbol{\beta} - \alpha = u_i + \bar{\varepsilon}_{i.}$$

But $\bar{\varepsilon}_{i.}$ converges to zero, which reveals u_i to us. Therefore, if T goes to infinity, u_i becomes the $\alpha_i \mathbf{d}_i$ we used earlier.

Unbalanced panels add a layer of difficulty in the random effects model. The first problem can be seen in (13-21). The matrix $\boldsymbol{\Omega}$ is no longer $\mathbf{I} \otimes \boldsymbol{\Sigma}$ because the diagonal blocks in $\boldsymbol{\Omega}$ are of different sizes. There is also groupwise heteroscedasticity, because the ith diagonal block in $\boldsymbol{\Omega}^{-1/2}$ is

$$\boldsymbol{\Omega}_i^{-1/2} = \mathbf{I}_{T_i} - \frac{\theta_i}{T_i}\mathbf{i}_{T_i}\mathbf{i}_{T_i}', \qquad \theta_i = 1 - \frac{\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2 + T_i\sigma_u^2}}.$$

In principle, estimation is still straightforward, since 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 (13-22).

13.4.2 FEASIBLE GENERALIZED LEAST SQUARES WHEN Σ 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:

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha + \varepsilon_{it} + u_i \tag{13-24}$$

and

$$\bar{y}_{i.} = \bar{\mathbf{x}}_{i.}'\boldsymbol{\beta} + \alpha + \bar{\varepsilon}_{i.} + u_i.$$

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

$$y_{it} - \bar{y}_{i.} = [\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.}]'\boldsymbol{\beta} + [\varepsilon_{it} - \bar{\varepsilon}_{i.}]. \tag{13-25}$$

Since

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

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

$$\hat{\sigma}_\varepsilon^2(i) = \frac{\sum_{t=1}^{T}(\varepsilon_{it} - \bar{\varepsilon}_{i.})^2}{T-1}. \tag{13-26}$$

Since $\boldsymbol{\beta}$ must be estimated—(13-25) implies that the LSDV estimator is consistent, indeed, unbiased in general—we make the degrees of freedom correction and use the LSDV residuals in

$$s_e^2(i) = \frac{\sum_{t=1}^{T}(e_{it} - \bar{e}_{i.})^2}{T-K-1}. \tag{13-27}$$

We have n such estimators, so we average them to obtain

$$\bar{s}_e^2 = \frac{1}{n}\sum_{i=1}^{n} s_e^2(i) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\sum_{t=1}^{T}(e_{it} - \bar{e}_{i.})^2}{T-K-1}\right] = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(e_{it} - \bar{e}_{i.})^2}{nT-nK-n}. \tag{13-28}$$

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 $\bar{y}_{i.}$ and the K slopes. Therefore, we propose the unbiased estimator[13]

$$\hat{\sigma}_\varepsilon^2 = s_{LSDV}^2 = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(e_{it} - \bar{e}_{i.})^2}{nT-n-K}. \tag{13-29}$$

This is the variance estimator in the LSDV model in (13-8), appropriately corrected for degrees of freedom.

It remains to estimate σ_u^2. Return to the original model specification in (13-24). 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

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

model with only a single overall constant, we have

$$\text{plim}\, s^2_{Pooled} = \text{plim}\, \frac{\mathbf{e}'\mathbf{e}}{nT - K - 1} = \sigma^2_\varepsilon + \sigma^2_u. \qquad \textbf{(13-30)}$$

This provides the two estimators needed for the variance components; the second would be $\hat{\sigma}^2_u = s^2_{Pooled} - s^2_{LSDV}$. A possible complication is that this second estimator could be negative. But, recall that for feasible generalized least squares, we do not need an unbiased estimator of the variance, only a consistent one. As such, we may drop the degrees of freedom corrections in (13-29) and (13-30). If so, then the two variance estimators must be nonnegative, since the sum of squares in the LSDV model cannot be larger than that in the simple regression with only one constant term. Alternative estimators have been proposed, all based on this principle of using two different sums of squared residuals.[14]

There is a remaining complication. If there are any regressors that do not vary within the groups, the LSDV estimator cannot be computed. For example, in a model of family income or labor supply, one of the regressors might be a dummy variable for location, family structure, or living arrangement. Any of these could be perfectly collinear with the fixed effect for that family, which would prevent computation of the LSDV estimator. In this case, it is still possible to estimate the random effects variance components. Let $[\mathbf{b}, a]$ be any consistent estimator of $[\boldsymbol{\beta}, \alpha]$, such as the ordinary least squares estimator. Then, (13-30) provides a consistent estimator of $m_{ee} = \sigma^2_\varepsilon + \sigma^2_u$. The mean squared residuals using a regression based only on the n group means provides a consistent estimator of $m_{**} = \sigma^2_u + (\sigma^2_\varepsilon / T)$, so we can use

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

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

where $\omega > 1$. As before, this estimator can produce a negative estimate of σ^2_u that, once again, calls the specification of the model into question. [Note, finally, that the residuals in (13-29) and (13-30) could be based on the same coefficient vector.]

13.4.3 TESTING FOR RANDOM EFFECTS

Breusch and Pagan (1980) have devised a Lagrange multiplier test for the random effects model based on the OLS residuals.[15] For

$$H_0\colon \sigma^2_u = 0 \quad (\text{or Corr}[\eta_{it}, \eta_{is}] = 0),$$
$$H_1\colon \sigma^2_u \neq 0,$$

[14]See, for example, Wallace and Hussain (1969), Maddala (1971), Fuller and Battese (1974), and Amemiya (1971).

[15]We have focused thus far 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 Balestra and Nerlove (1966, 2003) for this approach to estimation.

the test statistic is

$$\mathrm{LM} = \frac{nT}{2(T-1)}\left[\frac{\sum_{i=1}^{n}\left[\sum_{t=1}^{T} e_{it}\right]^2}{\sum_{i=1}^{n}\sum_{t=1}^{T} e_{it}^2} - 1\right]^2 = \frac{nT}{2(T-1)}\left[\frac{\sum_{i=1}^{n}(T\bar{e}_{i.})^2}{\sum_{i=1}^{n}\sum_{t=1}^{T} e_{it}^2} - 1\right]^2. \quad \textbf{(13-31)}$$

Under the null hypothesis, LM is distributed as chi-squared with one degree of freedom.

Example 13.3 Testing for Random Effects

The least squares estimates for the cost equation were given in Example 13.1. The firm specific means of the least squares residuals are

$$\bar{\mathbf{e}} = [0.068869, -0.013878, -0.19422, 0.15273, -0.021583, 0.0080906]'$$

The total sum of squared residuals for the least squares regression is $\mathbf{e}'\mathbf{e} = 1.33544$, so

$$\mathrm{LM} = \frac{nT}{2(T-1)}\left[\frac{T^2\bar{\mathbf{e}}'\bar{\mathbf{e}}}{\mathbf{e}'\mathbf{e}} - 1\right]^2 = 334.85.$$

Based on the least squares residuals, we obtain a Lagrange multiplier test statistic of 334.85, which far exceeds the 95 percent critical value for chi-squared with one degree of freedom, 3.84. At this point, we conclude that the classical regression model with a single constant 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 our earlier results for FGLS estimators apply here. It would also be possible to obtain the maximum likelihood estimator.[16] The likelihood function is complicated, but as we have seen repeatedly, the MLE of $\boldsymbol{\beta}$ will be GLS based on the maximum likelihood estimators of the variance components. It can be shown that the MLEs of σ_ε^2 and σ_u^2 are the unbiased estimators shown earlier, *without* the degrees of freedom corrections.[17] This model satisfies the requirements for the Oberhofer–Kmenta (1974) algorithm—see Section 11.7.2—so we could also use the iterated FGLS procedure to obtain the MLEs if desired. The initial consistent estimators can be based on least squares residuals. Still other estimators have been proposed. None will have better asymptotic properties than the MLE or FGLS estimators, but they may outperform them in a finite sample.[18]

Example 13.4 Random Effects Models

To compute the FGLS estimator, we require estimates of the variance components. The unbiased estimator of σ_ε^2 is the residual variance estimator in the within-units (LSDV) regression. Thus,

$$\hat{\sigma}_\varepsilon^2 = \frac{0.2926222}{90-9} = 0.0036126.$$

[16]See Hsiao (1986) and Nerlove (2003).

[17]See Berzeg (1979).

[18]See Maddala and Mount (1973).

Using the least squares residuals from the pooled regression we have

$$\widehat{\sigma_\varepsilon^2 + \sigma_u^2} = \frac{1.335442}{90 - 4} = 0.015528$$

so

$$\widehat{\sigma_u^2} = 0.015528 - 0.0036126 = 0.0199158.$$

For purposes of FGLS,

$$\hat{\theta} = 1 - \left[\frac{0.0036126}{15(0.0199158)}\right]^{1/2} = 0.890032.$$

The FGLS estimates for this random effects model are shown in Table 13.2, with the fixed effects estimates. The estimated within-groups variance is larger than the between-groups variance by a factor of five. Thus, by these estimates, over 80 percent of the disturbance variation is explained by variation within the groups, with only the small remainder explained by variation across groups.

None of the desirable properties of the estimators in the random effects model rely on T going to infinity.[19] Indeed, T is likely to be quite small. The maximum likelihood estimator of σ_ε^2 is exactly equal to an average of n estimators, each based on the T observations for unit i. [See (13-28).] Each component in this average is, in principle, consistent. That is, its variance is of order $1/T$ or smaller. Since 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 assume that T increases without bound or, in some cases, at all.[20] 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.

The LSDV model *does* rely on T increasing for consistency. To see this, we use the partitioned regression. The slopes are

$$\mathbf{b} = [\mathbf{X}'\mathbf{M_D}\mathbf{X}]^{-1}[\mathbf{X}'\mathbf{M_D}\mathbf{y}].$$

Since $\mathbf{X}$ is $nT \times K$, as long as the inverted moment matrix converges to a zero matrix, $\mathbf{b}$ is consistent as long as either n or T increases without bound. But the dummy variable coefficients are

$$a_i = \bar{y}_{i.} - \bar{\mathbf{x}}_{i.}'\mathbf{b} = \frac{1}{T}\sum_{t=1}^{T}(y_{it} - \mathbf{x}_{it}'\mathbf{b}).$$

We have already seen that $\mathbf{b}$ is consistent. Suppose, for the present, that $\bar{\mathbf{x}}_{i.} = 0$. Then $\text{Var}[a_i] = \text{Var}[y_{it}]/T$. Therefore, unless $T \to \infty$, the estimators of the unit-specific effects are not consistent. (They are, however, best linear unbiased.) This inconsistency is worth bearing in mind when analyzing data sets for which T is fixed and there is no intention

[19]See Nickell (1981).

[20]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).

to replicate the study and no logical argument that would justify the claim that it could have been replicated in principle.

The random effects model was developed by Balestra and Nerlove (1966). Their formulation included a time-specific component, κ_t, as well as the individual effect:

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

The extended formulation is rather complicated analytically. In Balestra and Nerlove's study, it was made even more so by the presence of a lagged dependent variable that causes all the problems discussed earlier in our discussion of autocorrelation. A full set of results for this extended model, including a method for handling the lagged dependent variable, has been developed.[21] We will turn to this in Section 13.7.

13.4.4 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.[22]

The specification test devised by Hausman (1978)[23] is used to test for orthogonality of the random effects and the regressors. The test is based on the idea that under the hypothesis of no correlation, both OLS in the LSDV model and GLS are consistent, but OLS is inefficient,[24] whereas under the alternative, OLS is consistent, but GLS 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, $[\mathbf{b} - \hat{\boldsymbol{\beta}}]$:

$$\text{Var}[\mathbf{b} - \hat{\boldsymbol{\beta}}] = \text{Var}[\mathbf{b}] + \text{Var}[\hat{\boldsymbol{\beta}}] - \text{Cov}[\mathbf{b}, \hat{\boldsymbol{\beta}}] - \text{Cov}[\mathbf{b}, \hat{\boldsymbol{\beta}}]. \quad \textbf{(13-32)}$$

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

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

or that

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

Inserting this result in (13-32) produces the required covariance matrix for the test,

$$\text{Var}[\mathbf{b} - \hat{\boldsymbol{\beta}}] = \text{Var}[\mathbf{b}] - \text{Var}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\Psi}. \quad \textbf{(13-33)}$$

[21] See Balestra and Nerlove (1966), Fomby, Hill, and Johnson (1984), Judge et al. (1985), Hsiao (1986), Anderson and Hsiao (1982), Nerlove (1971a, 2003), and Baltagi (1995).

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

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

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

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

$$W = \chi^2[K-1] = [\mathbf{b} - \hat{\boldsymbol{\beta}}]'\hat{\boldsymbol{\Psi}}^{-1}[\mathbf{b} - \hat{\boldsymbol{\beta}}]. \qquad \textbf{(13-34)}$$

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.

Example 13.5 Hausman Test

The Hausman test for the fixed and random effects regressions is based on the parts of the coefficient vectors and the asymptotic covariance matrices that correspond to the slopes in the models, that is, ignoring the constant term(s). The coefficient estimates are given in Table 13.2. The two estimated asymptotic covariance matrices are

$$\text{Est. Var}[\mathbf{b}_{FE}] = \begin{bmatrix} 0.0008934 & -0.0003178 & -0.001884 \\ -0.0003178 & 0.0002310 & -0.0007686 \\ -0.001884 & -0.0007686 & 0.04068 \end{bmatrix}$$

TABLE 13.2 Random and Fixed Effects Estimates

Specification	*Parameter Estimates* β_1	β_2	β_3	β_4	R^2	s^2
No effects	9.517	0.88274	0.45398	−1.6275	0.98829	0.015528
	(0.22924)	(0.013255)	(0.020304)	(0.34530)		
Firm effects	Fixed effects					
		0.91930	0.41749	−1.0704	0.99743	0.0036125
		(0.029890)	(0.015199)	(0.20169)		
	White(1)	(0.019105)	(0.013533)	(0.21662)		
	White(2)	(0.027977)	(0.013802)	(0.20372)		
	Fixed effects with autocorrelation $\hat{\rho} = 0.5162$					
		0.92975	0.38567	−1.22074		0.0019179
		(0.033927)	(0.0167409)	(0.20174)	$s^2/(1-\hat{\rho}^2) =$ 0.002807	
	Random effects					
	9.6106	0.90412	0.42390	−1.0646	$\hat{\sigma}_u^2 = 0.0119158$	
	(0.20277)	(0.02462)	(0.01375)	(0.1993)	$\hat{\sigma}_\varepsilon^2 = 0.00361262$	
	Random effects with autocorrelation $\hat{\rho} = 0.5162$					
	10.139	0.91269	0.39123	−1.2074	$\hat{\sigma}_u^2 = 0.0268079$	
	(0.2587)	(0.027783)	(0.016294)	(0.19852)	$\hat{\sigma}_\varepsilon^2 = 0.0037341$	
Firm and time effects	Fixed effects					
	12.667	0.81725	0.16861	−0.88281	0.99845	0.0026727
	(2.0811)	(0.031851)	(0.16348)	(0.26174)		
	Random effects					
	9.799	0.84328	0.38760	−0.92943	$\hat{\sigma}_u^2 = 0.0142291$	
	(0.87910)	(0.025839)	(0.06845)	(0.25721)	$\hat{\sigma}_\varepsilon^2 = 0.0026395$	
					$\hat{\sigma}_v^2 = 0.0551958$	

and

$$\text{Est. Var}[\mathbf{b}_{RE}] = \begin{bmatrix} 0.0006059 & -0.0002089 & -0.001450 \\ -0.0002089 & 0.00018897 & -0.002141 \\ -0.001450 & -0.002141 & 0.03973 \end{bmatrix}.$$

The test statistic is 4.16. The critical value from the chi-squared table with three degrees of freedom is 7.814, which is far larger than the test value. The hypothesis that the individual effects are uncorrelated with the other regressors in the model cannot be rejected. Based on the LM test, which is decisive that there are individual effects, and the Hausman test, which suggests that these effects are uncorrelated with the other variables in the model, we would conclude that of the two alternatives we have considered, the random effects model is the better choice.

13.5 INSTRUMENTAL VARIABLES ESTIMATION OF THE RANDOM EFFECTS MODEL

Recall the original specification of the linear model for panel data in (13-1)

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\alpha} + \varepsilon_{it}. \tag{13-35}$$

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:

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

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 (13-35) 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

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

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

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

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

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

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

$$y_{it} - \bar{y}_{i.} = (\mathbf{x}_{1it} - \bar{\mathbf{x}}_{1i})'\boldsymbol{\beta}_1 + (\mathbf{x}_{2it} - \bar{x}_{2i})'\boldsymbol{\beta}_2 + \varepsilon_{it} - \bar{\varepsilon}_i, \tag{13-36}$$

which implies that $\boldsymbol{\beta}$ 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 $\mathbf{x}_{2it}$. Now, 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 (13-36). Since $\mathbf{z}_1$ 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, $e_{it}^* = \bar{e}_{i.}, t = 1, \ldots, T, i = 1, \ldots, n$. 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

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

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,

$$\mathbf{w}_{it}^{*\prime} = \mathbf{w}_{it}' - (1-\hat{\theta})\bar{\mathbf{w}}_i' \quad \text{and} \quad y_{it}^* = y_{it} - (1-\hat{\theta})\bar{y}_i$$

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 $\hat{\theta}$. The instrumental variables are

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

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

$$(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\alpha}}')_{\text{IV}}' = [(\mathbf{W}^{*\prime}\mathbf{V})(\mathbf{V}'\mathbf{V})^{-1}(\mathbf{V}'\mathbf{W}^*)]^{-1}[(\mathbf{W}^{*\prime}\mathbf{V})(\mathbf{V}'\mathbf{V})^{-1}(\mathbf{V}'\mathbf{y}^*)].^{25} \qquad \textbf{(13-37)}$$

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.

Example 13.6 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 25–55, 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,
Bad Health = a dummy variable indicating general health,
Race = a dummy 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. [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?)]

The primary focus of the study is the coefficient on schooling in the log wage equation. Since schooling and, probably, Experience and Unemployed are correlated with the latent

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

TABLE 13.3 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
	$\rho = \sqrt{\sigma_u^2/(\sigma_u^2+\sigma_\varepsilon^2)}$		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.

effect, there is likely to be serious bias in conventional estimates of this equation. Table 13.3 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 0.067, 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 13.4.4, result (13-34). Since 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 percent 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 13.3 present the instrumental variable estimates. The specification test given with the fourth set of results suggests that the procedure has produced the desired 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 $\mathbf{z}_2$ rather than $\mathbf{z}_1$. 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, a difference of over 200 percent. As the authors note, at the same time, the measured effect of race nearly vanishes.

13.6 GMM ESTIMATION OF DYNAMIC PANEL DATA MODELS

Panel data are well suited for examining dynamic effects, as in the first-order model,

$$\begin{aligned} y_{it} &= \mathbf{x}_{it}'\boldsymbol{\beta} + \gamma y_{i,t-1} + \alpha_i + \varepsilon_{it} \\ &= \mathbf{w}_{it}'\boldsymbol{\delta} + \alpha_i + \varepsilon_{it}, \end{aligned}$$

where the set of right hand side variables, $\mathbf{w}_{it}$ now includes the lagged dependent variable, $y_{i,t-1}$. Adding dynamics to a model in this fashion is a major change in the interpretation of the equation. Without the lagged variable, the "independent variables" represent the full set of information that produce observed outcome y_{it}. With the lagged variable, we now have in the equation, the entire history of the right hand side variables, so that any measured influence is conditioned on this history; in this case, any impact of $\mathbf{x}_{it}$ represents the effect of *new* information. Substantial complications arise in estimation of such a model. In both the fixed and random effects settings, the difficulty is that the lagged dependent variable is correlated with the disturbance, even if it is assumed that ε_{it} is not itself autocorrelated. For the moment, consider the fixed effects model as an ordinary regression with a lagged dependent variable. We considered this case in Section 5.3.2 as a regression with a stochastic regressor that is dependent across observations. In that dynamic regression model, the estimator based on T observations is biased in finite samples, but it is consistent in T. That conclusion was the main result of Section 5.3.2. The finite sample bias is of order $1/T$. The same result applies here, but the difference is that whereas before we obtained our large sample results by allowing T to grow large, in this setting, T is assumed to be small and fixed, and large-sample results are obtained with respect to n growing large, not T. The fixed effects estimator of $\boldsymbol{\delta} = [\boldsymbol{\beta}, \gamma]$ can be viewed as an average of n such estimators. Assume for now that $T \geq K + 1$ where K is the number of variables in $\mathbf{x}_{it}$. Then, from (13-4),

$$\begin{aligned} \hat{\boldsymbol{\delta}} &= \left[\sum_{i=1}^{n} \mathbf{W}_i'\mathbf{M}^0\mathbf{W}_i\right]^{-1} \left[\sum_{i=1}^{n} \mathbf{W}_i'\mathbf{M}^0\mathbf{y}_i\right] \\ &= \left[\sum_{i=1}^{n} \mathbf{W}_i'\mathbf{M}^0\mathbf{W}_i\right]^{-1} \left[\sum_{i=1}^{n} \mathbf{W}_i'\mathbf{M}^0\mathbf{W}_i\mathbf{d}_i\right] \\ &= \sum_{i=1}^{n} \mathbf{F}_i\mathbf{d}_i \end{aligned}$$

where the rows of the $T \times (K+1)$ matrix $\mathbf{W}_i$ are $\mathbf{w}_{it}'$ and $\mathbf{M}^0$ is the $T \times T$ matrix that creates deviations from group means [see (13-5)]. Each group specific estimator, $\mathbf{d}_i$ is inconsistent, as it is biased in finite samples and its variance does not go to zero as n increases. This matrix weighted average of n inconsistent estimators will also be inconsistent. (This analysis is only heuristic. If $T < K + 1$, then the individual coefficient vectors cannot be computed.[26])

[26]Further discussion is given by Nickell (1981), Ridder and Wansbeek (1990), and Kiviet (1995).

The problem is more transparent in the random effects model. In the model

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

the lagged dependent variable is correlated with the compound disturbance in the model, since the same u_i enters the equation for every observation in group i.

Neither of these results renders the model inestimable, but they do make necessary some technique other than our familiar LSDV or FGLS estimators. The general approach, which has been developed in several stages in the literature,[27] relies on instrumental variables estimators and, most recently [by **Arellano and Bond** (1991) and **Arellano and Bover** (1995)] on a **GMM estimator.** For example, in either the fixed or random effects cases, the heterogeneity can be swept from the model by taking first differences, which produces

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

This model is still complicated by correlation between the lagged dependent variable and the disturbance (and by its first-order moving average disturbance). But without the group effects, there is a simple instrumental variables estimator available. Assuming that the time series is long enough, one could use the lagged differences, $(y_{i,t-2} - y_{i,t-3})$, or the lagged levels, $y_{i,t-2}$ and $y_{i,t-3}$, as one or two instrumental variables for $(y_{i,t-1} - y_{i,t-2})$. (The other variables can serve as their own instruments.) By this construction, then, the treatment of this model is a standard application of the instrumental variables technique that we developed in Section 5.4.[28] This illustrates the flavor of an instrumental variable approach to estimation. But, as Arellano et al. and Ahn and Schmidt (1995) have shown, there is still more information in the sample which can be brought to bear on estimation, in the context of a GMM estimator, which we now consider.

We extend the Hausman and Taylor (HT) formulation of the random effects model to include the lagged dependent variable;

$$\begin{aligned} y_{it} &= \gamma y_{i,t-1} + \mathbf{x}_{1it}'\boldsymbol{\beta}_1 + \mathbf{x}_{2it}'\boldsymbol{\beta}_2 + \mathbf{z}_{1i}'\boldsymbol{\alpha}_1 + \mathbf{z}_{2i}'\boldsymbol{\alpha}_2 + \varepsilon_{it} + u_i \\ &= \boldsymbol{\delta}'\mathbf{w}_{it} + \varepsilon_{it} + u_i \\ &= \boldsymbol{\delta}'\mathbf{w}_{it} + \eta_{it} \end{aligned}$$

where

$$\mathbf{w}_{it} = [y_{i,t-1}, \mathbf{x}_{1it}', \mathbf{x}_{2it}', \mathbf{z}_{1i}', \mathbf{z}_{2i}']'$$

is now a $(1 + K_1 + K_2 + L_1 + L_2) \times 1$ vector. The terms in the equation are the same as in the Hausman and Taylor model. Instrumental variables estimation of the model without the lagged dependent variable is discussed in the previous section on the HT estimator. Moreover, by just including $y_{i,t-1}$ in $\mathbf{x}_{2it}$, we see that the HT approach extends to this setting as well, essentially without modification. Arellano et al. suggest a GMM estimator, and show that efficiency gains are available by using a larger set of moment

[27]The model was first proposed in this form by Balestra and Nerlove (1966). See, for example, Anderson and Hsiao (1981, 1982), Bhargava and Sargan (1983), Arellano (1989), Arellano and Bond (1991), Arellano and Bover (1995), Ahn and Schmidt (1995), and Nerlove (2003).

[28]There is a question as to whether one should use differences or levels as instruments. Arellano (1989) gives evidence that the latter is preferable.

conditions. In the previous treatment, we used a GMM estimator constructed as follows: The set of moment conditions we used to formulate the instrumental variables were

$$E\left[\begin{pmatrix}\mathbf{x}_{1it}\\ \mathbf{x}_{2it}\\ \mathbf{z}_{1i}\\ \bar{\mathbf{x}}_{1i.}\end{pmatrix}(\eta_{it}-\bar{\eta}_i)\right]=E\left[\begin{pmatrix}\mathbf{x}_{1it}\\ \mathbf{x}_{2it}\\ \mathbf{z}_{1i}\\ \bar{\mathbf{x}}_{1i.}\end{pmatrix}(\varepsilon_{it}-\bar{\varepsilon}_i)\right]=\mathbf{0}.$$

This moment condition is used to produce the instrumental variable estimator. We could ignore the nonscalar variance of η_{it} and use simple instrumental variables at this point. However, by accounting for the random effects formulation and using the counterpart to feasible GLS, we obtain the more efficient estimator in (13-37). As usual, this can be done in two steps. The inefficient estimator is computed in order to obtain the residuals needed to estimate the variance components. This is Hausman and Taylor's steps 1 and 2. Steps 3 and 4 are the GMM estimator based on these estimated variance components.

Arellano et al. suggest that the preceding does not exploit all the information in the sample. In simple terms, within the T observations in group i, we have not used the fact that

$$E\left[\begin{pmatrix}\mathbf{x}_{1it}\\ \mathbf{x}_{2it}\\ \mathbf{z}_{1i}\\ \bar{\mathbf{x}}_{1i.}\end{pmatrix}(\eta_{is}-\bar{\eta}_i)\right]=\mathbf{0}\text{ for some } s\neq t.$$

Thus, for example, not only are disturbances at time t uncorrelated with these variables at time t, arguably, they are uncorrelated with the same variables at time $t-1, t-2$, possibly $t+1$, and so on. In principle, the number of valid instruments is potentially enormous. Suppose, for example, that the set of instruments listed above is strictly exogenous with respect to η_{it} in every period including current, lagged and future. Then, there are a total of $[T(K_1+K_2)+L_1+K_1)]$ moment conditions for every observation on this basis alone. Consider, for example, a panel with two periods. We would have for the two periods,

$$E\left[\begin{pmatrix}\mathbf{x}_{1i1}\\ \mathbf{x}_{2i1}\\ \mathbf{x}_{1i2}\\ \mathbf{x}_{2i2}\\ \mathbf{z}_{1i}\\ \bar{\mathbf{x}}_{1i.}\end{pmatrix}(\eta_{i1}-\bar{\eta}_i)\right]=E\left[\begin{pmatrix}\mathbf{x}_{1i1}\\ \mathbf{x}_{2i1}\\ \mathbf{x}_{1i2}\\ \mathbf{x}_{2i2}\\ \mathbf{z}_{1i}\\ \bar{\mathbf{x}}_{1i.}\end{pmatrix}(\eta_{i2}-\bar{\eta}_i)\right]=\mathbf{0}. \qquad \textbf{(13-38)}$$

How much useful information is brought to bear on estimation of the parameters is uncertain, as it depends on the correlation of the instruments with the included exogenous variables in the equation. The farther apart in time these sets of variables become the less information is likely to be present. (The literature on this subject contains reference to "strong" versus "weak" instrumental variables.[29]) In order to proceed, as noted, we can include the lagged dependent variable in $\mathbf{x}_{2i}$. This set of instrumental variables can be used to construct the estimator, actually whether the lagged variable is present or not. We note, at this point, that on this basis, Hausman and Taylor's estimator did not

[29]See West (2001).

actually use all the information available in the sample. We now have the elements of the Arellano et al. estimator in hand; what remains is essentially the (unfortunately, fairly involved) algebra, which we now develop.

Let

$$\mathbf{W}_i = \begin{bmatrix} \mathbf{w}'_{i1} \\ \mathbf{w}'_{i2} \\ \vdots \\ \mathbf{w}'_{iT_i} \end{bmatrix} = \text{the full set of rhs data for group } i, \quad \text{and} \quad \mathbf{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{bmatrix}.$$

Note that $\mathbf{W}_i$ is assumed to be, a $T \times (1 + K_1 + K_2 + L_1 + L_2)$ matrix. Since there is a lagged dependent variable in the model, it must be assumed that there are actually $T+1$ observations available on y_{it}. To avoid a cumbersome, cluttered notation, we will leave this distinction embedded in the notation for the moment. Later, when necessary, we will make it explicit. It will reappear in the formulation of the instrumental variables. A total of T observations will be available for constructing the IV estimators. We now form a matrix of instrumental variables. Different approaches to this have been considered by Hausman and Taylor (1981), Arellano et al. (1991, 1995, 1999), Ahn and Schmidt (1995) and Amemiya and MaCurdy (1986), among others. We will form a matrix $\mathbf{V}_i$ consisting of $T_i - 1$ rows constructed the same way for $T_i - 1$ observations and a final row that will be different, as discussed below. [This is to exploit a useful algebraic result discussed by Arellano and Bover (1995).] The matrix will be of the form

$$\mathbf{V}_i = \begin{bmatrix} \mathbf{v}'_{i1} & \mathbf{0}' & \cdots & \mathbf{0}' \\ \mathbf{0}' & \mathbf{v}'_{i2} & \cdots & \mathbf{0}' \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}' & \mathbf{0}' & \cdots & \mathbf{a}'_i \end{bmatrix}. \qquad \textbf{(13-39)}$$

The instrumental variable sets contained in $\mathbf{v}'_{it}$ which have been suggested might include the following from within the model:

$\mathbf{x}_{it}$ and $\mathbf{x}_{i,t-1}$ (i.e., current and one lag of all the time varying variables)
$\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}$ (i.e., all current, past and future values of all the time varying variables)
$\mathbf{x}_{i1}, \ldots, \mathbf{x}_{it}$ (i.e., all current and past values of all the time varying variables)

The time invariant variables that are uncorrelated with u_i, that is $\mathbf{z}_{1i}$, are appended at the end of the nonzero part of each of the first $T - 1$ rows. It may seem that including $\mathbf{x}_2$ in the instruments would be invalid. However, we will be converting the disturbances to deviations from group means which are free of the latent effects—that is, this set of moment conditions will ultimately be converted to what appears in (13-38). While the variables are correlated with u_i by construction, they are not correlated with $\varepsilon_{it} - \bar{\varepsilon}_i$. The final row of $\mathbf{V}_i$ is important to the construction. Two possibilities have been suggested:

$\mathbf{a}'_i = [\mathbf{z}'_{1i} \quad \bar{\mathbf{x}}_{i1}]$ (produces the Hausman and Taylor estimator)

$\mathbf{a}'_i = [\mathbf{z}'_{1i} \quad \mathbf{x}'_{1i1}, \mathbf{x}'_{1i2}, \ldots, \mathbf{x}_{1iT}]$ (produces Amemiya and MaCurdy's estimator).

Note that the **m** variables are exogenous time invariant variables, $\mathbf{z}_{1i}$ and the exogenous time varying variables, either condensed into the single group mean or in the raw form, with the full set of T observations.

To construct the estimator, we will require a transformation matrix, **H** constructed as follows. Let $\mathbf{M}^{01}$ denote the first $T-1$ rows of $\mathbf{M}^0$, the matrix that creates deviations from group means. Then,

$$\mathbf{H} = \begin{bmatrix} \mathbf{M}^{01} \\ \frac{1}{T}\mathbf{i}'_T \end{bmatrix}.$$

Thus, **H** replaces the last row of $\mathbf{M}^0$ with a row of $1/T$. The effect is as follows: if **q** is T observations on a variable, then $\mathbf{Hq}$ produces $\mathbf{q}^*$ in which the first $T-1$ observations are converted to deviations from group means and the last observation is the group mean. In particular, let the $T \times 1$ column vector of disturbances

$$\boldsymbol{\eta}_i = [\eta_{i1}, \eta_{i2}, \ldots, \eta_{iT}] = [(\varepsilon_{i1} + u_i), (\varepsilon_{i2} + u_i), \ldots, (\varepsilon_{iT} + u_i)]',$$

then

$$\mathbf{H}\eta = \begin{bmatrix} \eta_{i1} - \bar{\eta}_i \\ \vdots \\ \eta_{i,T-1} - \bar{\eta}_i \\ \bar{\eta}_i \end{bmatrix}.$$

We can now construct the moment conditions. With all this machinery in place, we have the result that appears in (13-40), that is

$$E[\mathbf{V}_i'\mathbf{H}\boldsymbol{\eta}_i] = E[\mathbf{g}_i] = \mathbf{0}.$$

It is useful to expand this for a particular case. Suppose $T = 3$ and we use as instruments the current values in Period 1, and the current and previous values in Period 2 and the Hausman and Taylor form for the invariant variables. Then the preceding is

$$E\left[\begin{pmatrix} \mathbf{x}_{1i1} & \mathbf{0} & \mathbf{0} \\ \mathbf{x}_{2i1} & \mathbf{0} & \mathbf{0} \\ \mathbf{z}_{1i} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{1i1} & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{2i1} & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{1i2} & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{2i2} & \mathbf{0} \\ \mathbf{0} & \mathbf{z}_{1i} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{z}_{1i} \\ \mathbf{0} & \mathbf{0} & \bar{\mathbf{x}}_{1i} \end{pmatrix} \begin{pmatrix} \eta_{i1} - \bar{\eta}_i \\ \eta_{i2} - \bar{\eta}_i \\ \bar{\eta}_i \end{pmatrix}\right] = \mathbf{0}. \qquad \textbf{(13-40)}$$

This is the same as (13-38).[30] The empirical moment condition that follows from this is

$$\text{plim}\frac{1}{n}\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}\boldsymbol{\eta}_i$$

$$= \text{plim}\frac{1}{n}\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}\begin{pmatrix} y_{i1} - \gamma y_{i0} - \mathbf{x}_{1i1}'\boldsymbol{\beta}_1 - \mathbf{x}_{2i1}'\boldsymbol{\beta}_2 - \mathbf{z}_{1i}'\boldsymbol{\alpha}_1 - \mathbf{z}_{2i}'\boldsymbol{\alpha}_2 \\ y_{i2} - \gamma y_{i1} - \mathbf{x}_{1i2}'\boldsymbol{\beta}_1 - \mathbf{x}_{2i2}'\boldsymbol{\beta}_2 - \mathbf{z}_{1i}'\boldsymbol{\alpha}_1 - \mathbf{z}_{2i}'\boldsymbol{\alpha}_2 \\ \vdots \\ y_{iT} - \gamma y_{i,T-1} - \mathbf{x}_{1iT}'\boldsymbol{\beta}_1 - \mathbf{x}_{2iT}'\boldsymbol{\beta}_2 - \mathbf{z}_{1i}'\boldsymbol{\alpha}_1 - \mathbf{z}_{2i}'\boldsymbol{\alpha}_2 \end{pmatrix} = \mathbf{0}.$$

Write this as

$$\text{plim}\frac{1}{n}\sum_{i=1}^{n}\mathbf{m}_i = \text{plim}\;\bar{\mathbf{m}} = \mathbf{0}.$$

The GMM estimator $\hat{\boldsymbol{\delta}}$ is then obtained by minimizing

$$q = \bar{\mathbf{m}}'\mathbf{A}\bar{\mathbf{m}}$$

with an appropriate choice of the weighting matrix, $\mathbf{A}$. The optimal weighting matrix will be the inverse of the asymptotic covariance matrix of $\sqrt{n}\,\bar{\mathbf{m}}$. With a consistent estimator of $\boldsymbol{\delta}$ in hand, this can be estimated empirically using

$$\text{Est.Asy. Var}[\sqrt{n}\,\bar{\mathbf{m}}] = \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{m}}_i\hat{\mathbf{m}}_i' = \frac{1}{n}\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}\hat{\boldsymbol{\eta}}_i\hat{\boldsymbol{\eta}}_i'\mathbf{H}'\mathbf{V}_i.$$

This is a robust estimator that allows an unrestricted $T \times T$ covariance matrix for the T disturbances, $\varepsilon_{it} + u_i$. But, we have assumed that this covariance matrix is the $\boldsymbol{\Sigma}$ defined in (13-20) for the random effects model. To use this information we would, instead, use the residuals in

$$\hat{\boldsymbol{\eta}}_i = \mathbf{y}_i - \mathbf{W}_i\hat{\boldsymbol{\delta}}$$

to estimate σ_u^2 and σ_ε^2 and then $\boldsymbol{\Sigma}$, which produces

$$\text{Est.Asy. Var}[\sqrt{n}\,\bar{\mathbf{m}}] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}\hat{\boldsymbol{\Sigma}}\mathbf{H}'\mathbf{V}_i.$$

We now have the full set of results needed to compute the GMM estimator. The solution to the optimization problem of minimizing q with respect to the parameter vector $\boldsymbol{\delta}$ is

$$\hat{\boldsymbol{\delta}}_{GMM} = \left[\left(\sum_{i=1}^{n}\mathbf{W}_i'\mathbf{H}\mathbf{V}_i\right)\left(\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}'\hat{\boldsymbol{\Sigma}}\mathbf{H}\mathbf{V}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}'\mathbf{W}_i\right)\right]^{-1}$$
$$\times\left(\sum_{i=1}^{n}\mathbf{W}_i'\mathbf{H}\mathbf{V}_i\right)\left(\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}'\hat{\boldsymbol{\Sigma}}\mathbf{H}\mathbf{V}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}'\mathbf{y}_i\right). \qquad \textbf{(13-41)}$$

The estimator of the asymptotic covariance matrix for $\hat{\boldsymbol{\delta}}$ is the inverse matrix in brackets.

[30] In some treatments [e.g., Blundell and Bond (1998)], an additional condition is assumed for the initial value, y_{i0}, namely $E[y_{i0}\,|\,\text{exogenous data}] = \mu_0$. This would add a row at the top of the matrix in (13-38) containing $[(y_{i0} - \mu_0), 0, 0]$.

The remaining loose end is how to obtain the consistent estimator of $\boldsymbol{\delta}$ to compute $\boldsymbol{\Sigma}$. Recall that the GMM estimator is consistent with any positive definite weighting matrix, $\mathbf{A}$ in our expression above. Therefore, for an initial estimator, we could set $\mathbf{A} = \mathbf{I}$ and use the simple instrumental variables estimator,

$$\hat{\boldsymbol{\delta}}_{\text{IV}} = \left[\left(\sum_{i=1}^{N} \mathbf{W}_i'\mathbf{H}\mathbf{V}_i\right)\left(\sum_{i=1}^{N} \mathbf{V}_i'\mathbf{H}\mathbf{W}_i\right)\right]^{-1}\left(\sum_{i=1}^{N} \mathbf{W}_i'\mathbf{H}\mathbf{V}_i\right)\left(\sum_{i=1}^{N} \mathbf{V}_i'\mathbf{H}\mathbf{y}_i\right).$$

It is more common to proceed directly to the "two stage least squares" estimator (see Chapter 15) which uses

$$\mathbf{A} = \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{V}_i'\mathbf{H}'\mathbf{H}\mathbf{V}_i\right)^{-1}.$$

The estimator is, then, the one given earlier in (13-41) with $\hat{\boldsymbol{\Sigma}}$ replace by $\mathbf{I}_T$. Either estimator is a function of the sample data only and provides the initial estimator we need.

Ahn and Schmidt (among others) observed that the IV estimator proposed here, as extensive as it is, still neglects quite a lot of information and is therefore (relatively) inefficient. For example, in the first differenced model,

$$E\left[y_{is}(\varepsilon_{it} - \varepsilon_{i,t-1})\right] = 0, \quad s = 0, \ldots, t-2, \quad t = 2, \ldots, T.$$

That is, the *level* of y_{is} is uncorrelated with the differences of disturbances that are at least two periods subsequent.[31] (The differencing transformation, as the transformation to deviations from group means, removes the individual effect.) The corresponding moment equations that can enter the construction of a GMM estimator are

$$\frac{1}{n}\sum_{i=1}^{n} y_{is}[(y_{it} - y_{i,t-1}) - \delta(y_{i,t-1} - y_{i,t-2}) - (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta}] = 0$$

$$s = 0, \ldots, t-2, \quad t = 2, \ldots, T.$$

Altogether, Ahn and Schmidt identify $T(T-1)/2 + T - 2$ such equations that involve mixtures of the levels and differences of the variables. The main conclusion that they demonstrate is that in the dynamic model, there is a large amount of information to be gleaned not only from the familiar relationships among the levels of the variables but also from the implied relationships between the levels and the first differences. The issue of correlation between the transformed y_{it} and the deviations of ε_{it} is discussed in the papers cited. (As Ahn and Schmidt show, there are potentially huge numbers of additional orthogonality conditions in this model owing to the relationship between first differences and second moments. We do not consider those. The matrix $\mathbf{V}_i$ could be huge. Consider a model with 10 time varying right-hand side variables and suppose T_i is 15. Then, there are 15 rows and roughly $15 \times (10 \times 15)$ or 2,250 columns. (The Ahn and Schmidt estimator, which involves potentially thousands of instruments in a model containing only a handful of parameters may become a bit impractical at this point. The common approach is to use only a small subset of the available instrumental

[31]This is the approach suggested by Holtz-Eakin (1988) and Holtz-Eakin, Newey, and Rosen (1988).

variables.) The order of the computation grows as the number of parameters times the square of T.)

The number of orthogonality conditions (instrumental variables) used to estimate the parameters of the model is determined by the number of variables in $\mathbf{v}_{it}$ and $\mathbf{a}_i$ in (13-39). In most cases, the model is vastly overidentified—there are far more orthogonality conditions than parameters. As usual in GMM estimation, a test of the overidentifying restrictions can be based on q, the estimation criterion. At its minimum, the limiting distribution of q is chi-squared with degrees of freedom equal to the number of instrumental variables in total minus $(1 + K_1 + K_2 + L_1 + L_2)$.[32]

Example 13.7 Local Government Expenditure

Dahlberg and Johansson (2000) estimated a model for the local government expenditure of several hundred municipalities in Sweden observed over the nine year period $t = 1979$ to 1987. The equation of interest is

$$S_{i,t} = \alpha_t + \sum_{j=1}^{m} \beta_j S_{i,t-j} + \sum_{j=1}^{m} \gamma_j R_{i,t-j} + \sum_{j=1}^{m} \delta_j G_{i,t-j} + f_i + \varepsilon_{it}.$$

(We have changed their notation slightly to make it more convenient.) $S_{i,t}$, $R_{i,t}$ and $G_{i,t}$ are municipal spending, receipts (taxes and fees) and central government grants, respectively. Analogous equations are specified for the current values of $R_{i,t}$ and $G_{i,t}$. The appropriate lag length, m, is one of the features of interest to be determined by the empirical study. Note that the model contains a municipality specific effect, f_i, which is not specified as being either "fixed" or "random." In order to eliminate the individual effect, the model is converted to first differences. The resulting equation has dependent variable $\Delta S_{i,t} = S_{i,t} - S_{i,t-1}$ and a moving average disturbance, $\Delta\varepsilon_{i,t} = \varepsilon_{i,t} - \varepsilon_{i,t-1}$. Estimation is done using the methods developed by Ahn and Schmidt (1995), Arellano and Bover (1995) and Holtz-Eakin, Newey, and Rosen (1988), as described previously. Issues of interest are the lag length, the parameter estimates, and Granger causality tests, which we will revisit (again using this application) in Chapter 19. We will examine this application in detail and obtain some estimates in the continuation of this example in Section 18.5 (GMM Estimation).

13.7 NONSPHERICAL DISTURBANCES AND ROBUST COVARIANCE ESTIMATION

Since the models considered here are extensions of the classical regression model, we can treat heteroscedasticity in the same way that we did in Chapter 11. 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. We will discuss this model under the heading of "**covariance structures**" in Section 13.9. In this section, we will consider robust estimation of the asymptotic covariance matrix for least squares.

13.7.1 ROBUST ESTIMATION OF THE FIXED EFFECTS MODEL

In the fixed effects model, the full regressor matrix is $\mathbf{Z} = [\mathbf{X}, \mathbf{D}]$. The White heteroscedasticity consistent covariance matrix for OLS—that is, for the fixed effects

[32]This is true generally in GMM estimation. It was proposed for the dynamic panel data model by Bhargava and Sargan (1983).

estimator—is the lower right block of the partitioned matrix

$$\text{Est.Asy. Var}[\mathbf{b}, \mathbf{a}] = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{E}^2\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1},$$

where $\mathbf{E}$ is a diagonal matrix of least squares (fixed effects estimator) residuals. This computation promises to be formidable, but fortunately, it works out very simply. The White estimator for the slopes is obtained just by using the data in group mean deviation form [see (13-4) and (13-8)] in the familiar computation of $\mathbf{S}_0$ [see (11-7) to (11-9)]. Also, the disturbance variance estimator in (13-8) is the counterpart to the one in (11-3), which we showed that after the appropriate scaling of $\boldsymbol{\Omega}$ was a consistent estimator of $\sigma^2 = \text{plim}[1/(nT)]\sum_{i=1}^{n}\sum_{t=1}^{T}\sigma_{it}^2$. The implication is that we may still use (13-8) to estimate the variances of the fixed effects.

A somewhat less general but useful simplification of this result can be obtained if we assume that the disturbance variance is constant within the ith group. If $E[\varepsilon_{it}^2] = \sigma_i^2$, then, with a panel of data, σ_i^2 is estimable by $\mathbf{e}_i'\mathbf{e}_i/T$ using the least squares residuals. (This heteroscedastic regression model was considered at various points in Section 11.7.2.) The center matrix in Est.Asy. Var[$\mathbf{b}, \mathbf{a}$] may be replaced with $\sum_i(\mathbf{e}_i'\mathbf{e}_i/T)$ $\mathbf{Z}_i'\mathbf{Z}_i$. Whether this estimator is preferable is unclear. If the groupwise model is correct, then it and the White estimator will estimate the same matrix. On the other hand, if the disturbance variances do vary within the groups, then this revised computation may be inappropriate.

Arellano (1987) has taken this analysis a step further. If one takes the ith group as a whole, then we can treat the observations in

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \alpha_i\mathbf{i}_T + \boldsymbol{\varepsilon}_i$$

as a generalized regression model with disturbance covariance matrix $\boldsymbol{\Omega}_i$. We saw in Section 11.4 that a model this general, with no structure on $\boldsymbol{\Omega}$, offered little hope for estimation, robust or otherwise. But the problem is more manageable with a panel data set. As before, let $\mathbf{X}_{i*}$ denote the data in group mean deviation form. The counterpart to $\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}$ here is

$$\mathbf{X}_*'\boldsymbol{\Omega}\mathbf{X}_* = \sum_{i=1}^{n}(\mathbf{X}_{i*}'\boldsymbol{\Omega}_i\mathbf{X}_{i*}).$$

By the same reasoning that we used to construct the White estimator in Chapter 12, we can consider estimating $\boldsymbol{\Omega}_i$ with the sample of one, $\mathbf{e}_i\mathbf{e}_i'$. As before, it is not consistent estimation of the individual $\boldsymbol{\Omega}_i$s that is at issue, but estimation of the sum. If n is large enough, then we could argue that

$$\begin{aligned}\text{plim}\,\frac{1}{nT}\mathbf{X}_*'\boldsymbol{\Omega}\mathbf{X}_* &= \text{plim}\,\frac{1}{nT}\sum_{i=1}^{n}\mathbf{X}_{i*}'\boldsymbol{\Omega}_i\mathbf{X}_{*i}\\ &= \text{plim}\,\frac{1}{n}\sum_{i=1}^{n}\frac{1}{T}\mathbf{X}_{*i}'\mathbf{e}_i\mathbf{e}_i'\mathbf{X}_{*i}\\ &= \text{plim}\,\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}e_{it}e_{is}\mathbf{x}_{*it}\mathbf{x}_{*is}'\right).\end{aligned}$$

The result is a combination of the White and Newey–West estimators. But the weights in the latter are 1 rather than $[1 - l/(L+1)]$ because there is no correlation across the groups, so the sum is actually just an average of finite matrices.

13.7.2 HETEROSCEDASTICITY IN THE RANDOM EFFECTS MODEL

Since 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 the preceding section 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.

A series of papers, notably Mazodier and Trognon (1978), Baltagi and Griffin (1988), and the recent monograph by Baltagi (1995, pp. 77–79) suggest how one might allow the group-specific component u_i to be heteroscedastic. But, empirically, there is an insurmountable problem with this approach. In the final analysis, all estimators of the variance components must be based on sums of squared residuals, and, in particular, an estimator of σ_{ui}^2 would be estimated using a set of residuals from the distribution of u_i. However, the data contain only a single observation on u_i repeated in each observation in group i. So, the estimators presented, for example, in Baltagi (1995), use, in effect, one residual in each case to estimate σ_{ui}^2. What appears to be a mean squared residual is only $(1/T)\sum_{t=1}^{T} \hat{u}_i^2 = \hat{u}_i^2$. The properties of this estimator are ambiguous, but efficiency seems unlikely. The estimators do not converge to any population figure as the sample size, even T, increases. Heteroscedasticity in the unique component, ε_{it} represents a more tractable modeling possibility.

In Section 13.4.1, we introduced heteroscedasticity into estimation of the random effects model by allowing the group sizes to vary. But the estimator there (and its feasible counterpart in the next section) would be the same if, instead of $\theta_i = 1 - \sigma_\varepsilon/(T_i\sigma_u^2 + \sigma_\varepsilon^2)^{1/2}$, we were faced with

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

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 above. One possible way to proceed is as follows: Since 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,

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

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

$$\widehat{\sigma_{\varepsilon i}^2} = \frac{\mathbf{e}_i'^{lsdv}\mathbf{e}_i^{lsdv}}{T}$$

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

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

We can now compute the FGLS estimator as before.

Example 13.8 Heteroscedasticity Consistent Estimation

The fixed effects estimates for the cost equation are shown in Table 13.2 on page 302. The row of standard errors labeled White (1) are the estimates based on the usual calculation. For two of the three coefficients, these are actually substantially smaller than the least squares results. The estimates labeled White (2) are based on the groupwise heteroscedasticity model suggested earlier. These estimates are essentially the same as White (1). As noted, it is unclear whether this computation is preferable. Of course, if it were known that the groupwise model were correct, then the least squares computation itself would be inefficient and, in any event, a two-step FGLS estimator would be better.

The estimators of $\sigma_{\varepsilon i}^2 + u_i^2$ based on the least squares residuals are 0.16188, 0.44740, 0.26639, 0.90698, 0.23199, and 0.39764. The six individual estimates of $\sigma_{\varepsilon i}^2$ based on the LSDV residuals are 0.0015352, 0.52883, 0.20233, 0.62511, 0.25054, and 0.32482, respectively. Two of the six implied estimates (the second and fifth) of u_i^2 are negative based on these results, which suggests that a groupwise heteroscedastic random effects model is not an appropriate specification for these data.

13.7.3 AUTOCORRELATION IN PANEL DATA MODELS

Autocorrelation in the fixed effects model is a minor extension of the model of the preceding chapter. With the LSDV estimator in hand, estimates of the parameters of a disturbance process and transformations of the data to allow FGLS estimation proceed exactly as before. The extension one might consider is to allow the autocorrelation coefficient(s) to vary across groups. But even if so, treating each group of observations as a sample in itself provides the appropriate framework for estimation.

In the random effects model, as before, there are additional complications. The regression model is

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

If ε_{it} is produced by an AR(1) process, $\varepsilon_{it} = \rho\varepsilon_{i,t-1} + v_{it}$, then the familiar partial differencing procedure we used before would produce[33]

$$\begin{aligned} y_{it} - \rho y_{i,t-1} &= \alpha(1-\rho) + (\mathbf{x}_{it} - \rho\mathbf{x}_{i,t-1})'\boldsymbol{\beta} + \varepsilon_{it} - \rho\varepsilon_{i,t-1} + u_i(1-\rho) \\ &= \alpha(1-\rho) + (\mathbf{x}_{it} - \rho\mathbf{x}_{i,t-1})'\boldsymbol{\beta} + v_{it} + u_i(1-\rho) \qquad \textbf{(13-42)} \\ &= \alpha(1-\rho) + (\mathbf{x}_{it} - \rho\mathbf{x}_{i,t-1})'\boldsymbol{\beta} + v_{it} + w_i. \end{aligned}$$

Therefore, if an estimator of ρ were in hand, then one could at least treat partially differenced observations two through T in each group as the same random effects model that we just examined. Variance estimators would have to be adjusted by a factor of $(1-\rho)^2$. Two issues remain: (1) how is the estimate of ρ obtained and (2) how does one treat the first observation? For the first of these, the first autocorrelation coefficient of

[33]See Lillard and Willis (1978).

the LSDV residuals (so as to purge the residuals of the individual specific effects, u_i) is a simple expedient. This estimator will be consistent in nT. It is in T alone, but, of course, T is likely to be small. The second question is more difficult. Estimation is simple if the first observation is simply dropped. If the panel contains many groups (large n), then omitting the first observation is not likely to cause the inefficiency that it would in a single time series. One can apply the Prais–Winsten transformation to the first observation in each group instead [multiply by $(1-\rho^2)^{1/2}$], but then an additional complication arises at the second (FGLS) step when the observations are transformed a second time. On balance, the Cochrane–Orcutt estimator is probably a reasonable middle ground. Baltagi (1995, p. 83) discusses the procedure. He also discusses estimation in higher-order AR and MA processes.

In the same manner as in the previous section, we could allow the autocorrelation to differ across groups. An estimate of each ρ_i is computable using the group mean deviation data. This estimator is consistent in T, which is problematic in this setting. In the earlier case, we overcame this difficulty by averaging over n such "weak" estimates and achieving consistency in the dimension of n instead. We lose that advantage when we allow ρ to vary over the groups. This result is the same that arose in our treatment of heteroscedasticity.

For the airlines data in our examples, the estimated autocorrelation is 0.5086, which is fairly large. Estimates of the fixed and random effects models using the Cochrane–Orcutt procedure for correcting the autocorrelation are given in Table 13.2. Despite the large value of r, the resulting changes in the parameter estimates and standard errors are quite modest.

13.8 RANDOM COEFFICIENTS MODELS

Thus far, the model $\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i$ has been analyzed within the familiar frameworks of heteroscedasticity and autocorrelation. Although the models in Sections 13.3 and 13.4 allow considerable flexibility, they do entail the not entirely plausible assumption that there is no parameter variation across firms (i.e., across the cross-sectional units). A fully general approach would combine all the machinery of the previous sections with a model that allows $\boldsymbol{\beta}$ to vary across firms.

Parameter heterogeneity across individuals or groups can be modeled as stochastic variation.[34] Suppose that we write

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i, \tag{13-43}$$

where

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{u}_i, \tag{13-44}$$

[34]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 (1986). The model bears some resemblance to the Bayesian approach of Section 16.2.2, but the similarity is only superficial. We maintain our classical approach to estimation.

and

$$E[\mathbf{u}_i \mid \mathbf{X}_i] = \mathbf{0},$$
$$E[\mathbf{u}_i\mathbf{u}_i' \mid \mathbf{X}_i] = \mathbf{\Gamma}. \tag{13-45}$$

(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 13.4.) Assume for now that there is no autocorrelation or cross-sectional correlation. 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 $\mathbf{\Gamma}$.[35] By inserting (13-44) in (13-43) and expanding the result, we find that $\mathbf{\Omega}$ is a block diagonal matrix with

$$\mathbf{\Omega}_{ii} = E[(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})' \mid \mathbf{X}_i] = \sigma^2\mathbf{I}_T + \mathbf{X}_i\mathbf{\Gamma}\mathbf{X}_i'.$$

We can write the GLS estimator as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{y} = \sum_{i=1}^{n} \mathbf{W}_i\mathbf{b}_i \tag{13-46}$$

where

$$\mathbf{W}_i = \left[\sum_{i=1}^{n}\left(\mathbf{\Gamma} + \sigma_i^2(\mathbf{X}_i'\mathbf{X}_i)^{-1}\right)^{-1}\right]^{-1}\left(\mathbf{\Gamma} + \sigma_i^2(\mathbf{X}_i'\mathbf{X}_i)^{-1}\right)^{-1}.$$

Empirical implementation of this model requires an estimator of $\mathbf{\Gamma}$. One approach [see, e.g., Swamy (1971)] is to use the empirical variance of the set of n least squares estimates, $\mathbf{b}_i$ minus the average value of $s_i^2(\mathbf{X}_i'\mathbf{X}_i)^{-1}$. This matrix may not be positive definite, however, in which case [as Baltagi (1995) suggests], one might drop the second term. The more difficult obstacle is that panels are often short and there may be too few observations to compute $\mathbf{b}_i$. More recent applications of random parameter variation have taken a completely different approach based on simulation estimation. [See Section 17.8, McFadden and Train (2000) and Greene (2001).]

Recent research in a number of fields have extended the random parameters model to a "multilevel" model or "**hierarchical regression**" model by allowing the means of the coefficients to vary with measured covariates. In this formulation, (13-44) becomes

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{\Delta}\mathbf{z}_i + \mathbf{u}_i.$$

This model retains the earlier stochastic specification, but adds the measurement equation to the generation of the random parameters. In principle, this is actually only a minor extension of the model used thus far, as the regression equation would now become

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{X}_i\mathbf{\Delta}\mathbf{z}_i + (\boldsymbol{\varepsilon}_i + \mathbf{X}_i\mathbf{u}_i)$$

which can still be fit by least squares. However, as noted, current applications have found this formulation to be useful in many settings that go beyond the linear model. We will examine an application of this approach in a nonlinear model in Section 17.8.

[35]Swamy and Tavlas (2001) label this the "first generation RCM." We'll examine the "second generation" extension at the end of this section.

13.9 COVARIANCE STRUCTURES FOR POOLED TIME-SERIES CROSS-SECTIONAL DATA

Many studies have analyzed data observed across countries or firms in which the number of cross-sectional units is relatively small and the number of time periods is (potentially) relatively large. The current literature in political science contains many applications of this sort. For example, in a cross-country comparison of economic performance over time, Alvarez, Garrett, and Lange (1991) estimated a model of the form

$$\text{performance}_{it} = f(\text{labor organization}_{it}, \text{political organization}_{it}) + \varepsilon_{it}. \quad \textbf{(13-47)}$$

The data set analyzed in Examples 13.1–13.5 is an example, in which the costs of six large firms are observed for the same 15 years. The modeling context considered here differs somewhat from the longitudinal data sets considered in the preceding sections. In the typical application to be considered here, it is reasonable to specify a common conditional mean function across the groups, with heterogeneity taking the form of different variances rather than shifts in the means. Another substantive difference from the longitudinal data sets is that the observational units are often large enough (e.g., countries) that correlation across units becomes a natural part of the specification, whereas in a "panel," it is always assumed away.

In the models we shall examine in this section, the data set consists of n cross-sectional units, denoted $i = 1, \ldots, n$, observed at each of T time periods, $t = 1, \ldots, T$. We have a total of nT observations. In contrast to the preceding sections, most of the asymptotic results we obtain here are with respect to $T \to \infty$. We will assume that n is fixed.

The framework for this analysis is the generalized regression model:

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}. \quad \textbf{(13-48)}$$

An essential feature of (13-48) is that we have assumed that $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_n$. It is useful to stack the n time series,

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n,$$

so that

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_n \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{bmatrix}. \quad \textbf{(13-49)}$$

Each submatrix or subvector has T observations. We also specify

$$E[\boldsymbol{\varepsilon}_i \mid \mathbf{X}] = \mathbf{0}$$

and

$$E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_j' \mid \mathbf{X}] = \sigma_{ij}\boldsymbol{\Omega}_{ij}$$

so that a generalized regression model applies to each block of T observations. One new element introduced here is the cross sectional covariance across the groups. Collecting the terms above, we have the full specification,

$$E[\boldsymbol{\varepsilon} \mid \mathbf{X}] = \mathbf{0}$$

and

$$E[\boldsymbol{\varepsilon\varepsilon}' \mid \mathbf{X}] = \boldsymbol{\Omega} = \begin{bmatrix} \sigma_{11}\boldsymbol{\Omega}_{11} & \sigma_{12}\boldsymbol{\Omega}_{12} & \cdots & \sigma_{1n}\boldsymbol{\Omega}_{1n} \\ \sigma_{21}\boldsymbol{\Omega}_{21} & \sigma_{22}\boldsymbol{\Omega}_{22} & \cdots & \sigma_{2n}\boldsymbol{\Omega}_{2n} \\ & \vdots & & \\ \sigma_{n1}\boldsymbol{\Omega}_{n1} & \sigma_{n2}\boldsymbol{\Omega}_{n2} & \cdots & \sigma_{nn}\boldsymbol{\Omega}_{nn} \end{bmatrix}.$$

A variety of models are obtained by varying the structure of $\boldsymbol{\Omega}$.

13.9.1 GENERALIZED LEAST SQUARES ESTIMATION

As we observed in our first encounter with the generalized regression model, the fully general covariance matrix in (13-49), which, as stated, contains $nT(nT+1)/2$ parameters is certainly inestimable. But, several restricted forms provide sufficient generality for empirical use. To begin, we assume that there is no correlation across periods, which implies that $\boldsymbol{\Omega}_{ij} = \mathbf{I}$.

$$\boldsymbol{\Omega} = \begin{bmatrix} \sigma_{11}\mathbf{I} & \sigma_{12}\mathbf{I} & \cdots & \sigma_{1n}\mathbf{I} \\ \sigma_{21}\mathbf{I} & \sigma_{22}\mathbf{I} & \cdots & \sigma_{2n}\mathbf{I} \\ & \vdots & & \\ \sigma_{n1}\mathbf{I} & \sigma_{n2}\mathbf{I} & \cdots & \sigma_{nn}\mathbf{I} \end{bmatrix}. \tag{13-50}$$

The generalized least squares estimator of $\boldsymbol{\beta}$ is based on a known $\boldsymbol{\Omega}$ would be

$$\hat{\boldsymbol{\beta}} = [\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}]^{-1}[\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}].$$

The matrix $\boldsymbol{\Omega}$ can be written as

$$\boldsymbol{\Omega} = \boldsymbol{\Sigma} \otimes \mathbf{I}, \tag{13-51}$$

where $\boldsymbol{\Sigma}$ is the $n \times n$ matrix $[\sigma_{ij}]$ (note the contrast to (13-21) where $\boldsymbol{\Omega} = \mathbf{I}_n \otimes \boldsymbol{\Sigma}$). Then,

$$\boldsymbol{\Omega}^{-1} = \boldsymbol{\Sigma}^{-1} \otimes \mathbf{I} = \begin{bmatrix} \sigma^{11}\mathbf{I} & \sigma^{12}\mathbf{I} & \cdots & \sigma^{1n}\mathbf{I} \\ \sigma^{21}\mathbf{I} & \sigma^{22}\mathbf{I} & \cdots & \sigma^{2n}\mathbf{I} \\ & & \vdots & \\ \sigma^{n1}\mathbf{I} & \sigma^{n2}\mathbf{I} & \cdots & \sigma^{nn}\mathbf{I} \end{bmatrix}. \tag{13-52}$$

where σ^{ij} denotes the ijth element of $\boldsymbol{\Sigma}^{-1}$. This provides a specific form for the estimator,

$$\hat{\boldsymbol{\beta}} = \left[\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma^{ij}\mathbf{X}_i'\mathbf{X}_j\right]^{-1}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma^{ij}\mathbf{X}_i'\mathbf{y}_j\right]. \tag{13-53}$$

The asymptotic covariance matrix of the GLS estimator is the inverse matrix in brackets.

13.9.2 FEASIBLE GLS ESTIMATION

As always in the generalized linear regression model, the slope coefficients, $\boldsymbol{\beta}$ can be consistently, if not efficiently estimated by ordinary least squares. A consistent estimator of σ_{ij} can be based on the sample analog to the result

$$E[\varepsilon_{it}\varepsilon_{jt}] = E\left[\frac{\boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_j}{T}\right] = \sigma_{ij}.$$

Using the least squares residuals, we have

$$\hat{\sigma}_{ij} = \frac{\mathbf{e}_i'\mathbf{e}_j}{T}. \tag{13-54}$$

Some treatments use $T-K$ instead of T in the denominator of $\hat{\sigma}_{ij}$.[36] There is no problem created by doing so, but the resulting estimator is not unbiased regardless. Note that this estimator is consistent in T. Increasing T increases the information in the sample, while increasing n increases the number of variance and covariance parameters to be estimated. To compute the FGLS estimators for this model, we require the full set of sample moments, $\mathbf{y}_i'\mathbf{y}_j$, $\mathbf{X}_i'\mathbf{X}_j$, and $\mathbf{X}_i'\mathbf{y}_j$ for all pairs of cross-sectional units. With $\hat{\sigma}_{ij}$ in hand, FGLS may be computed using

$$\hat{\hat{\boldsymbol{\beta}}} = [\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X}]^{-1}[\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{y}], \tag{13-55}$$

where $\mathbf{X}$ and $\mathbf{y}$ are the stacked data matrices in (13-49)—this is done in practice using (13-53) and (13-54) which involve only $K \times K$ and $K \times 1$ matrices. The estimated asymptotic covariance matrix for the FGLS estimator is the inverse matrix in brackets in (13-55).

There is an important consideration to note in feasible GLS estimation of this model. The computation requires inversion of the matrix $\hat{\boldsymbol{\Sigma}}$ where the ijth element is given by (13-54). This matrix is $n \times n$. It is computed from the least squares residuals using

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{e}_t\mathbf{e}_t' = \frac{1}{T}\mathbf{E}'\mathbf{E}$$

where $\mathbf{e}_t'$ is a $1 \times n$ vector containing all n residuals for the n groups at time t, placed as the tth row of the $T \times n$ matrix of residuals, $\mathbf{E}$. The rank of this matrix cannot be larger than T. Note what happens if $n > T$. In this case, the $n \times n$ matrix has rank T which is less than n, so it must be singular, and the FGLS estimator cannot be computed. For example, a study of 20 countries each observed for 10 years would be such a case. This result is a deficiency of the data set, not the model. The population matrix, $\boldsymbol{\Sigma}$ is positive definite. But, if there are not enough observations, then the data set is too short to obtain a positive definite estimate of the matrix. The heteroscedasticity model described in the next section can always be computed, however.

[36] See, for example, Kmenta (1986, p. 620). Elsewhere, for example, in Fomby, Hill, and Johnson (1984, p. 327), T is used instead.

13.9.3 HETEROSCEDASTICITY AND THE CLASSICAL MODEL

Two special cases of this model are of interest. The **groupwise heteroscedastic** model of Section 11.7.2 results if the off diagonal terms in $\boldsymbol{\Sigma}$ all equal zero. Then, the GLS estimator, as we saw earlier, is

$$\hat{\beta} = [\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}]^{-1}[\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}] = \left[\sum_{i=1}^{n} \frac{1}{\sigma_i^2}\mathbf{X}_i'\mathbf{X}_i\right]^{-1} \left[\sum_{i=1}^{n} \frac{1}{\sigma_i^2}\mathbf{X}_i'\mathbf{y}_i\right].$$

Of course, the disturbance variances, σ_i^2, are unknown, so the two-step FGLS method noted earlier, now based only on the diagonal elements of $\boldsymbol{\Sigma}$ would be used. The second special case is the classical regression model, which adds the further restriction $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2$. We would now stack the data in the pooled regression model in

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

For this simple model, the GLS estimator reduces to pooled ordinary least squares.

Beck and Katz (1995) suggested that the standard errors for the OLS estimates in this model should be corrected for the possible misspecification that would arise if $\sigma_{ij}\boldsymbol{\Omega}_{ij}$ were correctly specified by (13-49) instead of $\sigma^2\mathbf{I}$, as now assumed. The appropriate asymptotic covariance matrix for OLS in the general case is, as always,

$$\text{Asy. Var}[\mathbf{b}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

For the special case of $\boldsymbol{\Omega}_{ij} = \sigma_{ij}\mathbf{I}$,

$$\text{Asy. Var}[\mathbf{b}] = \left(\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{ij}\mathbf{X}_i'\mathbf{X}_j\right)\left(\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i\right)^{-1}. \qquad \textbf{(13-56)}$$

This estimator is straightforward to compute with estimates of σ_{ij} in hand. Since the OLS estimator is consistent, (13-54) may be used to estimate σ_{ij}.

13.9.4 SPECIFICATION TESTS

We are interested in testing down from the general model to the simpler forms if possible. Since the model specified thus far is distribution free, the standard approaches, such as likelihood ratio tests, are not available. We propose the following procedure. Under the null hypothesis of a common variance, σ^2 (i.e., the classical model) the Wald statistic for testing the null hypothesis against the alternative of the groupwise heteroscedasticity model would be

$$W = \sum_{i=1}^{n} \frac{(\hat{\sigma}_i^2 - \sigma^2)^2}{\text{Var}\left[\hat{\sigma}_i^2\right]}.$$

If the null hypothesis is correct,

$$W \xrightarrow{d} \chi^2[n].$$

By hypothesis,

$$\text{plim}\,\hat{\sigma}^2 = \sigma^2,$$

where $\hat{\sigma}^2$ is the disturbance variance estimator from the pooled OLS regression. We must now consider $\text{Var}[\hat{\sigma}_i^2]$. Since

$$\hat{\sigma}_i^2 = \frac{1}{T}\sum_{t=1}^{T} e_{it}^2,$$

is a mean of T observations, we may estimate $\text{Var}[\hat{\sigma}_i^2]$ with

$$f_{ii} = \frac{1}{T}\frac{1}{T-1}\sum_{t=1}^{T}\left(e_{it}^2 - \hat{\sigma}_i^2\right)^2.\text{[37]} \tag{13-57}$$

The modified Wald statistic is then

$$W' = \sum_{i=1}^{n}\frac{\left(\hat{\sigma}_i^2 - \hat{\sigma}^2\right)^2}{f_{ii}}.$$

A Lagrange multiplier statistic is also simple to compute and asymptotically equivalent to a likelihood ratio test—we consider these below. But, these assume normality, which we have not yet invoked. To this point, our specification is distribution free. White's general test[38] is an alternative. To use White's test, we would regress the squared OLS residuals on the P unique variables in $\mathbf{x}$ and the squares and cross products, including a constant. The chi-squared statistic, which has $P-1$ degrees of freedom, is $(nT)R^2$.

For the full model with nonzero off diagonal elements in $\boldsymbol{\Sigma}$, the preceding approach must be modified. One might consider simply adding the corresponding terms for the off diagonal elements, with a common $\sigma_{ij}=0$, but this neglects the fact that under this broader alternative hypothesis, the original n variance estimators are no longer uncorrelated, even asymptotically, so the limiting distribution of the Wald statistic is no longer chi-squared. Alternative approaches that have been suggested [see, e.g., Johnson and Wichern (1999, p. 424)] are based on the following general strategy: Under the alternative hypothesis of an unrestricted $\boldsymbol{\Sigma}$, the sample estimate of $\boldsymbol{\Sigma}$ will be $\hat{\boldsymbol{\Sigma}} = [\hat{\sigma}_{ij}]$ as defined in (13-54). Under any restrictive null hypothesis, the estimator of $\boldsymbol{\Sigma}$ will be $\hat{\boldsymbol{\Sigma}}_0$, a matrix that by construction will be larger than $\hat{\boldsymbol{\Sigma}}$ in the matrix sense defined in Appendix A. Statistics based on the "excess variation," such as $T(\hat{\boldsymbol{\Sigma}}_0 - \hat{\boldsymbol{\Sigma}})$ are suggested for the testing procedure. One of these is the likelihood ratio test that we will consider in Section 13.9.6.

13.9.5 AUTOCORRELATION

The preceding discussion dealt with heteroscedasticity and cross-sectional correlation. Through a simple modification of the procedures, it is possible to relax the assumption of nonautocorrelation as well. It is simplest to begin with the assumption that

$$\text{Corr}[\varepsilon_{it}, \varepsilon_{js}] = 0, \quad \text{if } i \neq j.$$

[37]Note that would apply strictly if we had observed the true disturbances, ε_{it}. We are using the residuals as estimates of their population counterparts. Since the coefficient vector is consistent, this procedure will obtain the desired results.

[38]See Section 11.4.1.

That is, the disturbances between cross-sectional units are uncorrelated. Now, we can take the approach of Chapter 12 to allow for autocorrelation within the cross-sectional units. That is,

$$\varepsilon_{it} = \rho_i \varepsilon_{i,t-1} + u_{it},$$
$$\text{Var}[\varepsilon_{it}] = \sigma_i^2 = \frac{\sigma_{ui}^2}{1-\rho_i^2}. \tag{13-58}$$

For FGLS estimation of the model, suppose that r_i is a consistent estimator of ρ_i. Then, if we take each time series $[\mathbf{y}_i, \mathbf{X}_i]$ separately, we can transform the data using the Prais–Winsten transformation:

$$\mathbf{y}_{*i} = \begin{bmatrix} \sqrt{1-r_i^2}\, y_{i1} \\ y_{i2} - r_i y_{i1} \\ y_{i3} - r_i y_{i2} \\ \vdots \\ y_{iT} - r_i\, y_{i,T-1} \end{bmatrix}, \quad \mathbf{X}_{*i} = \begin{bmatrix} \sqrt{1-r_i^2}\, \mathbf{x}_{i1} \\ \mathbf{x}_{i2} - r_i \mathbf{x}_{i1} \\ \mathbf{x}_{i3} - r_i \mathbf{x}_{i2} \\ \vdots \\ \mathbf{x}_{iT} - r_i \mathbf{x}_{i,T-1} \end{bmatrix}. \tag{13-59}$$

In terms of the transformed data $\mathbf{y}_{*i}$ and $\mathbf{X}_{*i}$, the model is now only heteroscedastic; the transformation has removed the autocorrelation. As such, the groupwise heteroscedastic model applies to the transformed data. We may now use weighted least squares, as described earlier. This requires a second least squares estimate. The first, OLS regression produces initial estimates of ρ_i. The transformed data are then used in a second least squares regression to obtain consistent estimators,

$$\hat{\sigma}_{ui}^2 = \frac{\mathbf{e}_{*i}'\mathbf{e}_{*i}}{T} = \frac{(\mathbf{y}_{*i} - \mathbf{X}_{*i}\hat{\boldsymbol{\beta}})'(\mathbf{y}_{*i} - \mathbf{X}_{*i}\hat{\boldsymbol{\beta}})}{T}. \tag{13-60}$$

[Note that both the initial OLS and the second round FGLS estimators of $\boldsymbol{\beta}$ are consistent, so either could be used in (13-60). We have used $\hat{\boldsymbol{\beta}}$ to denote the coefficient vector used, whichever one is chosen.] With these results in hand, we may proceed to the calculation of the groupwise heteroscedastic regression in Section 13.9.3. At the end of the calculation, the moment matrix used in the last regression gives the correct asymptotic covariance matrix for the estimator, now $\hat{\hat{\boldsymbol{\beta}}}$. If desired, then a consistent estimator of $\sigma_{\varepsilon i}^2$ is

$$\hat{\sigma}_{\varepsilon i}^2 = \frac{\hat{\sigma}_{ui}^2}{1-r_i^2}. \tag{13-61}$$

The remaining question is how to obtain the initial estimates r_i. There are two possible structures to consider. If each group is assumed to have its own autocorrelation coefficient, then the choices are the same ones examined in Chapter 12; the natural choice would be

$$r_i = \frac{\sum_{t=2}^{T} e_{it} e_{i,t-1}}{\sum_{t=1}^{T} e_{it}^2}.$$

If the disturbances have a common stochastic process with the same ρ_i, then several estimators of the common ρ are available. One which is analogous to that used in the single equation case is

$$r = \frac{\sum_{i=1}^{n}\sum_{t=2}^{T} e_{it}e_{i,t-1}}{\sum_{i=1}^{n}\sum_{t=1}^{T} e_{it}^2} \tag{13-62}$$

Another consistent estimator would be sample average of the group specific estimated autocorrelation coefficients.

Finally, one may wish to allow for cross-sectional correlation across units. The preceding has a natural generalization. If we assume that

$$\text{Cov}[u_{it}, u_{jt}] = \sigma_{uij},$$

then we obtain the original model in (13-49) in which the off-diagonal blocks of $\mathbf{\Omega}$, are

$$\sigma_{ij}\mathbf{\Omega}_{ij} = \frac{\sigma_{uij}}{1-\rho_i\rho_j}\begin{bmatrix} 1 & \rho_j & \rho_j^2 & \cdots & \rho_j^{T-1} \\ \rho_i & 1 & \rho_j & \cdots & \rho_j^{T-2} \\ \rho_i^2 & \rho_i & 1 & \cdots & \rho_j^{T-3} \\ & & \vdots & & \\ & & \vdots & & \\ \rho_i^{T-1} & \rho_i^{T-2} & \rho_i^{T-3} & \cdots & 1 \end{bmatrix}. \tag{13-63}$$

Initial estimates of ρ_i are required, as before. The Prais–Winsten transformation renders all the blocks in $\mathbf{\Omega}$ diagonal. Therefore, the model of cross-sectional correlation in Section 13.9.2 applies to the transformed data. Once again, the GLS moment matrix obtained at the last step provides the asymptotic covariance matrix for $\hat{\hat{\boldsymbol{\beta}}}$. Estimates of $\sigma_{\varepsilon ij}$ can be obtained from the least squares residual covariances obtained from the transformed data:

$$\hat{\sigma}_{\varepsilon ij} = \frac{\hat{\sigma}_{uij}}{1-r_ir_j}, \tag{13-64}$$

where $\hat{\sigma}_{uij} = \mathbf{e}_{*_i}'\mathbf{e}_{*_j}/T$.

13.9.6 MAXIMUM LIKELIHOOD ESTIMATION

Consider the general model with groupwise heteroscedasticity and cross group correlation. The covariance matrix is the $\mathbf{\Sigma}$ in (13-49). We now assume that the n disturbances at time t, $\boldsymbol{\varepsilon}_t$ have a multivariate normal distribution with zero mean and this $n \times n$ covariance matrix. Taking logs and summing over the T periods gives the log-likelihood for the sample,

$$\ln L(\boldsymbol{\beta}, \mathbf{\Sigma} \mid \text{data}) = -\frac{nT}{2}\ln 2\pi - \frac{T}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}\sum_{t=1}^{T}\boldsymbol{\varepsilon}_t'\mathbf{\Sigma}^{-1}\boldsymbol{\varepsilon}_t, \tag{13-65}$$

$$\varepsilon_{it} = y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta}, \quad i = 1, \ldots, n.$$

(This log-likelihood is analyzed at length in Section 14.2.4, so we defer the more detailed analysis until then.) The result is that the maximum likelihood estimator of $\boldsymbol{\beta}$ is the generalized least squares estimator in (13-53). Since the elements of $\mathbf{\Sigma}$ must be estimated, the FGLS estimator in (13-54) is used, based on the MLE of $\mathbf{\Sigma}$. As shown in

Section 14.2.4, the maximum likelihood estimator of $\boldsymbol{\Sigma}$ is

$$\hat{\sigma}_{ij} = \frac{(\mathbf{y}_i' - \mathbf{X}_i \hat{\hat{\boldsymbol{\beta}}}_{ML})'(\mathbf{y}_j - \mathbf{X}_j \hat{\hat{\boldsymbol{\beta}}}_{ML})}{T} = \frac{\hat{\boldsymbol{\varepsilon}}_i' \hat{\boldsymbol{\varepsilon}}_j}{T} \tag{13-66}$$

based on the MLE of $\boldsymbol{\beta}$. Since each MLE requires the other, how can we proceed to obtain both? The answer is provided by Oberhofer and Kmenta (1974) who show that for certain models, including this one, one can iterate back and forth between the two estimators. (This is the same estimator we used in Section 11.7.2.) Thus, the MLEs are obtained by iterating to convergence between (13-66) and

$$\hat{\hat{\boldsymbol{\beta}}} = [\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X}]^{-1}[\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{y}].$$

The process may begin with the (consistent) ordinary least squares estimator, then (13-66), and so on. The computations are simple, using basic matrix algebra. Hypothesis tests about $\boldsymbol{\beta}$ may be done using the familiar Wald statistic. The appropriate estimator of the asymptotic covariance matrix is the inverse matrix in brackets in (13-55).

For testing the hypothesis that the off-diagonal elements of $\boldsymbol{\Sigma}$ are zero—that is, that there is no correlation across firms—there are three approaches. The likelihood ratio test is based on the statistic

$$\lambda_{\text{LR}} = T(\ln|\hat{\boldsymbol{\Sigma}}_{heteroscedastic}| - \ln|\hat{\boldsymbol{\Sigma}}_{general}|) = T\left(\sum_{i=1}^{n} \ln \hat{\sigma}_i^2 - \ln|\hat{\boldsymbol{\Sigma}}|\right), \tag{13-67}$$

where $\hat{\sigma}_i^2$ are the estimates of σ_i^2 obtained from the maximum likelihood estimates of the groupwise heteroscedastic model and $\hat{\boldsymbol{\Sigma}}$ is the maximum likelihood estimator in the unrestricted model. (Note how the excess variation produced by the restrictive model is used to construct the test.) The large-sample distribution of the statistic is chi-squared with $n(n-1)/2$ degrees of freedom. The Lagrange multiplier test developed by Breusch and Pagan (1980) provides an alternative. The general form of the statistic is

$$\lambda_{\text{LM}} = T\sum_{i=2}^{n}\sum_{j=1}^{i-1} r_{ij}^2, \tag{13-68}$$

where r_{ij}^2 is the ijth residual correlation coefficient. If every individual had a different parameter vector, then individual specific ordinary least squares would be efficient (and ML) and we would compute r_{ij} from the OLS residuals (assuming that there are sufficient observations for the computation). Here, however, we are assuming only a single-parameter vector. Therefore, the appropriate basis for computing the correlations is the residuals from the iterated estimator in the groupwise heteroscedastic model, that is, the same residuals used to compute $\hat{\sigma}_i^2$. (An asymptotically valid approximation to the test can be based on the FGLS residuals instead.) Note that this is not a procedure for testing all the way down to the classical, homoscedastic regression model. That case, which involves different LM and LR statistics, is discussed next. If either the LR statistic in (13-67) or the LM statistic in (13-68) are smaller than the critical value from the table, the conclusion, based on this test, is that the appropriate model is the groupwise heteroscedastic model.

For the groupwise heteroscedasticity model, ML estimation reduces to groupwise weighted least squares. The maximum likelihood estimator of $\boldsymbol{\beta}$ is feasible GLS. The maximum likelihood estimator of the group specific variances is given by the diagonal element in (13-66), while the cross group covariances are now zero. An additional useful result is provided by the negative of the expected second derivatives matrix of the log-likelihood in (13-65) with diagonal $\boldsymbol{\Sigma}$,

$$-E[\mathbf{H}(\boldsymbol{\beta}, \sigma_i^2, i = 1, \ldots, n)] = \begin{bmatrix} \sum_{i=1}^{n} \left(\frac{1}{\sigma_i^2}\right) \mathbf{X}_i'\mathbf{X}_i & \mathbf{0} \\ \mathbf{0} & \operatorname{diag}\left(\frac{T}{2\sigma_i^4}, i = 1, \ldots, n\right) \end{bmatrix}.$$

Since the expected Hessian is block diagonal, the complete set of maximum likelihood estimates can be computed by iterating back and forth between these estimators for σ_i^2 and the feasible GLS estimator of $\boldsymbol{\beta}$. (This process is also equivalent to using a set of n group dummy variables in Harvey's model of heteroscedasticity in Section 11.7.1.)

For testing the heteroscedasticity assumption of the model, the full set of test strategies that we have used before is available. The Lagrange multiplier test is probably the most convenient test, since it does not require another regression after the pooled least squares regression. It is convenient to rewrite

$$\frac{\partial \log L}{\partial \sigma_i^2} = \frac{T}{2\sigma_i^2}\left[\frac{\hat{\sigma}_i^2}{\sigma_i^2} - 1\right],$$

where $\hat{\sigma}_i^2$ is the ith unit-specific estimate of σ_i^2 based on the true (but unobserved) disturbances. Under the null hypothesis of equal variances, regardless of what the common restricted estimator of σ_i^2 is, the first-order condition for equating $\partial \ln L/\partial \boldsymbol{\beta}$ to zero will be the OLS normal equations, so the restricted estimator of $\boldsymbol{\beta}$ is $\mathbf{b}$ using the pooled data. To obtain the restricted estimator of σ_i^2, return to the log-likelihood function. Under the null hypothesis $\sigma_i^2 = \sigma^2, i = 1, \ldots, n$, the first derivative of the log-likelihood function with respect to this common σ^2 is

$$\frac{\partial \log L_R}{\partial \sigma^2} = -\frac{nT}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n} \boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_i.$$

Equating this derivative to zero produces the restricted maximum likelihood estimator

$$\hat{\sigma}^2 = \frac{1}{nT}\sum_{i=1}^{n} \boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_i. = \frac{1}{n}\sum_{i=1}^{n} \hat{\sigma}_i^2,$$

which is the simple average of the n individual consistent estimators. Using the least squares residuals at the restricted solution, we obtain $\hat{\sigma}^2 = (1/nT)\mathbf{e}'\mathbf{e}$ and $\hat{\sigma}_i^2 = (1/T)\mathbf{e}_i'\mathbf{e}_i$. With these results in hand and using the estimate of the expected Hessian for the covariance matrix, the Lagrange multiplier statistic reduces to

$$\lambda_{\text{LM}} = \sum_{i=1}^{n}\left[\frac{T}{2\hat{\sigma}^2}\left(\frac{\hat{\sigma}_i^2}{\hat{\sigma}^2} - 1\right)\right]^2\left(\frac{2\hat{\sigma}^4}{T}\right) = \frac{T}{2}\sum_{i=1}^{n}\left[\frac{\hat{\sigma}_i^2}{\hat{\sigma}^2} - 1\right]^2.$$

The statistic has $n - 1$ degrees of freedom. (It has only $n - 1$ since the restriction is that the variances are all equal to each other, not a specific value, which is $n - 1$ restrictions.)

With the unrestricted estimates, as an alternative test procedure, we may use the Wald statistic. If we assume normality, then the asymptotic variance of each variance estimator is $2\sigma_i^4/T$ and the variances are asymptotically uncorrelated. Therefore, the Wald statistic to test the hypothesis of a common variance σ^2, using $\hat{\sigma}_i^2$ to estimate σ_i^2, is

$$W = \sum_{i=1}^{n} (\hat{\sigma}_i^2 - \sigma^2)^2 \left(\frac{2\sigma_i^4}{T}\right)^{-1} = \frac{T}{2}\sum_{i=1}^{n}\left(\frac{\sigma^2}{\hat{\sigma}_i^2} - 1\right)^2.$$

Note the similarity to the Lagrange multiplier statistic. The estimator of the common variance would be the pooled estimator from the first least squares regression. Recall, we produced a general counterpart for this statistic for the case in which disturbances are not normally distributed.

We can also carry out a likelihood ratio test using the test statistic in Section 12.3.4. The appropriate likelihood ratio statistic is

$$\lambda_{\text{LR}} = T(\ln|\hat{\boldsymbol{\Sigma}}_{homoscedastic}| - \ln|\hat{\boldsymbol{\Sigma}}_{heteroscedastic}|) = (nT)\ln\hat{\sigma}^2 - \sum_{i=1}^{n} T\ln\hat{\sigma}_i^2,$$

where

$$\hat{\sigma}^2 = \frac{\mathbf{e}'\mathbf{e}}{nT} \quad \text{and} \quad \hat{\sigma}_i^2 = \frac{\hat{\boldsymbol{\varepsilon}}_i'\hat{\boldsymbol{\varepsilon}}_i}{T},$$

with all residuals computed using the maximum likelihood estimators. This chi-squared statistic has $n - 1$ degrees of freedom.

13.9.7 APPLICATION TO GRUNFELD'S INVESTMENT DATA

To illustrate the techniques developed in this section, we will use a panel of data that has for several decades provided a useful tool for examining multiple equation estimators. Appendix Table F13.1 lists part of the data used in a classic study of investment demand.[39] The data consist of time series of 20 yearly observations for five firms (of 10 in the original study) and three variables:

I_{it} = gross investment,

F_{it} = market value of the firm at the end of the previous year,

C_{it} = value of the stock of plant and equipment at the end of the previous year.

All figures are in millions of dollars. The variables F_{it} and I_{it} reflect anticipated profit and the expected amount of replacement investment required.[40] The model to be estimated with these data is

$$I_{it} = \beta_1 + \beta_2 F_{it} + \beta_3 C_{it} + \varepsilon_{it},^{41}$$

[39]See Grunfeld (1958) and Grunfeld and Griliches (1960). The data were also used in Boot and deWitt (1960). Although admittedly not current, these data are unusually cooperative for illustrating the different aspects of estimating systems of regression equations.

[40]In the original study, the authors used the notation F_{t-1} and C_{t-1}. To avoid possible conflicts with the usual subscripting conventions used here, we have used the preceding notation instead.

[41]Note that we are modeling investment, a flow, as a function of two stocks. This could be a theoretical misspecification—it might be preferable to specify the model in terms of planned investment. But, 40 years after the fact, we'll take the specified model as it is.

TABLE 13.4 Estimated Parameters and Estimated Standard Errors

	β_1	β_2	β_3
Homoscedasticity			
Least squares	−48.0297	0.10509	0.30537
	$R^2 = 0.77886$, $\hat{\sigma}^2 = 15708.84$, log-likelihood $= -624.9928$		
OLS standard errors	(21.16)	(0.01121)	(0.04285)
White correction	(15.017)	(0.00915)	(0.05911)
Beck and Katz	(10.814)	(0.00832)	(0.033043)
Heteroscedastic			
Feasible GLS	−36.2537	0.09499	0.33781
	(6.1244)	(0.00741)	(0.03023)
Maximum likelihood	−23.2582	0.09435	0.33371
	(4.815)	(0.00628)	(0.2204)
	Pooled $\hat{\sigma}^2 = 15{,}853.08$, log-likelihood $= -564.535$		
Cross-section correlation			
Feasible GLS	−28.247	0.089101	0.33401
	(4.888)	(0.005072)	(0.01671)
Maximum likelihood	−2.217	0.02361	0.17095
	(1.96)	(0.004291)	(0.01525)
	log-likelihood $= -515.422$		
Autocorrelation model			
Heteroscedastic	−23.811	0.086051	0.33215
	(7.694)	(0.009599)	(0.03549)
Cross-section correlation	−15.424	0.07522	0.33807
	(4.595)	(0.005710)	(0.01421)

where i indexes firms and t indexes years. Different restrictions on the parameters and the variances and covariances of the disturbances will imply different forms of the model. By pooling all 100 observations and estimating the coefficients by ordinary least squares, we obtain the first set of results in Table 13.4. To make the results comparable all variance estimates and estimated standard errors are based on $\mathbf{e}'\mathbf{e}/(nT)$. There is no degrees of freedom correction. The second set of standard errors given are White's robust estimator [see (10-14) and (10-23)]. The third set of standard errors given above are the robust standard errors based on Beck and Katz (1995) using (13-56) and (13-54).

The estimates of σ_i^2 for the model of groupwise heteroscedasticity are shown in Table 13.5. The estimates suggest that the disturbance variance differs widely across firms. To investigate this proposition before fitting an extended model, we can use the tests for homoscedasticity suggested earlier. Based on the OLS results, the LM statistic equals 46.63. The critical value from the chi-squared distribution with four degrees of freedom is 9.49, so on the basis of the LM test, we reject the null hypothesis of homoscedasticity. To compute White's test statistic, we regress the squared least squares residuals on a constant, F, C, F^2, C^2, and FC. The R^2 in this regression is 0.36854, so the chi-squared statistic is $(nT)R^2 = 36.854$ with five degrees of freedom. The five percent critical value from the table for the chi-squared statistic with five degrees of freedom is 11.07, so the null hypothesis is rejected again. The likelihood ratio statistic, based on

TABLE 13.5 Estimated Group Specific Variances

	σ^2_{GM}	σ^2_{CH}	σ^2_{GE}	σ^2_{WE}	σ^2_{US}
Based on OLS	9,410.91	755.85	34,288.49	633.42	33,455.51
Heteroscedastic FGLS	8,612.14	409.19	36,563.24	777.97	32,902.83
	(2897.08)	(136.704)	(5801.17)	(323.357)	(7000.857)
Heteroscedastic ML	8,657.72	175.80	40,210.96	1,240.03	29,825.21
Cross Correlation FGLS	10050.52	305.61	34556.6	833.36	34468.98
Autocorrelation, $s^2_{u_i}(u_i)$	6525.7	253.104	14,620.8	232.76	8,683.9
Autocorrelation, $s^2_{e_i}(e_i)$	8453.6	270.150	16,073.2	349.68	12,994.2

the ML results in Table 13.4, is

$$\chi^2 = 100 \ln s^2 - \sum_{i=1}^{n} 20 \ln \hat{\sigma}_i^2 = 120.915.$$

This result far exceeds the tabled critical value. The Lagrange multiplier statistic based on all variances computed using the OLS residuals is 46.629. The Wald statistic based on the FGLS estimated variances and the pooled OLS estimate (15,708.84) is 17,676.25. We observe the common occurrence of an extremely large Wald test statistic. (If the test is based on the sum of squared FGLS residuals, $\hat{\sigma}^2 = 15{,}853.08$, then $W = 18{,}012.86$, which leads to the same conclusion.) To compute the modified Wald statistic absent the assumption of normality, we require the estimates of the variances of the FGLS residual variances. The square roots of f_{ii} are shown in Table 13.5 in parentheses after the FGLS residual variances. The modified Wald statistic is $W' = 14{,}681.3$, which is consistent with the other results. We proceed to reestimate the regression allowing for heteroscedasticity. The FGLS and maximum likelihood estimates are shown in Table 13.4. (The latter are obtained by iterated FGLS.)

Returning to the least squares estimator, we should expect the OLS standard errors to be incorrect, given our findings. There are two possible corrections we can use, the White estimator and direct computation of the appropriate asymptotic covariance matrix. The Beck et al. estimator is a third candidate, but it neglects to use the known restriction that the off-diagonal elements in $\boldsymbol{\Omega}$ are zero. The various estimates shown at the top of Table 13.5 do suggest that the OLS estimated standard errors have been distorted.

The correlation matrix for the various sets of residuals, using the estimates in Table 13.4, is given in Table 13.6.[42] The several quite large values suggests that the more general model will be appropriate. The two test statistics for testing the null hypothesis of a diagonal $\boldsymbol{\Sigma}$, based on the log-likelihood values in Table 13.4, are

$$\lambda_{\text{LR}} = -2(-565.535 - (-515.422)) = 100.226$$

and, based on the MLE's for the groupwise heteroscedasticity model, $\lambda_{\text{LM}} = 66.067$ (the MLE of $\boldsymbol{\Sigma}$ based on the coefficients from the heteroscedastic model is not shown).

For 10 degrees of freedom, the critical value from the chi-squared table is 23.21, so both results lead to rejection of the null hypothesis of a diagonal $\boldsymbol{\Sigma}$. We conclude that

[42]The estimates based on the MLEs are somewhat different, but the results of all the hypothesis tests are the same.

TABLE 13.6 Estimated Cross-Group Correlations Based on FGLS Estimates (Order is OLS, FGLS heteroscedastic, FGLS correlation, Autocorrelation)

Estimated and Correlations

	GM	*CH*	*GE*	*WE*	*US*
GM	1				
CH	−0.344 −0.185 −0.349 −0.225	1			
GE	−0.182 −0.185 −0.248 −0.287	0.283 0.144 0.158 0.105	1		
WE	−0.352 −0.469 −0.356 −0.467	0.343 0.186 0.246 0.166	0.890 0.881 0.895 0.885	1	
US	−0.121 −0.016 −0.716 −0.015	0.167 0.222 0.244 0.245	−0.151 −0.122 −0.176 −0.139	−0.085 −0.119 −0.040 −0.101	1

the simple heteroscedastic model is not general enough for these data.

If the null hypothesis is that the disturbances are both homoscedastic and uncorrelated across groups, then these two tests are inappropriate. A likelihood ratio test can be constructed using the OLS results and the MLEs from the full model; the test statistic would be

$$\lambda_{\text{LR}} = (nT)\ln(\mathbf{e}'\mathbf{e}/nT) - T\ln|\hat{\boldsymbol{\Sigma}}|.$$

This statistic is just the sum of the LR statistics for the test of homoscedasticity and the statistic given above. For these data, this sum would be $120.915 + 100.226 = 221.141$, which is far larger than the critical value, as might be expected.

FGLS and maximum likelihood estimates for the model with cross-sectional correlation are given in Table 13.4. The estimated disturbance variances have changed dramatically, due in part to the quite large off-diagonal elements. It is noteworthy, however, that despite the large changes in $\hat{\boldsymbol{\Sigma}}$, with the exceptions of the MLE's in the cross section correlation model, the parameter estimates have not changed very much. (This sample is moderately large and all estimators are consistent, so this result is to be expected.)

We shall examine the effect of assuming that all five firms have the same slope parameters in Section 14.2.3. For now, we note that one of the effects is to inflate the disturbance correlations. When the Lagrange multiplier statistic in (13-68) is recomputed with firm-by-firm separate regressions, the statistic falls to 29.04, which is still significant, but far less than what we found earlier.

We now allow for different AR(1) disturbance processes for each firm. The firm specific autocorrelation coefficients of the ordinary least squares residuals are

$$\mathbf{r}' = (0.478 \quad -0.251 \quad 0.301 \quad 0.578 \quad 0.576).$$

[An interesting problem arises at this point. If one computes these autocorrelations using the standard formula, then the results can be substantially affected because the group-specific residuals may not have mean zero. Since the population mean *is* zero if the model is correctly specified, then this point is only minor. As we will explore later, however, this model *is not* correctly specified for these data. As such, the nonzero residual mean for the group specific residual vectors matters greatly. The vector of autocorrelations computed without using deviations from means is $\mathbf{r}_0 = (0.478, 0.793, 0.905, 0.602, 0.868)$. Three of the five are very different. Which way the computations should be done now becomes a substantive question. The asymptotic theory weighs in favor of (13-62). As a practical matter, in small or moderately sized samples such as this one, as this example demonstrates, the mean deviations are preferable.]

Table 13.4 also presents estimates for the groupwise heteroscedasticity model and for the full model with cross-sectional correlation, with the corrections for first-order autocorrelation. The lower part of the table displays the recomputed group specific variances and cross-group correlations.

13.9.8 SUMMARY

The preceding sections have suggested a variety of different specifications of the generalized regression model. Which ones apply in a given situation depends on the setting. Homoscedasticity will depend on the nature of the data and will often be directly observable at the outset. Uncorrelatedness across the cross-sectional units is a strong assumption, particularly because the model assigns the same parameter vector to all units. Autocorrelation is a qualitatively different property. Although it does appear to arise naturally in time-series data, one would want to look carefully at the data and the model specification before assuming that it is present. The properties of all these estimators depend on an increase in T, so they are generally not well suited to the types of data sets described in Sections 13.2–13.8.

Beck et al. (1993) suggest several problems that might arise when using this model in small samples. If $T < n$, then with or without a correction for autocorrelation, the matrix $\hat{\boldsymbol{\Sigma}}$ is an $n \times n$ matrix of rank T (or less) and is thus singular, which precludes FGLS estimation. A preferable approach then might be to use pooled OLS and make the appropriate correction to the asymptotic covariance matrix. But in this situation, there remains the possibility of accommodating cross unit heteroscedasticity. One could use the groupwise heteroscedasticity model. The estimators will be consistent and more efficient than OLS, although the standard errors will be inappropriate if there is cross-sectional correlation. An appropriate estimator that extends (11-17) would be

$$\begin{aligned}
\text{Est. Var}[\mathbf{b}] &= [\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}]^{-1}[\mathbf{X}'\hat{\mathbf{V}}^{-1}\hat{\boldsymbol{\Omega}}\hat{\mathbf{V}}^{-1}\mathbf{X}][\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}]^{-1} \\
&= \left[\sum_{i=1}^{n}\left(\frac{1}{\hat{\sigma}_{ii}}\right)\mathbf{X}_i'\mathbf{X}_i\right]^{-1}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\hat{\sigma}_{ij}}{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}\right)\mathbf{X}_i'\mathbf{X}_j\right]\left[\sum_{i=1}^{n}\left(\frac{1}{\hat{\sigma}_{ii}}\right)\mathbf{X}_i'\mathbf{X}_i\right]^{-1} \\
&= \left[\sum_{i=1}^{n}\left(\frac{1}{\hat{\sigma}_{ii}}\right)\mathbf{X}_i'\mathbf{X}_i\right]^{-1}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{r_{ij}^2}{\hat{\sigma}_{ij}}\right)\mathbf{X}_i'\mathbf{X}_j\right]\left[\sum_{i=1}^{n}\left(\frac{1}{\hat{\sigma}_{ii}}\right)\mathbf{X}_i'\mathbf{X}_i\right]^{-1}.
\end{aligned}$$

(Note that this estimator bases all estimates on the model of groupwise heteroscedasticity, but it is "robust" to the possibility of cross-sectional correlation.) When n is large relative to T, the number of estimated parameters in the autocorrelation model becomes very large relative to the number of observations. Beck and Katz (1995) found that as a consequence, the estimated asymptotic covariance matrix for the FGLS slopes tends to underestimate the true variability of the estimator. They suggest two compromises. First, use OLS and the appropriate covariance matrix, and second, impose the restriction of equal autocorrelation coefficients across groups.

13.10 SUMMARY AND CONCLUSIONS

The preceding 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 the heterogeneity across groups that is typical in microeconomic data.

We will find in Chapter 14 that to some extent this model of heterogeneity can be misleading. What might have appeared at one level to be differences in the variances of the disturbances across groups may well be due to heterogeneity of a different sort, associated with the coefficient vectors. We will consider this possibility in the next chapter. We will also examine some additional models for disturbance processes that arise naturally in a multiple equations context but are actually more general cases of some of the models we looked at above, such as the model of groupwise heteroscedasticity.

Key Terms and Concepts

- Arellano, Bond, and Bover estimator
- Between-groups estimator
- Contrasts
- Covariance structures
- Dynamic panel data model
- Feasible GLS
- Fixed effects model
- Generalized least squares
- GMM estimator
- Group means
- Group means estimator
- Groupwise heteroscedasticity
- Hausman test
- Hausman and Taylor estimator
- Heterogeneity
- Hierarchical regression
- Individual effect
- Instrumental variables estimator
- Least squares dummy variable model
- LM test
- LR test
- Longitudinal data sets
- Matrix weighted average
- Maximum likelihood
- Panel data
- Pooled regression
- Random coefficients
- Random effects model
- Robust covariance matrix
- Unbalanced panel
- Wald test
- Weighted average
- Within-groups estimator

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 (13-2), and then test the hypothesis that the constant term is the same for all three firms.
   c. Estimate the random effects model of (13-18), 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 model of (13-2) 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.
3. Use the data in Section 13.9.7 (the Grunfeld data) to fit the random and fixed effect models. There are five firms and 20 years of data for each. Use the F, LM, and/or Hausman statistics to determine which model, the fixed or random effects model, is preferable for these data.
4. Derive the log-likelihood function for the model in (13-18), assuming that ε_{it} and u_i are normally distributed. [Hints: Write the log-likelihood function as $\ln L = \sum_{i=1}^{n} \ln L_i$, where $\ln L_i$ is the log-likelihood function for the T observations in group i. These T observations are joint normally distributed, with covariance matrix given in (13-20). The log-likelihood is the sum of the logs of the joint normal densities of the n sets of T observations,

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

This step will involve the inverse and determinant of $\boldsymbol{\Omega}$. Use (B-66) to prove that

$$\boldsymbol{\Omega}^{-1} = \frac{1}{\sigma_\varepsilon^2}\left[\mathbf{I} - \frac{\sigma_u^2}{\sigma_\varepsilon^2 + T\sigma_u^2}\mathbf{i}_T\mathbf{i}_T'\right].$$

To find the determinant, use the product of the characteristic roots. Note first that

$|\sigma_\varepsilon^2\mathbf{I} + \sigma_u^2\mathbf{ii}'| = (\sigma_\varepsilon^2)^T|\mathbf{I} + \frac{\sigma_u^2}{\sigma_\varepsilon^2}\mathbf{ii}'|$. The roots are determined by

$$\left[\mathbf{I} + \frac{\sigma_u^2}{\sigma_\varepsilon^2}\mathbf{ii}'\right]\mathbf{c} = \lambda\mathbf{c} \quad \text{or} \quad \frac{\sigma_u^2}{\sigma_\varepsilon^2}\mathbf{ii}'\mathbf{c} = (\lambda - 1)\mathbf{c}.$$

Any vector whose elements sum to zero is a solution. There are $T - 1$ such independent vectors, so $T - 1$ characteristic roots are $(\lambda - 1) = 0$ or $\lambda = 1$. Premultiply the expression by $\mathbf{i}'$ to obtain the remaining characteristic root. (Remember to add one to the result.) Now, collect terms to obtain the log-likelihood.]

5. *Unbalanced design for random effects.* Suppose that the random effects model of Section 13.4 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 in (13-11) is unbiased and consistent despite this complication.
   b. Show that the estimator in (13-29) 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.
6. What are the probability limits of $(1/n)$LM, where LM is defined in (13-31) under the null hypothesis that $\sigma_u^2 = 0$ and under the alternative that $\sigma_u^2 \neq 0$?
7. 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 $y_{it} = \alpha_i + \gamma_t + \boldsymbol{\beta}'\mathbf{x}_{it} + \varepsilon_{it}$. At every observation, the individual- and time-specific dummy variables sum to 1, so there are some redundant coefficients. The discussion in Section 13.3.3 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,

$$y_{it} = \mu + (\alpha_i - \alpha_1) + (\gamma_t - \gamma_1) + \boldsymbol{\beta}'\mathbf{x}_{it} + \varepsilon_{it}.$$

(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 (13-17) while $m = \bar{\bar{y}} - \mathbf{b}'\bar{\bar{\mathbf{x}}}$. 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.

8. *Two-way random effects model.* We modify the random effects model by the addition of a time specific disturbance. Thus,

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

where

$$\begin{aligned} E[\varepsilon_{it}] &= E[u_i] = E[v_t] = 0, \\ E[\varepsilon_{it}u_j] &= E[\varepsilon_{it}v_s] = E[u_i v_t] = 0 \quad \text{for all } i, j, t, s \\ \text{Var}[\varepsilon_{it}] &= \sigma^2, \quad \text{Cov}[\varepsilon_{it}, \varepsilon_{js}] = 0 \quad \text{for all } i, j, t, s \\ \text{Var}[u_i] &= \sigma_u^2, \quad \text{Cov}[u_i, u_j] = 0 \quad \text{for all } i, j \\ \text{Var}[v_t] &= \sigma_v^2, \quad \text{Cov}[v_t, v_s] = 0 \quad \text{for all } t, s. \end{aligned}$$

Write out the full covariance matrix for a data set with $n = 2$ and $T = 2$.

9. The model

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \beta + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix}$$

satisfies the groupwise heteroscedastic regression model of Section 11.7.2. All variables have zero means. The following sample second-moment matrix is obtained from a sample of 20 observations:

$$\begin{array}{c} \\ y_1 \\ y_2 \\ x_1 \\ x_2 \end{array} \begin{array}{c} \begin{array}{cccc} y_1 & y_2 & x_1 & x_2 \end{array} \\ \begin{bmatrix} 20 & 6 & 4 & 3 \\ 6 & 10 & 3 & 6 \\ 4 & 3 & 5 & 2 \\ 3 & 6 & 2 & 10 \end{bmatrix} \end{array}.$$

a. Compute the two separate OLS estimates of β, their sampling variances, the estimates of σ_1^2 and σ_2^2, and the R^2's in the two regressions.
b. Carry out the Lagrange multiplier test of the hypothesis that $\sigma_1^2 = \sigma_2^2$.
c. Compute the two-step FGLS estimate of β and an estimate of its sampling variance. Test the hypothesis that β equals 1.
d. Carry out the Wald test of equal disturbance variances.
e. Compute the maximum likelihood estimates of β, σ_1^2, and σ_2^2 by iterating the FGLS estimates to convergence.
f. Carry out a likelihood ratio test of equal disturbance variances.
g. Compute the two-step FGLS estimate of β, assuming that the model in (14-7) applies. (That is, allow for cross-sectional correlation.) Compare your results with those of part c.

10. Suppose that in the groupwise heteroscedasticity model of Section 11.7.2, $\mathbf{X}_i$ is the same for all i. What is the generalized least squares estimator of $\boldsymbol{\beta}$? How would you compute the estimator if it were necessary to estimate σ_i^2?
11. Repeat Exercise 10 for the cross sectionally correlated model of Section 13.9.1.

12. The following table presents a hypothetical panel of data:

	$i = 1$		$i = 2$		$i = 3$	
t	y	x	y	x	y	x
1	30.27	24.31	38.71	28.35	37.03	21.16
2	35.59	28.47	29.74	27.38	43.82	26.76
3	17.90	23.74	11.29	12.74	37.12	22.21
4	44.90	25.44	26.17	21.08	24.34	19.02
5	37.58	20.80	5.85	14.02	26.15	18.64
6	23.15	10.55	29.01	20.43	26.01	18.97
7	30.53	18.40	30.38	28.13	29.64	21.35
8	39.90	25.40	36.03	21.78	30.25	21.34
9	20.44	13.57	37.90	25.65	25.41	15.86
10	36.85	25.60	33.90	11.66	26.04	13.28

a. Estimate the groupwise heteroscedastic model of Section 11.7.2. Include an estimate of the asymptotic variance of the slope estimator. Use a two-step procedure, basing the FGLS estimator at the second step on residuals from the pooled least squares regression.
b. Carry out the Wald, Lagrange multiplier, and likelihood ratio tests of the hypothesis that the variances are all equal. For the likelihood ratio test, use the FGLS estimates.
c. Carry out a Lagrange multiplier test of the hypothesis that the disturbances are uncorrelated across individuals.