14 Introduction to Panel Data Models

14.1 Introduction

If the same units of observation in a cross-sectional sample are surveyed two or more times, the resulting observations are described as forming a **panel** or **longitudinal data set**. The National Longitudinal Survey of Youth that has provided data for many of the examples and exercises in this text is such a data set. The NLSY started with a baseline survey in 1979 and the same individuals have been reinterviewed many times since, annually until 1994 and biennially since then. However the unit of observation of a panel data set need not be individuals. It may be households, or enterprises, or geographical areas, or indeed any set of entities that retain their identities over time.

Because panel data have both cross-sectional and time series dimensions, the application of regression models to fit econometric models are more complex than those for simple cross-sectional data sets. Nevertheless, they are increasingly being used in applied work and the aim of this chapter is to provide a brief introduction. For comprehensive treatments see Hsiao (2003), Baltagi (2001), and Wooldridge (2002).

There are several reasons for the increasing interest in panel data sets. An important one is that their use may offer a solution to the problem of bias caused by unobserved heterogeneity, a common problem in the fitting of models with cross-sectional data sets. This will be discussed in the next section.

A second reason is that it may be possible to exploit panel data sets to reveal dynamics that are difficult to detect with cross-sectional data. For example, if one has cross-sectional data on a number of adults, it will be found that some are employed, some are unemployed, and the rest are economically inactive. For policy purposes, one would like to distinguish between frictional unemployment and long-term unemployment. Frictional unemployment is inevitable in a changing economy, but the long-term unemployment can indicate a social problem that needs to be addressed. To design an effective policy to counter long-term unemployment, one needs to know the characteristics of those affected or at risk. In principle the necessary information might be captured with a cross-sectional survey using retrospective questions about past emloyment status, but in practice

the scope for this is often very limited. The further back in the past one goes, the worse are the problems of a lack of records and fallible memories, and the greater becomes the problem of measurement error. Panel studies avoid this problem in that the need for recall is limited to the time interval since the previous interview, often no more than a year.

A third attraction of panel data sets is that they often have very large numbers of observations. If there are n units of observation and if the survey is undertaken in T time periods, there are potentially nT observations consisting of time series of length T on n parallel units. In the case of the NLSY, there were just over 6,000 individuals in the core sample. The survey has been conducted 19 times as of 2004, generating over 100,000 observations. Further, because it is expensive to establish and maintain them, such panel data sets tend to be well designed and rich in content.

A panel is described as **balanced** if there is an observation for every unit of observation for every time period, and as **unbalanced** if some observations are missing. The discussion that follows applies equally to both types. However, if one is using an unbalanced panel, one needs to take note of the possibility that the causes of missing observations are endogenous to the model. Equally, if a balanced panel has been created artificially by eliminating all units of observation with missing observations, the resulting data set may not be representative of its population.

Example of the use of a panel data set to investigate dynamics

In many studies of the determinants of earnings it has been found that married men earn significantly more than single men. One explanation is that marriage entails financial responsibilities—in particular, the rearing of children—that may encourage men to work harder or seek better paying jobs. Another is that certain unobserved qualities that are valued by employers are also valued by potential spouses and hence are conducive to getting married, and that the dummy variable for being married is acting as a proxy for these qualities. Other explanations have been proposed, but we will restrict attention to these two. With cross-sectional data it is difficult to discriminate between them. However, with panel data one can find out whether there is an uplift at the time of marriage or soon after, as would be predicted by the increased productivity hypothesis, or whether married men tend to earn more even before marriage, as would be predicted by the unobserved heterogeneity hypothesis.

In 1988 there were 1,538 NLSY males working 30 or more hours a week, not also in school, with no missing data. The respondents were divided into three categories: the 904 who were already married in 1988 (dummy variable $MARRIED = 1$); a further 212 who were single in 1988 but who married within the next four years (dummy variable $SOONMARR = 1$); and the remaining 422 who were single in 1988 and still single four years later (the omitted category). Divorced respondents were excluded from the sample. The following earnings

function was fitted (standard errors in parentheses):

$$\widehat{LGEARN} = 0.163\ MARRIED + 0.096\ SOONMARR + \text{constant} + \text{controls}$$
$$(0.028) \qquad (0.037) \qquad R^2 = 0.27. \tag{14.1}$$

The controls included years of schooling, *ASVABC* score, years of tenure with the current employer and its square, years of work experience and its square, age and its square, and dummy variables for ethnicity, region of residence, and living in an urban area.

The regression indicates that those who were married in 1988 earned 16.3 percent more than the reference category (strictly speaking, 17.7 percent, if the proportional increase is calculated properly as $e^{0.163} - 1$) and that the effect is highly significant. However, it is the coefficient of *SOONMARR* that is of greater interest here. Under the null hypothesis that the marital effect is dynamic and marriage encourages men to earn more, the coefficient of *SOONMARR* should be zero. The men in this category were still single as of 1988. The *t* statistic of the coefficient is 2.60 and so the coefficient is significantly different from zero at the 0.1 percent level, leading us to reject the null hypothesis at that level.

However, if the alternative hypothesis is true, the coefficient of *SOONMARR* should be equal to that of *MARRIED*, but it is lower. To test whether it is significantly lower, the easiest method is to change the reference category to those who were married by 1988 and to introduce a new dummy variable *SINGLE* that is equal to 1 if the respondent was single in 1988 and still single four years later. The omitted category is now those who were already married by 1988. The fitted regression is (standard errors in parentheses)

$$\widehat{LGEARN} = -0.163\ SINGLE - 0.066\ SOONMARR + \text{constant} + \text{controls}$$
$$(0.028) \qquad (0.034) \qquad R^2 = 0.27. \tag{14.2}$$

The coefficient of *SOONMARR* now estimates the difference between the coefficients of those married by 1988 and those married within the next four years, and if the second hypothesis is true, it should be equal to zero. The *t* statistic is −1.93, so we (just) do not reject the second hypothesis at the 5 percent level. The evidence seems to provide greater support for the first hypothesis, but it is possible that neither hypothesis is correct on its own and the truth might reside in some compromise.

In the foregoing example, we used data only from the 1988 and 1992 rounds of the NLSY. In most applications using panel data it is normal to exploit the data from all the rounds, if only to maximize the number of observations in the

sample. A standard specification is

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \sum_{p=1}^{s} \gamma_p Z_{pi} + \delta t + \varepsilon_{it} \tag{14.3}$$

where Y is the dependent variable, the X_j are observed explanatory variables, and the Z_p are unobserved explanatory variables. The index i refers to the unit of observation, t refers to the time period, and j and p are used to differentiate between different observed and unobserved explanatory variables. ε_{it} is a disturbance term assumed to satisfy the usual regression model conditions. A trend term t has been introduced to allow for a shift of the intercept over time. If the implicit assumption of a constant rate of change seems too strong, the trend can be replaced by a set of dummy variables, one for each time period except the reference period.

The X_j variables are usually the variables of interest, while the Z_p variables are responsible for unobserved heterogeneity and as such constitute a nuisance component of the model. The following discussion will be confined to the (quite common) special case where it is reasonable to assume that the unobserved heterogeneity is unchanging and accordingly the Z_p variables do not need a time subscript. Because the Z_p variables are unobserved, there is no means of obtaining information about the $\sum_{p=1}^{s} \gamma_p Z_{pi}$ component of the model and it is convenient to rewrite (14.3) as

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \alpha_i + \delta t + \varepsilon_{it} \tag{14.4}$$

where

$$\alpha_i = \sum_{p=1}^{s} \gamma_p Z_{pi}. \tag{14.5}$$

α_i, known as the **unobserved effect**, represents the joint impact of the Z_{pi} on Y_i. Henceforward it will be convenient to refer to the unit of observation as an individual, and to the α_i as the individual-specific unobserved effect, but it should be borne in mind that the individual in question may actually be a household or an enterprise, etc. If α_i is correlated with any of the X_j variables, the regression estimates from a regression of Y on the X_j variables will be subject to unobserved heterogeneity bias. Even if the unobserved effect is not correlated with any of the explanatory variables, its presence will in general cause OLS to yield inefficient estimates and invalid standard errors. We will now consider ways of overcoming these problems.

First, however, note that if the X_j controls are so comprehensive that they capture all the relevant characteristics of the individual, there will be no relevant unobserved characteristics. In that case the α_i term may be dropped and a **pooled**

OLS regression may be used to fit the model, treating all the observations for all of the time periods as a single sample.

14.2 Fixed effects regressions

The two main approaches to the fitting of models using panel data are known as **fixed effects regressions**, discussed in this section, and **random effects regressions**, discussed in the next. Three versions of the fixed effects approach will be described. In the first two, the model is manipulated in such a way that the unobserved effect is eliminated.

Within-groups fixed effects

In the first version, the mean values of the variables in the observations on a given individual are calculated and subtracted from the data for that individual. In view of (14.4), one may write

$$\overline{Y}_i = \beta_1 + \sum_{j=2}^{k} \beta_j \overline{X}_{ij} + \delta \overline{t} + \alpha_i + \overline{\varepsilon}_{it}. \tag{14.6}$$

Subtracting this from (14.4), one obtains

$$Y_{it} - \overline{Y}_i = \sum_{j=2}^{k} \beta_j \left(X_{ijt} - \overline{X}_{ij} \right) + \delta (t - \overline{t}) + \varepsilon_{it} - \overline{\varepsilon}_i \tag{14.7}$$

and the unobserved effect disappears. This is known as the **within-groups regression** model because it is explaining the variations about the mean of the dependent variable in terms of the variations about the means of the explanatory variables for the group of observations relating to a given individual. The possibility of tackling unobserved heterogeneity bias in this way is a major attraction of panel data for researchers.

However, there are some prices to pay. First, the intercept β_1 and any X variable that remains constant for each individual will drop out of the model. The elimination of the intercept may not matter, but the loss of the unchanging explanatory variables may be frustrating. Suppose, for example, that one is fitting an earnings function to data for a sample of individuals who have completed their schooling, and that the schooling variable for individual i in period t is S_{it}. If the education of the individual is complete by the time of the first time period, S_{it} will be the same for all t for that individual and $S_{it} = \overline{S}_i$ for all t. Hence $(S_{it} - \overline{S}_i)$ is zero for all time periods. If all individuals have completed their schooling by the first time period, S_{it} will be zero for all i and t. One cannot include a variable whose values are all zero in a regression model. Thus if the object of the exercise

were to obtain an estimate of the returns to schooling untainted by unobserved heterogeneity bias, one ends up with no estimate at all.

A second problem is the potential impact of the disturbance term. We saw in Chapter 3 that the precision of OLS estimates depends on the mean square deviations of the explanatory variables being large in comparison with the variance of the disturbance term. The analysis was in the context of the simple regression model, but it generalizes to multiple regression. The variation in $(X_j - \overline{X}_j)$ may well be much smaller than the variation in X_j. If this is the case, the impact of the disturbance term may be relatively large, giving rise to imprecise estimates. The situation is aggravated in the case of measurement error, since this will lead to bias, and the bias is the greater, the smaller the variation in the explanatory variable in comparison with the variance of the measurement error.

A third problem is that we lose a substantial number of degrees of freedom in the model when we manipulate the model to eliminate the unobserved effect: we lose one degree of freedom for every individual in the sample. If the panel is balanced, with nT observations in all, it may seem that there would be $nT - k$ degrees of freedom. However, in manipulating the model, the number of degrees of freedom is reduced by n, for reasons that will be explained later in this section. Hence the true number of degrees of freedom will be $n(T-1)-k$. If T is small, the impact can be large. (Regression applications with a fixed regression facility will automatically make the adjustment to the degrees of freedom when implementing the within-groups method.)

First differences fixed effects

In a second version of the fixed effects approach, the **first differences regression** model, the unobserved effect is eliminated by subtracting the observation for the previous time period from the observation for the current time period, for all time periods. For individual i in time period t the model may be written

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ijt} + \delta t + \alpha_i + \varepsilon_{it}. \tag{14.8}$$

For the previous time period, the relationship is

$$Y_{it-1} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ijt-1} + \delta(t-1) + \alpha_i + \varepsilon_{it-1}. \tag{14.9}$$

Subtracting (14.9) from (14.8), one obtains

$$\Delta Y_{it} = \sum_{j=2}^{k} \beta_j \Delta X_{ijt} + \delta + \varepsilon_{it} - \varepsilon_{it-1} \tag{14.10}$$

and again the unobserved heterogeneity has disappeared. However, the other problems remain. In particular, the intercept and any X variable that remains

fixed for each individual will disappear from the model and n degrees of freedom are lost because the first observation for each individual is not defined. In addition, this type of differencing gives rise to autocorrelation if ε_{it} satisfies the regression model conditions. The error term for ΔY_{it} is $(\varepsilon_{it} - \varepsilon_{it-1})$. That for the previous observation is $(\varepsilon_{it-1} - \varepsilon_{it-2})$. Thus the two error terms both have a component ε_{it-1} with opposite signs and negative moving average autocorrelation has been induced. However, if ε_{it} is subject to autocorrelation:

$$\varepsilon_{it} = \rho\varepsilon_{it-1} + \nu_{it} \tag{14.11}$$

where ν_{it} is a well behaved innovation, the moving average disturbance term is equal to $\nu_{it} - (1 - \rho)\varepsilon_{it-1}$. If the autocorrelation is severe, the $(1 - \rho)\varepsilon_{it-1}$ component could be small and so the first differences estimator could be preferable to the within-groups estimator.

Least squares dummy variable fixed effects

In the third version of the fixed effects approach, known as the **least squares dummy variable (LSDV) regression** model, the unobserved effect is brought explicitly into the model. If we define a set of dummy variables A_i, where A_i is equal to 1 in the case of an observation relating to individual i and 0 otherwise, the model can be rewritten

$$Y_{it} = \sum_{j=2}^{k} \beta_j X_{ijt} + \delta t + \sum_{i=1}^{n} \alpha_i A_i + \varepsilon_{it}. \tag{14.12}$$

Formally, the unobserved effect is now being treated as the coefficient of the individual-specific dummy variable, the $\alpha_i A_i$ term representing a fixed effect on the dependent variable Y_i for individual i (this accounts for the name given to the fixed effects approach). Having re-specified the model in this way, it can be fitted using OLS.

Note that if we include a dummy variable for every individual in the sample as well as an intercept, we will fall into the dummy variable trap described in Section 5.2. To avoid this, we could define one individual to be the reference category, so that β_1 is its intercept, and then treat the α_i as the shifts in the intercept for the other individuals. However, the choice of reference category is often arbitrary and accordingly the interpretation of the α_i in such a specification not particularly illuminating. Alternatively, we can drop the β_1 intercept and define dummy variables for all of the individuals, as has been done in (14.12). The α_i now become the intercepts for each of the individuals. Note that, in common with the first two versions of the fixed effects approach, the LSDV method requires panel data. With cross-sectional data, one would be defining a dummy variable for every observation, exhausting the degrees of freedom. The dummy variables on their own would give a perfect but meaningless fit.

If there are a large number of individuals, using the LSDV method directly is not a practical proposition, given the need for a large number of dummy

Table 14.1 Individual-specific dummy variables and an unchanging X variable

Individual	Time period	A_1	A_2	A_3	A_4	X_j
1	1	1	0	0	0	c_1
1	2	1	0	0	0	c_1
1	3	1	0	0	0	c_1
2	1	0	1	0	0	c_2
2	2	0	1	0	0	c_2
2	3	0	1	0	0	c_2
3	1	0	0	1	0	c_3
3	2	0	0	1	0	c_3
3	3	0	0	1	0	c_3
4	1	0	0	0	1	c_4
4	2	0	0	0	1	c_4
4	3	0	0	0	1	c_4

variables. However, it can be shown mathematically that the method is identical to the within-groups method. The only apparent difference is in the number of degrees of freedom. It is easy to see from (14.12) that there are $nT-k-n$ degrees of freedom if the panel is balanced. In the within-groups approach, it seemed at first that there were $nT-k$. However, n degrees of freedom are consumed in the manipulation that eliminates the α_i.

Given that it is equivalent to the within-groups approach, the LSDV method is subject to the same problems. In particular, we are unable to estimate coefficients for the X variables that are fixed for each individual. Suppose that X_{ij} is equal to c_i for all the observations for individual i. Then

$$X_j = \sum_{i=1}^{n} c_i A_i. \tag{14.13}$$

To see this, suppose that there are four individuals and three time periods, as in Table 14.1, and consider the observations for the first individual. X_j is equal to c_1 for each observation. A_1 is equal to 1. All the other A dummies are equal to 0. Hence both sides of the equation are equal to c_1. Similarly, both sides of the equation are equal to c_2 for the observations for individual 2, and similarly for individuals 3 and 4.

Thus there is an exact linear relationship linking X_j with the dummy variables and the model is subject to exact multicollinearity. Accordingly X_j cannot be included in the regression specification.

Example

To illustrate the use of a fixed effects model, we return to the example in Section 14.1 and use all the available data from 1980 to 1996, 20,343 observations in all. Table 14.2 shows the extra hourly earnings of married men and of men who are single but married within the next four years. The controls (not shown) are the

Table 14.2 Earnings premium for married and soon-to-be married men, NLSY 1980–96

	OLS	Fixed effects		Random effects	
Married	0.184	0.106	–	0.134	–
	(0.007)	(0.012)		(0.010)	
Single, married	0.096	0.045	−0.061	0.060	−0.075
within 4 years	(0.009)	(0.010)	(0.008)	(0.009)	(0.007)
Single, not married	–	–	−0.106	–	−0.134
within 4 years			(0.012)		(0.010)
R^2	0.358	0.268	0.268	0.346	0.346
DWH test	–	–	–	205.8	205.8
n	20,343	20,343	20,343	20,343	20,343

same as in Section 14.1. The first column gives the estimates obtained by simply pooling the observations and using OLS with robust standard errors. The second column gives the fixed effects estimates, using the within-groups method, with single men as the reference category. The third gives the fixed effects estimates with married men as the reference category. The fourth and fifth give the random effects estimates, discussed in the next section.

The OLS estimates are very similar to those in the wage equation for 1988 discussed in Section 14.1. The fixed effects estimates are considerably lower, suggesting that the OLS estimates were inflated by unobserved heterogeneity. Nevertheless, the pattern is the same. Soon-to-be-married men earn significantly more than single men who stay single. However, if we fit the specification corresponding to equation (14.2), shown in the third column, we find that soon-to-be married men earn significantly less than married men. Hence both hypotheses relating to the marriage premium appear to be partly true.

14.3 Random effects regressions

As we saw in the previous section, when the variables of interest are constant for each individual, a fixed effects regression is not an effective tool because such variables cannot be included. In this section we will consider an alternative approach, known as a random effects regression that may, subject to two conditions, provide a solution to this problem.

The first condition is that it is possible to treat each of the unobserved Z_p variables as being drawn randomly from a given distribution. This may well be the case if the individual observations constitute a random sample from a given population as, for example, with the NLSY where the respondents were randomly drawn from the US population aged 14 to 21 in 1979. If this is the case, the α_i may be treated as random variables (hence the name of this approach)

drawn from a given distribution and we may rewrite the model as

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \alpha_i + \delta t + \varepsilon_{it}$$

$$= \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \delta t + u_{it} \tag{14.14}$$

where

$$u_{it} = \alpha_i + \varepsilon_{it}. \tag{14.15}$$

We have thus dealt with the unobserved effect by subsuming it into the disturbance term.

The second condition is that the Z_p variables are distributed independently of all of the X_j variables. If this is not the case, α, and hence u, will not be uncorrelated with the X_j variables and the random effects estimation will be biased and inconsistent. We would have to use fixed effects estimation instead, even if the first condition seems to be satisfied.

If the two conditions are satisfied, we may use (14.14) as our regression specification, but there is a complication. u_{it} will be subject to a special form of autocorrelation and we will have to use an estimation technique that takes account of it.

First, we will check the other regression model conditions relating to the disturbance term. Given our assumption that ε_{it} satisfies the usual regression model conditions, we can see that u_{it} satisfies the condition that its expectation be zero, since

$$E(u_{it}) = E(\alpha_i + \varepsilon_{it}) = E(\alpha_i) + E(\varepsilon_{it}) = 0 \quad \text{for all } i \text{ and } t \tag{14.16}$$

Here we are assuming without loss of generality that $E(\alpha_i) = 0$, any nonzero component being absorbed by the intercept, β_1. u_{it} will also satisfy the condition that it should have constant variance, since

$$\sigma^2_{u_{it}} = \sigma^2_{\alpha_i + \varepsilon_{it}} = \sigma^2_\alpha + \sigma^2_\varepsilon + 2\sigma_{\alpha\varepsilon} = \sigma^2_\alpha + \sigma^2_\varepsilon \quad \text{for all } i \text{ and } t. \tag{14.17}$$

The $\sigma_{\alpha\varepsilon}$ term is zero on the assumption that α_i is distributed independently of ε_{it}. u_{it} will also satisfy the regression model condition that it be distributed independently of the values of X_j, since both α_i and ε_{it} are assumed to satisfy this condition.

However, there is a problem with the regression model condition that the value of u_{it} in any observation be generated independently of its value in all other observations. For all the observations relating to a given individual, α_i will have the same value, reflecting the unchanging unobserved characteristics of the individual. This is illustrated in Table 14.3 for the case where there are four individuals and three time periods.

Table 14.3 Example of disturbance term values in a random effects model

Individual	Time period	u
1	1	$\alpha_1 + \varepsilon_{11}$
1	2	$\alpha_1 + \varepsilon_{12}$
1	3	$\alpha_1 + \varepsilon_{13}$
2	1	$\alpha_2 + \varepsilon_{21}$
2	2	$\alpha_2 + \varepsilon_{22}$
2	3	$\alpha_2 + \varepsilon_{23}$
3	1	$\alpha_3 + \varepsilon_{31}$
3	2	$\alpha_3 + \varepsilon_{32}$
3	3	$\alpha_3 + \varepsilon_{33}$
4	1	$\alpha_4 + \varepsilon_{41}$
4	2	$\alpha_4 + \varepsilon_{42}$
4	3	$\alpha_4 + \varepsilon_{43}$

Since the disturbance terms for individual i have a common component α_i, they are correlated. For individual i in period t, the disturbance term is $(\alpha_i + \varepsilon_{it})$. For the same individual in any other period t' it is $(\alpha_i + \varepsilon_{it'})$. The population covariance between them is

$$\sigma_{u_{it},u_{it'}} = \sigma_{(\alpha_i+\varepsilon_{it}),(\alpha_i+\varepsilon_{it'})} = \sigma_{\alpha_i,\alpha_i} + \sigma_{\alpha_i,\varepsilon_{it'}} + \sigma_{\varepsilon_{it},\alpha_i} + \sigma_{\varepsilon_{it},\varepsilon_{it'}} = \sigma_\alpha^2. \qquad (14.18)$$

For observations relating to different individuals the problem does not arise because then the α components will be different and generated independently.

We have encountered a problem of the violation of this regression model condition once before, in the case of autocorrelated disturbance terms in a time series model. As in that case, OLS remains unbiased and consistent, but it is inefficient and the OLS standard errors are computed wrongly.

The solution then was to transform the model so that the transformed disturbance term satisfied the regression model condition, and a similar procedure is adopted in the present case. However, while the transformation in the case of autocorrelation was very straightforward, in the present case it is more complex. Known as feasible generalized least squares, its description requires the use of linear algebra and is therefore beyond the scope of this text. It yields consistent estimates of the coefficients and therefore depends on n being sufficiently large. For small n its properties are unknown.

Assessing the appropriateness of fixed effects and random effects estimation

When should you use fixed effects estimation rather than random effects estimation, or vice versa? In principle, random effects is more attractive because observed characteristics that remain constant for each individual are retained in the regression model. In fixed effects estimation, they have to be dropped.

Also, with random effects estimation we do not lose n degrees of freedom, as is the case with fixed effects.

However, if either of the preconditions for using random effects is violated, we should use fixed effects instead. One precondition is that the observations can be described as being drawn randomly from a given population. This is a reasonable assumption in the case of the NLSY because it was designed to be a random sample. By contrast, it would not be a reasonable assumption if the units of observation in the panel data set were countries and the sample consisted of those countries that are members of the Organization for Economic Cooperation and Development (OECD). These countries certainly cannot be considered to represent a random sample of the 200-odd sovereign states in the world.

The other precondition is that the unobserved effect be distributed independently of the X_j variables. How can we tell if this is the case? The standard procedure is yet another implementation of the **Durbin–Wu–Hausman test** used to help us choose between OLS and IV estimation in models where there is suspected measurement error (Section 8.5) or simultaneous equations endogeneity (Section 9.3). The null hypothesis is that the α_i are distributed independently of the X_j. If this is correct, both random effects and fixed effects are consistent, but fixed effects will be inefficient because, looking at it in its LSDV form, it involves estimating an unnecessary set of dummy variable coefficients. If the null hypothesis is false, the random effects estimates will be subject to unobserved heterogeneity bias and will therefore differ systematically from the fixed effects estimates.

As in its other applications, the DWH test determines whether the estimates of the coefficients, taken as a group, are significantly different in the two regressions. If any variables are dropped in the fixed effects regression, they are excluded from the test. Under the null hypothesis the test statistic has a chi-squared distribution. In principle this should have degrees of freedom equal to the number of slope coefficients being compared, but for technical reasons that require matrix algebra for an explanation, the actual number may be lower. A regression application that implements the test, such as Stata, should determine the actual number of degrees of freedom.

Example

The fixed effects estimates, using the within-groups approach, of the coefficients of married men and soon-to-be married men in Table 14.2 are 0.106 and 0.045, respectively. The corresponding random effects estimates are considerably higher, 0.134 and 0.060, inviting the suspicion that they may be inflated by unobserved heterogeneity. The DWH test involves the comparison of 13 coefficients (those of *MARRIED*, *SOONMARR*, and 11 controls). Stata reports that there are in fact only 12 degrees of freedom. The test statistic is 205.8. With 12 degrees of freedom the critical value of chi-squared at the 0.1 percent level is 32.9, so we definitely conclude that we should be using fixed effects estimation.

Our findings are the same as in the simpler example in Section 14.1. They confirm that married men earn more than single men. Part of the differential appears to be attributable to the characteristics of married men, since men who are soon-to-marry but still single also enjoy an earnings premium. However, part of the marriage premium appears to be attributable to the effect of marriage itself, since married men earn significantly more than those who are soon-to-marry but still single.

Random effects or OLS?

Suppose that the DWH test indicates that we can use random effects rather than fixed effects. We should then consider whether there are any unobserved effects at all. It is just possible that the model has been so well specified that the disturbance term

$$u_{it} = \alpha_i + \varepsilon_{it} \tag{14.19}$$

consists of only the purely random component ε_{it} and there is no individual-specific α_i term. In this situation we should use pooled OLS, with two advantages. There is a gain in efficiency because we are not attempting to allow for non-existent within-groups autocorrelation, and we will be able to take advantage of the finite-sample properties of OLS, instead of having to rely on the asymptotic properties of random effects.

Various tests have been developed to detect the presence of random effects. The most common, implemented in some regression applications, is the Breusch–Pagan Lagrange multiplier test, the test statistic having a chi-squared distribution with one degree of freedom under the null hypothesis of no random effects. In the case of the marriage effect example the statistic is very high indeed, 20,007, but in this case it is meaningless because we are not able to use random effects estimation.

A note on the random effects and fixed effects terminology

It is generally agreed that random effects/fixed effects terminology can be misleading, but that it is too late to change it now. It is natural to think that random effects estimation should be used when the unobserved effect can be characterized as being drawn randomly from a given population and that fixed effects should be used when the unobserved effect is considered to be non-random. The second part of that statement is correct. However, the first part is correct only if the unobserved effect is distributed independently of the X_j variables. If it is not, fixed effects should be used instead to avoid the problem of unobserved heterogeneity bias. Figure 14.1 summarizes the decision-making process for fitting a model with panel data.

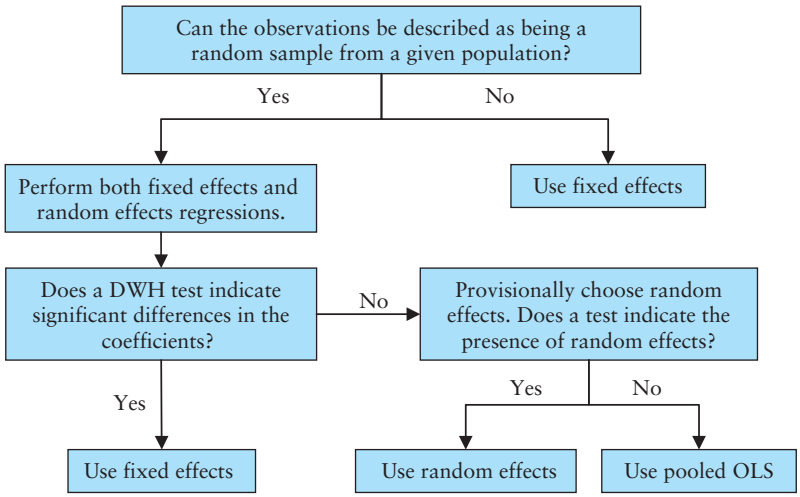

Figure 14.1 Choice of regression model for panel data

Key terms

balanced panel
Durbin–Wu–Hausman test
first differences regression
fixed effects
least squares dummy variable (LSDV) regression
longitudinal data set
panel data set
pooled OLS regression
random effects
unbalanced panel
unobserved effect
within-groups regression

Exercises

14.1 Download the OECD2000 data set from the website. See Appendix B for details The data set contains 32 variables:

ID This is the country identification, with 1 = Australia, 2 = Austria, 3 = Belgium, 4 = Canada, 5 = Denmark, 6 = Finland, 7 = France, 8 = Germany, 9 = Greece, 10 = Iceland, 11 = Ireland, 12 = Italy, 13 = Japan, 14 = Korea, 15 = Luxembourg, 16 = Mexico, 17 = Netherlands, 18 = New Zealand, 19 = Norway, 20 = Portugal, 21 = Spain,

22 = Sweden, 23 = Switzerland, 24 = Turkey, 25 = United Kingdom, 26 = United States. Four countries that have recently joined the OECD, the Czech Republic, Hungary, Poland, and Slovakia, are excluded because their data do not go back far enough.

ID01–26 These are individual country dummy variables. For example, *ID09* is the dummy variable for Greece.

E Average annual percentage rate of growth of employment for country i during time period t.

G Average annual percentage rate of growth of GDP for country i during time period t.

TIME There are three time periods, denoted 1, 2, and 3. They refer to average annual data for 1971–80, 1981–90, and 1991–2000.

TIME2 Dummy variable defined to be equal to 1 when $TIME = 2$, 0 otherwise.

TIME3 Dummy variable defined to be equal to 1 when $TIME = 3$, 0 otherwise.

Perform a pooled OLS regression of *E* on *G*. Regress *E* on *G*, *TIME2*, and *TIME3*. Perform appropriate statistical tests and give an interpretation of the regression results.

14.2 Using the OECD2000 data set, perform a (within-groups) fixed effects regression of *E* on *G*, *TIME2*, and *TIME3*. Perform appropriate statistical tests, give an interpretation of the regression coefficients, and comment on R^2.

14.3 Perform the corresponding LSDV regression, using OLS to regress *E* on *G*, *TIME2*, *TIME3*, and the country dummy variables (a) dropping the intercept, and (b) dropping one of the dummy variables. Perform appropriate statistical tests and give an interpretation of the coefficients in each case. Explain why either the intercept or one of the dummy variables must be dropped.

14.4 Perform a test for fixed effects in the OECD2000 regression by evaluating the explanatory power of the country dummy variables as a group.

14.5 Download the NLSY2000 data set from the website. See Appendix B for details. This contains the variables found in the *EAEF* data sets for the years 1980–94, 1996, 1998, and 2000 (there were no surveys in 1995, 1997, or 1999). Assuming that a random effects model is appropriate, investigate the apparent impact of unobserved heterogeneity on estimates of the coefficient of schooling by fitting the same earnings function, first using pooled OLS, then using random effects.

14.6 The *UNION* variable in the NLSY2000 data set is defined to be equal to 1 if the respondent was a member of a union in the year in question and 0 otherwise. Assuming that a random effects model is appropriate, add *UNION* to the earnings function specification and fit it using pooled OLS and random effects.

14.7 Using the NLSY2000 data set, perform a fixed effects regression of the earnings function specification used in Exercise 14.5 and compare the estimated coefficients

with those obtained using OLS and random effects. Perform a Durbin–Wu–Hausman test to discriminate between random effects and fixed effects.

14.8 Using the NLSY2000 data set, perform a fixed effects regression of the earnings function specification used in Exercise 14.6 and compare the estimated coefficients with those obtained using OLS and random effects. Perform a Durbin–Wu–Hausman test to discriminate between random effects and fixed effects.

14.9 The within-groups version of the fixed effects regression model involved subtracting the group mean relationship

$$\overline{Y}_i = \beta_1 + \sum_{j=2}^{k} (\beta_j \overline{X}_{ij}) + \delta \overline{t} + \alpha_i + \overline{\varepsilon}_{it}$$

from the original specification in order to eliminate the individual-specific effect α_i. Regressions using the group mean relationship are described as between effects regressions. Explain why the between effects model is in general inappropriate for estimating the parameters of a model using panel data. (Consider the two cases where the α_i are correlated and uncorrelated with the X_j controls.)