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 general attrition problems in panel data is complicated and beyond the scope of this text. [See, for example, Wooldridge (2010, Chapter 19).]

14-2 Random Effects Models

We begin with the same unobserved effects model as before,

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

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:

$$\mathrm{Cov}(x_{itj}, a_i) = 0, \quad t = 1, 2, \ldots, T;\, j = 1, 2, \ldots, k. \qquad [14.8]$$

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 or 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 also use the data 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

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

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,

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

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, and, 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. For the procedure to have good properties, we should 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 (2010, Chapter 10)]. But the transformation itself is simple. Define

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

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

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

where the overbar again denotes the time averages. This is a very interesting equation, as it involves **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. (See Problem 3.)

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 the explanatory variables 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 θ 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, $\hat{\theta}$ takes the form $\hat{\theta} = 1 - \{1/[1 + T(\hat{\sigma}_a^2/\hat{\sigma}_u^2)]\}^{1/2}$, 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 $\hat{\sigma}_a^2 = [NT(T-1)/2 - (k+1)]^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\hat{v}_{it}\hat{v}_{is}$, where the $\hat{v}_{it}$ are the residuals from estimating (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 (2010, Chapter 10) for additional discussion of these estimators.]

Many econometrics packages support estimation of random effects models and automatically compute some version of $\hat{\theta}$. The feasible GLS estimator that uses $\hat{\theta}$ in place of θ 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 random effects (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 $\theta = 0$, and FE is obtained when $\theta = 1$. In practice, the estimate $\hat{\theta}$ is never zero or one. But if $\hat{\theta}$ 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 (because 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 $\hat{\theta}$ will be closer to unity. As T gets large, $\hat{\theta}$ tends to one, and this makes the RE and FE estimates very similar.

We can gain more insight on the relative merits of random effects versus fixed effects by writing the quasi-demeaned error in equation (14.11) as $v_{it} - \theta\bar{v}_i = (1 - \theta)a_i + u_{it} - \theta\bar{u}_i$. This simple expression makes it clear that in the transformed equation the unobserved effect is weighted by $(1 - \theta)$. Although correlation between a_i and one or more x_{itj} causes inconsistency in the random effects estimation, we see that the correlation is attenuated by the factor $(1 - \theta)$. As $\theta \rightarrow 1$, the bias term goes to zero, as it must because the RE estimator tends to the FE estimator. If θ is close to zero, we are leaving a larger fraction of the unobserved effect in the error term, and, as a consequence, the asymptotic bias of the RE estimator will be larger.

In applications of FE and RE, it is usually informative also to compute the pooled OLS estimates. Comparing the three sets of estimates can help us determine the nature of the biases caused by leaving the unobserved effect, a_i, entirely in the error term (as does pooled OLS) or partially in the error term (as does the RE transformation). But we must remember that, even if a_i is uncorrelated with all explanatory variables in all time periods, the pooled OLS standard errors and test statistics are generally invalid: they ignore the often substantial serial correlation in the composite errors, $v_{it} = a_i + u_{it}$. As we mentioned in Chapter 13 (see Example 13.9), it is possible to compute standard errors and test statistics that are robust to arbitrary serial correlation (and heteroskedasticity) in v_{it}, and popular statistics packages often allow this option. [See, for example, Wooldridge (2010, Chapter 10).]

EXAMPLE 14.4 A Wage Equation Using Panel Data

We again use the data in WAGEPAN 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 (although *exper*^{2} remains). Each regression also contains a full set of year dummies. The estimation results are in Table 14.2.

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)

EXPLORING FURTHER 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?

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.

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

14-2a Random Effects or Fixed Effects?

Because fixed effects allows arbitrary correlation between a_i and the x_{itj}, while random effects does not, FE is widely thought to be a more convincing tool for estimating ceteris paribus effects. Still, random effects is applied in certain situations. Most obviously, if the key explanatory variable is constant over time, we cannot use FE to estimate its effect on y. For example, in Table 14.2, we must rely on the RE (or pooled OLS) estimate of the return to education. Of course, we can only use random effects because we are willing to assume the unobserved effect is uncorrelated with all explanatory variables. Typically, if one uses random effects, as many time-constant controls as possible are included among the explanatory variables. (With an FE analysis, it is not necessary to include such controls.) RE is preferred to pooled OLS because RE is generally more efficient.

If our interest is in a time-varying explanatory variable, is there ever a case to use RE rather than FE? Yes, but situations in which $\text{Cov}(x_{itj}, a_i) = 0$ should be considered the exception rather than the rule. If the key policy variable is set experimentally—say, each year, children are randomly assigned to classes of different sizes—then random effects would be appropriate for estimating the effect of class size on performance. Unfortunately, in most cases the regressors are themselves outcomes of choice processes and likely to be correlated with individual preferences and abilities as captured by a_i.

It is still fairly common to see researchers apply both random effects and fixed effects, and then formally test for statistically significant differences in the coefficients on the time-varying explanatory variables. (So, in Table 14.2, these would be the coefficients on $exper^2$, *married*, and *union*.) Hausman (1978) first proposed such a test, and some econometrics packages routinely compute the Hausman test under the full set of random effects assumptions listed in the appendix to this chapter. The idea is that one uses the random effects estimates unless the Hausman test rejects (14.8). In practice, a failure to reject means either that the RE and FE estimates are sufficiently close so that it does not matter which is used, or the sampling variation is so large in the FE estimates that one cannot conclude practically significant differences are statistically significant. In the latter case, one is left to wonder whether there is enough information in the data to provide precise estimates of the coefficients. A rejection using the Hausman test is taken to mean that the key RE assumption, (14.8), is false, and then the FE estimates are used. (Naturally, as in all applications of statistical inference, one should distinguish between a practically significant difference and a statistically significant difference.) Wooldridge (2010, Chapter 10) contains further discussion. In the next section we discuss an alternative, computationally simpler approach to choosing between the RE and FE approaches.

A final word of caution. In reading empirical work, you may find that some authors decide on FE versus RE estimation based on whether the a_i are properly viewed as parameters to estimate or as random variables. Such considerations are usually wrongheaded. In this chapter, we have treated the a_i as random variables in the unobserved effects model (14.7), regardless of how we decide to estimate the β_j. As we have emphasized, the key issue that determines whether we use FE or RE is whether we can plausibly assume a_i is uncorrelated with all x_{itj}. Nevertheless, in some applications of panel data methods, we cannot treat our sample as a random sample from a large population, especially when the unit of observation is a large geographical unit (say, states or provinces). Then, it often makes sense to think of each a_i as a separate intercept to estimate for each cross-sectional unit. In this case, we use fixed effects: remember, using FE is mechanically the same as allowing a different intercept for each cross-sectional unit. Fortunately, whether or not we engage in the philosophical debate about the nature of a_i, FE is almost always much more convincing than RE for policy analysis using aggregated data.

14-3 The Correlated Random Effects Approach

In applications where it makes sense to view the a_i (unobserved effects) as being random variables, along with the observed variables we draw, there is an alternative to fixed effects that still allows a_i to be correlated with the observed explanatory variables. To describe the approach, consider again the simple model in equation (14.1), with a single, time-varying explanatory variable x_{it}. Rather than assume a_i is uncorrelated with $\{x_{it}: t = 1, 2, \ldots, T\}$—which is the random effects approach—or take away time averages to remove a_i—the fixed effects approach—we might instead model correlation between a_i and $\{x_{it}: t = 1, 2, \ldots, T\}$. Because a_i is, by definition, constant over time, allowing it to be correlated with the average level of the x_{it} has a certain appeal. More specifically, let $\bar{x}_i = T^{-1}\sum_{t=1}^{T} x_{it}$ be the time average, as before. Suppose we assume the simple linear relationship

$$a_i = \alpha + \gamma\bar{x}_i + r_i, \qquad [14.12]$$

where we assume r_i is uncorrelated with each x_{it}. Because $\bar{x}_i$ is a linear function of the x_{it},

$$Cov(\bar{x}_i, r_i) = 0. \qquad [14.13]$$

Equations (14.12) and (14.13) imply that a_i and $\bar{x}_i$ are correlated whenever $\gamma \neq 0$.

The **correlated random effects** (CRE) approach uses (14.12) in conjunction with (14.1): substituting the former in the latter gives

$$y_{it} = \beta x_{it} + \alpha + \gamma\bar{x}_i + r_i + u_{it} = \alpha + \beta x_{it} + \gamma\bar{x}_i + r_i + u_{it}. \qquad [14.14]$$

Equation (14.14) is interesting because it still has a composite error term, $r_i + u_{it}$, consisting of a time-constant unobservable r_i and the idiosyncratic shocks, u_{it}. Importantly, assumption (14.8) holds when we replace a_i with r_i. Also, because u_{it} is assumed to be uncorrelated with x_{is}, all s and t, u_{it} is also uncorrelated with $\bar{x}_i$. All of these assumptions add up to random effects estimation of the equation

$$y_{it} = \alpha + \beta x_{it} + \gamma\bar{x}_i + r_i + u_{it}, \qquad [14.15]$$

which is like the usual equation underlying RE estimation with the important addition of the time-average variable, $\bar{x}_i$. It is the addition of $\bar{x}_i$ that controls for the correlation between a_i and the sequence $\{x_{it}: t = 1, 2, \ldots, T\}$. What is left over, r_i, is uncorrelated with the x_{it}.

In most econometrics packages it is easy to compute the unit-specific time averages, $\bar{x}_i$. Assuming we have done that for each cross-sectional unit i, what can we expect to happen if we apply RE to equation (14.15)? Notice that estimation of (14.15) gives $\hat{\alpha}_{CRE}, \hat{\beta}_{CRE}$, and $\hat{\gamma}_{CRE}$—the CRE estimators.