CHAPTER 15

Panel Data Models

LEARNING OBJECTIVES

Based on the material in this chapter, you should be able to

1. Explain how a data panel differs from either a cross section or a time series of data.
2. Explain the different ways in which individual heterogeneity can be modeled using panel data, and the assumptions underlying each approach.
3. Explain how the fixed effects model allows for differences in the parameter values for each individual cross section in a data panel.
4. Compare and contrast the least squares dummy variable estimator and the fixed effects estimator.
5. Compare and contrast the fixed effects model and the random effects model. Explain what leads us to consider individual differences to be random.
6. Explain the error assumptions in the random effects model, and what characteristic leads us to consider generalized least squares estimation.
7. Describe the steps required to obtain generalized least squares estimates for the random effects estimator.
8. Explain the meaning of cluster-robust standard errors, and describe how they can be used with pooled least squares, fixed effects, and random effects estimators.
9. Explain why endogeneity is a potential problem in random effects models, and how it affects our choice of estimator.
10. Test for the existence of fixed and/or random effects, and use the Hausman test to assess whether the random effects estimator is inconsistent.
11. Explain how the Hausman–Taylor estimator can be used to obtain consistent estimates of coefficients of time-invariant variables in a random effects model.
12. Use your software to estimate fixed effects models and random effects models for panel data.

KEYWORDS

Balanced panel
Cluster-robust standard errors
Deviations about the individual mean
Difference estimator
Endogeneity
Error components model
Fixed effects estimator
Fixed effects model
Hausman test
Hausman–Taylor estimator
Heterogeneity
Instrumental variables
Least squares dummy variable model
LM test
Pooled least squares
Pooled model
Random effects estimator
Random effects model
Time-invariant variables
Time-varying variables
Unbalanced panel
Within estimator

A panel of data consists of a group of cross-sectional units (people, households, firms, states, and countries) who are observed over time. We will often refer to such units as individuals, with the term "individual" being used generically, even when the unit of interest is not a person. Let us denote the number of cross-sectional units (individuals) by N, and number of time periods in which we observe them as T. Panel data come in several different "flavors," each of which introduces new challenges and opportunities. Peter Kennedy[1] describes the different types of panel data sets as

- "Long and narrow," with "long" describing the time dimension and "narrow" implying a relatively small number of cross-sectional units
- "Short and wide," indicating that there are many individuals observed over a relatively short period of time
- "Long and wide," indicating that both N and T are relatively large

A "long and narrow" panel may consist of data on several firms over a period of time. A classic example is a data set analyzed by Grunfeld and used subsequently by many authors.[2] These data track investment in plant and equipment by $N = 11$ large firms for $T = 20$ years. This panel is narrow because it consists of only $N = 11$ firms. It is relatively "long" because $T > N$.

Many microeconomic analyses are performed on panel data sets with thousands of individuals who are followed through time. For example, the Panel Study of Income Dynamics (PSID) has followed approximately 8,000 families since 1968.[3] The U.S. Department of Labor conducts National Longitudinal Surveys (NLS) such as NLSY79, "a nationally representative sample of 12,686 young men and women who were 14–22 years old when they were first surveyed in 1979.[4] These individuals were interviewed annually through 1994 and are currently interviewed on a biennial basis." Such data sets are "wide" and "short," because N is much, much larger than T. Using panel data sets of this kind we can account for unobserved individual differences, or **heterogeneity**. Furthermore, these data panels are becoming long enough so that dynamic factors, such as spells of employment and unemployment, can be studied. These very large data sets are rich in information, and require the use of considerable computing power.

Macroeconomists who study economic growth across nations employ data that is "long" and "wide." The Penn World Table[5] provides purchasing power parity and national income accounts converted to international prices for 182 countries for some or all of the years 1950–2014, which we may roughly characterize as having both large N and large T.

Finally, it is possible to have data that combines cross-sectional and time-series data which do not constitute a panel. We may collect a sample of data on individuals from a population at several points in time, but the individuals are not the same in each time period. Such data can be used to analyze a "natural experiment," for example, when a law affecting some individuals changes, such as a change in unemployment insurance in a particular state. Using data before and after the policy change, and on groups of affected and unaffected people, the effects of the policy change can be measured. Methods for estimating effects of this type were introduced in Section 7.5.

Our interest in this chapter is how to use all available data to estimate econometric models describing the behavior of the individual cross-sectional units over time. Such data allow us to control for individual differences and study dynamic adjustment, and to measure the effects of policy changes. For each type of data, we must take care not only with error assumptions, but also

[1] *A Guide to Econometrics*, 6th ed., Chapter 18, MIT Press, 2008.

[2] See Kleiber and Zeileis, "The Grunfeld Data at 50," *German Economic Review*, 2010, 11(4), pp. 404–417 and http://statmath.wu-wien.ac.at/∼zeileis/grunfeld/.

[3] See http://psidonline.isr.umich.edu/.

[4] See www.bls.gov/nls/.

[5] See http://cid.econ.ucdavis.edu/.

with our assumptions about whether, how, and when parameters may change across individuals and/or time.

EXAMPLE 15.1 | A Microeconometric Panel

Our first example is of a data set that is short and wide. It is typical of many microeconometric analyses that use large data sets with many individuals, coming from the NLS conducted by the U.S. Department of Labor, which has a database on women who were between 14 and 24 in 1968. To illustrate, we use a subsample of $N = 716$ women who were interviewed in 1982, 1983, 1985, 1987, and 1988. The sample consists of women who were employed, and whose schooling was completed, when interviewed. The data file is named *nls_panel* and contains 3,580 lines of data. Panel data observations are usually stacked, with all the time-series observations for one individual on top of the next. The observations on a few variables for the first three women in the NLS panel are shown in Table 15.1. The first column *ID* identifies the individual and *YEAR* represents the year in which the information was collected. These identifying variables must be present so that your software will properly identify the cross-section and time-series units. Then there are observations on each of the variables. In a typical panel, there are some observations with missing values, usually denoted as "·" or "NA." We have removed all the missing values in the data file *nls_panel*. In microeconomic panels, the individuals are not always interviewed the same number of times, leading to an **unbalanced panel** in which the number of time-series observations is different across individuals. The data file *nls_panel* is, however, a **balanced panel**; for each individual, we observe five time-series observations. A larger, unbalanced panel, is in the data file *nls*. Most modern software packages can handle both balanced and unbalanced panels.

TABLE 15.1 **Representative Observations from NLS Panel Data**

ID	*YEAR*	*LWAGE*	*EDUC*	*SOUTH*	*BLACK*	*UNION*	*EXPER*	*TENURE*
1	82	1.8083	12	0	1	1	7.6667	7.6667
1	83	1.8634	12	0	1	1	8.5833	8.5833
1	85	1.7894	12	0	1	1	10.1795	1.8333
1	87	1.8465	12	0	1	1	12.1795	3.7500
1	88	1.8564	12	0	1	1	13.6218	5.2500
2	82	1.2809	17	0	0	0	7.5769	2.4167
2	83	1.5159	17	0	0	0	8.3846	3.4167
2	85	1.9302	17	0	0	0	10.3846	5.4167
2	87	1.9190	17	0	0	1	12.0385	0.3333
2	88	2.2010	17	0	0	1	13.2115	1.7500
3	82	1.8148	12	0	0	0	11.4167	11.4167
3	83	1.9199	12	0	0	1	12.4167	12.4167
3	85	1.9584	12	0	0	0	14.4167	14.4167
3	87	2.0071	12	0	0	0	16.4167	16.4167
3	88	2.0899	12	0	0	0	17.8205	17.7500

15.1 The Panel Data Regression Function

A panel of data consists of a group of cross-sectional units (people, households, firms, states, or countries) who are observed over time. The sampling process we imagine is that (i) $i = 1, \ldots, N$ individuals are randomly selected from the population and (ii) each individual is observed for $t = 1, \ldots, T$ time periods. In the sampling process, we collect values y_{it} on an outcome, or dependent, variable of interest. Other characteristics concerning the individual will be used as

explanatory variables. Let $x_{1it} = 1$ be the intercept variable with $x_{2it}, \ldots, x_{Kit}$ being observations on $K - 1$ factors that vary across individual and time. Let $w_{1i}, w_{2i}, \ldots, w_{Mi}$ be observed data on M factors that do not change over time. Note that these variables **do not** have a time subscript and are said to be **time-invariant**. We cannot stress enough how important it is when using panel data to *examine the subscripts closely*, and recall that i is the indicator of the individual and t is the indicator of time.

In addition to the observed variables, there will be unobserved, omitted factors in each time period for each individual that will compose the regression's random error term. In panel data models, we can identify several types of unobserved effects. First, consider unobserved and/or unmeasurable, time-invariant individual characteristics. Let us denote these as u_{1i}, $u_{2i}, \ldots, u_{Si}$. Because we cannot observe them, we will simply refer to their combined effect as u_i, an unobserved, individual-specific random error component. Economists say that u_i represents **unobserved heterogeneity**, summarizing the unobserved factors leading to individual differences. Second, there are many, unobserved, and/or unmeasurable individual and time-varying factors $e_{1it}, e_{2it}, \ldots$ constituting the usual type of random errors in regression, and we refer to their combined effect as e_{it}. Econometricians call the random error e_{it} that varies across individual and time, an **idiosyncratic**[6] error. A third type of random error is time specific, an effect that varies over time but not individual. These factors $m_{1t}, m_{2t} \ldots$ have combined effect m_t and represent a third error component.

EXAMPLE 15.1 | Revisited

For example, in Table 15.1, the outcome variable of interest is $y_{it} = LWAGE_{it} = \ln(WAGE_{it})$. Explanatory variables include $x_{2it} = EXPER_{it}$, $x_{3it} = TENURE_{it}$, $x_{4it} = SOUTH_{it}$, and $x_{5it} = UNION_{it}$. These explanatory variables vary across both individual and time. For the indicator variables *SOUTH* and *UNION*, it means that at least some individuals moved into or out of the *SOUTH* during the 1982–1988 period, and at least some workers joined or quit a *UNION* over those years. The variables $w_{1i} = EDUC_i$ and $w_{2i} = BLACK_i$ do not change for the 716 individuals in our sample over the years 1982–1988. Two unobserved **time-invariant variables** are $u_{1i} = ABILITY_i$ and $u_{2i} = PERSEVERANCE_i$. Unobserved time-specific variables might be $m_{1t} = UNEMPLOYMENT\ RATE_t$ or $m_{2t} = INFLATION\ RATE_t$. Note that it is possible to have observable variables that change over time but not across individuals, like an indicator variable $D82_t = 1$ if the year is 1982 and $D82_t = 0$ otherwise.

A simple but representative panel data regression model is

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + (u_i + e_{it}) = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + v_{it} \tag{15.1}$$

In (15.1), the observable outcome variable of interest is y_{it}. On the right-hand side, we have a constant term, $x_{1it} = 1$. We include one observable variable, x_{2it}, that has variation across individuals and time. The variable w_{1i} is time-invariant and varies only across individuals. The population parameters β_1, β_2, and α_1 have no subscripts and are fixed in all time periods for all individuals. We have included only one x-variable and one w-variable to keep things simple, but there can be more of each type. In parentheses, we have the two random error components, one associated with the individual (u_i) and one associated with the individual and time (e_{it}). For simplicity, we are omitting the random time-specific error component. We define the combined error

$$v_{it} = u_i + e_{it} \tag{15.2}$$

Because the regression error in (15.2) has two components, one for the individual and one for the regression, it is often called an **error components model**.

[6] Jeffrey M. Wooldridge, *Econometric Analysis of Cross Section and Panel Data*, 2nd ed., MIT Press, 2010, p. 285.

The complicating factor in panel data modeling is that we observe each cross-sectional unit, individual i, for more than one time-period, t. If individuals are randomly sampled, then observations on the ith individual are statistically independent of observations on the jth individual. However, using panel data, we must consider dynamic, time-related effects, and model assumptions should take them into account, just as we did in Chapter 9. The regression function of interest in a panel data model is

$$E\left[y_{it}\left|\overbrace{x_{2i1}, x_{2i2}, \ldots, x_{2iT}}^{T \text{ terms}}, w_{1i}, u_i\right.\right] = E\left(y_{it}\left|\mathbf{x}_{2i}, w_{1i}, u_i\right.\right) = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + u_i \tag{15.3}$$

where $\mathbf{x}_{2i} = \left(x_{2i1}, x_{2i2}, \ldots, x_{2iT}\right)$ represents the values x_{2it} in all time periods. Equation (15.3) says that the population average value of the outcome variable is $\beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + u_i$, given (i) the values of x_{2it} in *all* time periods, past, present, and future; (ii) the observable individual-specific variable w_{1i}; and (iii) the unobservable individual heterogeneity term u_i. Our econometric challenge is to find a consistent and, if possible, efficient estimator for the parameters β_1, β_2, and α_1.

Equation (15.3) has several interesting features:

i. The model states that once we have controlled for x_{2it} in all time periods, and the individual-specific factors w_{1i} and u_i, only the current, contemporaneous value of x_{2it} has an effect on the expected outcome. The parameter β_2 measures the partial, or causal, effect of a change in x_{2it} on $E\left(y_{it}|\mathbf{x}_{2i}, w_{1i}, u_i\right)$, holding all else constant. Similarly, the causal effect of a change in w_{1i} on $E\left(y_{it}|\mathbf{x}_{2i}, w_{1i}, u_i\right)$ is α_1.

ii. The model conditions on the **unobservable** time-invariant error u_i. In Example 15.2, below, we examine the sales of chemical firms in China over several years using panel data. The observed explanatory variables include, for example, the amount of labor used by the firm in each year. A time-invariant variable is their location. The unobserved heterogeneity u_i might represent the ability of firm managers. The expected firm sales depend quite naturally on the unobserved managerial ability, as well as current production which depends on current labor input. However, what we are imagining is that, given managerial ability, the labor inputs of past years, or future years, have no impact on current sales.[7]

15.1.1 Further Discussion of Unobserved Heterogeneity

Every individual has unique characteristics. This is true for each of us as human beings and also for individual firms, farms, and geographic regions such as states, shires, or nations. Some individual characteristics can be observed and measured, such as an individual's height and weight, or the number of employees a firm has. Some characteristics of individuals are unmeasurable or unobservable, such as a person's ability, beauty, or fortitude. The ability of a firm's managers contributes to their revenues and profits, but just like individual ability, managerial skill is difficult or impossible to measure. Thus in a regression using cross-sectional data, these unobservable characteristics are by necessity excluded from the set of explanatory variables, and hence are included in the random error term. These unobservable individual differences are called **unobservable heterogeneity** in the economics and econometrics literature. When using panel data, it is important to separate out this component of the random error term from other components **if** we can argue that the factors causing the individual differences are unchanging over time. Such an argument is more feasible when the panel data set is wide and short, with large N and small T, as in many microeconomic panels. In a wage equation, for example, we would have to assume that unobservable factors such as ability and perseverance are constant over the period of the sample. If the panel

[7] For more discussion on this assumption, see Wooldridge (2010), p. 288.

data sample covers three or four years, we might be very comfortable with this assumption, but if the sample period covers 25 years, then we may worry about the validity of such an assumption.

Our concern with unobserved heterogeneity is exactly the same as with **omitted variables** discussed in Section 6.3.2. If omitted variables are correlated with any explanatory variables in the regression model, then the ordinary least squares (OLS) estimator suffers from **omitted variables bias**. And unfortunately, this bias does not disappear even in large samples so that the OLS estimator is inconsistent. In Chapter 10, we addressed this problem by finding a new estimator, the **instrumental variables** (IV), **two-stage least squares** (2SLS) estimator. As we will see, the beauty of having panel data is that we can control for the omitted variables bias, caused by time-invariant omitted variables, **without** having to find and use instrumental variables.

15.1.2 The Panel Data Regression Exogeneity Assumption

For the regression model (15.1), $y_{it} = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + (u_i + e_{it})$, to have the conditional expectation in (15.3), what must be true about the random error? A new exogeneity assumption takes into account the presence of the unobserved heterogeneity term. It is

$$E(e_{it}|\mathbf{x}_{2i}, w_{i1}, u_i) = 0 \tag{15.4}$$

The meaning of this **strict exogeneity** assumption is that given the values of the explanatory variable x_{2it} in all time periods, given w_{i1} and given the unobserved heterogeneity term u_i, the best prediction of the idiosyncratic errors is zero. Another way to say this is that there is no information in these factors about the value of the idiosyncratic random error e_{it}. One subtle but extremely important point about assumption (15.4) is that it **does not** require that the unobservable heterogeneity u_i be uncorrelated with the values of the explanatory variables. We will have much more discussion about this point as we go along. Two of the implications of assumption (15.4) are that

$$\text{cov}(e_{it}, x_{2is}) = 0, \text{ and } \text{cov}(e_{it}, w_{1i}) = 0 \tag{15.5a}$$

The first part, $\text{cov}(e_{it}, x_{2is}) = 0$, is much stronger than the usual sort of exogeneity assumption. It is stronger because it is more than just contemporaneous exogeneity $\text{cov}(e_{it}, x_{2it}) = 0$; it says e_{it} is uncorrelated with *all* the values $x_{2i1}, x_{2i2}, \ldots, x_{2iT}$. In thinking about whether (15.4) is valid in a specific application, ask yourself whether (15.5a) holds. If (15.5a) is not true, and if e_{it} is correlated with any $x_{2i1}, x_{2i2}, \ldots, x_{2iT}$ or w_{1i}, then assumption (15.4) fails and the regression function of interest (15.3) is not correct.

While we are being a bit lax about it, (15.4) should properly include the intercept variable $x_{1it} = 1$, so that really $E(e_{it}|\mathbf{x}_{1i}, \mathbf{x}_{2i}, w_{1i}, u_i) = 0$. This is important because it means that (15.5a) holds also for the intercept,

$$\text{cov}(e_{it}, x_{1is} = 1) = E(e_{it}x_{1is}) = E(e_{it}) = 0 \tag{15.5b}$$

Thus, the expected value of the idiosyncratic error is zero.

We are postponing new assumptions about error variances and covariances until Section 15.3.

15.1.3 Using OLS to Estimate the Panel Data Regression

Using our panel of data, can we consistently estimate the panel data regression function parameters in (15.3) using OLS? As we learned in Section 5.7.3 the answer is yes, if in (15.1) the **combined** error v_{it} is uncorrelated with the explanatory variable x_{2it} and with w_{1i}. That is, if

$$\text{cov}(x_{2it}, v_{it}) = E(x_{2it}v_{it}) = E(x_{2it}u_i) + E(x_{2it}e_{it}) = 0$$

and

$$\text{cov}(w_{1i}, v_{it}) = E(w_{1i}v_{it}) = E(w_{1i}u_i) + E(w_{1i}e_{it}) = 0$$

These equations say that the two random error components must be contemporaneously uncorrelated with the time-varying explanatory variables, and uncorrelated with the time-invariant explanatory variables. They in turn require

$$E(x_{2it}e_{it}) = 0, \quad E(w_{1i}e_{it}) = 0 \tag{15.6a}$$

$$E(x_{2it}u_i) = 0, \quad E(w_{1i}u_i) = 0 \tag{15.6b}$$

Equation (15.6a) says that the idiosyncratic error e_{it} is uncorrelated with the explanatory variables at time t. This is ensured by the key exogeneity assumption (15.4). On the other hand, (15.4) does not imply that (15.6b) is true, which requires the unobserved heterogeneity to be uncorrelated with the explanatory variables. The familiar example of *ABILITY* being absent from a wage equation is one case where this assumption is violated, as *ABILITY* is correlated with years of education. We should remember that if any explanatory variable is correlated with the random errors then the estimators of *all* model parameters are inconsistent. In the next section, we will introduce panel data estimation strategies that yield consistent estimators even when (15.6b) fails.

We note in passing that the model intercept variable $x_{1it} = 1$, which is exogenous, satisfies (15.6a) and (15.6b), implying that

$$E(e_{it}) = E(u_i) = E(v_{it}) = 0 \tag{15.6c}$$

Each of the random errors has mean zero. Finally, even if equations (15.6a) to (15.6c) hold, using the OLS estimator will require using a type of robust standard error which we explore in Section 15.3.

15.2 The Fixed Effects Estimator

In this section, we consider estimation procedures that employ a transformation to eliminate the individual heterogeneity from the estimation equation and thus solve the common **endogeneity** problem caused by correlation between unobservable individual characteristics and the explanatory variables. The methods achieve the same outcome using similar but different strategies. The estimators we will consider are (i) the **difference estimator**, (ii) the **within estimator**, and (iii) the **fixed effects estimator**. For each of the estimators to be consistent, the strict exogeneity assumption (15.4) must hold, but we **do not require** the unobserved heterogeneity u_i to be uncorrelated with the explanatory variables, that is, equation (15.6b) does not need to hold. The estimators successfully estimate parameters of variables that vary over time but they cannot estimate parameters of time-invariant variables. In equation (15.1), $y_{it} = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + v_{it}$, using these methods, we can consistently estimate β_2, but we cannot estimate β_1 or α_1.

15.2.1 The Difference Estimator: $T = 2$

It is easy to illustrate the power of having panel data with as few as $T = 2$ observations per individual, that is, when we observe each individual in two different time periods, $t = 1$ and $t = 2$. The two observations written out as in (15.1) are

$$y_{i1} = \beta_1 + \beta_2 x_{2i1} + \alpha_1 w_{1i} + u_i + e_{i1} \tag{15.7a}$$

$$y_{i2} = \beta_1 + \beta_2 x_{2i2} + \alpha_1 w_{1i} + u_i + e_{i2} \tag{15.7b}$$

Subtracting (15.7a) from (15.7b) creates a new equation

$$(y_{i2} - y_{i1}) = \beta_2 (x_{2i2} - x_{2i1}) + (e_{i2} - e_{i1}) \tag{15.8}$$

Note that (15.8) has no intercept, β_1, because it has been subtracted out. Also, $\alpha_1 w_{1i}$ subtracts out meaning that we cannot estimate the coefficient α_1 using this approach. Importantly, the unobservable individual differences u_i have dropped out due to the subtraction. Why? Because the terms β_1, $\alpha_1 w_{1i}$, and u_i are not different for time periods one and two; they are time-invariant and the subtraction removes them. We discussed variables such as $(y_{i2} - y_{i1})$ in Chapter 9. It is the change in the outcome variable's value for individual i from time period $t = 1$ to time period

$t = 2$. In the notation of Chapter 9, let "Δ" stand for "the change in," so that $\Delta y_i = (y_{i2} - y_{i1})$. Similarly, let $\Delta x_{i2} = (x_{2i2} - x_{2i1})$ and $\Delta e_i = (e_{i2} - e_{i1})$. Then, equation (15.8) becomes

$$\Delta y_i = \beta_2 \Delta x_{i2} + \Delta e_i \tag{15.9}$$

Note that a parameter of interest, β_2, is present in the transformed model (15.9). Do not be concerned about complicated data manipulations as econometric software has automatic commands to handle the differencing process.

The OLS estimator of β_2 in (15.9) is called the **first-difference estimator**, or simply the **difference estimator**. It is a consistent estimator if (i) Δe_i has zero mean and is uncorrelated with Δx_{i2}, and (ii) Δx_{i2} takes more than two values. The first condition holds if strict exogeneity, equation (15.4), holds. Recall that (15.4) implies that equations (15.5a) and (15.5b) are true. Then, Δe_i has zero mean using (15.5b). Also Δe_i is uncorrelated with Δx_{i2} because of (15.5a); the idiosyncratic error e_{it} is uncorrelated with x_{2is} in all time periods. In equation (15.8), this means that $\Delta x_{i2} = (x_{2i2} - x_{2i1})$ will be uncorrelated with $\Delta e_i = (e_{i2} - e_{i1})$.

In basic panel data analysis, the difference estimator is usually not used. We introduce it to illustrate that we can eliminate the unobserved heterogeneity through a transformation. In practice, we usually use the equivalent, but more flexible, fixed effects estimator, which we explain in Section 15.2.2.

EXAMPLE 15.2 | Using $T = 2$ Differenced Observations for a Production Function

The data file *chemical2* contains data on $N = 200$ chemical firms' sales in China for the years 2004–2006. We wish to estimate the log-log model

$$\ln(SALES_{it}) = \beta_1 + \beta_2 \ln(CAPITAL_{it}) + \beta_3 \ln(LABOR_{it}) + u_i + e_{it}$$

Using only data from 2005 and 2006, the OLS estimates with conventional, nonrobust, standard errors are

$$\widehat{\ln(SALES_{it})} = \underset{(\text{se})}{5.8745} + \underset{(0.2107)}{0.2536} \ln(CAPITAL_{it}) + \underset{(0.0354)}{0.4264} \ln(LABOR_{it}) \quad (0.0577)$$

We may be concerned that there are unobserved individual differences among the firms that are correlated with their usage of capital and labor in the production and sales process. The estimated first-difference model is

$$\widehat{\Delta \ln(SALES_{it})} = \underset{(\text{se})}{0.0384} \Delta \ln(CAPITAL_{it}) + \underset{(0.0507)}{0.3097} \Delta \ln(LABOR_{it}) \quad (0.0755)$$

There is a remarkable reduction in the estimated effect of the capital stock, which is no longer statistically significant. The estimated effect of labor is smaller but still significantly different from zero. The difference estimator is consistent when unobserved heterogeneity is correlated with the explanatory variables, but the OLS estimator is not. Given the substantial difference in the estimates we might suspect that the OLS estimates are unreliable.

EXAMPLE 15.3 | Using $T = 2$ Differenced Observations for a Wage Equation

Table 15.1 illustrates a panel data set with 5 years of data on 716 women. Consider only the final 2 years of data, 1987 and 1988, so that we have $N \times T = 716 \times 2 = 1{,}432$ observations. We wish to estimate

$$\ln(WAGE_{it}) = \beta_1 + \beta_2 EDUC_i + \beta_3 EXPER_{it} + u_i + e_{it}$$

for $i = 1, \ldots, N = 716$. Note that $EDUC_i$ has no time subscript. In this sample, all the women had completed their education by the time they were first interviewed, and therefore $EDUC_i$ is time-invariant. As usual we are concerned about omitted variable bias in this model because a person's ability is unobservable. In this panel, data model ability is captured in the individual heterogeneity term u_i. Subtracting the 1987 observation from the 1988 observation, we have

$$\Delta \ln(WAGE_i) = \beta_3 \Delta EXPER_i + \Delta e_i$$

The variable *EDUC* falls out of the model because it does not take at least two values. Using the first-difference

estimator eliminates any time-invariant variables and the intercept. The change in the log of wage is attributed to the change in experience. There is no omitted variable bias because the individual heterogeneity term, which includes ability, has subtracted out. It does not matter that ability might be correlated with years of education! Using data file *nls_panel2*, the OLS estimated first difference model is

$$\underset{\text{(se)}}{\widehat{\Delta \ln(WAGE_i)}} = \underset{(0.007141)}{0.0218} \Delta EXPER_i$$

15.2.2 The Within Estimator: $T = 2$

An alternative subtraction strategy is similar in spirit to that in equation (15.8). The advantage of the **within transformation** is that it generalizes nicely to situations when we have more than $T = 2$ time observations on each individual. We begin with the models for the two time periods in (15.7a) and (15.7b), then we find the time-average of the equations, that is,

$$\frac{1}{2}\sum_{t=1}^{2}\left(y_{it} = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + u_i + e_{it}\right)$$

On the left-hand side, we obtain $\bar{y}_{i\cdot} = \left(y_{i1} + y_{i2}\right)/2$. The "$\cdot$" is in the place of the second subscript t to remind us that it is an average over the time dimension. On the right-hand side, we obtain $\beta_1 + \beta_2 \bar{x}_{2i\cdot} + \alpha_1 w_{1i} + u_i + \bar{e}_{i\cdot}$, where the averaged variables are similarly defined: $\bar{x}_{2i\cdot} = \left(x_{2i1} + x_{2i2}\right)/2$ and $\bar{e}_{i\cdot} = \left(e_{i1} + e_{i2}\right)/2$. Note that the averaging does not affect the model parameters or the time-invariant terms β_1, w_{1i}, and u_i. The time-averaged model for $i = 1, \ldots, N$ is

$$\bar{y}_{i\cdot} = \beta_1 + \beta_2 \bar{x}_{2i\cdot} + \alpha_1 w_{1i} + u_i + \bar{e}_{i\cdot} \tag{15.10}$$

The within transformation subtracts (15.10) from the original observations to obtain

$$y_{it} - \bar{y}_{i\cdot} = \beta_2\left(x_{2it} - \bar{x}_{2i\cdot}\right) + \left(e_{it} - \bar{e}_{i\cdot}\right) \tag{15.11}$$

Instead of first-differenced variables, we have differences from the variable means. The time-invariant terms subtract out, including the unobservable heterogeneity term. Again do not be concerned about complicated data manipulations as econometric software has automatic commands to handle the process.

Let the transformed variables be denoted $\tilde{y}_{it} = y_{it} - \bar{y}_{i\cdot}$, $\tilde{x}_{2it} = \left(x_{2it} - \bar{x}_{2i\cdot}\right)$, with transformed error $\tilde{e}_{it} = \left(e_{it} - \bar{e}_{i\cdot}\right)$. The within-transformed model is

$$\tilde{y}_{it} = \beta_2 \tilde{x}_{2it} + \tilde{e}_{it} \tag{15.12}$$

The OLS estimator of β_2 using (15.12) is called the **within estimator**. It is a consistent estimator if (i) $\tilde{e}_{it}$ has zero mean and is uncorrelated with $\tilde{x}_{2it}$, and (ii) if $\tilde{x}_{2it}$ takes more than two values. The first condition is satisfied if (15.4) holds. Note that the variable $\tilde{x}_{2it} = \left(x_{2it} - \bar{x}_{2i\cdot}\right)$ incorporates the values of x_{2it} in all time periods because of the average term. Similarly $\tilde{e}_{it} = \left(e_{it} - \bar{e}_{i\cdot}\right)$ depends on the values of the idiosyncratic error in all time periods because of its average. Thus, strict exogeneity, equation (15.4) is required for consistent estimation of (15.12) by OLS. Once again there is no requirement that the unobserved heterogeneity u_i be uncorrelated with the explanatory variables.

EXAMPLE 15.4 | Using the Within Transformation with $T = 2$ Observations for a Production Function

Consider using the within transformation to the $T = 2$ sales observations in Example 15.2, to estimate the effect of changes in the capital stock and labor inputs on sales. To understand the within transformation precisely, examine the transformed data on *SALES* for the first two firms in Table 15.2. For 2005 the first difference of ln(*SALES*)

is missing, which is represented by a period, "·". The time-average of the 2 year $\ln(SALES)$ is $\overline{\ln(SALES_{it})}$, and the within transformation is $\widetilde{\ln(SALES_{it})}$. The **within** estimator uses only variation for each individual (within each individual) about the individual mean in order to estimate the parameters; it does not use variation across or between individuals in the estimation process.

There is no omitted variable bias using the within-transformed data because the time-invariant individual heterogeneity term, which includes any unmeasured characteristics of the firm, has subtracted out. Using the $N \times T = 200 \times 2 = 400$ observations the within estimates are

$$\begin{aligned}\widetilde{\ln(SALES_{it})} &= 0.0384\widetilde{\ln(CAPITAL_{it})} \\ \text{(se)} \quad & (0.0358) \\ \text{(se)} \quad & (0.0507) \\ &+ 0.3097\widetilde{\ln(LABOR_{it})} \\ & (0.0532) \qquad \text{(incorrect)} \\ & (0.0755) \qquad \text{(correct)}\end{aligned}$$

Notice that the within estimates are exactly the same as the first-difference estimates in Example 15.1. When $T = 2$, they will always be the same. Using OLS estimation software yields incorrect standard errors for the within estimator. The difference arises because the estimate of the error variance used by the OLS software uses the degrees of freedom $NT - 2 = 400 - 2 = 398$. The calculation ignores the loss of $N = 200$ degrees of freedom that occurs when the variables are corrected by their sample means. The correct divisor is $NT - N - 2 = 400 - 200 - 2 = 198$. Multiply the "incorrect" standard errors from the within estimates by the correction factor

$$\sqrt{(NT-2)/(NT-N-2)} = \sqrt{398/198} = 1.41778$$

The resulting "correct" standard errors are in fact identical to the standard errors from the first-difference estimator in Example 15.2. When using proper "within estimator" software this correction will automatically be done. In Section 15.2.4, we explain that most often software "within" estimator commands are called **fixed effects** estimation. The equality of the difference estimator and within estimator, and the correct standard errors, holds when $T = 2$, but not when $T > 2$.

TABLE 15.2 **Example 15.4: Transformed Sales Data**

FIRM	*YEAR*	$\ln(SALES_{it})$	$\Delta\ln(SALES_{it})$	$\overline{\ln(SALES_{it})}$	$\widetilde{\ln(SALES_{it})}$
1	2005	10.87933	·	11.08103	−0.2017047
1	2006	11.28274	0.40341	11.08103	0.2017053
2	2005	9.313799	·	9.444391	−0.1305923
2	2006	9.574984	0.261185	9.444391	0.1305927

Remark

In practice, there is no need to use the difference estimator, which was introduced as a pedagogical device to illustrate that it is possible to eliminate unobserved heterogeneity when panel data are available. Use the software option for "fixed effects" estimation.

15.2.3 The Within Estimator: $T > 2$

The advantage of the **within transformation** and use of the **within estimator** is that they generalize nicely to situations when we have more than $T = 2$ time observations on each individual. Suppose that we have T observations on each individual. So that

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + u_i + e_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T$$

Averaging over all time observations we have

$$\frac{1}{T}\sum_{t=1}^{T}(y_{it} = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + u_i + e_{it})$$

On the left-hand side, we obtain $\bar{y}_{i\cdot} = (y_{i1} + y_{i2} + \cdots + y_{iT})/T$. On the right-hand side, we obtain $\beta_1 + \beta_2\bar{x}_{2i\cdot} + \alpha_1 w_{1i} + u_i + \bar{e}_{i\cdot}$, where the averaged variables are similarly defined: $\bar{x}_{2i\cdot} = (x_{2i1} + \cdots + x_{2iT})/T$ and $\bar{e}_{i\cdot} = (e_{i1} + \cdots + e_{iT})/T$. Note that averaging does not affect the model parameters or the time-invariant terms w_{1i} and u_i. The time-averaged model, for $i = 1, \ldots, N$, is

$$\bar{y}_{i\cdot} = \beta_1 + \beta_2\bar{x}_{2i\cdot} + \alpha_1 w_{1i} + u_i + \bar{e}_{i\cdot} \tag{15.13}$$

The within transformation subtracts (15.13) from the original observations to obtain

$$y_{it} - \bar{y}_{i\cdot} = \beta_2(x_{2it} - \bar{x}_{2i\cdot}) + (e_{it} - \bar{e}_{i\cdot}) \tag{15.14}$$

Instead of first-differenced variables, we have differences from the variable means. The time-invariant variables subtract out, including the unobservable heterogeneity term.

Let the transformed variables be denoted $\tilde{y}_{it} = y_{it} - \bar{y}_{i\cdot}$, $\tilde{x}_{2it} = (x_{2it} - \bar{x}_{2i\cdot})$, with transformed error $\tilde{e}_{it} = (e_{it} - \bar{e}_{i\cdot})$. The within-transformed model is

$$\tilde{y}_{it} = \beta_2\tilde{x}_{2i} + \tilde{e}_{it} \tag{15.15}$$

The OLS estimator of β_2 in (15.15) is a consistent estimator if (i) $\tilde{e}_{it}$ has zero mean and is uncorrelated with $\tilde{x}_{2it}$, and (ii) if $\tilde{x}_{2it}$ takes more than two values. These conditions hold if the strict exogeneity assumption (15.4) holds. The usual OLS standard errors for (15.15) are not quite right but are easily corrected, as we explained in Example 15.4.

EXAMPLE 15.5 | Using the Within Transformation with $T = 3$ Observations for a Production Function

Consider using the within transformation to the $T = 3$ sales observations in the data file *chemical2*, from 2004 to 2006, for the 200 firms in Example 15.2, to estimate the effect of changes in the capital stock and labor inputs on sales. The within estimates are

$$\begin{aligned}
\widehat{\ln(SALES_{it})} &= 0.0889\widehat{\ln(CAPITAL_{it})} \\
\text{(se)} \quad & (0.0271) \\
\text{(se)} \quad & (0.0332) \\
&+ 0.3522\widehat{\ln(LABOR_{it})} \\
& (0.0413) \quad \text{(incorrect)} \\
& (0.0507) \quad \text{(correct)}
\end{aligned}$$

The incorrect standard errors are produced by OLS software using $NT - 2 = 598$ degrees of freedom when it should be $NT - N - 2 = 398$. Multiplying the incorrect standard errors by the correction factor

$$\sqrt{(NT-2)/(NT-N-2)} = \sqrt{598/398} = 1.22577$$

yields correct standard errors.

15.2.4 The Least Squares Dummy Variable Model

It turns out that the within estimator is numerically equivalent to another estimator that has long been used in empirical work and that is logically appealing. To be as general as possible, we expand our equation of interest to include more variables,

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \cdots + \beta_K x_{Kit} + \alpha_1 w_{1i} + \cdots + \alpha_M w_{Mi} + (u_i + e_{it}) \tag{15.16}$$

In this regression, there is a constant term, $x_{1it} = 1$, and $(K - 1) = K_S$ variables that vary across individuals and time, and also M variables that are time invariant. There is a new symbol, K_S, that can be thought of as the number of "slope" coefficients. This will be important below when we carry out a test for the existence of individual differences.

Unobserved heterogeneity is also controlled for by including in the panel data regression (15.16) an individual-specific indicator variable for each individual. That is, let

$$D_{1i} = \begin{cases} 1 & i = 1 \\ 0 & \text{otherwise} \end{cases}, \quad D_{2i} = \begin{cases} 1 & i = 2 \\ 0 & \text{otherwise} \end{cases}, \ldots\ldots, \quad D_{Ni} = \begin{cases} 1 & i = N \\ 0 & \text{otherwise} \end{cases}$$

Include these N indicator variables in the regression equation (15.16) to obtain

$$y_{it} = \beta_{11}D_{1i} + \beta_{12}D_{2i} + \cdots + \beta_{1N}D_{Ni} + \beta_1 + \beta_2 x_{2it} + \cdots + \beta_K x_{Kit} + \alpha_1 w_{1i} + \cdots + \alpha_M w_{Mi} + (u_i + e_{it})$$

In this equation there is exact collinearity. The time-invariant indicator variables sum to one, $D_{1i} + D_{2i} + \cdots + D_{Ni} = 1$. Including the indicator variables requires us to drop the now redundant constant term, $x_{1it} = 1$, the time-invariant variables, w_{1i}, w_{2i}, …, w_{Mi}, *and* the unobserved heterogeneity u_i. Doing so we are left with

$$y_{it} = \beta_{11}D_{1i} + \beta_{12}D_{2i} + \cdots + \beta_{1N}D_{Ni} + \beta_2 x_{2it} + \cdots + \beta_K x_{Kit} + e_{it} \tag{15.17}$$

Equation (15.17) is called the **fixed effects model**, or sometimes the **least squares dummy variable model**. The terminology **fixed effects** estimator, which is the most commonly used name in empirical work, arises because it is *as if* we are treating individual differences u_1, u_2, …, u_N, as fixed parameters, β_{11}, β_{12}, …, β_{1N}, that we can estimate. The fixed effects estimator is the OLS estimator of (15.17) using all NT observations.

Equation (15.17) is not estimated in practice unless N is small. Using the Frisch–Waugh–Lovell Theorem, Section 5.2.5 and Exercise 15.11, it can be shown that the OLS estimates of β_2, …, β_K in (15.17), and the sum of squared residuals, are identical to the within estimates of (15.16) and thus have the same consistency property under the same assumption (15.4). We remind you again that assumption (15.4) does not require that the unobserved heterogeneity term u_i be uncorrelated with $\mathbf{X}_i$ or $\mathbf{w}_i$, where $\mathbf{X}_i$ denotes all observations on the time-varying variables and $\mathbf{w}_i$ the observations on the time-invariant observations.

Remark

To summarize, the within estimator, the fixed effects estimator and the least squares dummy variable estimator are all names for the same estimators of β_2, …, β_K in (15.17). In practice, no choice is required. Use the computer software option for "fixed effects" estimation.

Because the fixed effects estimator is simply an OLS estimator, it has the usual OLS estimator variances and covariances. Including N indicator, dummy, variables means that the number of parameters is $N + K_S$, where $K_S = (K - 1)$ is the number of slope coefficients. The usual estimator of σ_e^2 is

$$\hat{\sigma}_e^2 = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{e}_{it}^2}{NT - N - K_S} \tag{15.18}$$

Testing for Unobserved Heterogeneity Testing for individual differences in the fixed effects model is a test of the joint hypothesis

$$H_0 : \beta_{11} = \beta_{12}, \ \beta_{12} = \beta_{13}, \ldots, \beta_{1,N-1} = \beta_{1N}$$
$$H_1 : \text{the } \beta_{1i} \text{ are not all equal} \tag{15.19}$$

If the null hypothesis is true, then $\beta_{11} = \beta_{12} = \beta_{13} = \cdots = \beta_{1N} = \beta_1$, where β_1 denotes the common value, and there are no individual differences and no unobserved heterogeneity. The null hypothesis is $J = N - 1$ separate equalities, $\beta_{11} = \beta_{12}$, $\beta_{12} = \beta_{13}$, and so on. If the null hypothesis is true, then the "restricted model" is

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \cdots + \beta_K x_{Kit} + e_{it}$$

Under the standard OLS assumptions, the F-test statistic is

$$F = \frac{(SSE_R - SSE_U)/(N-1)}{SSE_U/(NT - N - K_S)} \tag{15.20}$$

where SSE_U is the sum of squared residuals from the fixed effects model, and SSE_R is the sum of squared errors from the OLS regression that pools all the data, $y_{it} = \beta_1 + \beta_2 x_{2it} + \cdots + \beta_K x_{Kit} + e_{it}$. If the null hypothesis is true, the test statistic has the F-distribution with $J = N - 1$ numerator degrees of freedom and $NT - N - K_S$ denominator degrees of freedom. Using the α level of significance, we reject the null hypothesis if the test statistic value is greater than, or equal to, the $1 - \alpha$ percentile of the F-distribution, $F \geq F_{(1-\alpha,\ N-1,\ NT-N-K_S)}$. The test can be made "robust" to heteroskedasticity and serial correlation, topics that we consider in Section 15.3.

EXAMPLE 15.6 | Using the Fixed Effects Estimator with $T = 3$ Observations for a Production Function

For the Chinese chemical firm data file *chemical2*, the indicator variable model in (15.21) becomes

$$\ln(SALES_{it}) = \beta_{11}D_{1i} + \cdots + \beta_{1,200}D_{200,i} + \beta_2 \ln(CAPITAL_{it}) + \beta_3 \ln(LABOR_{it}) + e_{it}$$

The fixed effects estimates of β_2 and β_3 will be identical to the within estimates in Example 15.4, and the standard errors will be the correct ones because in this indicator variable model the degrees of freedom are the correct $NT - N - (K - 1) = 600 - 200 - 2 = 398$.

The $N = 200$ estimated indicator variable coefficients, $b_{11}, b_{12}, \ldots, b_{1N}$, may or may not be of specific interest. We include the indicator variables primarily to control for unobserved heterogeneity. If, however, we are interested in predicting the sales of a specific firm then the indicator variables become crucial. Given the estimates of β_2 and β_3, $b_{11}, b_{12}, \ldots, b_{1N}$ can be recovered using the fact that the fitted regression passes through the point of the means, just as it did in the simple regression model, that is, $\bar{y}_{i\cdot} = b_{1i} + b_2\bar{x}_{2i\cdot} + b_3\bar{x}_{3i\cdot}$, $i = 1, \ldots, N$. Reporting the estimates and their standard errors is inconvenient because N may be large. Software companies cope with this in different ways. Two popular econometric software programs, EViews and Stata, report a constant term C that is the average of the estimated coefficients on the cross-section indicator variables. For the Chinese chemical firm data, $C = N^{-1}\sum_{i=1}^{N} b_{1i} = 7.5782$.

To test the null hypothesis $H_0: \beta_{11} = \beta_{12}, \beta_{12} = \beta_{13}, \ldots, \beta_{1,N-1} = \beta_{1N}$, we use the sum of squared residuals from the fixed effects estimator, $SSE_U = 34.451469$, and from the pooled OLS regression

$$\widehat{\ln(SALES_{it})} = \underset{(\text{se})}{5.8797} + \underset{(0.1711)}{0.2732}\ln(CAPITAL_{it}) + \underset{(0.0467)}{0.3815}\ln(LABOR_{it})$$

with $SSE_R = 425.636557$. The F-statistic value is

$$F = \frac{(SSE_R - SSE_U)/(N-1)}{SSE_U/(NT - N - (K-1))} = \frac{(425.636557 - 34.451469)/199}{34.451469/(600 - 200 - 2)} = 22.71$$

Using the $\alpha = 0.01$ level of significance, $F_{(0.99,\ 199,\ 398)} = 1.32$. We reject the null hypothesis and conclude that there are individual differences in the fixed effects constant terms for these $N = 200$ firms.

15.3 Panel Data Regression Error Assumptions

In Section 15.2, we considered estimation strategies that eliminate unobservable heterogeneity, u_i, so that when it is correlated with the explanatory variables we can still consistently estimate the coefficients of variables, x_{kit}, that vary across individuals and time. In this section and the next, we

propose estimation methods for the cases in which unobservable heterogeneity, u_i, is not correlated with the explanatory variables, either the **time-varying variables**, x_{kit}, or the **time-invariant variables**, w_{mi}, so that we can use OLS estimation, or a more efficient generalized least squares estimator, GLS, called the random effects (RE) estimator. Because these estimators do not eliminate unobservable heterogeneity, u_i, from the estimation equation we must make a more complete set of assumptions than we did in Section 15.2.

Panel data model estimation and inference for the model $y_{it} = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + (u_i + e_{it})$ are complicated by the presence of two random errors. The first, u_i, accounts for time invariant unobserved heterogeneity across individuals. The second, e_{it}, is the "usual" regression error that varies across individuals and time. To be as general as possible, we return to equation (15.16), which we repeat here for your convenience,

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \cdots + \beta_K x_{Kit} + \alpha_1 w_{1i} + \cdots + \alpha_M w_{Mi} + (u_i + e_{it}) \tag{15.16}$$

As we have done in earlier chapters, let $\mathbf{x}_{it} = (1, x_{2it}, \ldots, x_{Kit})$ represent the tth observation on all time-varying variables, plus the intercept, for an individual, and let $\mathbf{X}_i$ represent all T observations on these variables for the ith individual. Let $\mathbf{w}_i = (w_{1i}, \ldots, w_{Mi})$ represent all the time-invariant variables for the ith individual. We discussed the important exogeneity assumption (15.4) that leads to the panel data regression function in (15.3). With the more complete model specification, assumption (15.4) becomes

$$E(e_{it}|\mathbf{X}_i, \mathbf{w}_i, u_i) = 0 \tag{15.21}$$

Recall that the strict exogeneity assumption in (15.21) means that neither $\mathbf{X}_i$, nor $\mathbf{w}_i$, nor u_i contain any information about the possible value of the idiosyncratic random error e_{it}.

The idiosyncratic random errors e_{it} and the unobservable heterogeneity random error u_i capture quite different effects and it is plausible to treat them as statistically independent, so that there is no correlation between them. In order for the OLS estimator of (15.16) to be unbiased a strong assumption, similar to (15.21), must hold for the unobserved heterogeneity term, u_i. If the explanatory variables $\mathbf{X}_i$ and $\mathbf{w}_i$ carry no information about random error component u_i then its best prediction is zero, meaning that

$$E(u_i|\mathbf{X}_i, \mathbf{w}_i) = 0 \tag{15.22}$$

Using the law of iterated expectations, it follows that

$$E(u_i) = 0, \quad \text{cov}(u_i, x_{kit}) = E(u_i x_{kit}) = 0, \quad \text{cov}(u_i, w_{mi}) = E(u_i w_{mi}) = 0 \tag{15.23}$$

The two assumptions (15.21) and (15.22) are sufficient to ensure that the OLS estimator is unbiased and consistent.

Remark

The verb "pool" means to combine or merge things. Consequently, econometricians talk about the combined data of all individuals in all time periods as a **pooled sample**. Then the regression equation (15.16) is a **pooled model** and if we apply OLS to this pooled model it is called **pooled least squares**, or **pooled OLS**. However, pooled OLS is nothing new; it is simply the OLS estimator applied to the combined data.

Now we ask about other assumptions, namely the random error conditional variances and covariances.

Conditional Homoskedasticity The usual homoskedasticity assumption for the idiosyncratic error e_{it} is that the conditional and unconditional variances are constant,

$$\text{var}(e_{it}|\mathbf{X}_i, \mathbf{w}_i, u_i) = \sigma_e^2 \tag{15.24a}$$

Using the variance decomposition discussed in Appendix B.1.8, and the law of iterated expectations, it also follows that

$$\text{var}(e_{it}) = E(e_{it}^2) = \sigma_e^2 \tag{15.24b}$$

Similarly, the unobserved heterogeneity random component u_i is conditionally and unconditionally homoskedastic,

$$\text{var}(u_i) = E(u_i^2) = \sigma_u^2 \tag{15.25}$$

If all individuals are drawn from one population, then homoskedasticity of u_i seems quite reasonable. However, the homoskedasticity of e_{it} is less likely to be true, for the usual reasons.

The variance of the combined error, $v_{it} = u_i + e_{it}$, is then

$$\text{var}(v_{it}|\mathbf{X}_i, \mathbf{w}_i) = \text{var}(u_i|\mathbf{X}_i, \mathbf{w}_i) + \text{var}(e_{it}|\mathbf{X}_i, \mathbf{w}_i) + 2\text{cov}(u_i, e_{it}|\mathbf{X}_i, \mathbf{w}_i)$$

Combining the two homoskedasticity assumptions and the statistical independence of u_i and e_{it}, we have

$$\text{var}(v_{it}) = E(v_{it}^2) = \sigma_v^2 = \sigma_u^2 + \sigma_e^2 \tag{15.26}$$

Conditionally Correlated When unobservable heterogeneity is recognized, the usual assumption that the errors are uncorrelated does not hold. To see this, find the covariance between the combined random errors in any two time periods,

$$\begin{aligned}\text{cov}(v_{it}, v_{is}) &= E(v_{it}v_{is}) = E\left[(u_i + e_{it})(u_i + e_{is})\right] \\ &= E(u_i^2 + u_i e_{it} + u_i e_{is} + e_{it}e_{is}) \\ &= E(u_i^2) + E(u_i e_{it}) + E(u_i e_{is}) + E(e_{it}e_{is}) \\ &= \sigma_u^2\end{aligned} \tag{15.27}$$

There is a covariance between the random errors for the ith individual for observations in any two different time periods. The correlation between the errors is

$$\rho = \text{corr}(v_{it}, v_{is}) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} \tag{15.28}$$

Interestingly, the covariance and correlation are constant and take the same value whether we are considering errors one period apart, or two periods apart, or more. As long as we have a random sample of individuals, we do not need to worry about any correlation *between* individuals, so that v_{it}, and v_{js} are uncorrelated for $i \neq j$.

Because of the intra individual error correlation, caused by the unobservable heterogeneity, the OLS estimator is not BLUE, and the usual standard errors are not correct. We will address how "robust" standard errors are calculated in Section 15.3.1 and how to carry out GLS in Section 15.4.

15.3.1 OLS Estimation with Cluster-Robust Standard Errors

In the panel data, multiple regression model (15.16), under the conventional homoskedasticity and serial correlation assumptions, equations (15.24a), (15.24b), (15.25), and (15.26), we have

$$\text{var}(v_{it}) = \sigma_u^2 + \sigma_e^2$$

and

$$\text{cov}(v_{it}, v_{is}) = \sigma_u^2$$

It is possible, however, that $\text{var}(e_{it})$ changes from individual to individual and perhaps also across time. In that case, $\text{var}(e_{it}) = \sigma_{it}^2$. We will introduce a new notation to handle this new possibility. Let

$$\text{var}(v_{it}) = \sigma_u^2 + \sigma_{it}^2 = \psi_{it}^2 \tag{15.29}$$

The variance ψ_{it}^2 (ψ is the Greek letter "psi") is potentially different for each individual in each time period. This might be true even if there is no unobserved heterogeneity, $\sigma_u^2 = 0$, or if the unobserved heterogeneity has a different variance for each individual. Assumption (15.29) is perfectly general and fits all possibilities.

Next, what about possible correlations among the error terms? The covariance between the random errors v_{it} and v_{is} is

$$\begin{aligned}\text{cov}(v_{it}, v_{is}) &= E(v_{it}v_{is}) = E\left[(u_i + e_{it})(u_i + e_{is})\right] \\ &= E(u_i^2) + E(e_{it}e_{is}) \\ &= \sigma_u^2 + \text{cov}(e_{it}, e_{is})\end{aligned} \tag{15.30}$$

where we have assumed u_i and e_{it} are statistically independent, or at least uncorrelated. The term $\text{cov}(e_{it}, e_{is})$ is the covariance between the usual random error, the idiosyncratic part, for the ith individual in time period t and time period s. If there is serial correlation, or autocorrelation, in this component of error then $\text{cov}(e_{it}, e_{is}) \neq 0$. The serial correlation may be of the AR(1) form we studied in Section 9.5.3, but it could be some other pattern as well. For now, we will make the most general possible assumption, that it may differ across individuals, and may differ for each pair of time periods as well, so that $\text{cov}(e_{it}, e_{is}) = \sigma_{its}$. Then (15.26) becomes

$$\text{cov}(v_{it}, v_{is}) = \sigma_u^2 + \sigma_{its} = \psi_{its} \tag{15.31}$$

Note that (15.31) is still valid even if there is no unobserved heterogeneity, so that $\sigma_u^2 = 0$.

What are the consequences of using pooled least squares in the presence of the heteroskedasticity and correlation described by (15.29) and (15.31)? The least squares estimator is still consistent, but its standard errors are incorrect, implying hypothesis tests and interval estimates based on these standard errors will be invalid. Typically, the standard errors will be too small, overstating the reliability of the least squares estimator. Fortunately, there is a way of correcting the standard errors. We had a similar situation in Chapters 8 and 9. In Chapter 8, we saw how White's heteroskedasticity-consistent standard errors could be used for assessing the reliability of least squares estimates in a regression model with heteroskedasticity of unknown form. Least squares is not efficient in these circumstances—the GLS estimator has lower variance—but using least squares avoids the need to specify the nature of the heteroskedasticity, and if the sample is large then using least squares with White standard errors provide a valid basis for interval estimation and hypothesis testing. The Newey-West standard errors introduced in Chapter 9 served a similar function in an autocorrelated-error model. They provide a valid basis for inference using least squares estimates without the need to specify the nature of the autocorrelated-error process.

In a similar way, standard errors that are valid for the pooled least squares estimator under the assumptions in (15.29) and (15.31) can be computed. These standard errors have various names, being referred to as **panel-robust standard errors** or **cluster-robust standard errors**. The T time-series observations on individuals form the clusters of data. Deriving cluster-robust standard errors requires some difficult and tedious algebra, which we briefly describe in Appendix 15A.

Two Important Notes Now for some good news and then some not so good news. First, the good news is that cluster-robust standard errors can be used in many contexts other than with panel data. Any data containing **groups** of observations can be treated as clusters if there are

within-group correlations but no across-group correlations. If we have a large sample of firms, then the firms within the same industry might define a cluster. If we have a survey of households, we may treat geographical neighborhoods as clusters. Second, the not so good news, is that while now easily obtained, using cluster-robust standard errors is not always appropriate. In order for them to be reliable, the number of individuals N must be large relative to T, so that the panel is "short and wide." For example, if there are $N = 1000$ individuals (cross sections) and we observed each for $T = 3$ time periods, then cluster-robust standard errors should work well. In situations with few individuals (few clusters) using cluster-robust standard errors may lead to inaccurate inferences. Naturally there is a great deal of discussion about what is meant by "few." In the U.S. there are $N = 50$ states. According to Cameron and Miller[8] (page 341), "Current consensus appears to be that … 50 is enough for state-year panel data." However, when carrying out tests, the number of clusters should be treated as the sample size.

EXAMPLE 15.7 | Using Pooled OLS with Cluster-Robust Standard Errors for a Production Function

In Example 15.6, we found that there is strong evidence in favor of using the fixed effects estimator rather than the pooled OLS estimator using the Chinese chemical firm data. However, for the purpose of giving a numerical illustration of pooled OLS with and without clustering, we examine the baseline model in Example 15.2 using $N = 1000$ firms using data file *chemical3*. Table 15.3 shows the OLS estimates with conventional, heteroskedasticity robust, and cluster-robust standard errors, and t-statistic values.

Note that while the heteroskedasticity-corrected standard errors are larger than the conventional standard errors, the cluster-corrected standard errors are larger yet. Of course, the t-values become smaller with the increased standard errors.

TABLE 15.3 **Example 15.7: OLS Estimates with Alternative Standard Errors**

		Conventional		Heteroskedastic		Cluster-Robust	
	Coefficient	Std. Error	t-Value	Std. Error	t-value	Std. Error	t-Value
C	5.5408	0.0828	66.94	0.0890	62.24	0.1424	38.90
ln(*CAPITAL*)	0.3202	0.0153	20.90	0.0179	17.87	0.0273	11.72
ln(*LABOR*)	0.3948	0.0225	17.56	0.0258	15.33	0.0390	10.12

15.3.2 Fixed Effects Estimation with Cluster-Robust Standard Errors

Consider now the fixed effects estimation procedure that employs the "within" transformation shown in (15.14). The within transformation removes the unobserved heterogeneity so that only the idiosyncratic error e_{it} remains. It is possible that within the cluster of observations defining each individual cross-sectional unit there remains serial correlation and/or heteroskedasticity. Cluster-robust standard errors[9] can be applied to the data in "deviation from the cluster-mean form," as in (15.14), or the least squares dummy variable model in (15.17).

[8]Cameron, A. C., and Miller, D. L., "A Practitioner's Guide to Cluster-Robust Inference," *Journal of Human Resources*, 2015, 50(2), 317–373.

[9]Interestingly, the usual White heteroskedasticity robust standard errors are not valid when $T > 2$ (Cameron and Miller, 2015, p. 352). Some panel data software will automatically use cluster-robust standard errors when any kind of robust standard errors are requested.

EXAMPLE 15.8 | Using Fixed Effects and Cluster-Robust Standard Errors for a Production Function

In Example 15.7, we estimated the production function by OLS with alternative standard errors. Here using data file *chemical3*, we obtain the fixed effects estimates using $N = 1000$ firms with conventional standard errors and cluster-robust standard errors. The cluster-robust standard errors are substantially larger than the usual standard errors. When this is the case, using the cluster-robust standard errors is recommended if N is large and T is small, like they are in this case (Table 15.4).

TABLE 15.4 **Example 15.8: Fixed Effects Estimates with Alternative Standard Errors**

		Conventional		Cluster-Robust	
	Coefficient	Std. Error	*t*-Value	Std. Error	*t*-Value
C	7.9463	0.2143	37.07	0.3027	26.25
ln(*CAPITAL*)	0.1160	0.0195	5.94	0.0273	4.24
ln(*LABOR*)	0.2689	0.0307	8.77	0.0458	5.87

15.4 The Random Effects Estimator

Panel data applications fall into one of two types. The first type of application is when the unobserved heterogeneity term u_i is correlated with one or more of the explanatory variables. In this case, we use the fixed effects (within) or difference estimators because these estimators are consistent and converge in probability to the true population parameter values as the sample size increases. These estimators deal with unobserved heterogeneity by eliminating it through a transformation, eliminating the potential endogeneity problem arising from a correlation between the unobserved heterogeneity and the explanatory variables.

The second type of application is when the unobserved heterogeneity term u_i is **not** correlated with any of the explanatory variables. In this case, we can simply use pooled OLS estimation, with robust-cluster standard errors. If for our purposes the OLS estimator is sufficiently precise, then we are done. Subsequent hypothesis tests and interval estimates are valid in large samples. If the OLS estimator is not sufficiently precise, then, providing the other assumptions hold, we can use an asymptotically more efficient feasible generalized least squares (FGLS) estimator.

The panel data regression model (15.1) with unobserved heterogeneity is sometimes called the **random effects model** because individual differences are random from the point of view of the researcher. The unobservable heterogeneity terms u_i are the **random effects**. The FGLS estimator is called the **random effects estimator**. It takes into account equation (15.27), the error covariance within the observations for each individual that arises from the unobserved heterogeneity. The use of this estimator also presumes the zero conditional mean assumptions, equations (15.4), and homoskedasticity, equation (15.26).

The minimum variance, efficient, estimator for the model is a GLS estimator. As was the case when we had heteroskedasticity or autocorrelation, we can obtain the GLS estimator in the random effects model by applying OLS to a transformed model. The transformed model, using $K = 2$ and $M = 1$ in (15.16), is

$$y_{it}^* = \beta_1 x_{1it}^* + \beta_2 x_{2it}^* + \alpha_1 w_{1i}^* + v_{it}^* \tag{15.32}$$

where the transformed variables are

$$y_{it}^* = y_{it} - \alpha \bar{y}_{i\cdot}, \qquad x_{1it}^* = 1 - \alpha, \qquad x_{2it}^* = x_{2it} - \alpha \bar{x}_{2i\cdot}, \qquad w_{1i}^* = w_{1i}(1 - \alpha) \tag{15.33}$$

The transformation parameter α is between zero and one, $0 < \alpha < 1$, and is given by

$$\alpha = 1 - \frac{\sigma_e}{\sqrt{T\sigma_u^2 + \sigma_e^2}} \tag{15.34}$$

The variables $\bar{y}_{i\cdot}$ and $\bar{x}_{2i\cdot}$ are the individual time-averaged means (15.13), and w_{1i}^* is a fraction of w_{1i}. A key feature of the random effects model is that *time-invariant variables are not eliminated*. The transformed error term is $v_{it}^* = v_{it} - \alpha\bar{v}_i$. It can be shown that the transformed error v_{it}^* has constant variance σ_e^2 and is serially uncorrelated. The proof is long and tedious, so we will not inflict it on you.[10] Because the transformation parameter α depends on the unknown variances σ_e^2 and σ_u^2, these variances need to be estimated before OLS can be applied to (15.32). Some details of how the estimates $\hat{\sigma}_e^2$ and $\hat{\sigma}_u^2$ are obtained can be found in Appendix 15B. The random effects, feasible GLS, estimates are obtained by applying least squares to (15.32) with σ_e^2 and σ_u^2 replaced by $\hat{\sigma}_e^2$ and $\hat{\sigma}_u^2$ in (15.34). From (15.33) we can see that if $\alpha = 1$ the random effects estimator is identical to the fixed effects estimator and if $\alpha = 0$ the random effects estimator is identical to the OLS estimator. When $0 < \hat{\alpha} < 1$ the random effects estimates may be closer to the OLS estimates or the fixed effects estimates depending on the magnitude of $\hat{\alpha}$.

EXAMPLE 15.9 | Random Effects Estimation of a Production Function

To illustrate the random effects estimator, we use the data file *chemical3* from $N = 1{,}000$ Chinese chemical firms using $T = 3$ time periods. The random effects estimates of the production function are

$$\begin{aligned}
\widehat{\ln(SALES_{it})} &= 6.1718 + 0.2393\ln(CAPITAL_{it}) \\
\text{(se_fgls)} \quad & (0.1142)\ (0.0147) \\
\text{(se_clus)} \quad & (0.1428)\ (0.0221) \\
&+ 0.4140\ln(LABOR_{it}) \\
& (0.0220) \\
& (0.0327)
\end{aligned}$$

These random effects estimates are obtained using the estimated "partial-demeaning coefficient"

$$\hat{\alpha} = 1 - \frac{\hat{\sigma}_e}{\sqrt{T\hat{\sigma}_u^2 + \hat{\sigma}_e^2}} = 1 - \frac{0.3722}{\sqrt{3(0.6127) + 0.1385}} = 0.7353$$

Because $\hat{\alpha} = 0.7353$ is not close to zero or one, we see that the random effects estimates are quite different from the fixed effects estimates in Example 15.8 and also quite different from the OLS estimates in Example 15.7. Note that the cluster-robust standard errors for the random effects estimates are slightly larger than the conventional FGLS standard errors, suggesting that there may be serial correlation and/or heteroskedasticity in the overall error component e_{it}.

EXAMPLE 15.10 | Random Effects Estimation of a Wage Equation

In Table 15.1, we introduced panel data using observations from a typical microeconomic data source, the National Longitudinal Surveys (NLS). In Example 15.3, we introduced a simple wage equation and noted that in the data file *nls_panel*, all the women when first surveyed had completed their education, so that the variable *EDUC*, years of education, did not vary. This resulted in it dropping out when we applied the difference estimator. All time-invariant variables are eliminated when using the difference estimator or the fixed effects estimator. In this example, we extend the model used in Example 15.3.

Because the women in our microeconomic data panel were randomly selected from a larger population, it seems sensible to treat individual differences between the 716

[10] The details can be found in Wooldridge (2010), pp. 326–328.

women as random effects. Let us specify the wage equation to have dependent variable ln(*WAGE*) and explanatory variables years of education (*EDUC*); total labor force experience (*EXPER*) and its square; tenure in current job (*TENURE*) and its square; and indicator variables *BLACK*, *SOUTH*, and *UNION*.

The fixed and random effects estimates are given in Table 15.5 along with conventional, nonrobust standard errors and *t*-values. For the random effects estimates, we use the estimated transformation parameter

$$\hat{\alpha} = 1 - \frac{\hat{\sigma}_e}{\sqrt{T\hat{\sigma}_u^2 + \hat{\sigma}_e^2}} = 1 - \frac{0.1951}{\sqrt{5 \times 0.1083 + 0.0381}}$$
$$= 0.7437$$

Using this value to transform the data as in (15.33), then applying least squares to the transformed regression model in (15.32) yields the random effects estimates. Because the random effects estimator only partially de-means the data the time-invariant variables, *EDUC* and *BLACK*, are not eliminated. We are able to estimate the effects of years of education and race on ln(*WAGE*). We estimate that the return to education is about 7.3%, and that blacks have wages about 12% lower than whites, everything else held constant. Living in the South leads to wages about 8% lower, and union membership leads to wages about 8% higher, everything else held constant.

TABLE 15.5 **Example 15.10: Fixed and Random Effects Estimates of a Wage Equation**

	Fixed Effects			Random Effects		
Variable	**Coefficient**	**Std. Error***	***t*-Value**	**Coefficient**	**Std. Error***	***t*-Value**
C	1.4500	0.0401	36.12	0.5339	0.0799	6.68
EDUC				0.0733	0.0053	13.74
EXPER	0.0411	0.0066	6.21	0.0436	0.0064	6.86
$EXPER^2$	−0.0004	0.0003	−1.50	−0.0006	0.0003	−2.14
TENURE	0.0139	0.0033	4.24	0.0142	0.0032	4.47
$TENURE^2$	−0.0009	0.0002	−4.35	−0.0008	0.0002	−3.88
BLACK				−0.1167	0.0302	−3.86
SOUTH	−0.0163	0.0361	−0.45	−0.0818	0.0224	−3.65
UNION	0.0637	0.0143	4.47	0.0802	0.0132	6.07

*Conventional standard errors.

15.4.1 Testing for Random Effects

The magnitude of the correlation ρ in (15.28) is an important feature of the random effects model. If $u_i = 0$ for every individual, then there are no individual differences and no heterogeneity to account for. In such a case, the pooled OLS linear regression model is appropriate, and there is no need for either a fixed or a random effects model. We are assuming the error component u_i has expectation zero, $E(u_i|\mathbf{X}_i, \mathbf{w}_i) = 0$. If in addition u_i has a conditional variance of *zero*, then it is said to be a degenerate random variable; it is a constant with value equal to zero. In this case, if $\sigma_u^2 = 0$, then the correlation $\rho = 0$ and there is no random individual heterogeneity present in the data. We can test for the presence of heterogeneity by testing the null hypothesis $H_0 : \sigma_u^2 = 0$ against the alternative hypothesis $H_1 : \sigma_u^2 > 0$. If the null hypothesis is rejected, then we conclude that there are random individual differences among sample members, and that the random effects model might be appropriate. On the other hand, if we fail to reject the null hypothesis, then we have no evidence to conclude that random effects are present.

The Lagrange multiplier (LM) principle for test construction is very convenient in this case, because **LM tests** require estimation of only the restricted model that assumes that the null

hypothesis is true. If the null hypothesis is true, then $u_i = 0$ and the random effects model reduces to the usual linear regression model

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + e_{it}$$

The test statistic is based on the OLS residuals

$$\hat{e}_{it} = y_{it} - b_1 - b_2 x_{2it} - a_1 w_{1i}$$

The test statistic for balanced panels is

$$\text{LM} = \sqrt{\frac{NT}{2(T-1)}} \left\{ \frac{\sum_{i=1}^{N} \left(\sum_{t=1}^{T} \hat{e}_{it} \right)^2}{\sum_{i=1}^{N} \sum_{t=1}^{T} \hat{e}_{it}^2} - 1 \right\} \tag{15.35}$$

The numerator of the first term in curly brackets differs from the denominator because it contains terms like $2\hat{e}_{i1}\hat{e}_{i2} + 2\hat{e}_{i1}\hat{e}_{i3} + 2\hat{e}_{i2}\hat{e}_{i3} + \cdots$ whose sum will not be significantly different from zero if there is no correlation over time for each individual and will reflect a positive correlation if there is one. If the sum of the cross product terms is not significant, the first term in the curly brackets is not significantly different from one, and the term in the curly brackets is not significantly different from zero. If the sum of the cross product terms is significant, then the first term in the curly brackets will be significantly greater than one and LM will be positive.

If the null hypothesis $H_0 : \sigma_u^2 = 0$ is true, that is, there are no random effects, then LM $\sim N(0, 1)$ in large samples. Thus, we reject H_0 at significance level α and accept the alternative $H_1 : \sigma_u^2 > 0$ if LM $> z_{(1-\alpha)}$, where $z_{(1-\alpha)}$ is the $100(1-\alpha)$ percentile of the standard normal $N(0, 1)$ distribution.[11] This critical value is 1.645 if $\alpha = 0.05$ and 2.326 if $\alpha = 0.01$. Rejecting the null hypothesis leads us to conclude that random effects are present.

EXAMPLE 15.11 | Testing for Random Effects in a Production Function

Using the $N = 1000$ Chinese chemical firms data from *chemical3*, the value of the test statistic in (15.35) is $LM = 44.0637$. This is far greater than the $\alpha = 0.01$ critical value 2.326, so we reject the null hypothesis $H_0 : \sigma_u^2 = 0$ and conclude that $\sigma_u^2 > 0$; there is evidence of unobserved heterogeneity, or random effects, in the data.

15.4.2 A Hausman Test for Endogeneity in the Random Effects Model

The random effects model has one critical assumption that is often violated. If the random error $v_{it} = u_i + e_{it}$ is correlated with any of the right-hand side explanatory variables in a random effects model, then the least squares and GLS estimators of the parameters are biased and inconsistent.

[11] The original LM test due to Breusch and Pagan used LM^2 with the distribution under H_0 as $\chi^2_{(1)}$. Subsequent authors pointed out that the alternative hypothesis for using LM^2 is $H_1 : \sigma_u^2 \neq 0$, and that we can do better by using LM as a one-sided $N(0, 1)$ test with alternative hypothesis $H_1 : \sigma_u^2 > 0$. Some software, for example Stata, reports LM^2. The danger from using LM^2 is that LM < 0 is possible and should not be taken as evidence that $\sigma_u^2 > 0$. The adjustment for a chi-square test at significance α is to use the $100(1-2\alpha)$ percentile of the χ^2-distribution. This critical value for an $\alpha = 0.05$ test is 2.706 which is 1.645^2. It should only be used for LM > 0.

The problem of **endogenous regressors** was first considered in a general context in Chapter 10. The problem is common in random effects models because the individual-specific error component u_i may well be correlated with some of the explanatory variables. Such a correlation will cause the random effects estimator to be inconsistent. Recall that a wonderful feature of having panel data is that we can consistently estimate the model parameters using fixed effects, within, or difference estimators, without having to find instrumental variables as we did in Chapter 10. The ability to **test** whether the random effect u_i is correlated with some of the explanatory variables is important.

To check for any correlation between the error component u_i and the regressors in a random effects model, we can use a **Hausman test**. While the basic concept underlying the test is the same, the mechanics of this Hausman test are different from the Hausman test introduced in Chapter 10. In this case, the test compares the coefficient estimates from the random effects model to those from the fixed effects model. The idea underlying Hausman's test is that both the random effects and fixed effects estimators are consistent if there is no correlation between u_i and the explanatory variables x_{kit}. If both estimators are consistent, then they should converge to the true parameter values β_k in large samples. That is, in large samples, the random effects and fixed effects estimates should be similar. On the other hand, if u_i is correlated with any of the explanatory variables, then the random effects estimator is inconsistent for all the model coefficients, while the fixed effects estimator remains consistent. Thus in large samples, the fixed effects estimator converges to the true parameter values, but the random effects estimator converges to some other values that are not the values of the true parameters. In this case, we expect to see differences between the fixed and random effects estimates.

The test can be carried out coefficient by coefficient using a t-test, or jointly, using a chi-square test. Let us consider the t-test first. Denote the fixed effects estimate of β_k as $b_{FE,k}$, and let the random effects estimate be $b_{RE,k}$. Then the t-statistic for testing that there is no difference between the estimators, and thus that there is no correlation between u_i and any of the explanatory variables, is

$$t = \frac{b_{FE,k} - b_{RE,k}}{\left[\widehat{\text{var}}\left(b_{FE,k}\right) - \widehat{\text{var}}\left(b_{RE,k}\right)\right]^{1/2}} = \frac{b_{FE,k} - b_{RE,k}}{\left[\text{se}\left(b_{FE,k}\right)^2 - \text{se}\left(b_{RE,k}\right)^2\right]^{1/2}} \tag{15.36}$$

The test can be carried out for each coefficient, and if any of the t-values are statistically different from zero, then we conclude that one or more of the explanatory variables are correlated with the unobserved heterogeneity term u_i. In this t-statistic, it is important that the denominator is the estimated variance of the fixed effects estimator minus the estimated variance of the random effects estimator. The reason is that under the null hypothesis that u_i is uncorrelated with any of the explanatory variables, the random effects estimator will have a smaller variance than the fixed effects estimator, at least in large samples. Consequently, we expect to find $\widehat{\text{var}}\left(b_{FE,k}\right) - \widehat{\text{var}}\left(b_{RE,k}\right) > 0$, which is necessary for a valid test. A second interesting feature of this test statistic is that

$$\begin{aligned} \text{var}\left(b_{FE,k} - b_{RE,k}\right) &= \text{var}\left(b_{FE,k}\right) + \text{var}\left(b_{RE,k}\right) - 2\text{cov}\left(b_{FE,k}, b_{RE,k}\right) \\ &= \text{var}\left(b_{FE,k}\right) - \text{var}\left(b_{RE,k}\right) \end{aligned} \tag{15.37}$$

The unexpected result in the last line occurs because Hausman proved that, in this particular case, $\text{cov}\left(b_{FE,k}, b_{RE,k}\right) = \text{var}\left(b_{RE,k}\right)$.

More commonly, the Hausman test is automated by software packages to contrast the complete set of estimates. That is, we carry out a test of a joint hypothesis comparing all the coefficients. The Hausman contrast[12] test jointly checks how close the differences between

[12] Details of the joint test are beyond the scope of this book. A reference that contains a careful exposition of the t-test, and the chi-square test, is Wooldridge (2010), pp. 328–334.

the pairs of coefficients are to zero. When testing all the coefficients except the intercept the resulting test statistic has the $\chi^2_{(K_S)}$-distribution, where K_S is the number of coefficients of variables that vary across time and individuals, if the null hypothesis of no endogeneity is true. The form of the Hausman test in (15.36) and its χ^2-distribution equivalent are not valid for cluster-robust standard errors because under these more general assumptions it is no longer true that $\text{var}(b_{FE,k} - b_{RE,k}) = \text{var}(b_{FE,k}) - \text{var}(b_{RE,k})$.

EXAMPLE 15.12 | Testing for Endogenous Random Effects in a Production Function

Intuitively it would seem quite likely that there are unobserved characteristics of the Chinese chemical firms that might be correlated with the amount of labor and capital they use to produce their products. Let us test the differences in the coefficient β_2 of ln(*CAPITAL*) using the fixed effects estimates in Example 15.8 and the random effects estimates in Example 15.9 with conventional, nonrobust standard errors.

$$t = \frac{b_{FE,2} - b_{RE,2}}{\left[\text{se}(b_{FE,2})^2 - \text{se}(b_{RE,2})^2\right]^{1/2}} = \frac{0.1160 - 0.2393}{\left[(0.0195)^2 - (0.0147)\right]^{1/2}}$$

$$= \frac{-0.1233}{0.0129} = -9.55$$

We reject the null hypothesis that the difference in the estimators is zero, and conclude that there is endogeneity in the random effects model. Using the joint hypothesis test on the $K_S = K - 1 = 2$ coefficients yields a Hausman contrast test statistic of 98.82, which is greater than $\chi^2_{(0.95,2)} = 5.991$, leading us to conclude that there is correlation between the unobserved heterogeneity term and some of the explanatory variables. Both of these tests support the notion that in this example the random effects estimator is inconsistent, so that we should choose the fixed effects estimator for the empirical analysis.

EXAMPLE 15.13 | Testing for Endogenous Random Effects in a Wage Equation

Using the Hausman contrast test to compare the fixed and random effects estimates of the wage equation in Table 15.5 is limited to the six common coefficients. Using the individual coefficient *t*-tests you will find significant differences at the 5% level for the coefficients of $TENURE^2$, *SOUTH*, and *UNION*. The joint test for the equality of the common coefficients yields a χ^2-statistic value of 20.73 while $\chi^2_{(0.95,6)} = 12.592$. Thus both approaches lead us to conclude that there is correlation between the individual heterogeneity term and one or more of the explanatory variables and therefore the random effects estimator should not be used.

15.4.3 A Regression-Based Hausman Test

The Hausman test described in Section 15.4.2 is based on assumptions of homoskedasticity and no serial correlation. In particular, it is not robust to heteroskedasticity and/or serial correlation. A second annoying problem is that the calculated χ^2-statistic can come out to be a negative number in samples that are not large. Such a result makes no sense theoretically and is due to features of a particular sample. These problems can be avoided by using a "regression-based" Hausman test.

The test is based on an idea by Yair Mundlak, so that it is sometimes called the **Mundlak approach**. Mundlak's notion was that if the unobservable heterogeneity is correlated with the explanatory variables then perhaps the random effects are correlated with the time averages of the explanatory variables. Consider the general model in (15.16) with $K = 3$ and $M = 2$,

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \beta_3 x_{3it} + \alpha_1 w_{1i} + \alpha_2 w_{2i} + u_i + e_{it}$$

Mundlak's suggestion is that we consider

$$u_i = \gamma_1 + \gamma_2 \bar{x}_{2i.} + \gamma_3 \bar{x}_{3i.} + c_i \tag{15.38}$$

where $E(c_i|\mathbf{X}_i) = 0$. Just as in the omitted variables problem, the solution is to take the relationship out of the error term and put it into the model, leaving the error with conditional expectation zero, that is, specify the panel data model

$$\begin{aligned} y_{it} &= \beta_1 + \beta_2 x_{2it} + \beta_3 x_{3it} + \alpha_1 w_{1i} + \alpha_2 w_{2i} + u_i + e_{it} \\ &= \beta_1 + \beta_2 x_{2it} + \beta_3 x_{3it} + \alpha_1 w_{1i} + \alpha_2 w_{2i} + (\gamma_1 + \gamma_2 \bar{x}_{2i.} + \gamma_3 \bar{x}_{3i.} + c_i) + e_{it} \\ &= (\beta_1 + \gamma_1) + \beta_2 x_{2it} + \beta_3 x_{3it} + \alpha_1 w_{1i} + \alpha_2 w_{2i} + \gamma_2 \bar{x}_{2i.} + \gamma_3 \bar{x}_{3i.} + c_i + e_{it} \\ &= \delta_1 + \beta_2 x_{2it} + \beta_3 x_{3it} + \alpha_1 w_{1i} + \alpha_2 w_{2i} + \gamma_2 \bar{x}_{2i.} + \gamma_3 \bar{x}_{3i.} + (c_i + e_{it}) \end{aligned} \tag{15.39}$$

Mundlak suggested testing $H_0: \gamma_2 = 0, \gamma_3 = 0$ against the alternative $H_1: \gamma_2 \neq 0$ or $\gamma_3 \neq 0$. The null hypothesis is that there is no endogeneity arising from a correlation between the unobserved heterogeneity and the explanatory variables. The asymptotically valid Wald test statistic has a $\chi^2_{(2)}$ distribution in this case. This test statistic will never be negative, and it can be made robust to heteroskedasticity and/or serial correlation using cluster-robust standard errors.

Equation (15.39) can be estimated by OLS, with cluster-robust standard errors, or by random effects, which should be more efficient. Interestingly, both OLS and random effects estimation of (15.39) yield fixed effects estimates of β_2 and β_3. Furthermore OLS and random effects estimates of γ_2 and γ_3 are identical. These outcomes are illustrated in the next two examples.

EXAMPLE 15.14 | The Mundlak Approach for a Production Function

For the production function data file *chemical3*, with $N = 1000$ firms, we create the time averages of ln(*CAPITAL*) and ln(*LABOR*) denoting them by adding a "*BAR*" over the name. The results are reported in Table 15.6. We give the estimates and standard errors to many decimal places to make the points in the previous paragraph. First, compare the OLS coefficient estimates to the random effects (RE) estimates. They are identical. Second, compare the coefficients of ln(*CAPITAL*) and ln(*LABOR*) to the fixed effects estimates in Example 15.8 and see that they are the same. Finally, note that the cluster-robust standard errors for OLS are identical to the random effects cluster-robust standard errors. The Wald test statistic value for the null hypothesis $H_0: \gamma_2 = 0, \gamma_3 = 0$ is 56.59 using cluster-robust standard errors and is 97.0 using the conventional RE standard errors. The test critical value is $\chi^2_{(0.99,2)} = 9.210$, thus using either test we reject the null hypothesis and conclude that the unobserved firm effects are correlated with the capital and/or labor inputs.

TABLE 15.6 Mundlak Regressions for a Production Function

	OLS Coefficient	Cluster Std. Error	RE Coefficient	Conventional Std. Error	RE Coefficient	Cluster Std. Error
C	5.45532814	0.14841700	5.45532814	0.13713197	5.45532814	0.14841700
ln(*CAPITAL*)	0.11603986	0.02735145	0.11603988	0.01954950	0.11603988	0.02735146
ln(*LABOR*)	0.26888033	0.04582462	0.26888041	0.03067342	0.26888041	0.04582462
$\overline{\ln(CAPITAL)}$	0.22232028	0.04125492	0.22232026	0.03338482	0.22232026	0.04125492
$\overline{\ln(LABOR)}$	0.10949491	0.06220441	0.10949483	0.05009737	0.10949483	0.06220441
Mundlak test	56.59		97.00		56.59	

EXAMPLE 15.15 | The Mundlak Approach for a Wage Equation

For the wage equation add the time averages of *EXPER* and its square, *TENURE* and its square, *SOUTH* and *UNION*. Note that we cannot use time averages of *EDUC* and *BLACK* because these variables do not change over time and are already in the model. In Table 15.7, we report the random effects estimates and both conventional and cluster-robust standard errors. The Mundlak test statistic of joint significance of the time average coefficients using the former is 20.44 and for the latter is 17.26. There are six coefficients being tested, and the test critical value is $\chi^2_{(0.99,6)} = 16.812$. Thus, we reject the null hypothesis and conclude that a woman's unobserved characteristics are correlated with some of the explanatory variables. We also for convenience provide the fixed effects (FE) estimates with cluster-robust standard errors. Note that for the time-varying variables the RE and FE coefficients are identical.

TABLE 15.7 Mundlak Regressions for a Wage Equation

	Random Effects			Fixed Effects	
	Coefficient	Conventional Std. Error	Cluster Std. Error	Coefficient	Cluster Std. Error
C	0.4167	0.1358	0.1101	1.4500	0.0550
EDUC	0.0708	0.0054	0.0056		
EXPER	0.0411	0.0066	0.0082	0.0411	0.0082
EXPER2	−0.0004	0.0003	0.0003	−0.0004	0.0003
TENURE	0.0139	0.0033	0.0042	0.0139	0.0042
TENURE2	−0.0009	0.0002	0.0002	−0.0009	0.0002
BLACK	−0.1216	0.0317	0.0284		
SOUTH	−0.0163	0.0361	0.0585	−0.0163	0.0585
UNION	0.0637	0.0143	0.0169	0.0637	0.0169
$\overline{EXPER}$	0.0251	0.0244	0.0223		
$\overline{EXPER2}$	−0.0012	0.0010	0.0010		
$\overline{TENURE}$	0.0026	0.0126	0.0137		
$\overline{TENURE2}$	0.0004	0.0007	0.0008		
$\overline{SOUTH}$	−0.0890	0.0464	0.0652		
$\overline{UNION}$	0.0920	0.0382	0.0415		
Mundlak test		20.44	17.26		

15.4.4 The Hausman–Taylor Estimator

The outcome from our comparison of the fixed and random effects estimates of the wage equation in Example 15.10 poses a dilemma. Correlation between the explanatory variables and the random effects means the random effects estimator will be inconsistent. We can overcome the inconsistency problem by using the fixed effects estimator, but doing so means we can no longer estimate the effects of the time-invariant variables *EDUC* and *BLACK*. The wage return for an extra year of education, and whether or not there is wage discrimination on the basis of race, might be two important questions that we would like to answer.

To solve this dilemma, we ask: How did we cope with the endogeneity problem in Chapter 10? We did so by using instrumental variable estimation. Variables known as instruments that are correlated with the endogenous variables but uncorrelated with the equation error were introduced, leading to an instrumental variables estimator which has the desirable property of consistency. The **Hausman–Taylor estimator** is an instrumental variables estimator applied

to the **random effects model**, to overcome the problem of inconsistency caused by correlation between the random effects and some of the explanatory variables. To explain how it works consider the regression model

$$y_{it} = \beta_1 + \beta_2 x_{it,exog} + \beta_3 x_{it,endog} + \beta_3 w_{i,exog} + \beta_4 w_{i,endog} + u_i + e_{it} \tag{15.40}$$

We have divided the explanatory variables into four categories:

$x_{it,exog}$: exogenous variables that vary over time and individuals

$x_{it,endog}$: endogenous variables that vary over time and individuals

$w_{i,exog}$: time-invariant exogenous variables

$w_{i,endog}$: time-invariant endogenous variables

Equation (15.40) is written as if there is one variable of each type, but in practice, there could be more than one. For the Hausman–Taylor estimator to work the number of exogenous time-varying variables $(x_{it,exog})$ must be at least as great as the number of endogenous time-invariant variables $(w_{i,endog})$. This is the necessary condition for there to be enough instrumental variables.

Following Chapter 10, we need instruments for $x_{it,endog}$ and $w_{i,endog}$. Since the fixed effects transformation $\tilde{x}_{it,endog} = x_{it,endog} - \bar{x}_{i,endog}$ eliminates correlation with u_i, we have $\tilde{x}_{it,endog}$ as a suitable instrument for $x_{it,endog}$. Also, the variables $\bar{x}_{i,exog}$ are suitable instruments for $w_{i,endog}$. The exogenous variables in (15.40) can be viewed as instruments for themselves, making the complete instrument set $x_{it,exog}$, $\tilde{x}_{it,endog}$, $w_{i,exog}$, $\bar{x}_{i,exog}$. Hausman and Taylor modify this set slightly using $\tilde{x}_{it,exog}$, $\tilde{x}_{it,endog}$, $w_{i,exog}$, $\bar{x}_{i,exog}$, which can be shown to yield the same results. Their estimator is applied to the transformed GLS model

$$y^*_{it} = \beta_1 + \beta_2 x^*_{it,exog} + \beta_3 x^*_{it,endog} + \beta_3 w^*_{i,exog} + \beta_4 w^*_{i,endog} + v^*_{it}$$

where, for example, $y^*_{it} = y_{it} - \hat{\alpha}\bar{y}_i$, and $\hat{\alpha} = 1 - \hat{\sigma}_e \Big/ \sqrt{T\hat{\sigma}_u^2 + \hat{\sigma}_e^2}$. The estimate $\hat{\sigma}_e^2$ is obtained from fixed effects residuals; an auxiliary instrumental variables regression[13] is needed to find $\hat{\sigma}_u^2$.

EXAMPLE 15.16 | The Hausman–Taylor Estimator for a Wage Equation

For the wage equation used in Example 15.10, we will make the following assumptions

$$x_{it,exog} = \{EXPER, EXPER2, TENURE, TENURE2, UNION\}$$
$$x_{it,endog} = \{SOUTH\}$$
$$w_{i,exog} = \{BLACK\}$$
$$w_{i,endog} = \{EDUC\}$$

The variable *EDUC* is chosen as an endogenous variable on the grounds that it will be correlated with personal attributes such as ability and perseverance. It is less clear why *SOUTH* should be endogenous, but we include it as endogenous because its fixed and random effects estimates were vastly different. Perhaps those living in the South have special attributes. The remaining variables, experience, tenure, *UNION*, and *BLACK*, are assumed uncorrelated with the random effects.

Estimates for the wage equation are presented in Table 15.8. Compared to the random effects estimates, there

TABLE 15.8 **Hausman–Taylor Estimates of Wage Equation**

Variable	Coefficient	Std. Error	*t*-Value	*p*-Value
C	−0.75077	0.58624	−1.28	0.200
EDUC	0.17051	0.04446	3.83	0.000
EXPER	0.03991	0.00647	6.16	0.000
EXPER2	−0.00039	0.00027	−1.46	0.144
TENURE	0.01433	0.00316	4.53	0.000
TENURE2	−0.00085	0.00020	−4.32	0.000
BLACK	−0.03591	0.06007	−0.60	0.550
SOUTH	−0.03171	0.03485	−0.91	0.363
UNION	0.07197	0.01345	5.35	0.000

[13] Details can be found in book, Jeffrey Wooldridge (2010), pp. 358–361.

has been a dramatic increase in the estimated wage returns to education from 7.3% to 17%. The estimated effects for experience and tenure are similar. The wage reduction for *BLACK* is estimated as 3.6% rather than 11.7%, and the penalty for being in the *SOUTH* is also less, 3.1% instead of 8.2%. The instrumental-variable standard errors are mostly larger, particularly for *EDUC* and *BLACK* where the biggest changes in estimates have been observed. Which set of estimates is better will depend on how successful we have been at making the partition into exogenous and endogenous variables in (15.40) and whether the gain from having consistent estimates is sufficiently large to compensate for the increased variance of the instrumental variables estimators.

15.4.5 Summarizing Panel Data Assumptions

It will be convenient to have a summary of the assumptions under which the random effects and the fixed effects estimators are appropriate.

Random Effects Estimation Assumptions

RE1. $y_{it} = \beta_1 + \beta_2 x_{2it} + \cdots + \beta_K x_{Kit} + \alpha_1 w_{1i} + \cdots + \alpha_M w_{Mi} + (u_i + e_{it})$. This is the population regression function. It may include (i) variables x_{kit} that vary across both time and individuals, (ii) time-invariant variables (w_{mi}), and (iii) variables that vary only across time, such as z_{gt}, although we have not included them explicitly. It includes unobserved idiosyncratic random errors, e_{it}, that vary across both time and individuals, and (ii) unobserved individual heterogeneity, u_i, that varies across individuals but not time.

RE2. (i) $E(e_{it}|\mathbf{X}_i, \mathbf{w}_i, u_i) = 0$ and (ii) $E(u_i|\mathbf{X}_i, \mathbf{w}_i) = E(u_i) = 0$. These are the exogeneity assumptions. Condition (i) says there is no information in the values of the explanatory variables or the unobserved heterogeneity that can be used to predict the values of e_{it}. Condition (ii) says there is no information in the values of the explanatory variables that can be used to predict u_i.

RE3. (i) $\text{var}(e_{it}|\mathbf{X}_i, \mathbf{w}_i, u_i) = \text{var}(e_{it}) = \sigma_e^2$ and (ii) $\text{var}(u_i|\mathbf{X}_i, \mathbf{w}_i) = \text{var}(u_i) = \sigma_u^2$. These are the homoskedasticity assumptions.

RE4. (i) Individuals are drawn randomly from the population, so that e_{it} is statistically independent of e_{js}; (ii) the random errors e_{it} and u_i are statistically independent; and (iii) $\text{cov}(e_{it}, e_{is}|\mathbf{X}_i, \mathbf{w}_i, u_i) = 0$ if $t \neq s$, the random errors e_{it} are serially uncorrelated.

RE5. There is no exact collinearity and all observable variables exhibit some variation.

Random Effects Estimator Notes

1. Under the assumptions RE1–RE5 the random effects (GLS) estimator is BLUE, assuming σ_e^2 and σ_u^2 are known.
2. Implementation of the random effects estimator requires the variance parameters to be estimated. The FGLS estimator is not BLUE, but it is consistent and asymptotically normal as N grows large if T is fixed, and it is asymptotically equivalent to the GLS estimator.
3. If the random errors are either heteroskedastic (RE3 fails) and/or serially correlated (RE4 (iii) fails), then the random effects estimator is consistent and asymptotically normal, but the usual standard errors are incorrect. Using cluster-robust standard errors provides a basis for valid asymptotic inference, including hypothesis tests and interval estimation.
4. Under RE1–RE5 the pooled OLS estimator is consistent and asymptotically normal.

5. Under RE1–RE5 the random effects, FGLS, estimator is more efficient asymptotically than the pooled OLS estimator with corrected cluster-robust standard errors.
6. The random effects estimator is more efficient in large samples than the fixed effects estimator for the coefficients of the variables that vary across individuals and time, x_{kit}.
7. The fixed effects estimator is, however, consistent for the coefficients of the variables that vary across individuals and time, x_{kit}, even if RE2 (ii) fails, and $E(u_i|\mathbf{X}_i, \mathbf{w}_i) \neq 0$.

Fixed Effects Estimation Assumptions

FE1. $y_{it} = \beta_1 + \beta_2 x_{2it} + \cdots + \beta_K x_{Kit} + (u_i + e_{it})$. This is the population regression function. It may include (i) variables x_{kit} that vary across both time and individuals and (ii) variables that vary only across time, such as z_{gt}, although we have not included them explicitly. It includes unobserved idiosyncratic random errors e_{it} that vary across both time and individuals, (ii) unobserved individual heterogeneity u_i that varies across individuals but not time. Note that we cannot include time-invariant variables.

FE2. $E(e_{it}|\mathbf{X}_i, u_i) = 0$. This is the (strict) exogeneity assumptions. There is no information in the values of the explanatory variables or the unobserved heterogeneity that can be used to predict the values of e_{it}. Note that we do not have to make any assumption about the relationship between the unobserved heterogeneity and the explanatory variables.

FE3. $\text{var}(e_{it}|\mathbf{X}_i, u_i) = \text{var}(e_{it}) = \sigma_e^2$. The random errors e_{it} are homoskedastic.

FE4. (i) Individuals are drawn randomly from the population, so that e_{it} is statistically independent of e_{js}, and (ii) $\text{cov}(e_{it}, e_{is}|\mathbf{X}_i, u_i) = 0$ if $t \neq s$, the random errors e_{it} are serially uncorrelated.

FE5. There is no exact collinearity and all observable variables exhibit some variation.

Fixed Effects Estimation Notes

1. Under FE1–FE5 the fixed effects estimator is BLUE.
2. The fixed effects estimator is consistent and asymptotically normal if N grows large and T is fixed.
3. If the random errors are either heteroskedastic (FE3 fails) and/or serially correlated (FE4 (ii) fails), then the fixed effects estimator is consistent and asymptotically normal, but the usual standard errors are incorrect. Using cluster-robust standard errors provides a basis for valid asymptotic inference, including hypothesis tests and interval estimation.

15.4.6 Summarizing and Extending Panel Data Model Estimation

The most common problem facing researchers using panel data is that unobservable characteristics of the cross-sectional unit, the "individual," are correlated with one or more of the explanatory variables. In this case, one or more of the explanatory variables are endogenous, so that OLS and the more efficient random effects estimator are inconsistent. Most of the time empirical researchers will use the fixed effects estimator because it eliminates the time-invariant unobserved heterogeneity term that causes the endogeneity problem. The fixed effects estimator is a consistent, but inefficient, estimator. Because of the major differences in the estimators, in each application using panel data, it is important to check for endogeneity using a Hausman or Mundlak test. Similarly, it is important to test for the presence of individual differences across individuals using the F-test with fixed effects estimation or the LM test for random effects.

Each of the estimators is subject to the usual problems of serial correlation and heteroskedasticity, but these problems are easily accounted for by using cluster-robust standard errors if the number of cross-sectional units N is much bigger than the time dimension T. A more perplexing problem for users of the fixed effects estimator is that time-invariant variables are eliminated from the model. In many applications, variables such as race and sex are vitally important. Using the Hausman–Taylor estimator solves the endogeneity problem by using instrumental variables estimation and does not eliminate the time-invariant variables. It can be a good choice if the IV are strong, and if there are enough time-varying exogenous variables. Another option is to use the Mundlak approach as a compromise, that is, assume that the unobserved heterogeneity depends on the time-averages of the variables varying over individual and time, as in (15.38). Once the time-averages are included in the model, if the remaining unobserved heterogeneity is not correlated with the included variables, then estimate an augmented model, like (15.39) by random effects.

Now, we briefly touch some other panel data issues.[14]

1. While we have not discussed it, panel data methods have been extended to **unbalanced panels**. These are cases when the number of time-series observations T_i differs across individuals.
2. In addition to unobserved heterogeneity associated with individuals, there can also be unobserved heterogeneity associated with time. Let m_t be a random time-specific error component. Note that the subscript is "t" only, so that it does not vary across individuals, only time. The combined error term has three terms, $v_{it} = u_i + m_t + e_{it}$. It is possible to carry out random effects estimation in this case with "two-way" error components models. A more common approach is to include a time-indicator variable in any model with relatively small T.
3. When $T = 2$, first-difference estimation is perfectly equivalent to fixed effects estimation. When $T > 2$, the first-difference random errors $\Delta v_{it} = \Delta e_{it}$ are serially correlated unless the idiosyncratic random errors e_{it} follow a random walk. This is diametrically opposite the usual fixed effects assumption that the idiosyncratic errors are serially uncorrelated. Using cluster-robust standard errors resolves the issue in both cases.
4. Dynamic panel data models that include a lagged dependent variable on the right-hand side have an endogeneity problem. To see this, let

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \beta_3 y_{i,t-1} + (u_i + e_{it})$$

   Note that y_{it} depends directly on u_i, and u_i is present in every time period including time $t - 1$. Therefore, $y_{i,t-1}$ also depends directly on u_i, causing a positive correlation, making $y_{i,t-1}$ endogenous. There is large literature on this difficult problem and many innovative IV estimators have been suggested. When T is large the dynamic, time-series data characteristics, must be taken into account. Using a difference estimator in this context is very common.
5. While we have focused on endogeneity resulting from the unobserved heterogeneity term, there can be endogeneity caused by simultaneous equations, such as supply and demand equations. There are IV/2SLS methods for estimating fixed effects, RE, and first-difference models.
6. In this edition, we have chosen to omit the section on "sets of regression equations" and "seemingly unrelated regressions." These topics arise when T is large and N is small, so that each cross-sectional unit, perhaps a firm, is modeled with its own equation.[15]

[14] You are encouraged to see Badi H. Baltagi (2013) *Econometric Analysis of Panel Data, Fifth Edition*, Wiley, along with previously cited textbooks by Greene (2018) and Wooldridge (2010) for more on these topics.

[15] See Greene, pp. 328–339, or the previous edition of this book, *Principles of Econometrics*, 4th ed., 2012, Chapter 15.7.

7. Unobserved heterogeneity can affect slope coefficients, that is, it is possible that each individual's response β_{ki} to a change in x_k is different. **Random coefficient models** recognize individual-specific slopes as a possibility.[16]
8. We have mentioned the **linear probability model** for situations in which individuals face binary choices. The panel data methods we have discussed can be used with linear probability models with the usual caveats. Looking forward to Chapter 16, we introduce new estimators, probit and logit, for handing binary outcome models. These too can be adapted for panel data methods.

15.5 Exercises

15.5.1 Problems

15.1 Consider the model

$$y_{it} = \beta_{1i} + \beta_2 x_{it} + e_{it}$$

a. Show that the fixed effects estimator for β_2 can be written as

$$\hat{\beta}_{2,FE} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)}{\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)^2}$$

b. Show that the random effects estimator for β_2 can be written as

$$\hat{\beta}_{2,RE} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\left[x_{it} - \hat{\alpha}\left(\bar{x}_i - \bar{\bar{x}}\right) - \bar{\bar{x}}\right]\left[y_{it} - \hat{\alpha}\left(\bar{y}_i - \bar{\bar{y}}\right) - \bar{\bar{y}}\right]}{\sum_{i=1}^{N}\sum_{t=1}^{T}\left[x_{it} - \hat{\alpha}\left(\bar{x}_i - \bar{\bar{x}}\right) - \bar{\bar{x}}\right]^2}$$

where $\bar{\bar{y}}$ and $\bar{\bar{x}}$ are the overall means.

c. Write down an expression for the pooled least squares estimator of β_2. Discuss the differences between the three estimators.

15.2 Consider the panel data regression model with unobserved heterogeneity, $y_{it} = \beta_1 + \beta_2 x_{it} + v_{it} = \beta_1 + \beta_2 x_{it} + u_i + e_{it}$. Given that assumptions RE1–RE5 hold, answer each of the following questions.

a. For the purpose of estimating the regression parameters precisely by OLS, the variance of the idiosyncratic error is more important than the variance of the unobserved heterogeneity error. True or False? Explain your choice.

b. For the purpose of estimating the regression parameters precisely by GLS, the variance of the idiosyncratic error is more important than the variance of the unobserved heterogeneity error. True or False? Explain your choice.

c. For the purpose of estimating the regression parameters precisely by fixed effects, the variance of the idiosyncratic error is more important than the variance of the unobserved heterogeneity error. True or False? Explain your choice.

15.3 In the random effects model, under assumptions RE1–RE5, suppose that the variance of the idiosyncratic error is $\sigma_e^2 = \text{var}(e_{it}) = 1$.

a. If the variance of the individual heterogeneity is $\sigma_u^2 = 1$, what is the correlation ρ between $v_{it} = u_i + e_{it}$ and $v_{is} = u_i + e_{is}$?

b. If the variance of the individual heterogeneity is $\sigma_u^2 = 1$, what is the value of the GLS transformation parameter α if $T = 2$? What is the value of the GLS transformation parameter α if $T = 5$?

[16] See, for example, Greene (2018), pp. 450–459, and Wooldridge (2010), pp. 374–387.

c. In general, for any given values of σ_u^2 and σ_e^2, as the time dimension T of the panel becomes larger, the transformation parameter α becomes smaller. Is this true, false, or are you uncertain? If you are uncertain, explain.

d. If $T = 2$ and $\sigma_e^2 = \text{var}(e_{it}) = 1$, what value of σ_u^2 will give the GLS transformation parameter $\alpha = 1/4$? What value of σ_u^2 will give the GLS transformation parameter $\alpha = 1/2$? What value of σ_u^2 will give the GLS transformation parameter $\alpha = 9/10$?

e. If we think of the random errors u_i and e_{it} as noise in the regression relationship, summarize how the relative variation of these noise components, the variances of error components, affects our ability to estimate the regression parameters.

15.4 Consider the regression model $y_{it} = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + u_i + e_{it}$, $i = 1, \ldots, N$, $t = 1, \ldots, T$, where x_{2it} and w_{1i} are explanatory variables. The time-averaged model is given in equation (15.13), $\bar{y}_{i\cdot} = \beta_1 + \beta_2 \bar{x}_{2i\cdot} + \alpha_1 w_{1i} + \bar{v}_{i\cdot}$ where $\bar{v}_{i\cdot} = u_i + \bar{e}_{i\cdot}$. The OLS estimator of the parameters in (15.13) is called the **between estimator**, because it uses variation between, or among, individuals to estimate the regression parameters.

a. Under assumptions RE1–RE5, derive the variance of the random error $\bar{v}_{i\cdot} = u_i + \bar{e}_{i\cdot}$.

b. Under assumptions RE1–RE5, find the covariance between $\bar{v}_{i\cdot}$ and $\bar{v}_{j\cdot}$, where $i \neq j$.

c. Under assumptions RE1–RE5, the between estimator is unbiased. Is this true or false? Explain the basis of your answer.

d. If assumptions RE1–RE5 hold except for RE2, part (ii), then the between estimator is biased and inconsistent. Is this true or false? Explain the basis of your answer.

15.5 Table 15.9 contains some simulated panel data, where *id* is the individual cross-section identifier, *t* is the time period, *x* is an explanatory variable, *e* is the idiosyncratic error, *y* is the outcome value. The data generating process is $y_{it} = 10 + 5x_{it} + u_i + e_{it}$, $i = 1, 2, 3$, $t = 1, 2$. The OLS residuals are $\hat{e}$, which we have rounded to two decimal places for convenience.

TABLE 15.9 **Simulated Data for Exercises 15.5 and 15.10**

id	*t*	*x*	*e*	*y*	$\hat{e}$
1	1	−0.51	−0.69	4.43	−3.21
1	2	−0.45	−1.70	1.70	−6.31
2	1	−2.44	−0.20	−2.29	2.20
2	2	−1.26	−0.41	2.98	0.06
3	1	−0.68	0.90	11.05	4.48
3	2	1.44	1.24	22.67	2.78

a. Using the true data generating process, calculate u_i, $i = 1, 2, 3$.

b. Calculate the value of the LM statistic in equation (15.35) and carry out a test for the presence of random effects at the 5% level of significance.

c. The fixed effects estimate of the coefficient of x_{it} is $b_{FE} = 5.21$ with standard error 0.94, while the random effects estimate is $b_{RE} = 5.31$ with standard error 0.81. Test for the presence of correlation between the unobserved heterogeneity u_i and the explanatory variable x_{it}. (*Note*: The sample is actually too small for this test to be valid.)

d. If estimates of the variance components are $\hat{\sigma}_u^2 = 34.84$ and $\hat{\sigma}_e^2 = 2.59$, calculate an estimated value of the GLS transformation parameter α. Based on its magnitude, would you expect the random effects estimates to be closer to the OLS estimates or the fixed effects estimates.

e. Using the estimates in (d), compute an estimate of the correlation between $v_{i1} = u_i + e_{i1}$ and $v_{i2} = u_i + e_{i2}$. Is this correlation relatively high, or relatively low?

15.6 Using the NLS panel data on $N = 716$ young women, we consider only years 1987 and 1988. We are interested in the relationship between ln(*WAGE*) and experience, its square, and indicator variables for living in the south and union membership. Some estimation results are in Table 15.10.

TABLE 15.10 **Estimation Results for Exercise 15.6**

	(1) OLS 1987	(2) OLS 1988	(3) FE	(4) FE Robust	(5) RE
C	0.9348	0.8993	1.5468	1.5468	1.1497
	(0.2010)	(0.2407)	(0.2522)	(0.2688)	(0.1597)
EXPER	0.1270	0.1265	0.0575	0.0575	0.0986
	(0.0295)	(0.0323)	(0.0330)	(0.0328)	(0.0220)
$EXPER^2$	−0.0033	−0.0031	−0.0012	−0.0012	−0.0023
	(0.0011)	(0.0011)	(0.0011)	(0.0011)	(0.0007)
SOUTH	−0.2128	−0.2384	−0.3261	−0.3261	−0.2326
	(0.0338)	(0.0344)	(0.1258)	(0.2495)	(0.0317)
UNION	0.1445	0.1102	0.0822	0.0822	0.1027
	(0.0382)	(0.0387)	(0.0312)	(0.0367)	(0.0245)
N	716	716	1432	1432	1432

(standard errors in parentheses)

a. The OLS estimates of the ln(*WAGE*) model for each of the years 1987 and 1988 are reported in columns (1) and (2). How do the results compare? For these individual year estimations, what are you assuming about the regression parameter values across individuals (heterogeneity)?

b. The ln(*WAGE*) equation specified as a panel data regression model is

$$\ln(WAGE_{it}) = \beta_1 + \beta_2 EXPER_{it} + \beta_3 EXPER_{it}^2 + \beta_4 SOUTH_{it} + \beta_5 UNION_{it} + (u_i + e_{it}) \quad \text{(XR15.6)}$$

Explain any differences in assumptions between this model and the models in part (a).

c. Column (3) contains the estimated fixed effects model specified in part (b). Compare these estimates with the OLS estimates. Which coefficients, apart from the intercepts, show the most difference?

d. The *F*-statistic for the null hypothesis that there are no individual differences, equation (15.20), is 11.68. What are the degrees of freedom of the *F*-distribution if the null hypothesis (15.19) is true? What is the 1% level of significance critical value for the test? What do you conclude about the null hypothesis.

e. Column (4) contains the fixed effects estimates with cluster-robust standard errors. In the context of this sample, explain the different assumptions you are making when you estimate with and without cluster-robust standard errors. Compare the standard errors with those in column (3). Which ones are substantially different? Are the robust ones larger or smaller?

f. Column (5) contains the random effects estimates. Which coefficients, apart from the intercepts, show the most difference from the fixed effects estimates? Use the Hausman test statistic (15.36) to test whether there are significant differences between the random effects estimates and the fixed effects estimates in column (3) (Why that one?). Based on the test results, is random effects estimation in this model appropriate?

15.7 Using the NLS panel data on $N = 716$ young women, we consider only years 1987 and 1988. We are interested in the relationship between ln(*WAGE*) and experience, its square, and indicator variables for living in the south and union membership. We form first differences of the variables, such as $\Delta\ln(WAGE) = \ln(WAGE_{i,1988}) - \ln(WAGE_{i,1987})$, and specify the regression

$$\Delta\ln(WAGE) = \beta_2 \Delta EXPER + \beta_3 \Delta EXPER^2 + \beta_4 \Delta SOUTH + \beta_5 \Delta UNION + \Delta e \quad \text{(XR15.7)}$$

Table 15.11 reports OLS estimates of equation (XR15.7) as Model (1), with conventional standard errors in parentheses.

TABLE 15.11 Estimates for Exercise 15.7

Model	C	$\Delta EXPER$	$\Delta EXPER^2$	$\Delta SOUTH$	$\Delta UNION$	$SOUTH_{i,1988}$	$UNION_{i,1988}$
(1)		0.0575	−0.0012	−0.3261	0.0822		
		(0.0330)	(0.0011)	(0.1258)	(0.0312)		
(2)	−0.0774	0.1187	−0.0014	−0.3453	0.0814		
	(0.0524)	(0.0530)	(0.0011)	(0.1264)	(0.0312)		
(3)		0.0668	−0.0012	−0.3157	0.0887	−0.0220	−0.0131
		(0.0338)	(0.0011)	(0.1261)	(0.0333)	(0.0185)	(0.0231)

a. The ability of first differencing to eliminate unobservable time-invariant heterogeneity is illustrated in equation (15.8). Explain why the strict form of exogeneity, FE2, is required for the difference estimator to be consistent. You may wish to reread the start of Section 15.1.2 to help clarify the assumption.

b. Equation (XR15.6) is the panel data regression specification at the base of the difference model. Suppose we define the indicator variable $D88_t = 1$ if the year is 1988 and $D88_t = 0$ otherwise, and add it to the specification in equation (XR15.6). What would its coefficient measure?

c. Model (2) in Table 15.11 is the difference model including an intercept term. Algebraically show that the constant term added to the difference model is the coefficient of the indicator variable discussed in part (b). Is the estimated coefficient statistically significant at the 5% level? What does this imply about the intercept parameter in equation (XR15.6) in 1987 versus 1988?

d. In the difference model, the assumption of strict exogeneity can be checked. Model (3) in Table 15.11 adds the variables *SOUTH* and *UNION* for year 1988 to the difference equation. As noted in equation (15.5a), the strict exogeneity assumption fails if the random error is correlated with the explanatory variables in any time period. We can check for such a correlation by including some, or all, of the explanatory variables for year t, or $t-1$ into the difference equation. If strict exogeneity holds these additional variables should not be significant. Based on the Model (3) result is there any evidence that the strict exogeneity assumption does not hold?

e. The F-test value for the joint significance of *SOUTH* and *UNION* from part (d), in Model (3), is 0.81. Are the variables jointly significant? What are the test degrees of freedom? What is the 5% critical value?

15.8 Using the NLS panel data on $N = 716$ young women, we are interested in the relationship between ln(*WAGE*) and experience, its square, and indicator variables for living in the south and union membership. The equation of interest is (XR15.6) in Exercise 15.6. Some estimation results are in Table 15.12. The estimates are based on 2864 observations covering the years 1982, 1983, 1985, and 1987. Standard errors are in parentheses.

TABLE 15.12 Estimates for Exercise 15.8

Model	C	$EXPER$	$EXPER^2$	$SOUTH$	$UNION$	$SOUTH_{1988}$	$UNION_{1988}$
(1)	1.3843	0.0565	−0.0011	0.0384	0.0459		
	(0.0487)	(0.0076)	(0.0003)	(0.0422)	(0.0160)		
(2)	1.3791	0.0564	−0.0011	0.0389	0.0478	0.0021	0.0160
	(0.0505)	(0.0076)	(0.0003)	(0.0451)	(0.0162)	(0.0481)	(0.0166)
robust	(0.0611)	(0.0084)	(0.0003)	(0.0636)	(0.0169)	(0.0581)	(0.0143)

a. Explain why the strict form of exogeneity, FE2, is required for the fixed effects estimator to be consistent. You may wish to reread the start of Section 15.1.2 to help clarify the assumption.

b. The fixed effects estimates of the regression coefficients and conventional standard errors are reported as Model (1). Are the coefficients significantly different from zero at the 5% level? What do the signs of the coefficients on experience and its square indicate about returns to experience?

c. In the fixed effects model, the assumption of strict exogeneity can be checked. Model (2) in Table 15.12 adds the variables *SOUTH* and *UNION* for year 1988 to the fixed effects equation and we report conventional standard and cluster-robust standard errors. As noted in equation (15.5a), the strict exogeneity assumption fails if the random error is correlated with the explanatory variables in any time period. We can check for such a correlation by including some, or all, of the explanatory variables for year $t + 1$ into the fixed effects model equation. If strict exogeneity holds these additional variables should not be significant. Based on the Model (2) result is there any evidence that the strict exogeneity assumption does not hold?

d. The joint F-test of $SOUTH_{1988}$ and $UNION_{1988}$ with conventional standard errors is 0.47. What are the degrees of freedom for the F-test? What is the 5% critical value? What do we conclude about strict exogeneity based on the joint test?

e. The joint F-test of $SOUTH_{1988}$ and $UNION_{1988}$ with robust-cluster standard errors is 0.63. When using a cluster-corrected covariance matrix the F-statistic used by some software has $M - 1$ denominator degrees of freedom, where M is the number of clusters. In this case, what is the 5% critical value? What do we conclude about strict exogeneity based on the robust joint test?

15.9 Examples 15.7 and 15.8 estimate a production function by OLS and fixed effects, respectively, with both conventional nonrobust standard errors and cluster-robust standard errors for $N = 1000$ Chinese chemical firms for 2004–2006.

a. Review the examples. What is the percent difference between the cluster-robust standard errors and the conventional standard errors?

b. Let $\hat{v}_{it}$ denote the OLS residuals from Example 15.7 and let $\hat{v}_{i,t-1}$ be the lagged residuals. Consider the regression $\hat{v}_{it} = \rho\hat{v}_{i,t-1} + r_{it}$, where r_{it} is an error term. Regressing the 2006 residuals on the 2005 residuals, we obtain $\hat{\rho} = 0.948$ with conventional OLS standard error 0.017 and White heteroskedasticity-consistent standard error 0.020. Do these results establish a time-series serial correlation in the idiosyncratic error component e_{it}? If not, what is the source of the strong correlation between $\hat{v}_{it}$ and $\hat{v}_{i,t-1}$?

c. Let $\hat{\tilde{e}}_{it}$ be the residuals from the within estimation, similar to Example 15.5, but using all 1000 firms. Let $\hat{\tilde{e}}_{i,t-1}$ be the lagged residuals. As noted in Exercise 15.10, part (e), we expect the errors in the "within" transformed model to be serially correlated with correlation $\text{corr}(\tilde{e}_{it}\tilde{e}_{is}) = -1/(T-1)$ under FE1-FE5. Here $T = 3$, thus we should find $\text{corr}(\tilde{e}_{it}\tilde{e}_{is}) = -1/2$. Consider the regression $\hat{\tilde{e}}_{it} = \rho\hat{\tilde{e}}_{i,t-1} + r_{it}$, where r_{it} is an error term. Using the 2006 data and $N = 1000$ observations, we estimate the value of ρ to be -0.233 with conventional standard error 0.046, and White heteroskedasticity robust standard error of 0.089. Test the null hypothesis $\rho = -1/2$ against the alternative $\rho \neq -1/2$ using a t-test at the 5% level, first with the conventional standard error and again with the heteroskedasticity robust standard error. Rejecting the null hypothesis implies that FE4, part (ii), does not hold, and time-series serial correlation exists in the idiosyncratic errors e_{it}. Such a finding justifies the use of cluster-robust standard errors in the fixed effects model regardless of any heteroskedasticity considerations.

d. Using the $N = 2000$ observations for 2005–2006, and the estimated regression $\hat{\tilde{e}}_{it} = \rho\hat{\tilde{e}}_{i,t-1} + r_{it}$, we estimate the value of ρ to be -0.270 with cluster-robust standard error, suggested by Wooldridge (2010, p. 311), of 0.017. Test the null hypothesis $\rho = -1/2$ against the alternative $\rho \neq -1/2$ using a t-test at the 5% level. Rejecting the null hypothesis implies that FE4, part (ii), does not hold, and time-series serial correlation exists in the idiosyncratic errors e_{it}.

15.10 This exercise uses the simulated data (y_{it}, x_{it}) in Table 15.9.

a. The fitted least squares dummy variable model, given in equation (15.17), is $\hat{y}_{it} = 5.57D_{1i} + 9.98D_{2i} + 14.88D_{3i} + 5.21x_{it}$. Compute the residuals from this estimated model for $id = 1$ and $id = 2$. What pattern do you observe in these residuals?

b. The same residual pattern occurs for $id = 3$. What is the correlation between the residuals for time periods $t = 1$ and $t = 2$?

c. The "within" model is given in equation (15.12). The transformed error is $\tilde{e}_{it} = (e_{it} - \bar{e}_{i\cdot})$. If the assumptions FE1–FE5 hold, then $\text{var}(\tilde{e}_{it}) = E\left[(e_{it} - \bar{e}_{i\cdot})^2\right]$, where $\bar{e}_{i\cdot} = (e_{i1} + e_{i2})/2$ because $T = 2$. Show that $\text{var}(\tilde{e}_{it}) = \sigma_e^2/2$.

d. Using the same approach, as in part (c), show that $\text{cov}(\tilde{e}_{i1}, \tilde{e}_{i2}) = E\left[(e_{i1} - \bar{e}_{i\cdot})(e_{i2} - \bar{e}_{i\cdot})\right] = -\sigma_e^2/2$.

e. Using the results in parts (c) and (d), it follows that $\text{corr}(\tilde{e}_{i1}, \tilde{e}_{i2}) = -1$. Relate this result to your answer in (b). In fact for $T > 1$, and assuming FE1-FE5 hold, $\text{corr}(\tilde{e}_{it}, \tilde{e}_{is}) = -1/(T-1)$ if $t \neq s$. We anticipate the within-transformed errors to be serially correlated.

15.11 Several software companies report fixed effects estimates with an estimated intercept. As explained in Example 15.6, the value they report is the average of the coefficients of the indicator variables in the least squares dummy variable model, given in equation (15.17). Using the data in Table 15.9, the fitted dummy variable model is $\hat{y}_{it} = 5.57D_{1i} + 9.98D_{2i} + 14.88D_{3i} + 5.21x_{it}$.

a. Compute the average of the dummy variable coefficients, calling it C.

b. The fitted fixed effects model, using the device from part (a), is $\hat{y}_{it} = C + 5.21x_{it}$. Calculate $\bar{y}_{i\cdot} - b_2\bar{x}_{2i\cdot}$ for $id = 1$ and $id = 2$. For your convenience, to two decimals, $\bar{y}_{1\cdot} = 3.07$, $\bar{y}_{2\cdot} = 0.34$ and $\bar{x}_{1\cdot} = -0.48$, $\bar{x}_{2\cdot} = -1.85$. Round the calculated values to two decimals and compare them to the dummy variable coefficients.

c. Given the fitted model $\hat{y}_{it} = C + 5.21x_{it}$, compute the residuals for $id = 1$ and $id = 2$.

d. What is the fitted within-model equation (15.17)?

e. Calculate the within-model residuals for $id = 1$ and $id = 2$.

f. Explain the relationship between the within model residuals in part (e) and the residuals calculated in part (c), apart from any error caused by the two decimal rounding.

15.12 Do larger universities have lower cost per student or a higher cost per student? A university is many things and here we only focus on the effect of undergraduate full-time student enrollment (*FTESTU*) on average total cost per student (*ACA*). Consider the regression model $ACA_{it} = \beta_1 + \beta_2 FTESTU_{it} + e_{it}$ where the subscripts i denote the university and t refers to the time period, and e_{it} is the usual random error term.

a. Using the 2010–2011 data on 141 public universities, we estimate the model above. The estimate of β_2 is $b_2 = 0.28$. The 95% interval estimate is [0.169, 0.392]. What is the estimated effect of increasing enrollment on average cost per student? Is there a statistically significant relationship?

b. There are many other factors affecting average cost per student besides enrollment. Some of them can be characterized as the university "identity" or "image." Let us denote these largely unobservable individual characteristics attributes as u_i. If we add this feature to the model, it becomes $ACA_{it} = \beta_1 + \beta_2 FTESTU_{it} + (u_i + e_{it}) = \beta_1 + \beta_2 FTESTU_{it} + v_{it}$. As long as v_{it} is statistically independent of full-time student enrollment, then the least squares estimator is BLUE. Is that true or false? Explain your answer.

c. The combined error is $v_{it} = u_i + e_{it}$. Let $\hat{v}_{it}$ be the least squares residual from the regression in (a). We then estimate a simple regression with dependent variable $\hat{v}_{i,2011}$ and explanatory variable $\hat{v}_{i,2010}$. The estimated coefficient is 0.93 and very significant. Is this evidence in support of the presence of unobservable individual attributes u_i, or against them? Explain your logic.

d. With our 2 years of data, we can take "first differences" of the model in (b). Subtracting the model in 2010 from the model in 2011, we have $\Delta ACA_i = \beta_2 \Delta FTESTU_i + \Delta v_i$, where

$$\Delta ACA_i = ACA_{i,2011} - ACA_{i,2010},$$

$$\Delta FTESTU_i = FTESTU_{i,2011} - FTESTU_{i,2010}$$

$$\text{and } \Delta v_i = v_{i,2011} - v_{i,2010}$$

Using the first-difference model, and given the results in (c), will there be serial correlation in the error Δv_i? Explain your reasoning.

e. Using OLS, we estimate the model in (d) and the resulting estimate of β_2 is $b_{FD} = -0.574$ with standard error $\text{se}(b_{FD}) = 0.107$. What now is the estimated effect of increasing enrollment on average cost per student? Explain why the result of this regression is so different from the pooled regression result in (a). Which set of estimates do you believe are more plausible? Why?

15.13 Consider the panel data regression in equation (15.1) for N cross-sectional units with $T = 3$ time-series observations. Assume that FE1–FE5 hold.

a. Apply the first-difference transformation to model (15.1). What is the resulting specification? Is there unobserved heterogeneity in this model? Explain.

b. Let $\Delta e_{it} = (e_{it} - e_{i,t-1})$. Find the variance of Δe_{it} for $t = 2$ and $t = 3$.

c. Assuming that the idiosyncratic error e_{it} is serially uncorrelated, show that the correlation between Δe_{i3} and Δe_{i2} is $-1/2$.

d. What must the serial correlation for e_{it} be in order for Δe_{i3} and Δe_{i2} to be uncorrelated?

15.14 Using the NLS panel data on $N = 716$ young women for years 1982, 1983, 1985, 1987, and 1988, we are interested in the relationship between ln(*WAGE*) and education, experience, its square, usual hours worked per week, and an indicator variable for black women. The equation is

$$\ln(WAGE_{it}) = \beta_1 + \beta_2 EDUC_i + \beta_3 EXPER_{it} + \beta_4 EXPER_{it}^2 + \beta_5 HOURS_{it} + \beta_6 BLACK_i + u_i + e_{it}$$

Table 15.13 contains OLS, random effects, and Hausman–Taylor model estimates for this model and includes conventional and cluster-robust standard errors for each. The Hausman–Taylor estimator treats *EDUC* and *HOURS* as endogenous and correlated with the unobserved heterogeneity.

TABLE 15.13 **Estimates for Exercise 15.14**

	C	*EDUC*	*EXPER*	$EXPER^2$	*HOURS*	*BLACK*
OLS	0.4509	0.0748	0.0631	−0.0012	−0.0008	−0.1347
(se)	(0.0617)	(0.0028)	(0.0080)	(0.0003)	(0.0008)	(0.0149)
(robust)	(0.1030)	(0.0055)	(0.0100)	(0.0004)	(0.0019)	(0.0290)
RE	0.6294	0.0769	0.0591	−0.0011	−0.0054	−0.1271
(se)	(0.0833)	(0.0055)	(0.0056)	(0.0002)	(0.0007)	(0.0298)
(robust)	(0.0999)	(0.0054)	(0.0069)	(0.0003)	(0.0017)	(0.0294)
HT	0.2153	0.1109	0.0583	−0.0011	−0.0063	−0.0910
(se)	(0.5536)	(0.0422)	(0.0057)	(0.0002)	(0.0007)	(0.0529)
(robust)	(0.4897)	(0.0381)	(0.0075)	(0.0003)	(0.0018)	(0.0494)

a. What is the interpretation of β_2? How much difference is there among the OLS, random effects, and Hausman–Taylor estimates of β_2? Construct a 95% interval estimate for β_2 using each estimator and cluster-robust standard errors. What differences do you observe?

b. For the Hausman–Taylor estimator, how many instrumental variables are required? How many instruments do we have? What are they?

c. For this model, why might we prefer the Hausman–Taylor estimator to the fixed effects estimator?

d. The fixed effects estimates of the coefficients of *EXPER*, $EXPER^2$, and *HOURS* and their conventional standard errors are 0.0584 (0.00574), −0.0011 (0.00023), and −0.0063 (0.00074), respectively. Comparing these estimates to the random effects estimates, with conventional standard errors, are we justified in worrying about endogeneity in this model?

e. By using cluster-robust standard errors for the random effects estimator, which of the assumptions RE1–RE5 are we relaxing?

f. Using the Hausman–Taylor model, $\hat{\sigma}_u = 0.35747$ and $\hat{\sigma}_e = 0.19384$. Given these estimates, which source of error variation is more important in this model? The variation in unobserved heterogeneity or the variation in the idiosyncratic error? What is the proportion of the combined variation that is accounted for by the unobserved heterogeneity?

15.15 Using 352 observations on 44 rice farmers in the Tarlac region of the Phillipines for 8 years from 1990 to 1997, we estimated the relationship between tonnes of freshly threshed rice produced (*PROD*), hectares planted (*AREA*), person-days of hired and family labor (*LABOR*), and kilograms of fertilizer (*FERT*). The log–log specification of the model, including the unobserved heterogeneity term, is

$$\ln(PROD_{it}) = \beta_1 + \beta_2 \ln(AREA_{it}) + \beta_3 \ln(LABOR_{it}) + \beta_4 \ln(FERT_{it}) + u_i + e_{it}$$

Table 15.14 contains various estimates of the model. Model (1) contains OLS estimates. Model (2) contains OLS estimates of the model including year dummy variables, which are not shown, such

as $D91 = 1$ for year 1991, $D91 = 0$ otherwise. Model (3) contains fixed effects estimates. Model (4) contains fixed effects estimates of the model including year dummy variables. In each case, conventional standard errors are reported, (se), and for Model (4), we also report cluster-robust standard errors (robust). For each model, we report the sum of squared residuals and the number of model parameters, apart from the intercept. The p-values are reported for the t-statistics computed using the conventional standard errors.

TABLE 15.14 Estimates for Exercise 15.15

Model		*C*	ln(*AREA*)	ln(*LABOR*)	ln(*FERT*)	*SSE*	*K*−*1*
(1)	OLS	−1.5468***	0.3617***	0.4328***	0.2095***	40.5654	3
	(se)	(0.2557)	(0.0640)	(0.0669)	(0.0383)		
(2)	OLS	−1.5549***	0.3759***	0.4221***	0.2075***	36.2031	10
	(se)	(0.2524)	(0.0618)	(0.0663)	(0.0380)		
(3)	FE	−0.3352	0.5841***	0.2586***	0.0952*	27.6623	46
	(se)	(0.3263)	(0.0802)	(0.0703)	(0.0432)		
(4)	FE	−0.3122	0.6243***	0.2412***	0.0890*	23.0824	53
	(se)	(0.3107)	(0.0755)	(0.0682)	(0.0415)		
	(robust)	(0.5748)	(0.0971)	(0.0968)	(0.0881)		

*$p < 0.05$
**$p < 0.01$
***$p < 0.001$

a. Comment on the sensitivity of the estimates of the input elasticities to the various models.
b. Which of the estimated models do you prefer? Perform a series of hypothesis tests to help you make your decision.
c. For Model (4), find 95% interval estimates for the input elasticities using (i) conventional standard errors and (ii) cluster-robust standard errors. Comment on any differences.
d. Calculate the p-value for the coefficient of ln(*FERT*) using the robust standard error.

15.5.2 Computer Exercises

15.16 The data file *liquor* contains observations on annual expenditure on liquor (*LIQUOR*) and annual income (*INCOME*), (both in thousands of dollars) for 40 randomly selected households for three consecutive years.

a. Using the data on *INCOME* for the first household, calculate the time average, within and differenced observations for *INCOME*. What is the sum of the within-transformed observations on *INCOME* for the first household?
b. Consider the panel data regression model $LIQUOR_{it} = \beta_1 + \beta_2 INCOME_{it} + u_i + e_{it}$ where $i = 1, 2, \ldots, 40$ refers to household and $t = 1, 2, 3$ refers to year. Obtain the OLS estimates of this model.
c. What are the fixed effects estimates of the parameters? What is the sum of squared residuals? Using the sum of squared residuals from the fixed effects estimates and the OLS estimation in (b), test for the presence of individual differences using an F-test. Show how the test statistic is computed. Using the 5% level of significance, what do we conclude?
d. Using OLS, regress *LIQUOR* on a constant term and 39 individual-specific indicator variables. Save the OLS residuals and call them *LIQUORW*. Regress *INCOME* on a constant term and 39 individual-specific indicator variables. Save the residuals and call them *INCOMEW*. Using OLS regress *LIQOURW* on *INCOMEW* without a constant term. What is the estimated coefficient? What is the sum of squared errors? How does this exercise illustrate the Frisch–Waugh–Lovell theorem discussed in Section 5.2.5?
e. Following Example 15.5, show how to correct the standard errors from the regression of *LIQOURW* on *INCOMEW* to make them match the fixed effects standard errors.

15.17 The data file *liquor* contains observations on annual expenditure on liquor (*LIQUOR*) and annual income (*INCOME*) (both in thousands of dollars) for 40 randomly selected households for three consecutive years.

a. Create the first-differenced observations on *LIQUOR* and *INCOME*. Call these new variables *LIQUORD* and *INCOMED*. Using OLS regress *LIQUORD* on *INCOMED* without a constant term. Construct a 95% interval estimate of the coefficient.

b. Estimate the model $LIQUOR_{it} = \beta_1 + \beta_2 INCOME_{it} + u_i + e_{it}$ using random effects. Construct a 95% interval estimate of the coefficient on *INCOME*. How does it compare to the interval in part (a)?

c. Test for the presence of random effects using the LM statistic in equation (15.35). Use the 5% level of significance.

d. For each individual, compute the time averages for the variable *INCOME*. Call this variable *INCOMEM*. Estimate the model $LIQUOR_{it} = \beta_1 + \beta_2 INCOME_{it} + \gamma INCOMEM_i + c_i + e_{it}$ using the random effects estimator. Test the significance of the coefficient γ at the 5% level. Based on this test, what can we conclude about the correlation between the random effect u_i and *INCOME*? Is it OK to use the random effects estimator for the model in (b)?

15.18 The data file *mexican* contains data collected in 2001 from the transactions of 754 female Mexican sex workers. There is information on four transactions per worker.[17] The labels *ID* and *TRANS* are used to describe a particular woman and a particular transaction. There are three categories of variables.

1. Sex worker characteristics: (i) *AGE*, (ii) an indicator variable *ATTRACTIVE* equal to 1 if the worker is attractive, and (iii) an indicator variable *SCHOOL* if she has completed secondary school or higher.
2. Client characteristics: (i) an indicator variable *REGULAR* equal to 1 if the client is a regular, (ii) an indicator variable *RICH* equal to 1 if the client is rich, and (iii) an indicator variable *ALCOHOL* if the client has consumed alcohol before the transaction.
3. Transaction characteristics: (i) the log of the price of the transaction *LNPRICE*, (ii) an indicator variable *NOCONDOM* equal to 1 if a condom was not used, and (iii) two indicator variables for location, *BAR* equal to 1 if the transaction originated in bar and *STREET* equal to 1 if the transaction originated in the street.

a. Using OLS, estimate a relationship with *LNPRICE* as the dependent variable, and as explanatory variables the sex worker characteristics, client characteristics, and transaction characteristics. Discuss the signs and significance of the estimated coefficients.

b. Gertler, Shah, and Bertozzi argue that the coefficient of *NOCONDOM* is a risk premium. Some sex workers are willing to take the risk of having unprotected sex because of the extra price some clients are willing to pay to avoid using a condom. What is your 95% interval estimate of the risk premium based on these OLS estimates?

c. What are some factors that might be included in an unobserved heterogeneity error component in this model? A crucial assumption for the consistency of the OLS estimator is that the unobserved heterogeneity term is uncorrelated with the explanatory variables. Without carrying out a formal test, what are your thoughts about this exogeneity assumption for the model in (a)?

d. Estimate the model in part (a) using the fixed effects estimator, omitting sex worker characteristics. (i) Why did we omit the sex worker characteristics? and (ii) Which coefficient estimates are significantly different from zero at a 5% level of significance?

e. Using the fixed effects estimation in (d), carry out an F-test for the presence of individual sex worker differences. Use the 1% level of significance.

f. Using the fixed effects estimates, how is the price affected when clients are rich, are regular, and have consumed alcohol? How does the location of the transaction influence the price?

g. What is your 95% interval estimate of the risk premium based on these fixed effects estimates? Compare this interval estimate to the one in part (b).

15.19 This exercise uses the data and model in Exercise 15.18.

a. Estimate the model assuming random effects and with the characteristics of the sex workers included in the model. Carry out a test of the joint significance of the sex worker characteristics at the 5% level. Are these coefficients jointly significant? Are they individually significant?

[17] These data are a subset of those used by Paul Gertler, Manisha Shah and Stefano Bertozzi in their study "Risky Business: The Market for Unprotected Sex", *Journal of Political Economy*, 2005, 113, 518–550.

b. What is your 95% interval estimate of the risk premium, the coefficient on *NOCONDOM*, based on these random effects estimates?

c. Test for the presence of random effects using the LM statistic in equation (15.35). Use the 5% level of significance.

d. Based on the random effects estimates, how much extra does a client have to pay to have unprotected sex with an attractive secondary-educated sex worker?

e. Using the *t*-test statistic in equation (15.36) and a 5% significance level, test whether there are any significant differences between the fixed effects and random effects estimates of the coefficients on *NOCONDOM*, *RICH*, *REGULAR*, *ALCOHOL*, *BAR*, and *STREET*. If there are significant differences between any of the coefficients, should we rely on the fixed effects estimates or the random effects estimates? Explain your choice.

f. Reconsider the random effects model from part (a), but assume *NOCONDOM* is correlated with the random effects. Reestimate the model using the Hausman–Taylor estimator with *NOCONDOM* treated as endogenous. Compare the results with those obtained in part (b). How much extra does a client have to pay to have unprotected sex with an attractive secondary-educated sex worker? What is your 95% interval estimate of the risk premium, the coefficient on *NOCONDOM*, based on the Hausman–Taylor estimates?

15.20 This exercise uses data from the STAR experiment introduced to illustrate fixed and random effects for grouped data. In the STAR experiment, children were randomly assigned within schools into three types of classes: small classes with 13–17 students, regular-sized classes with 22–25 students, and regular-sized classes with a full-time teacher aide to assist the teacher. Student scores on achievement tests were recorded as well as some information about the students, teachers, and schools. Data for the kindergarten classes are contained in the data file *star*.

a. Estimate a regression equation (with no fixed or random effects) where *READSCORE* is related to *SMALL*, *AIDE*, *TCHEXPER*, *BOY*, *WHITE_ASIAN*, and *FREELUNCH*. Discuss the results. Do students perform better in reading when they are in small classes? Does a teacher's aide improve scores? Do the students of more experienced teachers score higher on reading tests? Does the student's sex or race make a difference?

b. Reestimate the model in part (a) with school fixed effects. Compare the results with those in part (a). Have any of your conclusions changed? [*Hint*: specify *SCHID* as the cross-section identifier and *ID* as the "time" identifier.]

c. Test for the significance of the school fixed effects. Under what conditions would we expect the inclusion of significant fixed effects to have little influence on the coefficient estimates of the remaining variables?

d. Reestimate the model in part (a) with school random effects. Compare the results with those from parts (a) and (b). Are there any variables in the equation that might be correlated with the school effects? Use the LM test for the presence of random effects.

e. Using the *t*-test statistic in equation (15.36) and a 5% significance level, test whether there are any significant differences between the fixed effects and random effects estimates of the coefficients on *SMALL*, *AIDE*, *TCHEXPER*, *WHITE_ASIAN*, and *FREELUNCH*. What are the implications of the test outcomes? What happens if we apply the test to the fixed and random effects estimates of the coefficient on *BOY*?

f. Create school-averages of the variables and carry out the Mundlak test for correlation between them and the unobserved heterogeneity.

15.21 This exercise uses data from the STAR experiment introduced to illustrate fixed and random effects for grouped data. It replicates Exercise 15.20 with teachers (*TCHID*) being chosen as the cross section of interest. In the STAR experiment, children were randomly assigned within schools into three types of classes: small classes with 13–17 students, regular-sized classes with 22–25 students, and regular-sized classes with a full-time teacher aide to assist the teacher. Student scores on achievement tests were recorded as well as some information about the students, teachers, and schools. Data for the kindergarten classes are contained in the data file *star*.

a. Estimate a regression equation (with no fixed or random effects) where *READSCORE* is related to *SMALL*, *AIDE*, *TCHEXPER*, *TCHMASTERS*, *BOY*, *WHITE_ASIAN*, and *FREELUNCH*. Discuss the results. Do students perform better in reading when they are in small classes? Does a teacher's aide improve scores? Do the students of more experienced teachers score higher on reading tests? Does gender or race make a difference?

b. Repeat the estimation in (a) using cluster-robust standard errors, with the cluster defined by individual teachers, *TCHID*. Are the robust standard errors larger or smaller. Compare the 95% interval estimate for the coefficient of *SMALL* using conventional and robust standard errors.

c. Reestimate the model in part (a) with teacher random effects and using both conventional and cluster-robust standard errors. Compare these results with those from parts (a) and (b).

d. Are there any variables in the equation that might be correlated with the teacher effects? Recall that teachers were randomly assigned within schools, but not across schools. Create teacher-level averages of the variables *BOY*, *WHITE_ASIAN*, and *FREELUNCH* and carry out the Mundlak test for correlation between them and the unobserved heterogeneity.

e. Suppose that we treat *FREELUNCH* as endogenous. Use the Hausman–Taylor estimator for this model. Compare the results to the OLS estimates in (a) and the random effects estimates in part (d). Do you find any substantial differences?

15.22 What is the relationship between crime and punishment? This important question has been examined by Cornwell and Trumbull[18] using a panel of data from North Carolina. The cross sections are 90 counties, and the data are annual for the years 1981–1987. The data are in the data file *crime*. In these models, the crime rate is explained by variables describing the deterrence effect of the legal system, wages in the private sector (which represents returns to legal activities), socioeconomic conditions such as population density and the percentage of young males in the population, and annual dummy variables to control for time effects. The authors argue that there may be heterogeneity across counties (unobservable county-specific characteristics).

a. What do you expect will happen to the crime rate if (i) deterrence increases, (ii) wages in the private sector increase, (iii) population density increases, and (iv) the percentage of young males increases?

b. Consider a model in which the log of crime rate (*LCRMRTE*) is a function of the log of the probability of arrest (*LPRBARR*), the log of probability of conviction (*LPRBCONV*), the log of the probability of a prison sentence (*LPRBPRIS*), the log of average prison sentence (*LAVGSEN*), and the log of average weekly wage in the manufacturing sector (*LWMFG*). Estimate this model by OLS. (i) Discuss the signs of the estimated coefficients and their significance. Are they as you expected? (ii) Interpret the coefficient on *LPRBARR*.

c. Estimate the model in (b) using a fixed effects estimator. (i) Discuss the signs of the estimated coefficients and their significance. Are they as you expected? (ii) Interpret the coefficient on *LPRBARR* and compare it to the estimate in (b). What do you conclude about the deterrent effect of the probability of arrest? (iii) Interpret the coefficient on *LAVGSEN*. What do you conclude about the severity of punishment as a deterrent?

d. In the fixed effects estimation from part (c), test whether the county level effects are all equal.

e. Based on these results, what public policies would you advocate to deal with crime in the community?

15.23 Macroeconomists are interested in factors that explain economic growth. An aggregate production function specification was studied by Duffy and Papageorgiou.[19] The data are in the data file *ces*. They consist of cross-sectional data on 82 countries for 28 years, 1960–1987.

a. Estimate a Cobb–Douglas production function

$$LY_{it} = \beta_1 + \beta_2 LK_{it} + \beta_3 LL_{it} + e_{it}$$

where *LY* is the log of GDP, *LK* is the log of capital, and *LL* is the log of labor. Interpret the coefficients on *LK* and *LL*. Test the hypothesis that there are constant returns to scale, $\beta_2 + \beta_3 = 1$.

b. Add a time trend variable $t = 1, 2, \ldots, 28$, to the specification in (a). Interpret the coefficient of this variable. Test its significance at the 5% level. What effect does this addition have on the estimates of β_2 and β_3?

c. Assume $\beta_2 + \beta_3 = 1$. Solve for β_3 and substitute this expression into the model in (b). Show that the resulting model is $LYL_{it} = \beta_1 + \beta_2 LKL_{it} + \lambda t + e_{it}$ where *LYL* is the log of the output–labor ratio, and *LKL* is the log of the capital–labor ratio. Estimate this restricted, constant returns to

[18] "Estimating the Economic Model of Crime with Panel Data," *Review of Economics and Statistics*, 1994, 76, 360–366.

[19] "A Cross-Country Empirical Investigation of the Aggregate Production Function Specification," *Journal of Economic Growth*, 2000, 5, 83–116.

scale, version of the Cobb–Douglas production function. Compare the estimate of β_2 from this specification to that in part (b).

d. Estimate the model in (b) using a fixed effects estimator. Test the hypothesis that there are no cross-country differences. Compare the estimates to those in part (b).

e. Using the results in (d), test the hypothesis that $\beta_2 + \beta_3 = 1$. What do you conclude about constant returns to scale?

f. Estimate the restricted version of the Cobb–Douglas model in (c) using the fixed effects estimator. Compare the results to those in part (c). Which specification do you prefer? Explain your choice.

g. Using the specification in (b), replace the time trend variable t with dummy variables *D2–D28*. What is the effect of using this dummy variable specification rather than the single time trend variable?

15.24 This exercise illustrates the transformation that is necessary to produce GLS estimates for the random effects model. It utilizes the data on investment (*INV*), value (*V*) and capital (*K*) in the data file *grunfeld11*. The model is

$$INV_{it} = \beta_1 + \beta_2 V_{it} + \beta_3 K_{it} + u_i + e_{it}$$

We assume the random effects assumptions RE1–RE5 hold.

a. Find fixed effects estimates of β_2 and β_3. Check that the variance estimate that you obtain is $\hat{\sigma}_e^2 = 2530.042$.

b. Compute the sample means $\overline{INV}_i$, $\overline{V}_i$, and $\overline{K}_i$ for each of the 11 firms. [*Hint*: one way to do this to regress each of the variables (*INV*, then *V*, then *K*) on 11 indicator variables, 1 for each firm, and in each case save the predictions.]

c. Estimate β_1, β_2, and β_3 from the between regression

$$\overline{INV}_i = \beta_1 + \beta_2 \overline{V}_i + \beta_3 \overline{K}_i + u_i + \overline{e}_i.$$

Check that the variance estimate for $\sigma_*^2 = \text{var}(u_i + \overline{e}_i.)$ is $\hat{\sigma}_*^2 = 6328.554$. [*Hint*: use the predictions obtained in (b) to run the regression. If you do so, you will be using each of the N observations repeated T times. The coefficient estimates will be unaffected, but the sum of squared errors will be $T = 20$ times bigger than it should be, and the divisor used to estimate the error variance will be $NT - K$ instead of $N - K$. You will need to make adjustments accordingly.]

d. Show that

$$\hat{\alpha} = 1 - \sqrt{\frac{\hat{\sigma}_e^2}{T\hat{\sigma}_*^2}} = 0.85862$$

e. Apply least squares to the regression model

$$INV_{it}^* = \beta_1 x_1^* + \beta_2 V_{it}^* + \beta_3 K_{it}^* + v_{it}^*$$

where the transformed variables are given by $INV_{it}^* = INV_{it} - \hat{\alpha}\overline{INV}_i$, $x_1^* = 1 - \hat{\alpha}$, $V_{it}^* = V_{it} - \hat{\alpha}\overline{V}_i$, and $K_{it}^* = K_{it} - \hat{\alpha}\overline{K}_i$.

f. Use your software to obtain random effects estimates of the original equation. Compare those estimates with those you obtained in part (e).

15.25 Consider the production relationship on Chinese firms used in several chapter examples. We now add another input, *MATERIALS*. Use the data set from the data file *chemical3* for this exercise. (The data file *chemical* includes many more firms.)

$$\ln(SALES_{it}) = \beta_1 + \beta_2 \ln(CAPITAL_{it}) + \beta_3 \ln(LABOR_{it}) + \beta_4 \ln(MATERIALS_{it}) + u_i + e_{it}$$

a. Estimate this model using OLS. Compute conventional, heteroskedasticity robust, and cluster-robust standard errors. Using each type of standard error construct a 95% interval estimate for the elasticity of *SALES* with respect to *MATERIALS*. What do you observe about these intervals?

b. Using each type of standard error in part (a), test at the 5% level the null hypothesis of constant returns to scale, $\beta_2 + \beta_3 + \beta_4 = 1$ versus the alternative $\beta_2 + \beta_3 + \beta_4 \neq 1$. Are the results consistent?

c. Use the OLS residuals from (a) and carry out the $N \times R^2$ test from Chapter 9 to test for AR(1) serial correlation in the errors using the 2005 and 2006 data. Is there evidence of serial correlation? What factors might be causing it?

d. Estimate the model using random effects. How do these estimates compare to the OLS estimates? Test the null hypothesis $\beta_2 + \beta_3 + \beta_4 = 1$ versus the alternative $\beta_2 + \beta_3 + \beta_4 \neq 1$. What do you conclude. Is there evidence of unobserved heterogeneity? Carry out the LM test for the presence of random effects at the 5% level of signficance.

e. Estimate the model using fixed effects. How do the estimates compare to those in (d)? Use the Hausman test for the significance of the difference in the coefficients. Is there evidence that the unobserved heterogeneity is correlated with one or more of the explanatory variables? Explain.

f. Obtain the fixed effects residuals, $\tilde{e}_{it}$. Using OLS with cluster-robust standard errors estimate the regression $\tilde{e}_{it} = \rho \tilde{e}_{i,t-1} + r_{it}$, where r_{it} is a random error. As noted in Exercise 15.10, if the idiosyncratic errors e_{it} are uncorrelated we expect $\rho = -1/2$. Rejecting this hypothesis implies that idiosyncratic errors e_{it} are serially correlated. Using the 5% level of significance, what do you conclude?

g. Estimate the model by fixed effects using cluster-robust standard errors. How different are these standard errors from the conventional ones in part (e)?

15.26 The data file *collegecost* contains data on cost per student and related factors at four-year colleges in the U.S., covering the period 1987 to 2011. In this exercise, we explore a minimalist model predicting cost per student. Specify the model to be

$$\ln(TC_{it}) = \beta_1 + \beta_2 FTESTU_{it} + \beta_3 FTGRAD_{it} + \beta_4 TT_{it} + \beta_5 GA_{it} + \beta_6 CF_{it} + \sum_{t=2}^{8} \delta_t D_t + u_i + e_{it}$$

where *TC* is the total cost per student, *FTESTU* is number of full-time equivalent students, *FTGRAD* is number of full-time graduate students, *TT* is number of tenure track faculty per 100 students, *GA* is number of graduate assistants per 100 students, and *CF* is the number of contract faculty per 100 students, which are hired on a year to year basis. The D_t are indicator variables for the years 1989, 1991, 1999, 2005, 2008, 2010, and 2011. The base year is 1987. Only use data on public universities in this exercise.

a. Calculate the summary statistics for the model variables for the years 1987 and 2011. What do you observe about the sample averages of these variables?

b. Estimate the model by random effects. Discuss the signs and significance of the estimated coefficients. What is the predicted percentage cost per student change if one additional tenure track faculty is hired, per 100 students? What does the estimated value of δ_8 suggest?

c. Using the random effects estimates, test the following hypotheses at the 5% level: (i) $H_0: \beta_2 \geq \beta_3$, $H_1: \beta_2 < \beta_3$; (ii) $H_0: \beta_4 \leq \beta_6$, $H_1: \beta_4 > \beta_6$; and (iii) $H_0: \beta_5 \geq \beta_6$, $H_1: \beta_5 < \beta_6$. What do these tests imply about the relative costs of undergraduate students versus graduate students, tenure track faculty relative to contract faculty, and contract faculty relative to graduate assistants?

d. Calculate the time averages of the explanatory variables other than the indicator variables, for example, $\overline{FTESTU}_{i\bullet}$. Add these variables to the model and test their joint significance at the 1% level. What does the test result tell us about using the random effects estimator in this case? Which assumption is being tested?

e. Obtain the fixed effects estimates of the model. Discuss the signs and significance of the estimated coefficients. What is the predicted percentage cost per student change if one additional tenure track faculty is hired, per 100 students? What does the estimated value of δ_8 suggest? How do these estimates compare to the random effects estimates?

f. Using the fixed effects estimates, test the following hypotheses at the 5% level: (i) $H_0: \beta_2 \geq \beta_3$, $H_1: \beta_2 < \beta_3$; (ii) $H_0: \beta_4 \leq \beta_6$, $H_1: \beta_4 > \beta_6$; and (iii) $H_0: \beta_5 \geq \beta_6$, $H_1: \beta_5 < \beta_6$. What do these tests imply about the relative costs of undergraduate students versus graduate students, tenure track faculty relative to contract faculty, and contract faculty relative to graduate assistants?

15.27 The data file *collegecost* contains data on cost per student and related factors at four-year colleges in the U.S., covering the period from 1987 to 2011. In this exercise, we explore a minimalist model predicting cost per student. Specify the model to be

$$\ln(TC_{it}) = \beta_1 + \beta_2 FTESTU_{it} + \beta_3 FTGRAD_{it} + \beta_4 TT_{it} + \beta_5 GA_{it} + \beta_6 CF_{it} + \sum_{t=2}^{8} \delta_t D_t + u_i + e_{it}$$

where *TC* is the total cost per student, *FTESTU* is number of full-time equivalent students, *FTGRAD* is number of full-time graduate students, *TT* is number of tenure track faculty per 100 students, *GA* is number of graduate assistants per 100 students, and *CF* is the number of contract faculty per 100 students, which are hired on a year to year basis. The D_t are indicator variables for the years 1989, 1991, 1999, 2005, 2008, 2010, and 2011. The base year is 1987.

a. Calculate the summary statistics for the model variables for the years 1987 and 2011 separately for public and private universities. What do you observe about the sample averages of these variables? In particular, what is the increase in *TC* between 1987 and 2011 for each type of university. What has happened to the number of tenure track faculty and the number of contract faculty?

b. Using OLS, estimate the model for public universities using conventional and cluster-robust standard errors. Are the standard errors noticeably different?

c. Using OLS, estimate the model for private universities using conventional and cluster-robust standard errors. Are the standard errors noticeably different? How do the coefficient estimates for the private universities compare to those for the public universities?

d. Estimate the model using fixed effects with cluster-robust standard errors for the public universities. How do these estimates compare to the OLS estimates in (b)? What are the important differences?

e. Estimate the model using fixed effects with cluster-robust standard errors for the private universities. How do these estimates compare to the estimates for the public universities in part (d)? What are the important differences?

15.28 The data file *collegecost* contains data on cost per student and related factors at four-year colleges in the U.S., covering the period 1987 to 2011. In this exercise, we explore a minimalist model predicting cost per student. Specify the model to be

$$\ln(TC_{it}) = \beta_1 + \beta_2 FTESTU_{it} + \beta_3 FTGRAD_{it} + \beta_4 TT_{it} + \beta_5 GA_{it} + \beta_6 CF_{it} + \sum_{t=2}^{8} \delta_t D_t + u_i + e_{it}$$

where *TC* is the total cost per student, *FTESTU* is number of full-time equivalent students, *FTGRAD* is number of full-time graduate students, *TT* is number of tenure track faculty per 100 students, *GA* is number of graduate assistants per 100 students, and *CF* is the number of contract faculty, which are hired on a year to year basis. The D_t are indicator variables for the years 1989, 1991, 1999, 2005, 2008, 2010, and 2011. The base year is 1987. Use data only on public universities for this question.

a. Create first differences of the variables. Using the 2011 data, estimate by OLS the first-difference model

$$\Delta\ln(TC_{it}) = \beta_2 \Delta FTESTU_{it} + \beta_3 \Delta FTGRAD_{it} + \beta_4 \Delta TT_{it} + \beta_5 \Delta GA_{it} + \beta_6 \Delta CF_{it} + \Delta e_{it}$$

b. Repeat the estimation in (a) adding an intercept term. What is the interpretation of the constant?

c. Repeat the estimation in (a) adding an intercept plus the 2011 observations on the variables *FTESTU*, *FTGRAD*, *TT*, *GA*, and *CF*. If the assumption of strict exogeneity holds none of the coefficients on these variables should be significant, and they should be jointly insignificant as well. What do you conclude? Why is this assumption important for the estimation of panel data regression models?

d. Create the one period future, or forward, value for each variable, x_{t+1}. That is, for example, in year *t* create a new variable $FTESTU_{i,t+1}$. Using data from 2008 and 2010, estimate the panel data regression model by fixed effects, including the forward values of *FTESTU*, *FTGRAD*, *TT*, *GA*, and *CF*. If the assumption of strict exogeneity holds none of the coefficients on these variables should be significant, and they should be jointly insignificant as well. What do you conclude?

15.29 In this exercise, we re-examine the data in Exercise 15.22, a panel of data from North Carolina. Consider a model in which the log of crime rate (*LCRMRTE*) is a function of the log of police per capita (*LPOLPC*), the log of the probability of arrest (*LPRBARR*), the log of the probability of conviction (*LPRBCONV*), the log of average prison sentence (*LAVGSEN*), and the log of average weekly wage in the manufacturing sector (*LWMFG*) and indicator variables for the western region (*WEST*) and urban counties (*URBAN*).

a. It is possible that the crime rate and police per capita are jointly determined and that *LPOLPC* might be endogenous. Hence we consider estimating the model by 2SLS. As instruments we use the log of tax revenue per capita (*LTAXPC*) and the log of the ratio of face-to-face crimes relative to other types of crimes (*LMIX*). Estimate the first-stage regression of *LPOLPC* on the other

variables, except *LCRMRTE*, and the two instruments. Test the joint significance of the IV. Can we reject the null hypothesis that the IV are weak?

b. Using the instruments in (a), estimate the model by 2SLS. Are the deterrence variables significant?

c. Test for the endogeneity of *LPOLPC* and test the validity of the surplus instrument. What do you conclude in each case?

d. The estimation in (b) ignores unobserved county heterogeneity. For each variable, except the time-invariant variables *WEST* and *URBAN*, obtain the variables in the deviation about the county mean form, that is, apply the within transformation to each variable. Estimate the first-stage model with the variables in deviation from the mean form. Test the joint significance of the two transformed instruments.

e. Using the transformed instruments and other variables, estimate the model by 2SLS. What differences do you observe between these estimates and those in part (b)? Recall that you must adjust the standard errors for the correct degrees of freedom, as in Example 15.5. (*Note*: You may investigate whether your software has an automatic command to do 2SLS with panel data as a check.)

f. Using the transformed instruments and other variables, test for the endogeneity of *LPOLPC* and test the validity of the surplus instrument. What do you conclude in each case?

15.30 In this exercise, we extend Exercise 15.29 by also considering the possibility that the probability of arrest is jointly determined with the crime rate and the number of police per capita. The idea is that when the crime rate is high, the police may intensify their efforts to reduce crime by increasing the arrest rate. Consider the same model as in Exercise 15.29.

a. It is possible that the crime rate and police per capita are jointly determined and that *LPOLPC* and *LPRBARR* might be endogenous. Hence we consider estimating the model by 2SLS. As instruments we use the log of tax revenue per capita (*LTAXPC*) and the log of the ratio of face-to-face crimes relative to other types of crimes (*LMIX*). Estimate the first-stage regression of *LPOLPC* on the other variables, except *LCRMRTE*, and the two instruments. Test the joint significance of the IV. Can we reject the null hypothesis that the IV are weak? Estimate the first-stage regression of *LPRBARR* on the other variables, except *LCRMRTE*, and the two instruments. Test the joint significance of the IV. Can we reject the null hypothesis that the IV are weak?

b. Using the instruments in (a), estimate the model, treating both *LPOLPC* and *LPRBARR* as endogenous, by 2SLS. Are the deterrence variables significant?

c. Test for the endogeneity of *LPOLPC* and *LPRBARR* using the regression-based Hausman test. What do you conclude in each case?

d. The estimation in (b) ignores unobserved county heterogeneity. For each variable, except the time-invariant variables *WEST* and *URBAN*, obtain the variables in the deviation about the county mean form, that is, apply the within transformation to each variable. Estimate the first-stage model for both *LPOLPC* and *LPRBARR* with the variables in deviation from the mean form. Test the joint significance of the two transformed instruments.

e. Using the transformed instruments and other variables, estimate the model, treating both *LPOLPC* and *LPRBARR* as endogenous, by 2SLS. What differences do you observe between these estimates and those in part (b)? Recall that you must adjust the standard errors for the correct degrees of freedom, as in Example 15.5. (*Note*: You may investigate whether your software has an automatic command to do 2SLS with panel data as a check.)

f. Test for the endogeneity of *LPOLPC* and *LPRBARR* using the regression-based Hausman test. What do you conclude in each case?

Appendix 15A Cluster-Robust Standard Errors: Some Details

To appreciate the nature of cluster-robust standard errors, we return momentarily to a simple regression model for cross-sectional data

$$y_i = \beta_1 + \beta_2 x_i + e_i$$

Using the result $b_2 = \beta_2 + \sum_{i=1}^{N} w_i e_i$, where $w_i = (x_i - \bar{x}) \Big/ \sum_{i=1}^{N} (x_i - \bar{x})^2$, in Appendix 8A, we showed that the variance of the least squares estimator b_2, in the presence of heteroskedasticity, is given by

$$\operatorname{var}(b_2|\mathbf{x}) = \operatorname{var}\left(\sum_{i=1}^{N} w_i e_i|\mathbf{x}\right) = \sum_{i=1}^{N} w_i^2 \operatorname{var}(e_i|\mathbf{x}) + \sum_{i=1}^{N}\sum_{j=i+1}^{N} 2w_i w_j \operatorname{cov}(e_i, e_j|\mathbf{x})$$

$$= \sum_{i=1}^{N} w_i^2 \operatorname{var}(e_i|\mathbf{x}) = \sum_{i=1}^{N} w_i^2 \sigma_i^2$$

Because we are assuming a random sample of cross-sectional individuals, $\operatorname{cov}(e_i, e_j|\mathbf{x}) = 0$ for $i \neq j$, leading to the simplification in the second line of the above equation.

Now suppose we have a panel simple regression model

$$y_{it} = \beta_1 + \beta_2 x_{it} + e_{it} \tag{15A.1}$$

with the assumptions $\operatorname{cov}(e_{it}, e_{is}|\mathbf{x}) = \psi_{its}$ and $\operatorname{cov}(e_{it}, e_{js}|\mathbf{x}) = 0$ for $i \neq j$. In equation (15.29) we denoted $\operatorname{var}(v_{it}) = \sigma_u^2 + \sigma_{it}^2 = \psi_{it}^2$. In this appendix we use an alternative notation, to simplify the double summations. Let $\operatorname{var}(v_{it}) = \psi_{itt} = \operatorname{cov}(v_{it}, v_{it})$. The pooled least squares estimator for β_2 is given by

$$b_2 = \beta_2 + \sum_{i=1}^{N}\sum_{t=1}^{T} w_{it} e_{it} \tag{15A.2}$$

where

$$w_{it} = \frac{x_{it} - \bar{\bar{x}}}{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it} - \bar{\bar{x}}\right)^2}$$

with $\bar{\bar{x}} = \sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}/NT$. The variance of the pooled least squares estimator b_2 is given by

$$\operatorname{var}(b_2|\mathbf{x}) = \operatorname{var}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} w_{it} e_{it}|\mathbf{x}\right) = \operatorname{var}\left(\sum_{i=1}^{N} g_i|\mathbf{x}\right) \tag{15A.3}$$

where $g_i = \sum_{t=1}^{T} w_{it} e_{it}$ is a weighted sum of the errors for individual i. Because we have a random sample, the errors for different individuals are uncorrelated, implying that g_i is uncorrelated with g_j for $i \neq j$. Thus,

$$\operatorname{var}(b_2|\mathbf{x}) = \operatorname{var}\left(\sum_{i=1}^{N} g_i|\mathbf{x}\right) = \sum_{i=1}^{N} \operatorname{var}(g_i|\mathbf{x}) + \sum_{i=1}^{N}\sum_{j=i+1}^{N} 2\operatorname{cov}(g_i, g_j|\mathbf{x}) = \sum_{i=1}^{N} \operatorname{var}(g_i|\mathbf{x}) \tag{15A.4}$$

To find $\operatorname{var}(g_i|\mathbf{x})$ suppose for the moment that $T = 2$, then

$$\operatorname{var}(g_i|\mathbf{x}) = \operatorname{var}\left(\sum_{t=1}^{2} w_{it} e_{it}|\mathbf{x}\right) = w_{i1}^2 \operatorname{var}(e_{i1}|\mathbf{x}) + w_{i2}^2 \operatorname{var}(e_{i2}|\mathbf{x}) + 2w_{i1}w_{i2}\operatorname{cov}(e_{i1}, e_{i2}|\mathbf{x})$$

$$= w_{i1}^2 \psi_{i11} + w_{i2}^2 \psi_{i22} + 2w_{i1}w_{i2}\psi_{i12}$$

$$= \sum_{t=1}^{2}\sum_{s=1}^{2} w_{it} w_{is} \psi_{its}$$

For $T > 2$, $\text{var}(g_i|\mathbf{x}) = \sum_{t=1}^{T}\sum_{s=1}^{T} w_{it}w_{is}\psi_{its}$. Substituting this expression into (15A.4), we have

$$\begin{aligned}\text{var}(b_2|\mathbf{x}) &= \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T} w_{it}w_{is}\psi_{its} \\ &= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\left(x_{it}-\bar{\bar{x}}\right)\left(x_{is}-\bar{\bar{x}}\right)\psi_{its}}{\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it}-\bar{\bar{x}}\right)^2\right)^2}\end{aligned} \tag{15A.5}$$

Recall that $\text{cov}(e_{it}, e_{is}|\mathbf{x}) = E(e_{it}e_{is}|\mathbf{x}) = \psi_{its}$. A cluster-robust variance estimate is obtained from (15A.5) by replacing ψ_{its} with $\hat{e}_{it}\hat{e}_{is}$. Thus, a cluster-robust standard error for b_2 is given by the square root of

$$\widehat{\text{var}}(b_2|\mathbf{x}) = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\left(x_{it}-\bar{\bar{x}}\right)\left(x_{is}-\bar{\bar{x}}\right)\hat{e}_{it}\hat{e}_{is}}{\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it}-\bar{\bar{x}}\right)^2\right)^2} \tag{15A.6}$$

The above description of how cluster-robust standard errors are calculated and the logic behind them was done in terms of a model with just one explanatory variable. To describe the robust variance estimator for models with more than one explanatory variable, matrix algebra is required, but the principle is the same.

Finally, you will find that the cluster-robust standard errors produced by most software packages apply a degrees of freedom correction to the expression in (15A.6). Unfortunately, they do not all use the same correction factor. When using a cluster-robust standard error, the effective number of observations is G, the number of clusters.[20]

Appendix 15B Estimation of Error Components

The RE model is

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + (u_i + e_{it}) \tag{15B.1}$$

where u_i is the individual-specific error and e_{it} is the usual regression error. We will discuss the case for a balanced panel, with T time-series observations for each of N individuals. To implement GLS estimation we need to consistently estimate σ_u^2, the variance of the individual-specific error component, and σ_e^2, the variance of the regression error.

The regression error variance σ_e^2 comes from the fixed effects estimator. In (15.14), we transform the panel data regression into "**deviation about the individual mean**" form

$$y_{it} - \bar{y}_i = \beta_2(x_{2it} - \bar{x}_{2i}) + (e_{it} - \bar{e}_i) \tag{15B.2}$$

The least squares estimator of this equation yields the same estimates and sum of squared errors (denoted here by SSE_{DV}) as least squares applied to a model that includes a dummy variable for each individual in the sample. A consistent estimator of σ_e^2 is obtained by dividing SSE_{DV} by the

[20] See Carter, et al. "Asymptotic Behavior of a *t*-Test Robust to Cluster Heterogeneity," *The Review of Economics and Statistics*, 2017, 99(4), 698–709.

appropriate degrees of freedom, which is $NT - N - K_S$, where K_S is the number of parameters that are present in the transformed model (15B.2)

$$\hat{\sigma}_e^2 = \frac{SSE_{DV}}{NT - N - K_S} \tag{15B.3}$$

The estimator of σ_u^2 requires a bit more work. We begin with the time-averaged observations in (15.13)

$$\bar{y}_i = \beta_1 + \beta_2 \bar{x}_{2i} + \alpha_1 w_{1i} + u_i + \bar{e}_i, \quad i = 1, 2, \ldots, N \tag{15B.4}$$

The least squares estimator of (15B.4) is called the **between estimator**, as it uses variation between individuals as a basis for estimating the regression parameters. This estimator is unbiased and consistent, but not minimum variance under the error assumptions of the random effects model. The error term in this model is $u_i + \bar{e}_i$; it is uncorrelated across individuals, and has homoskedastic variance

$$\begin{aligned} \operatorname{var}(u_i + \bar{e}_i) &= \operatorname{var}(u_i) + \operatorname{var}(\bar{e}_i) = \operatorname{var}(u_i) + \operatorname{var}\left(\sum_{t=1}^{T} e_{it}/T\right) \\ &= \sigma_u^2 + \frac{1}{T^2}\operatorname{var}\left(\sum_{t=1}^{T} e_{it}\right) = \sigma_u^2 + \frac{T\sigma_e^2}{T^2} \\ &= \sigma_u^2 + \frac{\sigma_e^2}{T} \end{aligned} \tag{15B.5}$$

We can estimate the variance in (15B.5) by estimating the between regression in (15B.4), and dividing the sum of squared errors, SSE_{BE}, by the degrees of freedom $N - K_{BE}$, where K_{BE} is the total number of parameters in the between regression, including the intercept parameter. Then

$$\widehat{\sigma_u^2 + \frac{\sigma_e^2}{T}} = \frac{SSE_{BE}}{N - K_{BE}} \tag{15B.6}$$

With this estimate in hand, we can estimate σ_u^2 as

$$\hat{\sigma}_u^2 = \widehat{\sigma_u^2 + \frac{\sigma_e^2}{T}} - \frac{\hat{\sigma}_e^2}{T} = \frac{SSE_{BE}}{N - K_{BE}} - \frac{SSE_{DV}}{T(NT - N - K_S)} \tag{15B.7}$$

We have obtained the estimates of σ_u^2 and σ_e^2 using what is called the Swamy–Arora method. This method is implemented in software packages and is well established. We note, however, that it is possible in finite samples to obtain an estimate $\hat{\sigma}_u^2$ in (15B.7) that is negative, which is obviously infeasible. If this should happen, one option is simply to set $\hat{\sigma}_u^2 = 0$, which implies that there are no random effects. Alternatively, your software may offer other options for estimating the variance components, which you might try.