Chapter 18

Panel Data

18.1 Introduction

Modern econometrics is divided into two branches: microeconometrics and time series analysis. The latter is covered in chapter 19. The former has many elements, of which we have discussed several examples, such as qualitative dependent variables, duration models, count data, and limited dependent variables, all of which primarily involve different types of cross-sectional data. In light of this it would seem natural to call microeconometrics cross-sectional data analysis. We do not, however, because a major category of microeconometrics involves longitudinal or panel data in which a cross-section (of people, firms, countries, etc.) is observed over time. Thanks to the computer revolution, such data sets, in which we have observations on the same units in several different time periods, are more common and have become more amenable to analysis.

Two prominent examples of panel data are the PSID (Panel Study of Income Dynamics) data and the NLS (National Longitudinal Surveys of Labor Market Experience) data, both of which were obtained by interviewing several thousand people over and over again through time. These data sets were designed to enable examination of the causes and nature of poverty in the United States, by collecting information on such things as employment, earnings, mobility, housing, and consumption behavior. Indeed, thousands of variables were recorded. These data are typical of panel data in that they are short and wide, consisting of a very large number of cross-sectional units observed over a small number of time periods. Such data are expensive to obtain, involving tracking large numbers of people over extended time periods. Is this extra expense warranted?

Panel data have several attractive features that justify this extra cost, four of which are noted below.

1. Panel data can be used to deal with heterogeneity in the micro units. In any cross-section there is a myriad of unmeasured explanatory variables that affect

the behavior of the people (firms, countries, etc.) being analyzed. (Heterogeneity means that these micro units are all different from one another in fundamental unmeasured ways.) Omitting these variables causes bias in estimation. The same holds true for omitted time series variables that influence the behavior of the micro units uniformly, but differently in each time period. Panel data enable correction of this problem. Indeed, some would claim that the ability to deal with this omitted variable problem is the main attribute of panel data.

2. Panel data create more variability, through combining variation across micro units with variation over time, alleviating multicollinearity problems. With this more informative data, more efficient estimation is possible.
3. Panel data can be used to examine issues that cannot be studied using time series or cross-sectional data alone. As an example, consider the problem of separating economies of scale from technological change in the analysis of production functions. Cross-sectional data can be used to examine economies of scale, by comparing the costs of small and large firms, but because all the data come from one time period there is no way to estimate the effect of technological change. Things are worse with time series data on a single firm; we cannot separate the two effects because we cannot tell if a change in that firm's costs over time is due to technological change or due to a change in the size of the firm. As a second example, consider the distinction between temporary and long-term unemployment. Cross-sectional data tell us who is unemployed in a single year, and time series data tell us how the unemployment level changed from year to year. But neither can tell us if the same people are unemployed from year to year, implying a low turnover rate, or if different people are unemployed from year to year, implying a high turnover rate. Analysis using panel data can address the turnover question because these data track a common sample of people over several years.
4. Panel data allow better analysis of dynamic adjustment. Cross-sectional data can tell us nothing about dynamics. Time series data need to be very lengthy to provide good estimates of dynamic behavior, and then typically relate to aggregate dynamic behavior. Knowledge of individual dynamic reactions can be crucial to understanding economic phenomena. Panel data avoid the need for a lengthy time series by exploiting information on the dynamic reactions of each of several individuals.

18.2 Allowing for Different Intercepts

Suppose an individual's consumption y is determined linearly by his or her income x and we have observations on a thousand individuals ($N = 1000$) in each of four time periods ($T = 4$). A plot of all the data produces a scatter shown in simplified form (only a few observations are shown, not all 4000 observations!) in Figure 18.1. (Ignore the ellipses for the moment.) If we were to run ordinary least squares (OLS), we would produce a slope estimate shown by the line AA drawn through these data. But now suppose we identify these data by the cross-sectional unit (person, firm, or country, for example) to which they belong, in this case a person. This is shown in Figure 18.1 by drawing an ellipse for each person, surrounding all four time series observations

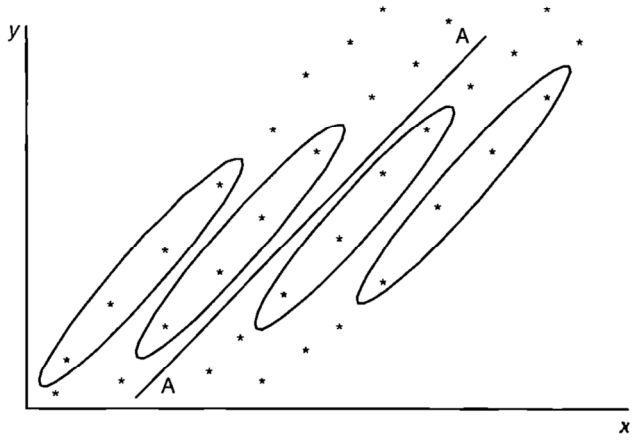

Figure 18.1 Panel data showing four observations on each of four individuals.

on that person. (There would be a thousand such ellipses in the actual data scatterplot, with roughly half above and half below *AA*; only four are drawn in Figure 18.1.) This way of viewing the data reveals that although each person in this example has the same slope, these people all have different intercepts. Most researchers would agree that this cross-sectional heterogeneity is the normal state of affairs – there are so many unmeasured variables that determine y that their influence gives rise to a different intercept for each individual. This phenomenon suggests that OLS is biased unless the influence of these omitted variables (embodied in different intercepts) is uncorrelated with the included explanatory variables. Two ways of improving estimation have been suggested, associated with two different ways of modeling the presence of a different intercept for each cross-sectional unit.

The first way is to put in a dummy for each individual (and omit the intercept). Doing this allows each individual to have a different intercept, and so OLS including all these dummies should guard against the bias discussed above. This "fixed effect" model gives rise to what is called the *fixed effects estimator* – OLS applied to the fixed effects model. At first glance this seems as though it would be difficult to estimate because (in our example above) we would require a thousand dummies. It turns out that a computational trick avoids this problem via an easy transformation of the data. This transformation consists of subtracting from each observation the average of the values within its ellipse – the observations for each individual have subtracted from them the averages of all the observations for that individual. OLS on these transformed data produces the desired slope estimate.

The fixed effects model has two major drawbacks:

1. By implicitly including a thousand dummy variables we lose 999 degrees of freedom (by dropping the intercept we save one degree of freedom). If we could find some way of avoiding this loss, we could produce a more efficient estimate of the common slope.

2. The transformation involved in this estimation process wipes out all explanatory variables that do not vary within an individual. This means that any explanatory variable that is time-invariant, such as gender, race, or religion, disappears, and so we are unable to estimate a slope coefficient for that variable. (This happens because within the ellipse in Figure 18.1, the values of these variables are all the same so that when we subtract their average they all become zero.)

The second way of allowing for different intercepts, the "random effects" model, is designed to overcome these two drawbacks of the fixed effects model. This model is similar to the fixed effects model in that it postulates a different intercept for each individual, but it interprets these differing intercepts in a novel way. This procedure views the different intercepts as having been drawn from a bowl of possible intercepts, so they may be interpreted as random (usually assumed to be normally distributed) and treated as though they were a part of the error term. As a result, we have a specification in which there is an overall intercept, a set of explanatory variables with coefficients of interest, and a composite error term. This composite error has two parts. For a particular individual, one part is the "random intercept" term, measuring the extent to which this individual's intercept differs from the overall intercept. The other part is just the traditional random error with which we are familiar, indicating a random deviation for that individual in that time period. For a particular individual the first part is the same in all time periods; the second part is different in each time period.

The trick to estimation using the random effects model is to recognize that the variance–covariance matrix of this composite error is nonspherical (i.e., not all off-diagonal elements are zero). In the example above, for all four observations on a specific individual, the random intercept component of the composite error is the same, so these composite errors will be correlated in a special way. Observations on different individuals are assumed to have zero correlation between their composite errors. This creates a variance–covariance matrix with a special pattern. The *random effects estimator* estimates this variance–covariance matrix and performs estimated generalized least squares (EGLS). The EGLS calculation is done by finding a transformation of the data that creates a spherical variance–covariance matrix and then performing OLS on the transformed data. In this respect it is similar to the fixed effects estimator except that it uses a different transformation.

18.3 Fixed Versus Random Effects

By saving on degrees of freedom, the random effects model produces a more efficient estimator of the slope coefficients than the fixed effects model. Furthermore, the transformation used for the random effects estimation procedure does not wipe out the explanatory variables that are time-invariant, allowing estimation of coefficients on variables such as gender, race, and religion. These results suggest that the random effects model is superior to the fixed effects model. So should we always use the

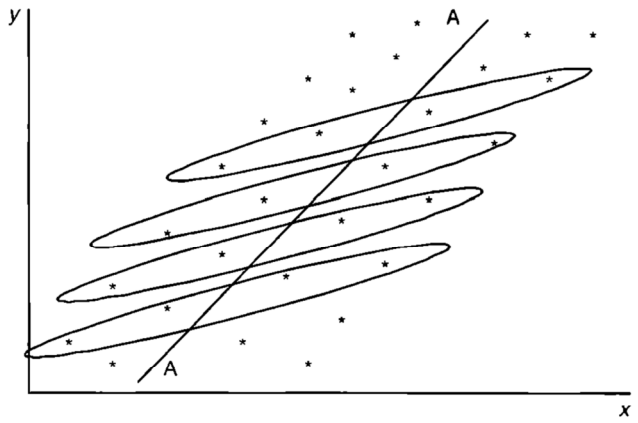

Figure 18.2 Panel data showing four observations on each of four individuals, with positive correlation between x and the intercept.

random effects model? Unfortunately, the random effects model has a major qualification that makes it applicable only in special circumstances.

This qualification is illustrated in Figure 18.2, where the data look exactly the same as in Figure 18.1, but the ellipses are drawn differently, to reflect a different allocation of observations to individuals. All persons have the same slope and different intercepts, just as before, but there is a big difference now – the common slope is not the same as the slope of the *AA* line, as it was in Figure 18.1. The main reason for this is that *the intercept for an individual is larger the larger is that individual's x value.* (Lines drawn through the observations in ellipses associated with higher x values cut the y axis at larger values.) This causes the OLS estimate using all the data to produce the *AA* line, clearly an overestimate of the common slope. This happens because as we move toward a higher x value, the y value increases for two reasons. First, it increases because the x value increases, and second, because there is likely to be a higher intercept. OLS estimation is biased upward because when x changes, OLS gives it credit for both of these y changes.

This bias does not characterize the fixed effects estimator because as described earlier the different intercepts are explicitly recognized by putting in dummies for them. But it is a problem for the random effects estimator because rather than being explicitly recognized, the intercepts are incorporated into the (composite) error term. As a consequence, the composite error term will tend to be bigger whenever the x value is bigger, creating correlation between x and the composite error term. Correlation between the error and an explanatory variable creates bias. As an example, suppose that wages are being regressed on schooling for a large set of individuals, and that a missing variable, ability, is thought to affect the intercept. Since schooling and ability are likely to be correlated, modeling this as a random effect will create correlation between the composite error and the regressor schooling, causing the random effects estimator to be biased. The bottom line here is that the random effects estimator should only be used

whenever we are confident that its composite error is uncorrelated with the explanatory variables. A test for this, a variant of the *Hausman test* (discussed in the general notes), is based on seeing if the random effects estimate is insignificantly different from the unbiased fixed effects estimate.

Here is a summary of the discussion above. Estimation with panel data begins by testing the null that the intercepts are equal. If this null is accepted the data are pooled. If this null is rejected, a Hausman test is applied to test if the random effects estimator is unbiased. If this null is not rejected, the random effects estimator is used; if this null is rejected, the fixed effects estimator is used. For the example shown in Figure 18.1, OLS, fixed effects, and random effects estimators are all unbiased, but random effects is most efficient. For the example shown in Figure 18.2, OLS and random effects estimators are biased, but the fixed effects estimator is not.

There are two kinds of variation in the data pictured in Figures 18.1 and 18.2. One kind is variation from observation to observation within a single ellipse (i.e., variation *within* a single individual). The other kind is variation in observations from ellipse to ellipse (i.e., variation *between* individuals). The fixed effects estimator uses the first type of variation (in all the ellipses), ignoring the second type. Because this first type of variation is variation *within* each cross-sectional unit, the fixed effects estimator is sometimes called the "within" estimator. An alternative estimator can be produced by using the second type of variation, ignoring the first type. This is done by finding the average of the values within each ellipse and then running OLS on these average values. This is called the "between" estimator because it uses variation between individuals (ellipses). Remarkably, the OLS estimator on the pooled data is an unweighted average of the within and between estimators. The random effects estimator is a (matrix-) weighted average of these two estimators. Three implications of this are of note.

1. This is where the extra efficiency of the random effects estimator comes from – it uses information from both the within and the between estimators.
2. This is how the random effects estimator can produce estimates of coefficients of time-invariant explanatory variables – these variables vary between ellipses, but not within ellipses.
3. This is where the bias of the random effects estimator comes from when the explanatory variable is correlated with the composite error – the between estimator is biased. The between estimator is biased because a higher x value gives rise to a higher y value both because x is higher and because the composite error is higher (because the intercept is higher) – the estimating formula gives the change in x all the credit for the change in y.

18.4 Short Run Versus Long Run

Suppose that an individual's consumption (y) is determined in the long run by his or her level of income (x), producing a data plot such as that in Figure 18.2. But suppose that due to habit persistence, in the short run the individual adjusts consumption only