Chapter Fourteen

Advanced Panel Data Methods

In this chapter, we cover two methods for estimating unobserved effects panel data models that are at least as common as first differencing. Although these methods are somewhat harder to describe and implement, several econometrics packages support them.

In Section 14.1, we discuss the fixed effects estimator, which, like first differencing, uses a transformation to remove the unobserved effect a_i prior to estimation. Any time-constant explanatory variables are removed along with a_i.

The random effects estimator is attractive when we think the unobserved effect is uncorrelated with all the explanatory variables. If we have good controls in our equation, we might believe that any leftover neglected heterogeneity only induces serial correlation in the composite error term, but it does not cause correlation between the composite errors and the explanatory variables. Estimation of random effects models by generalized least squares is fairly easy and is routinely done by many econometrics packages.

In Section 14.3, we show how panel data methods can be applied to other data structures, including matched pairs and cluster samples.

14.1 FIXED EFFECTS ESTIMATION

First differencing is just one of the many ways to eliminate the fixed effect, a_i. An alternative method, which works better under certain assumptions, is called the **fixed effects transformation**. To see what this method involves, consider a model with a single explanatory variable: for each i,

$$y_{it} = \beta_1 x_{it} + a_i + u_{it}, \; t = 1,2,\ldots,T. \tag{14.1}$$

Now, for each i, average this equation over time. We get

$$\bar{y}_i = \beta_1 \bar{x}_i + a_i + \bar{u}_i, \tag{14.2}$$

where $\bar{y}_i = T^{-1}\sum_{t=1}^{T} y_{it}$, and so on. Because a_i is fixed over time, it appears in both (14.1) and (14.2). If we substract (14.2) from (14.1) for each t, we wind up with

$$y_{it} - \bar{y}_i = \beta_1(x_{it} - \bar{x}_i) + u_{it} - \bar{u}_i, \; t = 1,2,\ldots,T,$$

or

$$\ddot{y}_{it} = \beta_1\ddot{x}_{it} + \ddot{u}_{it}, \; t = 1,2,\ldots,T, \qquad \textbf{(14.3)}$$

where $\ddot{y}_{it} = y_{it} - \bar{y}_i$ is the **time-demeaned data** on y, and similarly for $\ddot{x}_{it}$ and $\ddot{u}_{it}$. The fixed effects transformation is also called the **within transformation**. The important thing about equation (14.3) is that the unobserved effect, a_i, has disappeared. This suggests that we estimate (14.3) by pooled OLS. A pooled OLS estimator that is based on the time-demeaned variables is called the **fixed effects estimator** or the **within estimator**. The latter name comes from the fact that OLS on (14.3) uses the time variation in y and x *within* each cross-sectional observation.

The *between estimator* is obtained as the OLS estimator on the cross-sectional equation (14.2) (where we include an intercept, β_0): we use the time-averages for both y and x and then run a cross-sectional regression. We will not study the between estimator in detail because it is biased when a_i is correlated with x_i (see Problem 14.2). If we think a_i is uncorrelated with x_{it}, it is better to use the random effects estimator, which we cover in Section 14.2. The between estimator ignores important information on how the variables change over time.

Adding more explanatory variables to the equation causes few changes. The original model is

$$y_{it} = \beta_1 x_{it1} + \beta_2 x_{it2} + \ldots + \beta_k x_{itk} + a_i + u_{it}, \; t = 1,2,\ldots,T. \qquad \textbf{(14.4)}$$

We simply use the time-demeaning on each explanatory variable—including things like time period dummies—and then do a pooled OLS regression using all time-demeaned variables. The general time-demeaned equation for each i is

$$\ddot{y}_{it} = \beta_1\ddot{x}_{it1} + \beta_2\ddot{x}_{it2} + \ldots + \beta_k\ddot{x}_{itk} + \ddot{u}_{it}, \; t = 1,2,\ldots,T, \qquad \textbf{(14.5)}$$

which we estimate by pooled OLS.

Under a strict exogeneity assumption on the explanatory variables, the fixed effects estimator is unbiased: roughly, the idiosyncratic error u_{it} should be uncorrelated with each explanatory variable across *all* time periods. (See the chapter appendix for precise statements of the assumptions.) The fixed effects estimator allows for arbitrary correlation between a_i and the explanatory variables in any time period, just as with first differencing. Because of this, any explanatory variable that is constant over time for all i gets swept away by the fixed effects transformation: $\ddot{x}_{it} = 0$ for all i and t, if x_{it} is constant across t. Therefore, we cannot include variables such as gender or whether a city is located near a river.

The other assumptions needed for a straight OLS analysis to be valid are that the errors u_{it} are homoskedastic and serially uncorrelated (across t); see the appendix to this chapter.

QUESTION 14.1

Suppose that in a family savings equation, for the years 1990, 1991, and 1992, we let $kids_{it}$ denote the number of children in family i for year t. If the number of kids is constant over this three-year period for most families in the sample, what problems might this cause for estimating the effect that the number of kids has on savings?

There is one subtle point in determining the degrees of freedom for the fixed effects estimator. When we estimate the time-demeaned equation (14.5) by pooled OLS, we have NT total observations and k independent variables. [Notice that there is no intercept in (14.5); it is eliminated by the fixed effects transformation.] Therefore, we should apparently have $NT - k$ degrees of freedom. This calculation is incorrect. For each cross-sectional observation i, we lose one df because of the time-demeaning. In other words, for each i, the demeaned errors $\ddot{u}_{it}$ add up to zero when summed across t, so we lose one degree of freedom. (There is no such constraint on the original idiosyncratic errors u_{it}.) Therefore, the appropriate degrees of freedom is $df = NT - N - k = N(T - 1) - k$. Fortunately, modern regression packages that have a fixed effects estimation feature properly compute the df. But if we have to do the time-demeaning and the estimation by pooled OLS ourselves, we need to correct the standard errors and test statistics.

EXAMPLE 14.1

(Effect of Job Training on Firm Scrap Rates)

We use the data for the three years, 1987, 1988, and 1989 on the 54 firms that reported scrap rates in each year. No firms received grants prior to 1988; in 1988, 19 firms received grants; in 1989, 10 different firms received grants. Therefore, we must also allow for the possibility that the additional job training in 1988 made workers more productive in 1989. This is easily done by including a lagged value of the grant indicator. We also include year dummies for 1988 and 1989. The results are given in Table 14.1:

Table 14.1

Fixed Effects Estimation of the Scrap Rate Equation

Dependent Variable: log*(scrap)*	
Independent Variables	
d88	−.080 (.109)
d89	−.247 (.133)
grant	−.252 (.151)
$grant_{-1}$	−.422 (.210)
Observations Degrees of Freedom *R*-Squared	162 104 .201

We have reported the results in a way that emphasizes the need to interpret the estimates in light of the unobserved effects model, (14.4). We are explicitly controlling for the unobserved, time-constant effects in a_i. The time-demeaning allows us to estimate the β_j, but (14.5) is not the best equation for interpreting the estimates.

Interestingly, the estimated lagged effect of the training grant is substantially larger than the contemporaneous effect: job training has an effect at least one year later. Because the dependent variable is in logarithmic form, obtaining a grant in 1988 is predicted to lower the firm scrap rate in 1989 by about 34.4% [exp(−.422) − 1 ≈ −.344]; the coefficient on $grant_{-1}$ is significant at the 5% level against a two-sided alternative. The coefficient on *grant* is significant at the 10% level, and the size of the coefficient is hardly trivial. Notice the *df* is obtained as $N(T - 1) - k = 54(3 - 1) - 4 = 104$.

The coefficient on *d89* indicates that the scrap rate was substantially lower in 1989 than in the base year, 1987, even in the absence of job training grants. Thus, it is important to allow for these aggregate effects. If we omitted the year dummies, the secular increase in worker productivity would be attributed to the job training grants. Table 14.1 shows that, even after controlling for aggregate trends in productivity, the job training grants had a large estimated effect.

QUESTION 14.2

Under the Michigan program, if a firm received a grant in one year, it was not eligible for a grant the following year. What does this imply about the correlation between *grant* and $grant_{-1}$?

Finally, it is crucial to allow for the lagged effect in the model. If we omit $grant_{-1}$, then we are assuming that the effect of job training does not last into the next year. The estimate on *grant* when we drop $grant_{-1}$ is −.082 ($t = -.65$); this is much smaller and statistically insignificant.

When estimating an unobserved effects model by fixed effects, it is not clear how we should compute a goodness-of-fit measure. The *R*-squared given in Table 14.1 is based on the within transformation: it is the *R*-squared obtained from estimating (14.5). Thus, it is interpreted as the amount of time variation in the y_{it} that is explained by the time variation in the explanatory variables. Other ways of computing *R*-squared are possible, one of which we discuss later.

Although time-constant variables cannot be included by themselves in a fixed effects model, they *can* be interacted with variables that change over time and, in particular, with year dummy variables. For example, in a wage equation where education is constant over time for each individual in our sample, we can interact education with each year dummy to see how the return to education has changed over time. But we cannot use fixed effects to estimate the return to education in the base period—which means we cannot estimate the return to education in any period—we can only see how the return to education in each year differs from that in the base period.

When we include a full set of year dummies—that is, year dummies for all years but the first—we cannot estimate the effect of any variable whose *change* across time is constant. An example is years of experience in a panel data set where each person works in every year, so that experience always increases by one in each year, for every person in the sample. The presence of a_i accounts for differences across people in their

years of experience in the initial time period. But then the effect of a one-year increase in experience cannot be distinguished from the aggregate time effects (because experience increases by the same amount for everyone). This would also be true if, in place of separate year dummies, we used a linear time trend: for each person, experience cannot be distinguished from a linear trend.

EXAMPLE 14.2
(Has the Return to Education Changed Over Time?)

The data in WAGEPAN.RAW are from Vella and Verbeek (1998). Each of the 545 men in the sample worked in every year from 1980 through 1987. Some variables in the data set change over time: experience, marital status, and union status are the three important ones. Other variables do not change: race and education are the key examples. If we use fixed effects (or first differencing), we cannot include race, education, or experience in the equation. However, we can include interactions of *educ* with year dummies for 1981 through 1987 to test whether the return to education was constant over this time period. We use log(*wage*) as the dependent variable, a quadratic in experience, dummy variables for marital and union status, a full set of year dummies, and the interaction terms *d81·educ*, *d82·educ*, …, *d87·educ*.

The estimates on these interaction terms are all positive, and they generally get larger for more recent years. The largest coefficient of .030 is on *d87·educ*, with $t = 2.48$. In other words, the return to education is estimated to be about 3 percentage points larger in 1987 than in the base year, 1980. (We do not have an estimate of the return to education in the base year for the reasons given earlier.) The other significant interaction term is *d86·educ* (coefficient = .027, $t = 2.23$). The estimates on the earlier years are smaller and insignificant at the 5% level against a two-sided alternative. If we do a joint F test for significance of all seven interaction terms, we get p-value = .28: this gives an example where a set of variables is jointly insignificant even though some variables are individually significant. [The *df* for the F test are 7 and 3,799; the second of these comes from $N(T - 1) - k = 545(8 - 1) - 16 = 3{,}799$.] Generally, the results are consistent with an increase in the return to education over this period.

The Dummy Variable Regression

A traditional view of the fixed effects model is to assume that the unobserved effect, a_i, is a parameter to be estimated for each i. Thus, in equation (14.4), a_i is the intercept for person i (or firm i, city i, and so on) that is to be estimated along with the β_j. (Clearly we cannot do this with a single cross section: there would be $N + k$ parameters to estimate with only N observations. We need at least two time periods.) The way we estimate an intercept for each i is to put in a dummy variable for each cross-sectional observation, along with the explanatory variables (and probably dummy variables for each time period). This method is usually called the **dummy variable regression**. Even when N is not very large (say, $N = 54$ as in Example 14.1), this results in many explanatory variables—in most cases, too many to explicitly carry out the regression. Thus, the

dummy variable method is not very practical for panel data sets with many cross-sectional observations.

Nevertheless, the dummy variable regression has some interesting features. Most importantly, it gives us *exactly* the same estimates of the β_j that we would obtain from the regression on time-demeaned data, and the standard errors and other major statistics are identical. Therefore, the fixed effects estimator can be obtained by the dummy variable regression. One benefit of the dummy variable regression is that it properly computes the degrees of freedom directly. This is a minor advantage now that many econometrics packages have programmed fixed effects options.

The *R*-squared from the dummy variable regression is usually rather high. This is because we are including a dummy variable for each cross-sectional unit, which explains much of the variation in the data. For example, if we estimate the unobserved effects model in Example 13.8 by fixed effects using the dummy variable regression (which is possible with $N = 22$), then $R^2 = .933$. We should not get too excited about this large *R*-squared: it is not surprising that we can explain much of the variation in unemployment claims using both year and city dummies. Just as in Example 13.8, the estimate on the EZ dummy variable is more important than R^2.

The *R*-squared from the dummy variable regression can be used to compute *F* tests in the usual way, assuming of course that the classical linear model assumptions hold (see the chapter appendix). In particular, we can test the joint significance of all of the cross-sectional dummies ($N - 1$, since one unit is chosen as the base group). The unrestricted *R*-squared is obtained from the regression with all of the cross-sectional dummies; the restricted *R*-squared omits these. In the vast majority of applications, the dummy variables will be jointly significant.

Occasionally, the estimated intercepts, say $\hat{a}_i$, are of interest. This is the case if we want to study the distribution of the $\hat{a}_i$ across *i*, or if we want to pick a particular firm or city to see whether its $\hat{a}_i$ is above or below the average value in the sample. These estimates are directly available from the dummy variable regression, but they are rarely reported by packages that have fixed effects routines (for the practical reason that there are so many $\hat{a}_i$). After fixed effects estimation with *N* of any size, the $\hat{a}_i$ are pretty easy to compute:

$$\hat{a}_i = \bar{y}_i - \hat{\beta}_1\bar{x}_{i1} - \ldots - \hat{\beta}_k\bar{x}_{ik},\ i = 1, \ldots, N, \qquad \textbf{(14.6)}$$

where the overbar refers to the time averages and the $\hat{\beta}_j$ are the fixed effects estimates. For example, if we have estimated a model of crime while controlling for various time-varying factors, we can obtain $\hat{a}_i$ for a city to see whether the unobserved fixed effects that contribute to crime are above or below average.

In most studies, the $\hat{\beta}_j$ are of interest, and so the time-demeaned equations are used to obtain these estimates. Further, it is usually best to view the a_i as omitted variables that we control for through the within transformation. The sense in which the a_i can be estimated is generally weak. In fact, even though $\hat{a}_i$ is unbiased (under assumptions FE.1 through FE.4 in the chapter appendix), it is not consistent with a fixed *T* as $N \to \infty$. The reason is that, as we add each additional cross-sectional observation, we add a new a_i. No information accumulates on each a_i when *T* is fixed. With larger *T*, we can get better estimates of the a_i, but most panel data sets are of the large *N* and small *T* variety.

Fixed Effects or First Differencing?

So far, we have seen two methods for estimating unobserved effects models. One involves differencing the data, and the other involves time-demeaning. How do we know which one to use?

We can eliminate one case immediately: when $T = 2$, the FE and FD estimates and all test statistics are *identical*, and so it does not matter which we use. First differencing has the advantage of being straightforward in virtually any econometrics package, and it is easy to compute heteroskedasticity-robust statistics in the FD regression.

When $T \geq 3$, the FE and FD estimators are not the same. Since both are unbiased under Assumptions FE.1 through FE.4, we cannot use unbiasedness as a criterion. Further, both are consistent (with T fixed as $N \rightarrow \infty$) under FE.1 through FE.4. For large N and small T, the choice between FE and FD hinges on the relative efficiency of the estimators, and this is determined by the serial correlation in the idiosyncratic errors, u_{it}. (We will assume homoskedasticity of the u_{it}, since efficiency comparisons require homoskedastic errors.)

When the u_{it} are serially uncorrelated, fixed effects is more efficient than first differencing (and the standard errors reported from fixed effects are valid). Since the fixed effects model is almost always stated with serially uncorrelated idiosyncratic errors, the FE estimator is used more often. But we should remember that this assumption can be false. In many applications, we can expect the unobserved factors that change over time to be serially correlated. If u_{it} follows a random walk—which means that there is very substantial, positive serial correlation—then the difference Δu_{it} is serially uncorrelated, and first differencing is better. In many cases, the u_{it} exhibit some positive serial correlation, but perhaps not as much as a random walk. Then, we cannot easily compare the efficiency of the FE and FD estimators.

It is difficult to test whether the u_{it} are serially uncorrelated after FE estimation: we can estimate the time-demeaned errors, $\ddot{u}_{it}$, but not the u_{it}. However, in Section 13.3, we showed how to test whether the differenced errors, Δu_{it}, are serially uncorrelated. If this seems to be the case, FD can be used. If there is substantial negative serial correlation in the Δu_{it}, FE is probably better. It is often a good idea to try both: if the results are not sensitive, so much the better.

When T is large, and especially when N is not very large (for example, $N = 20$ and $T = 30$), we must exercise caution in using the fixed effects estimator. While exact distributional results hold for any N and T under the classical fixed effects assumptions, they are extremely sensitive to violations of the assumptions when N is small and T is large. In particular, if we are using unit root processes—see Chapter 11—the spurious regression problem can arise. As we saw in Chapter 11, differencing an integrated process results in a weakly dependent process, and we must appeal to the central limit approximations. In this case, using differences is favorable.

On the other hand, fixed effects turns out to be less sensitive to violation of the strict exogeneity assumption, especially with large T. Some authors even recommend estimating fixed effects models with lagged dependent variables (which clearly violates Assumption FE.3 in the chapter appendix). When the processes are weakly dependent over time and T is large, the bias in the fixed effects estimator can be small [see, for example, Wooldridge (1999, Chapter 11)].

It is difficult to choose between FE and FD when they give substantively different results. It makes sense to report both sets of results and to try to determine why they differ.

Fixed Effects with Unbalanced Panels

Some panel data sets, especially on individuals or firms, have missing years for at least some cross-sectional units in the sample. In this case, we call the data set an **unbalanced panel**. The mechanics of fixed effects estimation with an unbalanced panel are not much more difficult than with a balanced panel. If T_i is the number of time periods for cross-sectional unit i, we simply use these T_i observations in doing the time-demeaning. The total number of observations is then $T_1 + T_2 + \ldots + T_N$. As in the balanced case, one degree of freedom is lost for every cross-sectional observation due to the time-demeaning. Any regression package that does fixed effects makes the appropriate adjustment for this loss. The dummy variable regression also goes through in exactly the same way as with a balanced panel, and the *df* is appropriately obtained.

It is easy to see that units for which we have only a single time period play no role in a fixed effects analysis. The time-demeaning for such observations yields all zeros, which are not used in the estimation. (If T_i is at most two for all i, we can use first differencing: if $T_i = 1$ for any i, we do not have two periods to difference.)

The more difficult issue with an unbalanced panel is determining why the panel is unbalanced. With cities and states, for example, data on key variables are sometimes missing for certain years. Provided the reason we have missing data for some i is not correlated with the idiosyncratic errors, u_{it}, the unbalanced panel causes no problems. When we have data on individuals, families, or firms, things are trickier. Imagine, for example, that we obtain a random sample of manufacturing firms in 1990, and we are interested in testing how unionization affects firm profitability. Ideally, we can use a panel data analysis to control for unobserved worker and management characteristics that affect profitability and might also be correlated with the fraction of the firm's work force that is unionized. If we collect data again in subsequent years, some firms may be lost because they have gone out of business or have merged with other companies. If so, we probably have a nonrandom sample in subsequent time periods. The question is: If we apply fixed effects to the unbalanced panel, when will the estimators be unbiased (or at least consistent)?

If the reason a firm leaves the sample (called *attrition*) is correlated with the idiosyncratic error—those unobserved factors that change over time and affect profits—then the resulting sample section problem (see Chapter 9) can cause biased estimators. This is a serious consideration in this example. Nevertheless, one useful thing about a fixed effects analysis is that it *does* allow attrition to be correlated with a_i, the unobserved effect. The idea is that, with the initial sampling, some units are more likely to drop out of the survey, and this is captured by a_i.

EXAMPLE 14.3
(Effect of Job Training on Firm Scrap Rates)

We add two variables to the analysis in Table 14.1: log($sales_{it}$) and log($employ_{it}$), where *sales* is annual firm sales and *employ* is number of employees. Three of the 54 firms drop out of

the analysis entirely because they do not have sales or employment data. Five additional observations are lost due to missing data on one or both of these variables for some years, leaving us with $n = 148$. Using fixed effects on the unbalanced panel does not change the basic story, although the estimated grant effect gets larger: $\hat{\beta}_{grant} = -.297$, $t_{grant} = -1.89$; $\hat{\beta}_{grant_{-1}} = -.536$, $t_{grant_{-1}} = -2.389$.

Solving attrition problems in panel data is complicated and beyond the scope of this text. [See, for example, Wooldridge (1999, Chapter 17).]

14.2 RANDOM EFFECTS MODELS

We begin with the same unobserved effects model as before,

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it}, \tag{14.7}$$

where we explicitly include an intercept so that we can make the assumption that the unobserved effect, a_i, has zero mean (without loss of generality). We would usually allow for time dummies among the explanatory variables as well. In using fixed effects or first differencing, the goal is to eliminate a_i because it is thought to be correlated with one or more of the x_{itj}. But suppose we think a_i is *uncorrelated* with each explanatory variable in all time periods? Then, using a transformation to eliminate a_i results in inefficient estimators.

Equation (14.7) becomes a **random effects model** when we assume that the unobserved effect a_i is uncorrelated with each explanatory variable:

$$\text{Cov}(x_{itj}, a_i) = 0, \; t = 1,2,\ldots,T; \; j = 1,2,\ldots,k. \tag{14.8}$$

In fact, the ideal random effects assumptions include all of the fixed effects assumptions plus the additional requirement that a_i is independent of all explanatory variables in all time periods. (See the chapter appendix for the actual assumptions used.) If we think the unobserved effect a_i is correlated with any explanatory variables, we should use first differencing for fixed effects.

Under (14.8) and along with the random effects assumptions, how should we estimate the β_j? It is important to see that, if we believe that a_i is uncorrelated with the explanatory variables, the β_j can be consistently estimated by using a single cross section: there is no need for panel data at all. But using a single cross section disregards much useful information in the other time periods. We can use this information in a pooled OLS procedure: just run OLS of y_{it} on the explanatory variables and probably the time dummies. This, too, produces consistent estimators of the β_j under the random effects assumption. But it ignores a key feature of the model. If we define the **composite error term** as $v_{it} = a_i + u_{it}$, then (14.7) can be written as

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + v_{it}. \tag{14.9}$$

Because a_i is in the composite error in each time period, the v_{it} are serially correlated across time. In fact, under the random effects assumptions,

$$\text{Corr}(v_{it},v_{is}) = \sigma_a^2/(\sigma_a^2 + \sigma_u^2),\ t \neq s,$$

where $\sigma_a^2 = \text{Var}(a_i)$ and $\sigma_u^2 = \text{Var}(u_{it})$. This (necessarily) positive serial correlation in the error term can be substantial: because the usual pooled OLS standard errors ignore this correlation, they will be incorrect, as will the usual test statistics. In Chapter 12, we showed how generalized least squares can be used to estimate models with autoregressive serial correlation. We can also use GLS to solve the serial correlation problem here. In order for the procedure to have good properties, it must have large N and relatively small T. We assume that we have a balanced panel, although the method can be extended to unbalanced panels.

Deriving the GLS transformation that eliminates serial correlation in the errors requires sophisticated matrix algebra [see, for example, Wooldridge (1999) Chapter 10]. But the transformation itself is simple. Define

$$\lambda = 1 - [\sigma_u^2/(\sigma_u^2 + T\sigma_a^2)]^{1/2}, \tag{14.10}$$

which is between zero and one. Then, the transformed equation turns out to be

$$\begin{aligned} y_{it} - \lambda\bar{y}_i = \beta_0(1 - \lambda) + \beta_1(x_{it1} - \lambda\bar{x}_{i1}) + \ldots \\ + \beta_k(x_{itk} - \lambda\bar{x}_{ik}) + (v_{it} - \lambda\bar{v}_i), \end{aligned} \tag{14.11}$$

where the overbar again denotes the time averages. This is a very interesting equation, as it involves a **quasi-demeaned data** on each variable. The fixed effects estimator subtracts the time averages from the corresponding variable. The random effects transformation subtracts a fraction of that time average, where the fraction depends on σ_u^2, σ_a^2, and the number of time periods, T. The GLS estimator is simply the pooled OLS estimator of equation (14.11). It is hardly obvious that the errors in (14.11) are serially uncorrelated, but they are.

The transformation in (14.11) allows for explanatory variables that are constant over time, and this is one advantage of random effects (RE) over either fixed effects or first differencing. This is possible because RE assumes that the unobserved effect is uncorrelated with all explanatory variables, whether they are fixed over time or not. Thus, in a wage equation, we can include a variable such as education even if it does not change over time. But we are assuming that education is uncorrelated with a_i, which contains ability and family background. In many applications, the whole reason for using panel data is to allow the unobserved effect to be correlated with the explanatory variables.

The parameter λ is never known in practice, but it can always be estimated. There are different ways to do this, which may be based on pooled OLS or fixed effects, for example. Generally, $\hat{\lambda}$ takes the form $\hat{\lambda} = 1 - \{1/[1 + T(\hat{\sigma}_a^2/\hat{\sigma}_u^2)]\}^{1/2}$, where $\hat{\sigma}_a^2$ is a consistent estimator of σ_a^2 and $\hat{\sigma}_u^2$ is a consistent estimator of σ_u^2. These estimators can be based on the pooled OLS or fixed effects residuals. One possibility is that $\hat{\sigma}_a^2 = [NT(T-1)/2 - k]^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} \hat{v}_{it}\hat{v}_{is}$, where the $\hat{v}_{it}$ are the residuals from esti-

mating (14.9) by pooled OLS. Given this, we can estimate σ_u^2 by using $\hat{\sigma}_u^2 = \hat{\sigma}_v^2 - \hat{\sigma}_a^2$, where $\hat{\sigma}_v^2$ is the square of the usual standard error of the regression from pooled OLS. [See Wooldridge (1999, Chapter 10) for additional discussion of these estimators.]

Many econometrics packages support estimation of random effects models and automatically compute some version of $\hat{\lambda}$. The feasible GLS estimator that uses $\hat{\lambda}$ in place of λ is called the **random effects estimator**. Under the random effects assumptions in the chapter appendix, the estimator is consistent (not unbiased) and asymptotically normally distributed as N gets large with fixed T. The properties of the RE estimator with small N and large T are largely unknown, although it has certainly been used in such situations.

Equation (14.11) allows us to relate the RE estimator to both pooled OLS and fixed effects. Pooled OLS is obtained when $\lambda = 0$, and FE is obtained when $\lambda = 1$. In practice, the estimate $\hat{\lambda}$ is never zero or one. But if $\hat{\lambda}$ is close to zero, the RE estimates will be close to the pooled OLS estimates. This is the case when the unobserved effect, a_i, is relatively unimportant (since it has small variance relative to σ_u^2). It is more common for σ_a^2 to be large relative to σ_u^2, in which case $\hat{\lambda}$ will be closer to unity. As T gets large, $\hat{\lambda}$ tends to one, and this makes the RE and FE estimates very similar.

EXAMPLE 14.4
(A Wage Equation Using Panel Data)

We again use the data in WAGEPAN.RAW to estimate a wage equation for men. We use three methods: pooled OLS, random effects, and fixed effects. In the first two methods, we can include *educ* and race dummies (*black* and *hispan*), but these drop out of the fixed effects analysis. The time-varying variables are *exper*, *exper*^{2}, *union*, and *married*. As we discussed in Section 14.1, *exper* is dropped in the FE analysis (but *exper*^{2} remains). Each regression also contains a full set of year dummies. The estimation results are in Table 14.2.

The coefficients on *educ*, *black*, and *hispan* are similar for the pooled OLS and random effects estimations. The pooled OLS standard errors are the usual OLS standard errors, and these underestimate the true standard errors because they ignore the positive serial correlation; we report them here for comparison only. The experience profile is somewhat different, and both the marriage and union premiums fall notably in the random effects estimation. When we eliminate the unobserved effect entirely by using fixed effects, the marriage premium falls to about 4.7%, although it is still statistically significant. The drop in the marriage premium is consistent with the idea that men who are more able—as captured by a higher unobserved effect, a_i—are more likely to be married. Therefore, in the pooled OLS estimation, a large part of the marriage premium reflects the fact that men who are married would earn more even if they were not married. The remaining 4.7% has at least two possible explanations: (1) marriage really makes men more productive or (2) employers pay married men a premium because marriage is a signal of stability. We cannot distinguish between these two hypotheses.

QUESTION 14.3

The union premium estimated by fixed effects is about 10 percentage points lower than the OLS estimate. What does this strongly suggest about the correlation between *union* and the unobserved effect?

Table 14.2

Three Different Estimators of a Wage Equation

Dependent Variable: log*(wage)*			
Independent Variables	**Pooled OLS**	**Random Effects**	**Fixed Effects**
educ	.091 (.005)	.092 (.011)	———
black	−.139 (.024)	−.139 (.048)	———
hispan	.016 (.021)	.022 (.043)	———
exper	.067 (.014)	.106 (.015)	———
$exper^2$	−.0024 (.0008)	−.0047 (.0007)	−.0052 (.0007)
married	.108 (.016)	.064 (.017)	.047 (.018)
union	.182 (.017)	.106 (.018)	.080 (.019)

The estimate of λ for the random effects estimation is $\hat{\lambda} = .643$, which explains why, on the time-varying variables, the RE estimates lie closer to the FE estimates than to the pooled OLS estimates.

Random Effects or Fixed Effects?

In reading empirical work, you may find that authors decide between fixed and random effects based on whether the a_i (or whatever notation the authors use) are best viewed as parameters to be estimated or as outcomes of a random variable. When we cannot consider the observations to be random draws from a large population—for example, if we have data on states or provinces—it often makes sense to think of the a_i as parameters to estimate, in which case we use fixed effects methods. Remember that using fixed effects is the same as allowing a different intercept for each observation, and we can estimate these intercepts by including dummy variables or by (14.6).

Even if we decide to treat the a_i as random variables, we must decide whether the a_i are uncorrelated with the explanatory variables. People sometimes mistakenly believe that assuming a_i is random automatically means that random effects is the appropriate estimation strategy. *If* we can assume the a_i are uncorrelated with all x_{it}, then the random effects method is appropriate. But if the a_i are correlated with some explanatory variables, the fixed effects method (or first differencing) is needed; if RE is used, then the estimators are generally inconsistent.

Comparing the FE and RE estimates can be a test for whether there is correlation between the a_i and the x_{itj}, assuming that the idiosyncratic errors and explanatory variables are uncorrelated across all time periods. Hausman (1978) first suggested this test. Some econometrics packages routinely compute the test under the ideal random effects assumptions listed in the chapter appendix. Details on this statistic can be found in Wooldridge (1999, Chapter 10).

14.3 APPLYING PANEL DATA METHODS TO OTHER DATA STRUCTURES

Differencing, fixed effects, and random effects methods can be applied to data structures that do not involve time. For example, in demography, it is common to use siblings (sometimes twins) to control for unobserved family and background characteristics. Differencing across siblings or, more generally, using the within transformation within a family, removes family effects that may be correlated with the explanatory variables.

As an example, Geronimus and Korenman (1992) use pairs of sisters to study the effects of teen childbearing on future economic outcomes. When the outcome is income relative to needs—something that depends on the number of children—the model is

$$\log(incneeds_{fs}) = \beta_0 + \delta_0 sister2_s + \beta_1 teenbrth_{fs} + \beta_2 age_{fs} + other\ factors + a_f + u_{fs}, \quad \textbf{(14.12)}$$

where f indexes family and s indexes a sister within the family. The intercept for the first sister is β_0, and the intercept for the second sister is $\beta_0 + \delta_0$. The variable of interest is $teenbrth_{fs}$, which is a binary variable equal to one if sister s in family f had a child while a teenager. The variable age_{fs} is the current age of sister s in family f; Geronimus and Korenman also use some other controls. The unobserved variable a_f, which changes only across family, is an *unobserved family effect* or a *family fixed effect*. The main concern in the analysis is that *teenbrth* is correlated with the family effect. If so, an OLS analysis that pools across families and sisters gives a biased estimator of the effect of teenage motherhood on economic outcomes. Solving this problem is simple: within each family, difference (14.12) across sisters to get

$$\Delta\log(incneeds) = \delta_0 + \beta_1 \Delta teenbrth + \beta_2 \Delta age + \ldots + \Delta u; \quad \textbf{(14.13)}$$

this removes the family effect, a_f, and the resulting equation can be estimated by OLS. Notice that there is no time element here: the differencing is across sisters within a family.

Using 129 sister pairs from the 1982 National Longitudinal Survey of Young Women, Geronimus and Korenman first estimate β_1 by pooled OLS to obtain $-.33$ or $-.26$, where the second estimate comes from controlling for family background variables (such as parents' education); both estimates are very statistically significant [see Table 3 in Geronimus and Korenman (1992)]. Therefore, teenage motherhood has a rather large impact on future family income. However, when the differenced equation is estimated, the coefficient on *teenbrth* is $-.08$, which is small and statistically insignificant. This suggests that it is largely a woman's family background that affects her future income, rather than teenage childbearing.

QUESTION 14.4

When using the differencing method, does it make sense to include dummy variables for the mother and father's race in (14.12)? Explain.

Geronimus and Korenman look at several other outcomes and two other data sets; in some cases, the within family estimates are economically large and statistically significant. They also show how the effects disappear entirely when the sisters' education levels are controlled for.

Ashenfelter and Krueger (1994) used the differencing methodology to estimate the return to education. They obtained a sample of 149 identical twins and collected information on earnings, education, and other variables. The reason for using identical twins is that they should have the same underlying ability. This can be differenced away by using twin differences, rather than OLS on the pooled data. Because identical twins are the same in age, gender, and race, these factors all drop out of the differenced equation. Therefore, Ashenfelter and Krueger regressed the difference in log(*earnings*) on the difference in education and estimated the return to education to be about 9.2% ($t = 3.83$). Interestingly, this is actually *larger* than the pooled OLS estimate of 8.4% (which controls for gender, age, and race). Ashenfelter and Krueger also estimated the equation by random effects and obtained 8.7% as the return to education. (See Table 5 in their paper.) The random effects analysis is mechanically the same as the panel data case with two time periods.

The samples used by Geronimus and Korenman (1992) and Ashenfelter and Krueger (1994) are examples of **matched pair samples**. Generally, fixed and random effects methods can be applied to a **cluster sample**. These are cross-sectional data sets, but each observation belongs to a well-defined cluster. In the previous examples, each family is a cluster. As another example, suppose we have participation data on various pension plans, where firms offer more than one plan. We can then view each firm as a cluster, and it is pretty clear that unobserved firm effects would be an important factor in determining participation rates in pension plans within the firm.

Educational data on students sampled from many schools form a cluster sample, where each school is a cluster. Since the outcomes within a cluster are likely to be correlated, allowing for an unobserved cluster effect is typically important. Fixed effects estimation is preferred when we think the unobserved **cluster effect**—an example of which is a_f in (14.12)—is correlated with one or more of the explanatory variables. Then, we can only include explanatory variables that vary, at least somewhat, within clusters. The cluster sizes are rarely the same, so fixed effects methods for unbalanced panels are usually required.

Random effects methods can also be used with unbalanced clusters, provided the cluster effect is uncorrelated with all the explanatory variables. We can also use pooled OLS in this case, but the usual standard errors are incorrect unless there is no correlation within clusters. Some regression packages have simple commands to correct standard errors and the usual test statistics for general within cluster correlation (as well as heteroskedasticity). These are the same corrections that work for pooled OLS on panel data sets, which we reported in Example 13.9. As an example, Papke (1999) estimates linear probability models for the continuation of defined benefit pension plans based on whether firms adopted defined contribution plans. Because there is likely to be a firm effect that induces correlation across different plans within the same firm, Papke corrects the usual OLS standard errors for cluster sampling, as well as for heteroskedasticity in the linear probability model.

SUMMARY

We have studied two common methods for estimating panel data models with unobserved effects. Compared with first differencing, the fixed effects estimator is efficient when the idiosyncratic errors are serially uncorrelated (as well as homoskedastic), and we make no assumptions about correlation between the unobserved effect a_i and the explanatory variables. As with first differencing, any time-constant explanatory variables drop out of the analysis. Fixed effects methods apply immediately to unbalanced panels, but we must assume that the reasons some time periods are missing are not systematically related to the idiosyncratic errors.

The random effects estimator is appropriate when the unobserved effect is thought to be uncorrelated with all the explanatory variables. Then, a_i can be left in the error term, and the resulting serial correlation over time can be handled by generalized least squares estimation. Conveniently, feasible GLS can be obtained by a pooled regression on quasi-demeaned data. The value of the estimated transformation parameter, $\hat{\lambda}$, indicates whether the estimates are likely to be closer to the pooled OLS or the fixed effects estimates. If the full set of random effects assumptions hold, the random effects estimator is asymptotically—as N gets large with T fixed—more efficient than pooled OLS, first differencing, or fixed effects (which are all unbiased, consistent, and asymptotically normal).

Finally, the panel data methods studied in Chapters 13 and 14 can be used when working with matched pairs or cluster samples. Differencing or the within transformation eliminates the cluster effect. If the cluster effect is uncorrelated with the explanatory variables, pooled OLS can be used, but the standard errors and test statistics should be adjusted for cluster correlation. Random effects estimation is also a possibility.

KEY TERMS

Cluster Effect
Cluster Sample
Composite Error Term
Dummy Variable Regression
Fixed Effects Estimator
Fixed Effects Transformation
Matched Pair Samples
Quasi-Demeaned Data
Random Effects Estimator
Random Effects Model

Time-Demeaned Data
Unbalanced Panel
Within Estimator
Within Transformation

PROBLEMS

14.1 Suppose that the idiosyncratic errors in (14.4), $\{u_{it}: t = 1,2,\ldots,T\}$, are serially uncorrelated with constant variance, σ_u^2. Show that the correlation between adjacent differences, Δu_{it} and $\Delta u_{i,t+1}$, is $-.5$. Therefore, under the ideal FE assumptions, first differencing induces negative serial correlation of a known value.

14.2 With a single explanatory variable, the equation used to obtain the between estimator is

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_i + a_i + \bar{u}_i,$$

where the overbar represents the average over time. We can assume that $\mathrm{E}(a_i) = 0$ because we have included an intercept in the equation. Suppose that $\bar{u}_i$ is uncorrelated with $\bar{x}_i$, but $\mathrm{Cov}(x_{it},a_i) = \sigma_{xa}$ for all t (and i because of random sampling in the cross section).

(i) Letting $\tilde{\beta}_1$ be the between estimator, that is, the OLS estimator using the time averages, show that

$$\text{plim } \tilde{\beta}_1 = \beta_1 + \sigma_{xa}/\mathrm{Var}(\bar{x}_i),$$

where the probability limit is defined as $N \to \infty$. [*Hint*: See equations (5.5) and (5.6).]

(ii) Assume further that the x_{it}, for all $t = 1,2,\ldots,T$, are uncorrelated with constant variance σ_x^2. Show that plim $\tilde{\beta}_1 = \beta_1 + T(\sigma_{xa}/\sigma_x^2)$.

(iii) If the explanatory variables are not very highly correlated across time, what does part (ii) suggest about whether the inconsistency in the between estimator is smaller when there are more time periods?

14.3 In a random effects model, define the composite error $v_{it} = a_i + u_{it}$, where a_i is uncorrelated with u_{it} and the u_{it} have constant variance σ_u^2 and are serially uncorrelated. Define $e_{it} = v_{it} - \lambda \bar{v}_i$, where λ is given in (14.10). Show that the e_{it} have mean zero, constant variance, and are serially uncorrelated.

14.4 In order to determine the effects of collegiate athletic performance on applicants, you collect data on applications for a sample of Division I colleges for 1985, 1990, and 1995.

(i) What measures of athletic success would you include in an equation? What are some of the timing issues?

(ii) What other factors might you control for in the equation?

(iii) Write an equation that allows you to estimate the effects of athletic success on the percentage change in applications. How would you estimate this equation? Why would you choose this method?

14.5 Suppose that, for one semester, you can collect the following data on a random sample of college juniors and seniors for each class taken: a standardized final exam

score, percentage of lectures attended, a dummy variable indicating whether the class is within the student's major, cumulative grade point average prior to the start of the semester, and SAT score.

(i) Why would you classify this data set as a cluster sample? Roughly how many observations would you expect for the typical student?

(ii) Write a model, similar to equation (14.12), that explains final exam performance in terms of attendance and the other characteristics. Use s to subscript student and c to subscript class. Which variables do not change within a student?

(iii) If you pool all of the data together and use OLS, what are you assuming about unobserved student characteristics that affect performance and attendance rate? What roles do SAT score and prior GPA play in this regard?

(iv) If you think SAT score and prior GPA do not adequately capture student ability, how would you estimate the effect of attendance on final exam performance?

COMPUTER EXERCISES

14.6 Use the data in RENTAL.RAW for this exercise. The data on rental prices and other variables for college towns are for the years 1980 and 1990. The idea is to see whether a stronger presence of students affects rental rates. The unobserved effects model is

$$\log(rent_{it}) = \beta_0 + \delta_0 y90_t + \beta_1 \log(pop_{it}) + \beta_2 \log(avginc_{it}) + \beta_3 pctstu_{it} + a_i + u_{it},$$

where *pop* is city population, *avginc* is average income, and *pctstu* is student population as a percentage of city population (during the school year).

(i) Estimate the equation by pooled OLS and report the results in standard form. What do you make of the estimate on the 1990 dummy variable? What do you get for $\hat{\beta}_{pctstu}$?

(ii) Are the standard errors you report in part (i) valid? Explain.

(iii) Now, difference the equation and estimate by OLS. Compare your estimate of β_{pctstu} with that from part (i). Does the relative size of the student population appear to affect rental prices?

(iv) Estimate the model by fixed effects to verify that you get identical estimates and standard errors to those in part (iii).

14.7 Use CRIME4.RAW for this exercise.

(i) Reestimate the unobserved effects model for crime in Example 13.9 but use fixed effects rather than differencing. Are there any notable sign or magnitude changes in the coefficients? What about statistical significance?

(ii) Add the logs of each wage variable in the data set and estimate the model by fixed effects. How does including these variables affect the coefficients on the criminal justice variables in part (i)?

(iii) Do the wage variables in part (ii) all have the expected sign? Explain. Are they jointly significant?

14.8 For this exercise, we use JTRAIN.RAW to determine the effect of the job training grant on hours of job training per employee. The basic model for the three years is

$$hrsemp_{it} = \beta_0 + \delta_1 d88_t + \delta_2 d89_t + \beta_1 grant_{it} + \beta_2 grant_{i,t-1} + \beta_3 \log(employ_{it}) + a_i + u_{it}.$$

(i) Estimate the equation using fixed effects. How many firms are used in the FE estimation? How many total observations would be used if each firm had data on all variables (in particular, *hrsemp*) for all three years?
(ii) Interpret the coefficient on *grant* and comment on its significance.
(iii) Is it surprising that $grant_{-1}$ is insignificant? Explain.
(iv) Do larger firms provide their employees with more or less training, on average? How big are the differences? (For example, if a firm has 10% more employees, what is the change in average hours of training?)

14.9 In Example 13.8, we used the unemployment claims data from Papke (1994) to estimate the effect of enterprise zones on unemployment claims. Papke also uses a model that allows each city to have its own time trend:

$$\log(uclms_{it}) = a_i + c_i t + \beta_1 ez_{it} + u_{it},$$

where a_i and c_i are both unobserved effects. This allows for more heterogeneity across cities.

(i) Show that, when the previous equation is first differenced, we obtain

$$\Delta\log(uclms_{it}) = c_i + \beta_1 \Delta ez_{it} + \Delta u_{it}, \; t = 2, \ldots, T.$$

Notice that the differenced equation contains a fixed effect, c_i.
(ii) Estimate the differenced equation by fixed effects. What is the estimate of β_1? Is it very different from the estimate obtained in Example 13.8? Is the effect of enterprise zones still statistically significant?
(iii) Add a full set of year dummies to the estimation in part (ii). What happens to the estimate of β_1?

14.10 (i) In the wage equation in Example 14.4, explain why dummy variables for occupation might be important omitted variables for estimating the union wage premium.
(ii) Using the data in WAGEPAN.RAW, include eight of the occupation dummy variables in the equation and estimate the equation using fixed effects. Does the coefficient on *union* change by much? What about its statistical significance?

14.11 Add the interaction term $union_{it} \cdot t$ to the equation estimated in Table 14.2 to see if wage *growth* depends on union status. Estimate the equation by random and fixed effects and compare the results.

APPENDIX 14A

Assumptions for Fixed and Random Effects

In this appendix, we provide statements of the assumptions for fixed and random effects estimation. We also provide a discussion of the properties of the estimators under different sets of assumptions. Verification of these claims is somewhat involved, but it can be found in Wooldridge (1999, Chapter 10).

ASSUMPTION FE.1
For each i, the model is

$$y_{it} = \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it}, t = 1, \ldots, T,$$

where the β_j are the parameters to estimate.

ASSUMPTION FE.2
We have a random sample in the cross-sectional dimension.

ASSUMPTION FE.3
For each t, the expected value of the idiosyncratic error given the explanatory variables in *all* time periods and the unobserved effect is zero: $\mathrm{E}(u_{it}|\boldsymbol{X}_i, a_i) = 0$.

ASSUMPTION FE.4
Each explanatory variable changes over time (for at least some i), and there are no perfect linear relationships among the explanatory variables.

Under these first four assumptions—which are identical to the assumptions for the first-differencing estimator—the fixed effects estimator is unbiased. Again, the key is the strict exogeneity assumption, FE.3. Under these same assumptions, the FE estimator is consistent with a fixed T as $N \to \infty$.

ASSUMPTION FE.5
$\mathrm{Var}(u_{it}|\boldsymbol{X}_i, a_i) = \mathrm{Var}(u_{it}) = \sigma_u^2$, for all $t = 1, \ldots, T$.

ASSUMPTION FE.6
For all $t \neq s$, the idiosyncratic errors are uncorrelated (conditional on all explanatory variables and a_i): $\mathrm{Cov}(u_{it}, u_{is}|\boldsymbol{X}_i, a_i) = 0$.

Under Assumptions FE.1 through FE.6, the fixed effects estimator of the β_j is the best linear unbiased estimator. Since the FD estimator is linear and unbiased, it is necessarily worse than the FE estimator. The assumption that makes FE better than FD is FE.6, which implies that the idiosyncratic errors are serially uncorrelated.

ASSUMPTION FE.7
Conditional on $\boldsymbol{X}_i$ and a_i, the u_{it} are independent and identically distributed as Normal$(0,\sigma_u^2)$.

Assumption FE.7 implies FE.3, FE.5, and FE.6, but it is stronger because it assumes a normal distribution for the idiosyncratic errors. If we add FE.7, the FE estimator is normally distributed, and t and F statistics have exact t and F distributions. Without FE.7, we can rely on asymptotic approximations. But, without making special assumptions, these approximations require large N and small T.

The ideal random effects assumptions include FE.1, FE.2, FE.3, FE.5, and FE.6. We can now allow for time-constant variables. (FE.7 could be added, but it gains us little in practice.) However, we need to add assumptions about how a_i is related to the explanatory variables. Thus, the third assumption is strengthened as follows.

ASSUMPTION RE.3
In addition to FE.3, the expected value of a_i given all explanatory variables is zero: $E(a_i|\boldsymbol{X}_i) = 0$.

This is the assumption that rules out correlation between the unobserved effect and the explanatory variables. Because the RE transformation does not completely remove the time average, we can allow explanatory variables that are constant across time for all i.

ASSUMPTION RE.4
There are no perfect linear relationships among the explanatory variables.

We also need to impose homoskedasticity on a_i as follows:

ASSUMPTION RE.5
In addition to FE.5, the variance of a_i given all explanatory variables is constant: $\text{Var}(a_i|\boldsymbol{X}_i) = \sigma_a^2$.

Under the six random effects assumptions (FE.1, FE.2, RE.3, RE.4, RE.5, and FE.6), the random effects estimator is consistent as N gets large for fixed T. (Actually, only the first four assumptions are needed for consistency.) The RE estimator is not unbiased unless we know λ, which keeps up from having to estimate it. The RE estimator is also approximately normally distributed with large N, and the usual standard errors, t statistics, and F statistics obtained from the quasi-demeaned regression are valid with large N. [For more information, see Wooldridge (1999, Chapter 10).]