Chapter Thirteen

Pooling Cross Sections Across Time. Simple Panel Data Methods

Up until now, we have covered multiple regression analysis using pure cross-sectional or pure time series data. While these two cases arise often in applications, data sets that have both cross-sectional and time series dimensions are being used more and more often in empirical research. Multiple regression methods can still be used on such data sets. In fact, data with cross-sectional and time series aspects can often shed light on important policy questions. We will see several examples in this chapter.

We will analyze two kinds of data sets in this chapter. An **independently pooled cross section** is obtained by sampling randomly from a large population at different points in time (usually, but not necessarily, different years). For instance, in each year, we can draw a random sample on hourly wages, education, experience, and so on, from the population of working people in the United States. Or, in every other year, we draw a random sample on the selling price, square footage, number of bathrooms, and so on, of houses sold in a particular metropolitan area. From a statistical standpoint, these data sets have an important feature: they consist of *independently* sampled observations. This was also a key aspect in our analysis of cross-sectional data: among other things, it rules out correlation in the error terms for different observations.

An independently pooled cross section differs from a single random sample in that sampling from the population at different points in time likely leads to observations that are not identically distributed. For example, distributions of wages and education have changed over time in most countries. As we will see, this is easy to deal with in practice by allowing the intercept in a multiple regression model, and in some cases the slopes, to change over time. We cover such models in Section 13.1. In Section 13.2, we discuss how pooling cross sections over time can be used to evaluate policy changes.

A **panel data** set, while having both a cross-sectional and a time series dimension, differs in some important respects from an independently pooled cross section. To collect panel data—sometimes called **longitudinal data**—we follow (or attempt to follow) the *same* individuals, families, firms, cities, states, or whatever, across time. For example, a panel data set on individual wages, hours, education, and other factors is collected by randomly selecting people from a population at a given point in time. Then, these *same* people are reinterviewed at several subsequent points in time. This

7/14/99 7:25 PM Page 409

gives us data on wages, hours, education, and so on, for the same group of people in different years.

Panel data sets are fairly easy to collect for school districts, cities, counties, states, and countries, and policy analysis is greatly enhanced by using panel data sets; we will see some examples in the following discussion. For the econometric analysis of panel data, we cannot assume that the observations are independently distributed across time. For example, unobserved factors (such as ability) that affect someone's wage in 1990 will also affect that person's wage in 1991; unobserved factors that affect a city's crime rate in 1985 will also affect that city's crime rate in 1990. For this reason, special models and methods have been developed to analyze panel data. In Sections 13.3, 13.4, and 13.5, we describe the straightforward method of differencing to remove time-constant, unobserved attributes of the units being studied. Because panel data methods are somewhat more advanced, we will rely mostly on intuition in describing the statistical properties of the estimation procedures, leaving details to the chapter appendix. We follow the same strategy in Chapter 14, which covers more complicated panel data methods.

13.1 POOLING INDEPENDENT CROSS SECTIONS ACROSS TIME

Many surveys of individuals, families, and firms are repeated at regular intervals, often each year. An example is the *Current Population Survey* (or CPS), which randomly samples households each year. (See, for example, CPS78_85.RAW, which contains data from the 1978 and 1985 CPS.) If a random sample is drawn at each time period, pooling the resulting random samples gives us an independently pooled cross section.

One reason for using independently pooled cross sections is to increase the sample size. By pooling random samples drawn from the same population, but at different points in time, we can get more precise estimators and test statistics with more power. Pooling is helpful in this regard only insofar as the relationship between the dependent variable and at least some of the independent variables remains constant over time.

As mentioned in the introduction, using pooled cross sections raises only minor statistical complications. Typically, to reflect the fact that the population may have different distributions in different time periods, we allow the intercept to differ across periods, usually years. This is easily accomplished by including dummy variables for all but one year, where the earliest year in the sample is usually chosen as the base year. It is also possible that the error variance changes over time, something we discuss later.

Sometimes, the pattern of coefficients on the year dummy variables is itself of interest. For example, a demographer may be interested in the following question: *After* controlling for education, has the pattern of fertility among women over age 35 changed between 1972 and 1984? The following example illustrates how this question is simply answered by using multiple regression analysis with **year dummy variables**.

EXAMPLE 13.1
(Women's Fertility Over Time)

The data set in FERTIL1.RAW, which is similar to that used by Sander (1994), comes from the National Opinion Research Center's *General Social Survey* for the even years from 1972

7/14/99 7:25 PM Page 410

to 1984, inclusively. We use these data to estimate a model explaining the total number of kids born to a woman (*kids*).

One question of interest is: After controlling for other observable factors, what has happened to fertility rates over time? The factors we control for are years of education, age, race, region of the country where living at age 16, and living environment at age 16. The estimates are given in Table 13.1.

The base year is 1972. The coefficients on the year dummy variables show a sharp drop in fertility in the early 1980s. For example, the coefficient on *y82* implies that, holding education, age, and other factors fixed, a woman had on average .52 less children, or about one-half a child, in 1982 than in 1972. This is a very large drop: holding *educ*, *age*, and the other factors fixed, 100 women in 1982 are predicted to have about 52 fewer children than 100 comparable women in 1972. Since we are controlling for education, this drop is separate from the decline in fertility that is due to the increase in average education levels. (The average years of education are 12.2 for 1972 and 13.3 for 1984.) The coefficients on *y82* and *y84* represent drops in fertility for reasons that are not captured in the explanatory variables.

Given that the 1982 and 1984 year dummies are individually quite significant, it is not surprising that as a group the year dummies are jointly very significant: the R-squared for the regression without the year dummies is .1019, and this leads to $F_{6,1111} = 5.87$ and p-value ≈ 0.

Women with more education have fewer children, and the estimate is very statistically significant. Other things being equal, 100 women with a college education will have about 51 fewer children on average than 100 women with only a high school education: .128(4) = .512. Age has a diminishing effect on fertility. (The turning point in the quadratic is at about *age* = 46, by which time most women have finished having children.)

The model estimated in Table 13.1 assumes that the effect of each explanatory variable, particularly education, has remained constant. This may or may not be true; you will be asked to explore this issue in Problem 13.7.

Finally, there may be heteroskedasticity in the error term underlying the estimated equation. This can be dealt with using the methods in Chapter 8. There is one interesting difference here: now, the error variance may change over time even if it does not change with the values of *educ*, *age*, *black*, and so on. The heteroskedasticity-robust standard errors and test statistics are nevertheless valid. The Breusch-Pagan test would be obtained by regressing the squared OLS residuals on *all* of the independent variables in Table 13.1, including the year dummies. (For the special case of the White statistic, the fitted values $\hat{kids}$ and the squared fitted values are used as the independent variables, as always.) A weighted least squares procedure should account for variances that possibly change over time. In the procedure discussed in Section 8.4, year dummies would be included in equation (8.32).

QUESTION 13.1

In reading Table 13.1, someone claims that, if everything else is equal in the table, a black woman is expected to have one more child than a nonblack woman. Do you agree with this claim?

We can also interact a year dummy variable with key explanatory variables to see if the effect of that variable has changed over a certain time period. The next example examines how the return to education and the gender gap have changed from 1978 to 1985.

7/14/99 7:25 PM Page 411

Table 13.1

Determinants of Women's Fertility

Dependent Variable: *kids*		
Independent Variables	**Coefficients**	**Standard Errors**
educ	−.128	.018
age	.532	.138
age^2	−.0058	.0016
black	1.076	.174
east	.217	.133
northcen	.363	.121
west	.198	.167
farm	−.053	.147
othrural	−.163	.175
town	.084	.124
smcity	.212	.160
y74	.268	.173
y76	−.097	.179
y78	−.069	.182
y80	−.071	.183
y82	−.522	.172
y84	−.545	.175
constant	−7.742	3.052
$n = 1{,}129$ $R^2 = .1295$ $\bar{R}^2 = .1162$		

7/14/99 7:25 PM Page 412

EXAMPLE 13.2
(Changes in the Return to Education and the Gender Wage Gap)

A log(*wage*) equation (where *wage* is hourly wage) pooled across the years 1978 (the base year) and 1985 is

$$\log(wage) = \beta_0 + \delta_0 y85 + \beta_1 educ + \delta_1 y85 \cdot educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 union + \beta_5 female + \delta_5 y85 \cdot female + u, \quad (13.1)$$

where most explanatory variables should by now be familiar. The variable *union* is a dummy variable equal to one if the person belongs to a union, and zero otherwise. The variable *y85* is a dummy variable equal to one if the observation comes from 1985 and zero if it comes from 1978. There are 550 people in the sample in 1978 and a different set of 534 people in 1985.

The intercept for 1978 is β_0, and the intercept for 1985 is $\beta_0 + \delta_0$. The return to education in 1978 is β_1, and the return to education in 1985 is $\beta_1 + \delta_1$. Therefore, δ_1 measures how the return to another year of education has changed over the seven-year period. Finally, in 1978, the log(*wage*) differential between women and men is β_5; the differential in 1985 is $\beta_5 + \delta_5$. Thus, we can test the null hypothesis that nothing has happened to the gender differential over this seven-year period by testing H_0: $\delta_5 = 0$. The alternative that the gender differential has been *reduced* is H_1: $\delta_5 < 0$. For simplicity, we have assumed that experience and union membership have the same effect on wages in both time periods.

Before we present the estimates, there is one other issue we need to address; namely, hourly wage here is in nominal (or current) dollars. Since nominal wages grow simply due to inflation, we are really interested in the effect of each explanatory variable on real wages. Suppose that we settle on measuring wages in 1978 dollars. This requires deflating 1985 wages to 1978 dollars. (Using the consumer price index for the 1997 *Economic Report of the President*, the deflation factor is 107.6/65.2 $\approx$ 1.65.) While we can easily divide each 1985 wage by 1.65, it turns out that this is not necessary, *provided* a 1985 year dummy is included in the regression *and* log(*wage*) (as opposed to *wage*) is used as the dependent variable. Using real or nominal wage in a logarithmic functional form only affects the coefficient on the year dummy, *y85*. To see this, let *P85* denote the deflation factor for 1985 wages (1.65, if we use the CPI). Then, the log of the real wage for each person *i* in the 1985 sample is

$$\log(wage_i/P85) = \log(wage_i) - \log(P85).$$

Now, while $wage_i$ differs across people, *P85* does not. Therefore, log(*P85*) will be absorbed into the intercept for 1985. (This conclusion would change if, for example, we used a different price index for people in various parts of the country.) The bottom line is that, for studying how the return to education or the gender gap has changed, we do not need to turn nominal wages into real wages in equation (13.1). Problem 13.8 asks you to verify this for the current example.

If we forget to allow different intercepts in 1978 and 1985, the use of nominal wages can produce seriously misleading results. If we use *wage* rather than log(*wage*) as the dependent variable, it is important to use the real wage and to include a year dummy.

7/14/99 7:25 PM Page 413

The previous discussion generally holds when using dollar values for either the dependent or independent variables. Provided the dollar amounts appear in logarithmic form and dummy variables are used for all time periods (except, of course, the base period), the use of aggregate price deflators will only affect the intercepts; none of the slope estimates will change.

Now, we use the data in CPS78_85.RAW to estimate the equation:

$$\begin{aligned}\widehat{\log(wage)} = \underset{(.093)}{.459} + \underset{(.123)}{.118}\, y85 + \underset{(.0067)}{.0747}\, educ + \underset{(.0094)}{.0185}\, y85{\cdot}educ \\ + \underset{(.0036)}{.0296}\, exper - \underset{(.00008)}{.00040}\, exper^2 + \underset{(.030)}{.202}\, union \\ - \underset{(.037)}{.317}\, female + \underset{(.051)}{.085}\, y85{\cdot}female \\ n = 1{,}084,\ R^2 = .426,\ \bar{R}^2 = .422.\end{aligned} \tag{13.2}$$

The return to education in 1978 is estimated to be about 7.5%; the return to education in 1985 is about 1.85 percentage points *higher*, or about 9.35%. Because the *t* statistic on the interaction term is .0185/.0094 ≈ 1.97, the difference in the return to education is statistically significant at the 5% level against a two-sided alternative.

What about the gender gap? In 1978, other things being equal, a woman earned about 31.7% less than a man (27.2% is the more accurate estimate). In 1985, the gap in log(*wage*) is −.317 + .085 = −.232. Therefore, the gender gap appears to have fallen from 1978 to 1985 by about 8.5 percentage points. The *t* statistic on the interaction term is about 1.67, which means it is significant at the 5% level against the positive one-sided alternative.

What happens if we interact *all* independent variables with *y85* in equation (13.2)? This is identical to estimating two separate equations, one for 1978 and one for 1985. Sometimes this is desirable. For example, in Chapter 7, we discussed a study by Krueger (1993), where he estimated the return to using a computer on the job. Krueger estimates two separate equations, one using the 1984 CPS and the other using the 1989 CPS. By comparing how the return to education changes across time and whether or not computer usage is controlled for, he estimates that one-third to one-half of the observed increase in the return to education over the five-year period can be attributed to increased computer usage. [See Tables VIII and IX in Krueger (1993).]

The Chow Test for Structural Change Across Time

In Chapter 7, we discussed how the Chow test—which is simply an *F* test—can be used to determine whether a multiple regression function differs across two groups. We can apply that test to two different time periods as well. One form of the test obtains the sum of squared residuals from the pooled estimation as the restricted SSR. The unrestricted SSR is the sum of the SSRs for the two separately estimated time periods. The

7/14/99 7:25 PM Page 414

mechanics of computing the statistic are exactly as they were in Section 7.4. A heteroskedasticity-robust version is also available (see Section 8.2).

Example 13.2 suggests another way to compute the Chow test for two time periods by interacting each variable with a year dummy for one of the two years and testing for joint significance of the year dummy and all of the interaction terms. Since the intercept in a regression model often changes over time (due to, say, inflation in the housing price example), this full-blown Chow test can detect such changes. It is usually more interesting to allow for an intercept difference and then to test whether certain slope coefficients change over time (as we did in Example 13.2).

A Chow test can be computed for more than two time periods, but the calculations can be tedious. Usually, after an allowance for intercept difference, certain slope coefficients are tested for constancy by interacting the variable of interest with year dummies. (See Problems 13.7 and 13.8 for examples.)

13.2 POLICY ANALYSIS WITH POOLED CROSS SECTIONS

Pooled cross sections can be very useful for evaluating the impact of a certain event or policy. The following example of an event study shows how two cross-sectional data sets, collected before and after the occurrence of an event, can be used to determine the effect on economic outcomes.

EXAMPLE 13.3
(Effect of a Garbage Incinerator's Location on Housing Prices)

Kiel and McClain (1995) studied the effect that a new garbage incinerator had on housing values in North Andover, Massachusetts. They used many years of data and a fairly complicated econometric analysis. We will use two years of data and some simplified models, but our analysis is similar.

The rumors that a new incinerator would be built in North Andover began after 1978, and construction began in 1981. The incinerator was expected to be in operation soon after the start of construction; the incinerator actually began operating in 1985. We will use data on prices of houses that sold in 1978 and another sample on those that sold in 1981. The hypothesis is that the price of houses located near the incinerator would fall below the price of more distant houses.

For illustration, we define a house to be near the incinerator if it is within three miles. [In the problems, you are instead asked to use the actual distance from the house to the incinerator, as in Kiel and McClain (1995).] We will start by looking at the dollar effect on housing prices. This requires us to measure price in constant dollars. We measure all housing prices in 1978 dollars, using the Boston housing price index. Let *rprice* denote the house price in real terms.

A naive analyst would use only the 1981 data and estimate a very simple model:

$$rprice = \gamma_0 + \gamma_1 nearinc + u, \qquad (13.3)$$

7/14/99 7:25 PM Page 415

where *nearinc* is a binary variable equal to one if the house is near the incinerator, and zero otherwise. Estimating this equation using the data in KIELMC.RAW gives

$$\widehat{rprice} = 101{,}307.5 - 30{,}688.27\ nearinc \quad (3{,}093.0) \quad (5{,}827.71) \quad n = 142,\ R^2 = .165. \tag{13.4}$$

Since this is a simple regression on a single dummy variable, the intercept is the average selling price for homes not near the incinerator, and the coefficient on *nearinc* is the difference in the average selling price between homes near the incinerator and those that are not. The estimate shows that the average selling price for the former group was $30,688.27 less than for the latter group. The *t* statistic is greater than five in absolute value, so we can strongly reject the hypothesis that the average value for homes near to and far from the incinerator are not the same.

Unfortunately, equation (13.4) does *not* imply that the siting of the incinerator is causing the lower housing values. In fact, if we run the same regression for 1978 (before the incinerator was even rumored), we obtain

$$\widehat{rprice} = 82{,}517.23 - 18{,}824.37\ nearinc \quad (2{,}653.79) \quad (5{,}827.71) \quad n = 179,\ R^2 = .082. \tag{13.5}$$

Therefore, even *before* there was any talk of an incinerator, the average value of a home near the site was $18,824.37 less than the average value of a home not near the site ($82,517.23); the difference is statistically significant, as well. This is consistent with the view that the incinerator was built in an area with lower housing values.

How, then, can we tell whether building a new incinerator depresses housing values? The key is to look at how the coefficient on *nearinc* changed between 1978 and 1981. The difference in average housing value was much larger in 1981 than in 1978 ($30,688.27 versus $18,824.37), even as a percentage of the average value of homes not near the incinerator site. The difference in the two coefficients on *nearinc* was

$$\hat{\delta}_1 = -30{,}688.27 - (-18{,}824.37) = -11{,}863.9.$$

This is our estimate of the effect of the incinerator on values of homes near the incinerator site. In empirical economics, $\hat{\delta}_1$ has become known as the **difference-in-differences estimator** because it can be expressed as

$$\hat{\delta}_1 = (\overline{rprice}_{81,nr} - \overline{rprice}_{81,fr}) - (\overline{rprice}_{78,nr} - \overline{rprice}_{78,fr}), \tag{13.6}$$

where "nr" stands for "near the incinerator site" and "fr" stands for "farther away from the site." In other words, $\hat{\delta}_1$ is the difference over time in the average difference of housing prices in the two locations.

To test whether $\hat{\delta}_1$ is statistically different from zero, we need to find its standard error by using a regression analysis. In fact, $\hat{\delta}_1$ can be obtained by estimating

$$rprice = \beta_0 + \delta_0 y81 + \beta_1 nearinc + \delta_1 y81 \cdot nearinc + u, \tag{13.7}$$

7/14/99 7:25 PM Page 416

using the data pooled over both years. The intercept, β_0, is the average price of a home not near the incinerator in 1978. The parameter, δ_0 captures changes in *all* housing values in North Andover from 1978 to 1981. [A comparison of equations (13.4) and (13.5) showed that housing values in North Andover, relative to the Boston housing price index, increased sharply over this period.] The coefficient on *nearinc*, β_1, measures the location effect that is *not* due to the presence of the incinerator: as we saw in equation (13.5), even in 1978, homes near the incinerator site sold for less than homes farther away from the site.

The parameter of interest is on the interaction term *y81·nearinc*: δ_1 measures the decline in housing values due to the new incinerator, provided we assume that houses both near and far from the site did not appreciate at different rates for other reasons.

The estimates of equation (13.7) are given in column (1) of Table 13.2.

Table 13.2

Dependent Variable: *rprice*

Independent Variable	(1)	(2)	(3)
constant	82,517.23 (2,726.91)	89,116.54 (2,406.05)	13,807.67 (11,166.59)
y81	18,790.29 (4,050.07)	21,321.04 (3,443.63)	13,928.48 (2,798.75)
nearinc	−18,824.37 (4,875.32)	9,397.94 (4,812.22)	3,780.34 (4,453.42)
y81·nearinc	−11,863.90 (7,456.65)	−21,920.27 (6,359.75)	−14,177.93 (4,987.27)
Other Controls	No	*age*, age^2	Full Set
Observations *R*-Squared	321 .174	321 .414	321 .660

The only number we could not obtain from equations (13.4) and (13.5) is the standard error of $\hat{\delta}_1$. The *t* statistic on $\hat{\delta}_1$ is about −1.59, which is marginally significant against a one-sided alternative (*p*-value ≈ .057).

Kiel and McClain (1995) included various housing characteristics in their analysis of the incinerator siting. There are two good reasons for doing this. First, the kinds of houses selling in 1981 might have been systematically different than those selling in 1978; if so, it is important to control for characteristics that might have been different. But just as important, even if the average housing characteristics are the same for both years, including them can greatly reduce the error variance, which can then shrink the standard error of $\hat{\delta}_1$. (See Section 6.3 for discussion.) In column (2), we control for the age of the houses, using a qua-

7/14/99 7:25 PM Page 417

dratic. This substantially increases the *R*-squared (by reducing the residual variance). The coefficient on *y81·nearinc* is now much larger in magnitude, and its standard error is lower.

In addition to the age variables in column (2), column (3) controls for distance to the interstate in feet (*intst*), land area in feet (*land*), house area in feet (*area*), number of rooms (*rooms*), and number of baths (*baths*). This produces an estimate on *y81·nearinc* closer to that without any controls, but it yields a much smaller standard error: the *t* statistic for $\hat{\delta}_1$ is about -2.84. Therefore, we find a much more significant effect in column (3) than in column (1). The column (3) estimates are preferred because they control for the most factors and have the smallest standard errors (except in the constant, which is not important here). The fact that *nearinc* has a much smaller coefficient and is insignificant in column (3) indicates that the characteristics included in column (3) largely capture the housing characteristics that are most important for determining housing prices.

For the purpose of introducing the method, we used the level of real housing prices in Table 13.2. It makes more sense to use log(*price*) [or log(*rprice*)] in the analysis in order to get an approximate percentage effect. The basic model becomes

$$\log(price) = \beta_0 + \delta_0 y81 + \beta_1 nearinc + \delta_1 y81 \cdot nearinc + u. \qquad \textbf{(13.8)}$$

Now, $100 \cdot \delta_1$ is the approximate percentage reduction in housing value due to the incinerator. [Just as in Example 13.2, using log(*price*) versus log(*rprice*) only affects the coefficient on *y81*.] Using the same 321 pooled observations gives

$$\widehat{\log(price)} = \underset{(0.31)}{11.29} + \underset{(.045)}{.457\ y81} - \underset{(.055)}{.340\ nearinc} - \underset{(.083)}{.063\ y81 \cdot nearinc} \qquad \textbf{(13.9)}$$
$$n = 321,\ R^2 = .409.$$

The coefficient on the interaction term implies that, because of the new incinerator, houses near the incinerator lost about 6.3% in value. However, this estimate is not statistically different from zero. But when we use a full set of controls, as in column (3) of Table 13.2 (but with *intst*, *land*, and *area* appearing in logarithmic form), the coefficient on *y81·nearinc* becomes $-.132$ with a *t* statistic of about -2.53. Again, controlling for other factors turns out to be important. Using the logarithmic form, we estimate that houses near the incinerator were devalued by about 13.2%.

The methodology applied to the previous example has numerous applications, especially when the data arise from a **natural experiment** (or a **quasi-experiment**). A natural experiment occurs when some exogenous event—often a change in government policy—changes the environment in which individuals, families, firms, or cities operate. A natural experiment always has a control group, which is not affected by the policy change, and a treatment group, which is thought to be affected by the policy change. Unlike with a true experiment, where treatment and control groups are randomly and explicitly chosen, the control and treatment groups in natural experiments arise from the particular policy change. In order to control for systematic differences between the control and treatment groups, we need two years of data, one before the policy change

7/14/99 7:25 PM Page 418

and one after the change. Thus, our sample is usefully broken down into four groups: the control group before the change, the control group after the change, the treatment group before the change, and the treatment group after the change.

Call A the control group and B the treatment group, letting *dB* equal unity for those in the treatment group B, and zero otherwise. Then, letting *d2* denote a dummy variable for the second (postpolicy change) time period, the equation of interest is

$$y = \beta_0 + \delta_0 d2 + \beta_1 dB + \delta_1 d2 \cdot dB + other\ factors, \tag{13.10}$$

where y is the outcome variable of interest. As in Example 13.3, δ_1 measures the effect of the policy. Without other factors in the regression, $\hat{\delta}_1$ will be the difference-in-differences estimator:

$$\hat{\delta}_1 = (\bar{y}_{2,B} - \bar{y}_{2,A}) - (\bar{y}_{1,B} - \bar{y}_{1,A}), \tag{13.11}$$

where the bar denotes average, the first subscript denotes the year, and the second subscript denotes the group. When explanatory variables are added to equation (13.10) (to control for the fact that the populations sampled may differ systematically over the two periods), the OLS estimate of δ_1 no longer has the simple form of (13.11), but its interpretation is similar.

EXAMPLE 13.4

(Effect of Worker Compensation Laws on Duration)

Meyer, Viscusi, and Durbin (1995) (hereafter, MVD) studied the length of time (in weeks) that an injured worker receives workers' compensation. On July 15, 1980, Kentucky raised the cap on weekly earnings that were covered by workers' compensation. An increase in the cap has no effect on the benefit for low-income workers, but it makes it less costly for a high-income worker to stay on workers' compensation. Therefore, the control group is low-income workers, and the treatment group is high-income workers; high-income workers are defined as those who are subject to the prepolicy change cap. Using random samples both before and after the policy change, MVD were able to test whether more generous workers' compensation causes people to stay out of work longer (everything else fixed). They started with a difference-in-differences analysis, using log(*durat*) as the dependent variable. Let *afchnge* be the dummy variable for observations after the policy change and *highearn* the dummy variable for high earners. The estimated equation, with standard errors in parentheses, is

$$\begin{aligned} \widehat{\log(durat)} = {} & 1.126 + .0077\ afchnge + .256\ highearn \\ & \quad (0.031) \quad (.0447) \qquad\quad (.047) \\ & + .191\ afchnge \cdot highearn \\ & \quad (.069) \\ & n = 5{,}626,\ R^2 = .021. \end{aligned} \tag{13.12}$$

Therefore, $\hat{\delta}_1 = .191$ ($t = 2.77$), which implies that the average length of time on workers' compensation increased by about 19% due to the higher earnings cap. The coefficient on

7/14/99 7:25 PM Page 419

afchnge is small and statistically insignificant: as is expected, the increase in the earnings cap has no effect on duration for low-income workers.

This is a good example of how we can get a fairly precise estimate of the effect of a policy change, even though we cannot explain much of the variation in the dependent variable. The dummy variables in (13.12) explain only 2.1% of the variation in log(*durat*). This makes sense: there are clearly many factors, including severity of the injury, that affect how long someone is on workers' compensation. Fortunately, we have a very large sample size, and this allows us to get a significant *t* statistic.

MVD also added a variety of controls for gender, marital status, age, industry, and type of injury. This allows for the fact that the kinds of people and types of injuries differ systematically in the two years. Controlling for these factors turns out to have little effect on the estimate of δ_1. (See Problem 13.10.)

Sometimes, the two groups consist of people living in two neighboring states in the United States. For example, to assess the impact of changing cigarette taxes on cigarette consumption, we can obtain random samples from two states for two years. In State A, the control group, there was no change in the cigarette tax. In State B, the tax increased (or decreased) between the two years. The outcome variable would be a measure of cigarette consumption, and equation (13.10) can be estimated to determine the effect of the tax on cigarette consumption.

QUESTION 13.2

What do you make of the coefficient and *t* statistic on *highearn* in equation (13.12)?

For an interesting survey on natural experiment methodology and several additional examples, see Meyer (1995).

13.3 TWO-PERIOD PANEL DATA ANALYSIS

We now turn to the analysis of the simplest kind of panel data: for a cross section of individuals, schools, firms, cities, or whatever, we have two years of data; call these $t = 1$ and $t = 2$. These years need not be adjacent, but $t = 1$ corresponds to the earlier year. For example, the file CRIME2.RAW contains data on (among other things) crime and unemployment rates for 46 cities for 1982 and 1987. Therefore, $t = 1$ corresponds to 1982, and $t = 2$ corresponds to 1987.

What happens if we use the 1987 cross section and run a simple regression of *crmrte* on *unem*? We obtain

$$\widehat{crmrte} = 128.38 - 4.16\ unem$$
$$(20.76)\quad (3.42)$$
$$n = 46,\ R^2 = .033.$$

If we interpret the estimated equation causally, it implies that an increase in the unemployment rate *lowers* the crime rate. This is certainly not what we expect. The coefficient on *unem* is not statistically significant at standard significance levels: at best, we have found no link between crime and unemployment rates.

7/14/99 7:25 PM Page 420

As we have emphasized throughout this text, this simple regression equation likely suffers from omitted variable problems. One possibile solution is to try to control for more factors, such as age distribution, gender distribution, education levels, law enforcement efforts, and so on, in a multiple regression analysis. But many factors might be hard to control for. In Chapter 9, we showed how including the *crmrte* from a previous year—in this case, 1982—can help to control for the fact that different cities have historically different crime rates. This is one way to use two years of data for estimating a causal effect.

An alternative way to use panel data is to view the unobserved factors affecting the dependent variable as consisting of two types: those that are constant and those that vary over time. Letting i denote the cross-sectional unit and t the time period, we can write a model with a single observed explanatory variable as

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + a_i + u_{it}, \quad t = 1,2. \tag{13.13}$$

In the notation y_{it}, i denotes the person, firm, city, and so on, and t denotes the time period. The variable $d2_t$ is a dummy variable that equals zero when $t = 1$ and one when $t = 2$; it does not change across i, which is why it has no i subscript. Therefore, the intercept for $t = 1$ is β_0, and the intercept for $t = 2$ is $\beta_0 + \delta_0$. Just as in using independently pooled cross sections, allowing the intercept to change over time is important in most applications. In the crime example, secular trends in the United States will cause crime rates in all U.S. cities to change, perhaps markedly, over a five-year period.

The variable a_i captures all unobserved, time-constant factors that affect y_{it}. (The fact that a_i has no t subscript tells us that it does not change over time.) Generically, a_i is called an **unobserved effect**. It is also common in applied work to find a_i referred to as a **fixed effect**, which helps us to remember that a_i is fixed over time. The model in (13.13) is called an **unobserved effects model** or a **fixed effects model**. In applications, you might see a_i referred to as **unobserved heterogeneity** as well (or *individual heterogeneity*, *firm heterogeneity*, *city heterogeneity*, and so on).

The error u_{it} is often called the **idiosyncratic error** or time-varying error, because it represents unobserved factors that change over time and affect y_{it}. These are very much like the errors in a straight time series regression equation.

A simple unobserved effects model for city crime rates for 1982 and 1987 is

$$crmrte_{it} = \beta_0 + \delta_0 d87_t + \beta_1 unem_{it} + a_i + u_{it}, \tag{13.14}$$

where *d87* is a dummy variable for 1987. Since i denotes different cities, we call a_i an *unobserved city effect* or a *city fixed effect*: it represents all factors affecting city crime rates that do not change over time. Geographical features, such as the city's location in the United States, are included in a_i. Many other factors may not be exactly constant, but they might be roughly constant over a five-year period. These might include certain demographic features of the population (age, race, and education). Different cities may have their own methods for reporting crimes, and the people living in the cities might have different attitudes toward crime; these are typically slow to change. For historical reasons, cities can have very different crime rates, which are at least partially captured by the unobserved effect a_i.

7/14/99 7:25 PM Page 421

How should we estimate the parameter of interest, β_1, given two years of panel data? One possibility is to just pool the two years and use OLS, essentially as in Section 13.1. This method has two drawbacks. The most important of these is that, in order for pooled OLS to produce a consistent estimator of β_1, we would have to assume that the unobserved effect, a_i, is uncorrelated with x_{it}. We can easily see this by writing (13.13) as

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + v_{it}, \quad t = 1,2, \tag{13.15}$$

where $v_{it} = a_i + u_{it}$ is often called the **composite error**. From what we know about OLS, we must assume that v_{it} is uncorrelated with x_{it}, where $t = 1$ or 2, for OLS to consistently estimate β_1 (and the other parameters). This is true whether we use a single cross section or pool the two cross sections. Therefore, even if we assume that the idiosyncratic error u_{it} is uncorrelated with x_{it}, pooled OLS is biased and inconsistent if a_i and x_{it} are correlated. The resulting bias in pooled OLS is sometimes called **heterogeneity bias**, but it is really just bias caused from omitting a time-constant variable.

QUESTION 13.3

Suppose that a_i, u_{i1}, and u_{i2} have zero means and are pairwise uncorrelated. Show that $\text{Cov}(v_{i1},v_{i2}) = \text{Var}(a_i)$, so that the composite errors are positively serially correlated across time, unless $a_i = 0$. What does this imply about the usual OLS standard errors from pooled OLS estimation?

To illustrate what happens, we use the data in CRIME2.RAW to estimate (13.14) by pooled OLS. Since there are 46 cities and two years for each city, there are 92 total observations:

$$\widehat{crmrte} = 93.42 + 7.94\, d87 + .427\, unem \underset{(12.74)\quad (7.98)\qquad (1.188)}{} \tag{13.16}$$
$$n = 92, R^2 = .012.$$

(When reporting the estimated equation, we usually drop the i and t subscripts.) The coefficient on *unem*, though positive in (13.16), has a very small t statistic. Thus, using pooled OLS on the two years has not substantially changed anything from using a single cross section. This is not surprising since using pooled OLS does not solve the omitted variables problem. (The standard errors in this equation are incorrect because of the serial correlation noted earlier, but we ignore this since pooled OLS is not the focus here.)

In most applications, the main reason for collecting panel data is to allow for the unobserved effect, a_i, to be correlated with the explanatory variables. For example, in the crime equation, we want to allow the unmeasured city factors in a_i that affect the crime rate to also be correlated with the unemployment rate. It turns out that this is simple to allow: because a_i is constant over time, we can difference the data across the two years. More precisely, for a cross-sectional observation i, write the two years as

$$y_{i2} = (\beta_0 + \delta_0) + \beta_1 x_{i2} + a_i + u_{i2} \quad (t = 2)$$
$$y_{i1} = \beta_0 + \beta_1 x_{i1} + a_i + u_{i1} \quad (t = 1).$$

If we subtract the *second* equation from the *first*, we obtain

$$(y_{i2} - y_{i1}) = \delta_0 + \beta_1 (x_{i2} - x_{i1}) + (u_{i2} - u_{i1}),$$

7/14/99 7:25 PM Page 422

or

$$\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_i, \tag{13.17}$$

where "Δ" denotes the change from $t = 1$ to $t = 2$. The unobserved effect, a_i, does not appear in (13.17): it has been "differenced away." Also, the intercept in (13.17) is actually the *change* in the intercept from $t = 1$ to $t = 2$.

Equation (13.17), which we call the **first-differenced equation**, is very simple. It is just a single cross-sectional equation, but each variable is differenced over time. We can analyze (13.17) using the methods we developed in Part 1, provided the key assumptions are satisfied. The most important of these is that Δu_i is uncorrelated with Δx_i. This assumption holds if the idiosyncratic error at each time t, u_{it}, is uncorrelated with the explanatory variable in *both* time periods. This is another version of the **strict exogeneity** assumption that we encountered in Chapter 10 for time series models. In particular, this assumption rules out the case where x_{it} is the lagged dependent variable, $y_{i,t-1}$. Unlike in Chapter 10, we allow x_{it} to be correlated with unobservables that are constant over time. When we obtain the OLS estimator of β_1 from (13.17) we call the resulting estimator the **first-differenced estimator**.

In the crime example, assuming that Δu_i and $\Delta unem_i$ are uncorrelated may be reasonable, but it can also fail. For example, suppose that law enforcement effort (which is in the idiosyncratic error) increases more in cities where the unemployment rate decreases. This can cause negative correlation between Δu_i and $\Delta unem_i$, which would then lead to bias in the OLS estimator. Naturally, this problem can be overcome to some extent by including more factors in the equation, something we will cover later. As usual, it is always possible that we have not accounted for enough time-varying factors.

Another crucial condition is that Δx_i must have some variation across i. This qualification fails if the explanatory variable does not change over time for any cross-sectional observation, or if it changes by the same amount for every observation. This is not an issue in the crime rate example because the unemployment rate changes across time for almost all cities. But, if i denotes an individual and x_{it} is a dummy variable for gender, $\Delta x_i = 0$ for all i; we clearly cannot estimate (13.17) by OLS in this case. This actually makes perfectly good sense: since we allow a_i to be correlated with x_{it}, we cannot hope to separate the effect of a_i on y_{it} from the effect of any variable that does not change over time.

The only other assumption we need to apply to the usual OLS statistics is that (13.17) satisfies the homoskedasticity assumption. This is reasonable in many cases, and, if it does not hold, we know how to test and correct for heteroskedasticity using the methods in Chapter 8. It is sometimes fair to assume that (13.17) fulfills all of the classical linear model assumptions. The OLS estimators are unbiased and all statistical inference is exact in such cases.

When we estimate (13.17) for the crime rate example, we get

$$\begin{aligned} \widehat{\Delta crmrte} &= 15.40 + 2.22\ \Delta unem \\ &\quad\ (4.70)\quad (0.88) \\ n &= 46,\ R^2 = .127, \end{aligned} \tag{13.18}$$

7/14/99 7:25 PM Page 423

which now gives a positive, statistically significant relationship between the crime and unemployment rates. Thus, differencing to eliminate time-constant effects makes a big difference in this example. The intercept in (13.18) also reveals something interesting. Even if $\Delta unem = 0$, we predict an increase in the crime rate (crimes per 1,000 people) of 15.40. This reflects a secular increase in crime rates throughout the United States from 1982 to 1987.

Even if we do not begin with the unobserved effects model (13.13), using differences across time makes intuitive sense. Rather than estimating a standard cross-sectional relationship—which may suffer from omitted variables, thereby making ceteris paribus conclusions difficult—equation (13.17) explicitly considers how changes in the explanatory variable over time affect the change in y over the same time period. Nevertheless, it is still very useful to have (13.13) in mind: it explicitly shows that we can estimate the effect of x_{it} on y_{it}, holding a_i fixed.

While differencing two years of panel data is a powerful way to control for unobserved effects, it is not without cost. First, panel data sets are harder to collect than a single cross section, especially for individuals. We must use a survey and keep track of the individual for a follow-up survey. It is often difficult to locate some people for a second survey. For units such as firms, some firms will go bankrupt or merge with other firms. Panel data are much easier to obtain for schools, cities, counties, states, and countries.

Even if we have collected a panel data set, the differencing used to eliminate a_i can greatly reduce the variation in the explanatory variables. While x_{it} frequently has substantial variation in the cross section for each t, Δx_i may not have much variation. We know from Chapter 3 that little variation in Δx_i can lead to large OLS standard errors. We can combat this by using a large cross section, but this is not always possible. Also, using longer differences over time is sometimes better than using year-to-year changes.

As an example, consider the problem of estimating the return to education, now using panel data on individuals for two years. The model for person i is

$$\log(wage_{it}) = \beta_0 + \delta_0 d2_t + \beta_1 educ_{it} + a_i + u_{it}, \quad t = 1,2,$$

where a_i contains unobserved ability—which is probably correlated with $educ_{it}$. Again, we allow different intercepts across time to account for aggregate productivity gains (and inflation, if $wage_{it}$ is in nominal terms). Since, by definition, innate ability does not change over time, panel data methods seem ideally suited to estimate the return to education. The equation in first differences is

$$\Delta\log(wage_i) = \delta_0 + \beta_1 \Delta educ_i + \Delta u_i, \tag{13.19}$$

and we can estimate this by OLS. The problem is that we are interested in working adults, and for most employed individuals, education does not change over time. If only a small fraction of our sample has $\Delta educ_i$ different from zero, it will be difficult to get a precise estimator of β_1 from (13.19), unless we have a rather large sample size. In theory, using a first differenced equation to estimate the return to education is a good idea, but it does not work very well with most currently available panel data sets.

7/14/99 7:25 PM Page 424

Adding several explanatory variables causes no difficulties. We begin with the unobserved effects model

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it1} + \beta_2 x_{it2} + \ldots + \beta_k x_{itk} + a_i + u_{it}, \quad \textbf{(13.20)}$$

for $t = 1$ and 2. This equation looks more complicated than it is because each explanatory variable has three subscripts. The first denotes the cross-sectional observation number, the second denotes the time period, and the third is just a variable label.

EXAMPLE 13.5
(Sleeping Versus Working)

We use the two years of panel data in SLP75_81.RAW, from Biddle and Hamermesh (1990), to estimate the tradeoff between sleeping and working. In Problem 3.3, we used just the 1975 cross section. The panel data set for 1975 and 1981 has 239 people, which is much smaller than the 1975 cross section that includes over 700 people. An unobserved effects model for total minutes of sleeping per week is

$$slpnap_{it} = \beta_0 + \delta_0 d81_t + \beta_1 totwrk_{it} + \beta_2 educ_{it} + \beta_3 marr_{it} + \beta_4 yngkid_{it} + \beta_5 gdhlth_{it} + a_i + u_{it}, \quad t = 1,2.$$

The unobserved effect, a_i, would be called an *unobserved individual effect* or an *individual fixed effect*. It is potentially important to allow a_i to be correlated with $totwrk_{it}$: the same factors (some biological) that cause people to sleep more or less (captured in a_i) are likely correlated with the amount of time spent working. Some people just have more energy, and this causes them to sleep less and work more. The variable *educ* is years of education, *marr* is a marriage dummy variable, *yngkid* is a dummy variable indicating the presence of a small child, and *gdhlth* is a "good health" dummy variable. Notice that we do not include gender or race (as we did in the cross-sectional analysis), since these do not change over time; they are part of a_i. Our primary interest is in β_1.

Differencing across the two years gives the estimable equation

$$\Delta slpnap_i = \delta_0 + \beta_1 \Delta totwrk_i + \beta_2 \Delta educ_i + \beta_3 \Delta marr_i + \beta_4 \Delta yngkid_i + \beta_5 \Delta gdhlth_i + \Delta u_i.$$

Assuming that the change in the idiosyncratic error, Δu_i, is uncorrelated with the changes in all explanatory variables, we can get consistent estimators using OLS. This gives

$$\begin{aligned} \widehat{\Delta slpnap} &= \underset{}{-92.63} - \underset{(45.87)}{.227}\,\Delta totwrk - \underset{(48.759)}{.024}\,\Delta educ \\ &+ \underset{(92.86)}{104.21}\,\Delta marr + \underset{(87.65)}{94.67}\,\Delta yngkid + \underset{(76.60)}{87.58}\,\Delta gdhlth \\ n &= 239,\ R^2 = .150. \end{aligned} \quad \textbf{(13.21)}$$

The coefficient on $\Delta totwrk$ indicates a tradeoff between sleeping and working: holding other factors fixed, one more hour of work is associated with .227(60) = 13.62 less minutes of sleeping. The t statistic (−6.31) is very significant. No other estimates, except the

7/14/99 7:25 PM Page 425

intercept, are statistically different from zero. The *F* test for joint significance of all variables except $\Delta totwrk$ gives *p*-value = .49, which means they are jointly insignificant at any reasonable significance level and could be dropped from the equation.

The standard error on $\Delta educ$ is especially large relative to the estimate. This is the phenomenon described earlier for the wage equation. In the sample of 239 people, 183 (76.6%) have no change in education over the six-year period; 90% of the people have a change in education of at most one year. As reflected by the extremely large standard error of $\hat{\beta}_2$, there is not nearly enough variation in education to estimate β_2 with any precision. Anyway, $\hat{\beta}_2$ is practically very small.

Panel data can also be used to estimate finite distributed lag models. Even if we specify the equation for only two years, we need to collect more years of data to obtain the lagged explanatory variables. The following is a simple example.

EXAMPLE 13.6
(Distributed Lag of Crime Rate on Clear-up Rate)

Eide (1994) uses panel data from police districts in Norway to estimate a distributed lag model for crime rates. The single explanatory variable is the "clear-up percentage" (*clrprc*)—the percentage of crimes that led to a conviction. The crime rate data are from the years 1972 and 1978. Following Eide, we lag *clrprc* for one and two years: it is likely that past clear-up rates have a deterrent effect on current crime. This leads to the following unobserved effects model for the two years:

$$\log(crime_{it}) = \beta_0 + \delta_0 d78_t + \beta_1 clrprc_{i,t-1} + \beta_2 clrprc_{i,t-2} + a_i + u_{it}.$$

When we difference the equation and estimate it using the data in CRIME3.RAW, we get

$$\begin{aligned} \Delta\widehat{\log(crime)} &= .086 - .0040\ \Delta clrprc_{-1} - .0132\ \Delta clrprc_{-2} \\ &\quad (.064)\quad (.0047)\qquad\quad (.0052) \\ n &= 53,\ R^2 = .193,\ \bar{R}^2 = .161. \end{aligned} \tag{13.22}$$

The second lag is negative and statistically significant, which implies that a higher clear-up percentage two years ago would deter crime this year. In particular, a 10 percentage point increase in *clrprc* two years ago would lead to an estimated 13.2% drop in the crime rate this year. This suggests that using more resources for solving crimes and obtaining convictions can reduce crime in the future.

Organizing Panel Data

In using panel data in an econometric study, it is important to know how the data should be stored. We must be careful to arrange the data so that the different time periods for the same cross-sectional unit (person, firm, city, and so on) are easily linked. For concreteness, suppose that the data set is on cities for two different years. For most pur-

7/14/99 7:25 PM Page 426

poses, the best way to enter the data is to have *two* records for each city, one for each year: the first record for each city corresponds to the early year, and the second record is for the later year. These two records should be adjacent. Therefore, a data set for 100 cities and two years will contain 200 records. The first two records are for the first city in the sample, the next two records are for the second city, and so on. (See Table 1.5 in Chapter 1 for an example.) This makes it easy to construct the differences to store these in the second record for each city, and to do a pooled cross-sectional analysis, which can be compared with the differencing estimation.

Most of the two-period panel data sets accompanying this text are stored in this way (for example, CRIME2.RAW, CRIME3.RAW, GPA3.RAW, LOWBRTH.RAW, and RENTAL.RAW). We use a direct extension of this scheme for panel data sets with more than two time periods.

A second way of organizing two periods of panel data is to have only one record per cross-sectional unit. This requires two entries for each variable, one for each time period. The panel data in SLP75_81.RAW are organized in this way. Each individual has data on the variables *slpnap75*, *slpnap81*, *totwrk75*, *totwrk81*, and so on. Creating the differences from 1975 to 1981 is easy. Other panel data sets with this structure are TRAFFIC1.RAW and VOTE2.RAW. A drawback to putting the data in one record is that it does not allow a pooled OLS analysis using the two time periods on the original data. Also, this organizational method does not work for panel data sets with more than two time periods, a case we will consider in Section 13.5.

13.4 POLICY ANALYSIS WITH TWO-PERIOD PANEL DATA

Panel data sets are very useful for policy analysis and, in particular, progam evaluation. In the simplest program evaluation setup, a sample of individuals, firms, or cities, and so on, is obtained in the first time period. Some of these units then take part in a particular program in a later time period; the ones that do not are the control group. This is similar to the natural experiment literature discussed earlier, with one important difference: the *same* cross-sectional units appear in each time period.

As an example, suppose we wish to evaluate the effect of a Michigan job training program on worker productivity of manufacturing firms (see also Problem 9.8). Let $scrap_{it}$ denote the scrap rate of firm i during year t (the number of items, per 100, that must be scrapped due to defects). Let $grant_{it}$ be a binary indicator equal to one if firm i in year t received a job training grant. For the years 1987 and 1988, the model is

$$scrap_{it} = \beta_0 + \delta_0 y88_t + \beta_1 grant_{it} + a_i + u_{it}, \quad t = 1,2, \tag{13.23}$$

where $y88_t$ is a dummy variable for 1988 and a_i is the *unobserved firm effect* or the *firm fixed effect*. The unobserved effect contains things such as average employee ability, capital, and managerial skill; these are roughly constant over a two-year period. We are concerned about a_i being systematically related to whether a firm receives a grant. For example, administrators of the program might give priority to firms whose workers have lower skills. Or, the opposite problem could occur: in order to make the job train-

7/14/99 7:25 PM Page 427

ing program appear effective, administrators may give the grants to employers with more productive workers. Actually, in this particular program, grants were awarded on a first-come, first-serve basis. But whether a firm applied early for a grant could be correlated with worker productivity. In any case, an analysis using a single cross section or just a pooling of the cross sections will produce biased and inconsistent estimators.

Differencing to remove a_i gives

$$\Delta scrap_i = \delta_0 + \beta_1 \Delta grant_i + \Delta u_i. \tag{13.24}$$

Therefore, we simply regress the change in the scrap rate on the change in the grant indicator. Because no firms received grants in 1987, $grant_{i1} = 0$ for all i, and so $\Delta grant_i = grant_{i2} - grant_{i1} = grant_{i2}$, which simply indicates whether the firm received a grant in 1988. However, it is generally important to difference all variables (dummy variables included) because this is necessary for removing a_i in the unobserved effects model (13.23).

Estimating the first-differenced equation using the data in JTRAIN.RAW gives

$$\widehat{\Delta scrap} = -.564 - .739\ \Delta grant$$
$$(.405)\quad (.683)$$
$$n = 54,\ R^2 = .022.$$

Therefore, we estimate that having a job training grant lowered the scrap rate on average by $-.739$. But the estimate is not statistically different from zero.

We get stronger results by using log(*scrap*) and estimating the percentage effect:

$$\widehat{\Delta \log(scrap)} = -.057 - .317\ \Delta grant$$
$$(.097)\quad (.164)$$
$$n = 54,\ R^2 = .067.$$

Having a job training grant is estimated to lower the scrap rate by about 27.2% [because $\exp(-.317) - 1 \approx -.272$]. The t statistic is about -1.93, which is marginally significant. By contrast, using pooled OLS of log(*scrap*) on *y88* and *grant* gives $\hat{\beta}_1 = .057$ (standard error = .431). Thus, we find no significant relationship between the scrap rate and the job training grant. Since this differs so much from the first-difference estimates, it suggests that firms that have lower ability workers are more likely to receive a grant.

It is useful to study the program evaluation model more generally. Let y_{it} denote an outcome variable and let $prog_{it}$ be a program participation dummy variable. The simplest unobserved effects model is

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 prog_{it} + a_i + u_{it}. \tag{13.25}$$

If program participation only occurred in the second period, then the OLS estimator of β_1 in the differenced equation has a very simple representation:

$$\hat{\beta}_1 = \overline{\Delta y}_{treat} - \overline{\Delta y}_{control}. \tag{13.26}$$

7/14/99 7:25 PM Page 428

That is, we compute the average change in y over the two time periods for the treatment and control groups. Then, $\hat{\beta}_1$ is the difference of these. This is the panel data version of the difference-in-differences estimator in equation (13.11) for two pooled cross sections. With panel data, we have a potentially important advantage: we can difference y across time for the *same* cross-sectional units. This allows us to control for person, firm, or city specific effects, as the model in (13.25) makes clear.

If program participation takes place in both periods, $\hat{\beta}_1$ cannot be written as in (13.26), but we interpret it in the same way: it is the change in the average value of y due to program participation.

Controlling for time-varying factors does not change anything of significance. We simply difference those variables and include them along with $\Delta prog$. This allows us to control for time-varying variables that might be correlated with program designation.

The same differencing method works for analyzing the effects of any policy that varies across city or state. The following is a simple example.

EXAMPLE 13.7

(Effect of Drunk Driving Laws on Traffic Fatalities)

Many states in the United States have adopted different policies in an attempt to curb drunk driving. Two types of laws that we will study here are *open container laws*—which make it illegal for passengers to have open containers of alcoholic beverages—and *administrative per se laws*—which allow courts to suspend licenses after a driver is arrested for drunk driving but before the driver is convicted. One possible analysis is to use a single cross section of states to regress driving fatalities (or those related to drunk driving) on dummy variable indicators for whether each law is present. This is unlikely to work well because states decide, through legislative processes, whether they need such laws. Therefore, the presence of laws is likely to be related to the average drunk driving fatalities in recent years. A more convincing analysis uses panel data over a time period where some states adopted new laws (and some states may have repealed existing laws). The file TRAFFIC1.RAW contains data for 1985 and 1990 for all 50 states and the District of Columbia. The dependent variable is the number of traffic deaths per 100 million miles driven (*dthrte*). In 1985, 19 states had open container laws, while 22 states had such laws in 1990. In 1985, 21 states had per se laws; the number had grown to 29 by 1990.

Using OLS after first differencing gives

$$\widehat{\Delta dthrte} = -.497 - .420\,\Delta open - .151\,\Delta admn \quad \textbf{(13.27)}$$
$$(.052)\quad (.206)\qquad\qquad (.117)$$
$$n = 51,\ R^2 = .119.$$

The estimates suggest that adopting an open container law lowered the traffic fatality rate by .42, a nontrivial effect given that the average death rate in 1985 was 2.7 with a standard deviation of about .6. The estimate is statistically significant at the 5% level against a two-sided alternative. The administrative per se law has a smaller effect, and its t statistic is only -1.29; but the estimate is the sign we expect. The intercept in this equation shows that traffic fatalities fell substantially for all states over the five-year period, whether or not

7/14/99 7:25 PM Page 429

there were any law changes. The states that adopted an open container law over this period saw a further drop, on average, in fatality rates.

QUESTION 13.4

In Example 13.7, $\Delta admn = -1$ for the state of Washington. Explain what this means.

Other laws might also affect traffic fatalities, such as seat belt laws, motorcycle helmet laws, and maximum speed limits. In addition, we might want to control for age and gender distributions, as well as measures of how influential an organization such as Mothers Against Drunk Driving is in each state.

13.5 DIFFERENCING WITH MORE THAN TWO TIME PERIODS

We can also use differencing with more than two time periods. For illustration, suppose we have N individuals and $T = 3$ time periods for each individual. A general fixed effects model is

$$y_{it} = \delta_1 + \delta_2 d2_t + \delta_3 d3_t + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it}, \quad \textbf{(13.28)}$$

for $t = 1$, 2, and 3. (The total number of observations is therefore $3N$.) Notice that we now include two time period dummies in addition to the intercept. It is a good idea to allow a separate intercept for each time period, especially when we have a small number of them. The base period, as always, is $t = 1$. The intercept for the second time period is $\delta_1 + \delta_2$, and so on. We are primarily interested in $\beta_1, \beta_2, \ldots, \beta_k$. If the unobserved effect a_i is correlated with any of the explanatory variables, then using pooled OLS on the three years of data results in biased and inconsistent estimates.

The key assumption is that the idiosyncratic errors are uncorrelated with the explanatory variable in each time period:

$$\text{Cov}(x_{itj}, u_{is}) = 0, \text{ for all } t, s, \text{ and } j. \quad \textbf{(13.29)}$$

That is, the explanatory variables are *strictly exogenous* after we take out the unobserved effect, a_i. (The strict exogeneity assumption stated in terms of a zero conditional expectation is given in the chapter appendix.) Assumption (13.29) rules out cases where future explanatory variables react to current changes in the idiosyncratic errors, as must be the case if x_{itj} is a lagged dependent variable. If we have omitted an important time-varying variable, then (13.29) is generally violated. Measurement error in one or more explanatory variables can cause (13.29) to be false, just as in Chapter 9. In Chapters 15 and 16, we will discuss what can be done in such cases.

If a_i is correlated with x_{itj}, then x_{itj} will be correlated with the *composite* error, $v_{it} = a_i + u_{it}$, under (13.29). We can eliminate a_i by differencing adjacent periods. In the $T = 3$ case, we subtract time period one from time period two and time period two from time period three. This gives

$$\Delta y_{it} = \delta_2 \Delta d2_t + \delta_3 \Delta d3_t + \beta_1 \Delta x_{it1} + \ldots + \beta_k \Delta x_{itk} + \Delta u_{it}, \quad \textbf{(13.30)}$$

7/14/99 7:25 PM Page 430

for $t = 2$ and 3. We do not have a differenced equation for $t = 1$ because there is nothing to subtract from the $t = 1$ equation. Now, (13.30) represents *two* time periods for each individual in the sample. If this equation satisfies the classical linear model assumptions, then pooled OLS gives unbiased estimators, and the usual t and F statistics are valid for hypothesis. We can also appeal to asymptotic results. The important requirement for OLS to be consistent is that Δu_{it} is uncorrelated with Δx_{itj} for all j and $t = 2$ and 3. This is the natural extension from the two time period case.

Notice how (13.30) contains the differences in the year dummies, $d2_t$ and $d3_t$. For $t = 2$, $\Delta d2_t = 1$ and $\Delta d3_t = 0$; for $t = 3$, $\Delta d2_t = -1$ and $\Delta d3_t = 1$. Therefore, (13.30) does not contain an intercept. This is inconvenient for certain purposes, including the computation of R-squared. Unless the time intercepts in the original model (13.28) are of direct interest—they rarely are—it is better to estimate the first-differenced equation with an intercept and a single time period dummy, usually for the third period. In other words, the equation becomes

$$\Delta y_{it} = \alpha_0 + \alpha_3 d3_t + \beta_1 \Delta x_{it1} + \ldots + \beta_k \Delta x_{itk} + \Delta u_{it}, \quad \text{for } t = 2 \text{ and } 3.$$

The estimates of the β_j are identical in either formulation.

With more than three time periods, things are similar. If we have the same T time periods for each of N cross-sectional units, we say that the data set is a **balanced panel**: we have the same time periods for all individuals, firms, cities, and so on. When T is small relative to N, we should include a dummy variable for each time period to account for secular changes that are not being modeled. Therefore, after first differencing, the equation looks like

$$\Delta y_{it} = \alpha_0 + \alpha_3 d3_t + \alpha_4 d4_t + \ldots + \alpha_T dT_t + \beta_1 \Delta x_{it1} + \ldots + \beta_k \Delta x_{itk} + \Delta u_{it}, \quad t = 2,3,\ldots,T, \tag{13.31}$$

where we have $T - 1$ time periods on each unit i for the first-differenced equation. The total number of observations is $N(T - 1)$.

It is simple to estimate (13.31) by pooled OLS, provided the observations have been properly organized and the differencing carefully done. To facilitate first differencing, the data file should consist of NT records. The first T records are for the first cross-sectional observation, arranged chronologically; the second T records are for the second cross-sectional observations, arranged chronologically; and so on. Then, we compute the differences, with the change from $t - 1$ to t stored in the time t record. Therefore, the differences for $t = 1$ should be missing values for all N cross-sectional observations. Without doing this, you run the risk of using bogus observations in the regression analysis. An invalid observation is created when the last observation for, say, person $i - 1$ is substracted from the first observation for person i. If you do the regression on the differenced data, and NT or $NT - 1$ observations are reported, then you forgot to set the $t = 1$ observations as missing.

When using more than two time periods, we must assume that Δu_{it} is uncorrelated over time for the usual standard errors and test statistics to be valid. This assumption is sometimes reasonable, but it does not follow if we assume that the original idiosyncratic errors, u_{it}, are uncorrelated over time (an assumption we will use in Chapter 14). In fact, if we assume the u_{it} are serially uncorrelated with constant variance, then the correla-

7/14/99 7:25 PM Page 431

tion between Δu_{it} and $\Delta u_{i,t+1}$ can be shown to be $-.5$. If u_{it} follows a stable AR(1) model, then Δu_{it} will be serially correlated. Only when u_{it} follows a random walk will Δu_{it} be serially uncorrelated.

It is easy to test for serial correlation in the first-differenced equation. Let $r_{it} = \Delta u_{it}$ denote the first difference of the original error. If r_{it} follows the AR(1) model $r_{it} = \rho r_{i,t-1} + e_{it}$, then we can easily test H_0: $\rho = 0$. First, we estimate (13.31) by pooled OLS and obtain the residuals, $\hat{r}_{it}$. Then, we run the regression again with $\hat{r}_{i,t-1}$ as an additional explanatory variable. The coefficient on $\hat{r}_{i,t-1}$ is an estimate of ρ, and so we can use the usual t statistic on $\hat{r}_{i,t-1}$ to test H_0: $\rho = 0$. Because we are using the lagged OLS residual, we lose another time period. For example, if we originally had $T = 3$, the differenced equation has $T = 2$. The test for serial correlation is just a cross-sectional regression on first differences, using the third time period, with the lagged OLS residual included. This is similar to the test we covered in Section 12.2 for pure time series models. We give an example later.

We can correct for the presence of AR(1) serial correlation by quasi-differencing equation (13.31). [We can also use the Prais-Winsten transformation for the first time period in (13.31).] Unfortunately, standard packages that perform AR(1) corrections for time series regressions will not work. Standard Cochrane-Orcutt or Prais-Winsten methods will treat the observations as if they followed an AR(1) process across i and t; this makes no sense, as we are assuming the observations are independent across i. Corrections to the OLS standard errors that allow arbitrary forms of serial correlation (and heteroskedasticity) can be computed when N is large (and N should be notably larger than T). A detailed treatment of these topics is beyond the scope of this text [see Wooldridge (1999, Chapter 10)], but they are easy to compute in certain regression packages.

QUESTION 13.5

Does serial correlation in Δu_{it} cause the first-differenced estimator to be biased and inconsistent? Why is serial correlation a concern?

If there is no serial correlation in the errors, the usual methods for dealing with heteroskedasticity are valid. We can use the Breusch-Pagan and White tests for heteroskedasticity from Chapter 8, and we can also compute robust standard errors.

Differencing more than two years of panel data is very useful for policy analysis, as shown by the following example.

EXAMPLE 13.8

(Effect of Enterprise Zones on Unemployment Claims)

Papke (1994) studied the effect of the Indiana enterprise zone (EZ) program on unemployment claims. She analyzed 22 cities in Indiana over the period from 1980 to 1988. Six enterprise zones were designated in 1984, and four more were assigned in 1985. Twelve of the cities in the sample did not receive an enterprise zone over this period; they served as the control group.

A simple policy evaluation model is

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

7/14/99 7:25 PM Page 432

where $uclms_{it}$ is the number of unemployment claims filed during year t in city i. The parameter θ_t just denotes a different intercept for each time period. Generally, unemployment claims were falling statewide over this period, and this should be reflected in the different year intercepts. The binary variable ez_{it} is equal to one if city i at time t was an enterprise zone; we are interested in β_1. The unobserved effect a_i represents fixed factors that affect the economic climate in city i. Because enterprise zone designation was not determined randomly—enterprise zones are usually economically depressed areas—it is likely that ez_{it} and a_i are positively correlated (high a_i means higher unemployment claims, which lead to a higher chance of being given an EZ). Thus, we should difference the equation to eliminate a_i:

$$\Delta \log(uclms_{it}) = \alpha_0 + \alpha_1 d82_t + \ldots + \alpha_7 d88_t + \beta_1 \Delta ez_{it} + \Delta u_{it}. \quad \textbf{(13.32)}$$

The dependent variable in this equation, the change in $\log(uclms_{it})$, is the approximate annual growth rate in unemployment claims from year $t-1$ to t. We can estimate this equation for the years 1981 to 1988 using the data in EZUNEM.RAW; the total sample size is $22 \cdot 8 = 176$. The estimate of β_1 is $\hat{\beta}_1 = -.182$ (standard error = .078). Therefore, it appears that the presence of an EZ causes about a 16.6% [$\exp(-.182) - 1 \approx -.166$] fall in unemployment claims. This is an economically large and statistically significant effect.

There is no evidence of heteroskedasticity in the equation: the Breusch-Pagan F test yields $F = .85$, p-value = .557. However, when we add the lagged OLS residuals to the differenced equation (and lose the year 1981), we get $\hat{\rho} = -.197$ ($t = -2.44$), so there is evidence of minimal negative serial correlation in the first-differenced errors. Unlike with positive serial correlation, the usual OLS standard errors may not greatly understate the correct standard errors when the errors are negatively correlated (see Section 12.1). Thus, the significance of the enterprise zone dummy variable will probably not be affected.

EXAMPLE 13.9
(County Crime Rates in North Carolina)

Cornwell and Trumbull (1994) used data on 90 counties in North Carolina, for the years 1981 through 1987, to estimate an unobserved effects model of crime; the data are contained in CRIME4.RAW. Here, we estimate a simpler version of their model, and we difference the equation over time to eliminate a_i, the unobserved effect. (Cornwell and Trumbull use a different transformation, which we will cover in Chapter 14.) Various factors including geographical location, attitudes toward crime, historical records, and reporting conventions might be contained in a_i. The crime rate is number of crimes per person, *prbarr* is the estimated probability of arrest, *prbconv* is the estimated probability of conviction (given an arrest), *prbpris* is the probability of serving time in prison (given a conviction), *avgsen* is the average sentence length served, and *polpc* is the number of police officers per capita. As is standard in criminometric studies, we use the logs of all variables in order to estimate elasticities. We also include a full set of year dummies to control for state trends in crime rates. We can use the years 1982 through 1987 to estimate the differenced equation. The quan-

7/14/99 7:25 PM Page 433

tities in parentheses are the usual OLS standard errors; the quantities in brackets are standard errors robust to both serial correlation and heteroskedasticity:

$$\begin{aligned}\widehat{\Delta\log(crmrte)} &= .008 - .100\,d83 - .048\,d84 - .005\,d85 \\ &\qquad (.017)\ (.024)\ (.024)\ (.023) \\ &\qquad [.014]\ [.022]\ [.020]\ [.025] \\ &\quad + .028\,d86 + .041\,d87 - .327\,\Delta\log(prbarr) \\ &\qquad (.024)\ (.024)\ (.030) \\ &\qquad [.021]\ [.024]\ [.056] \\ &\quad - .238\,\Delta\log(prbconv) - .165\,\Delta\log(prbpris) \\ &\qquad (.018)\ (.026) \\ &\qquad [.039]\ [.045] \\ &\quad - .022\,\Delta\log(avgsen) + .398\,\Delta\log(polpc) \\ &\qquad (.022)\ (.027) \\ &\qquad [.025]\ [.101] \\ &\qquad n = 540,\ R^2 = .433,\ \bar{R}^2 = .422.\end{aligned} \tag{13.33}$$

The three probability variables—of arrest, conviction, and serving prison time—all have the expected sign, and all are statistically significant. For example, a 1% increase in the probability of arrest is predicted to lower the crime rate by about .33%. The average sentence variable shows a modest deterrent effect, but it is not statistically significant.

The coefficient on the police per capita variable is somewhat surprising and is a feature of most studies that seek to explain crime rates. Interpreted causally, it says that a 1% increase in police per capita *increases* crime rates by about .4%. (The usual t statistic is very large, almost 15.) It is hard to believe that having more police officers causes more crime. What is going on here? There are at least two possibilities. First, the crime rate variable is calculated from *reported* crimes. It might be that, when there are additional police, more crimes are reported. The police variable might be endogenous in the equation for other reasons: counties may enlarge the police force when they expect crime rates to increase. In this case, (13.33) cannot be interpreted in a causal fashion. In Chapters 15 and 16, we will cover models and estimation methods that can account for this additional form of endogeneity.

The special case of the White test for heteroskedasticity in Section 8.3 gives $F = 75.48$ and p-value $= .0000$, so there is strong evidence of heteroskedasticity. (Technically, this test is not valid if there is also serial correlation, but it is strongly suggestive.) Testing for AR(1) serial correlation yields $\hat{\rho} = -.233$, $t = -4.77$, so negative serial correlation exists. The standard errors in brackets adjust for serial correlation and heteroskedasticity. [We will not give the details of this; the calculations are similar to those described in Section 12.5 and are carried out by many econometric packages. See Wooldridge (1999, Chapter 10) for more discussion.] No variables lose statistical significance, but the t statistics on the significant deterrent variables get notably smaller. For example, the t statistic on the probability of conviction variable goes from -13.22 using the usual OLS standard error to -6.10 using the fully robust standard error. Equivalently, the confidence intervals constructed using the robust standard errors will, appropriately, be much wider than those based on the usual OLS standard errors.

7/14/99 7:25 PM Page 434

SUMMARY

We have studied methods for analyzing independently pooled cross-sectional and panel data sets. Independent cross sections arise when different random samples are obtained in different time periods (usually years). OLS using pooled data is the leading method of estimation, and the usual inference procedures are available, including corrections for heteroskedasticity. (Serial correlation is not an issue because the samples are independent across time.) Because of the time series dimension, we often allow different time intercepts. We might also interact time dummies with certain key variables to see how they have changed over time. This is especially important in the policy evaluation literature for natural experiments.

Panel data sets are being used more and more in applied work, especially for policy analysis. These are data sets where the same cross-sectional units are followed over time. Panel data sets are most useful when controlling for time-constant unobserved features—of people, firms, cities, and so on—which we think might be correlated with the explanatory variables in our model. One way to remove the unobserved effect is to difference the data in adjacent time periods. Then, a standard OLS analysis on the differences can be used. Using two periods of data results in a cross-sectional regression of the differenced data. The usual inference procedures are asymptotically valid under homoskedasticity; exact inference is available under normality.

For more than two time periods, we can use pooled OLS on the differenced data; we lose the first time period because of the differencing. In addition to homoskedasticity, we must assume that the *differenced* errors are serially uncorrelated in order to apply the usual t and F statistics. (The chapter appendix contains a careful listing of the assumptions.) Naturally, any variable that is constant over time drops out of the analysis.

KEY TERMS

Balanced Panel
Composite Error
Difference-in-Differences Estimator
First Differenced Equation
First-Differenced Estimator
Fixed Effect
Fixed Effects Model
Heterogeneity Bias
Idiosyncratic Error
Independently Pooled Cross Section
Longitudinal Data
Natural Experiment
Panel Data
Quasi-Experiment
Strict Exogeneity
Unobserved Effect
Unobserved Effects Model
Unobserved Heterogeneity
Year Dummy Variables

PROBLEMS

13.1 In Example 13.1, assume that the average of all factors other than *educ* have remained constant over time and that the average level of education is 12.2 for the 1972 sample and 13.3 in the 1984 sample. Using the estimates in Table 13.1, find the estimated change in average fertility between 1972 and 1984. (Be sure to account for the intercept change and the change in average education.)

7/14/99 7:25 PM Page 435

13.2 Using the data in KIELMC.RAW, the following equations were estimated using the years 1978 and 1981:

$$\widehat{\log(price)} = \underset{(0.26)}{11.49} - \underset{(.058)}{.547}\, nearinc + \underset{(.080)}{.394}\, y81{\cdot}nearinc$$
$$n = 321,\ R^2 = .220$$

and

$$\widehat{\log(price)} = \underset{(0.27)}{11.18} + \underset{(.044)}{.563}\, y81 - \underset{(.067)}{.403}\, y81{\cdot}nearinc$$
$$n = 321,\ R^2 = .337.$$

Compare the estimates on the interaction term *y81·nearinc* with those from equation (13.9). Why are the estimates so different?

13.3 Why can we not use first differences when we have independent cross sections in two years (as opposed to panel data)?

13.4 If we think that β_1 is positive in (13.14) and that Δu_i and $\Delta unem_i$ are negatively correlated, what is the bias in the OLS estimator of β_1 in the first-differenced equation? (*Hint*: Review Table 3.2.)

13.5 Suppose that we want to estimate the effect of several variables on annual saving and that we have a panel data set on individuals collected on January 31, 1990 and January 31, 1992. If we include a year dummy for 1992 and use first differencing, can we also include age in the original model? Explain.

13.6 In 1985, neither Florida nor Georgia had laws banning open alcohol containers in vehicle passenger compartments. By 1990, Florida had passed such a law, but Georgia had not.

(i) Suppose you can collect random samples of the driving-age population in both states, for 1985 and 1990. Let *arrest* be a binary variable equal to unity if a person was arrested for drunk driving during the year. Without controlling for any other factors, write down a linear probability model that allows you to test whether the open container law reduced the probability of being arrested for drunk driving. Which coefficient in your model measures the effect of the law?

(ii) Why might you want to control for other factors in the model? What might some of these factors be?

COMPUTER EXERCISES

13.7 Use the data in FERTIL1.RAW for this exercise.

(i) In the equation estimated in Example 13.1, test whether living environment at age 16 has an effect on fertility. (The base group is large city.) Report the value of the *F* statistic and the *p*-value.

(ii) Test whether region of the country at age 16 (south is the base group) has an effect on fertility.

7/14/99 7:25 PM Page 436

(iii) Let u be the error term in the population equation. Suppose you think that the variance of u changes over time (but not with *educ*, *age*, and so on). A model that captures this is

$$u^2 = \gamma_0 + \gamma_1 y74 + \gamma_2 y76 + \ldots + \gamma_6 y84 + v.$$

Using this model, test for heteroskedasticity in u. [*Hint*: Your F test should have 6 and 1122 degrees of freedom.]

(iv) Add the interaction terms $y74 \cdot educ$, $y76 \cdot educ, \ldots, y84 \cdot educ$ to the model estimated in Table 13.1. Explain what these terms represent. Are they jointly significant?

13.8 Use the data in CPS78_85.RAW for this exercise.

(i) How do you interpret the coefficient on *y85* in equation (13.2)? Does it have an interesting interpretation? (Be careful here; you must account for the interaction terms $y85 \cdot educ$ and $y85 \cdot female$.)

(ii) Holding other factors fixed, what is the estimated percent increase in nominal wage for a male with twelve years of education? Propose a regression to obtain a confidence interval for this estimate. [*Hint*: To get the confidence interval, replace $y85 \cdot educ$ with $y85 \cdot (educ - 12)$; refer to Example 6.3.]

(iii) Reestimate equation (13.2) but let all wages be measured in 1978 dollars. In particular, define the real wage as $rwage = wage$ for 1978 and as $rwage = wage/1.65$ for 1985. Now use $\log(rwage)$ in place of $\log(wage)$ in estimating (13.2). Which coefficients differ from those in equation (13.2)?

(iv) Explain why the R-squared from your regression in part (iii) is not the same as in equation (13.2). (*Hint*: The residuals, and therefore the sum of squared residuals, from the two regressions *are* identical.)

(v) Describe how union participation has changed from 1978 to 1985.

(vi) Starting with equation (13.2), test whether the union wage differential has changed over time. (This should be a simple t test.)

(vii) Do your findings in parts (v) and (vi) conflict? Explain.

13.9 Use the data in KIELMC.RAW for this exercise.

(i) The variable *dist* is the distance from each home to the incinerator site, in feet. Consider the model

$$\log(price) = \beta_0 + \delta_0 y81 + \beta_1 \log(dist) + \delta_1 y81 \cdot \log(dist) + u.$$

If building the incinerator reduces the value of homes closer to the site, what is the sign of δ_1? What does it mean if $\beta_1 > 0$?

(ii) Estimate the model from part (i) and report the results in the usual form. Interpret the coefficient on $y81 \cdot \log(dist)$. What do you conclude?

(iii) Add *age*, age^2, *rooms*, *baths*, $\log(intst)$, $\log(land)$, and $\log(area)$ to the equation. Now what do you conclude about the effect of the incinerator on housing values?

13.10 Use the data in INJURY.RAW for this exercise.

7/14/99 7:25 PM Page 437

(i) Using the data for Kentucky, reestimate equation (13.12) adding as explanatory variables *male*, *married*, and a full set of industry and injury type dummy variables. How does the estimate on *afchnge·highearn* change when these other factors are controlled for? Is the estimate still statistically significant?

(ii) What do you make of the small *R*-squared from part (i)? Does this mean the equation is useless?

(iii) Estimate equation (13.12) using the data for Michigan. Compare the estimates on the interaction term for Michigan and Kentucky. Is the Michigan estimate statistically significant? What do you make of this?

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

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

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

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

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

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

(iv) Obtain the heteroskedasticity-robust standard errors for the first-differenced equation in part (iii). Does this change your conclusions?

13.12 Use CRIME3.RAW for this exercise.

(i) In the model of Example 13.6, test the hypothesis H_0: $\beta_1 = \beta_2$. (*Hint*: Define $\theta_1 = \beta_1 - \beta_2$ and write β_1 in terms of θ_1 and β_2. Substitute this into the equation and then rearrange. Do a *t* test on θ_1.)

(ii) If $\beta_1 = \beta_2$, show that the differenced equation can be written as

$$\Delta\log(crime_i) = \delta_0 + \delta_1 \Delta avgclr_i + \Delta u_i,$$

where $\delta_1 = 2\beta_1$ and $avgclr_i = (clrprc_{i,-1} + clrprc_{i,-2})/2$ is the average clear-up percentage over the previous two years.

(iii) Estimate the equation from part (ii). Compare the adjusted *R*-squared with that in (13.22). Which model would you finally use?

13.13 Use GPA3.RAW for this exercise. The data set is for 366 student athletes from a large university for fall and spring semesters. (A similar analysis is in Maloney and McCormick (1993), but here we use a true panel data set). Because you have two terms of data for each student, an unobserved effects model is appropriate. The primary question of interest is this: Do athletes perform more poorly in school during the semester their sport is in season?

7/14/99 7:25 PM Page 438

(i) Use pooled OLS to estimate a model with term GPA (*trmgpa*) as the dependent variable. The explanatory variables are *spring*, *sat*, *hsperc*, *female*, *black*, *white*, *frstsem*, *tothrs*, *crsgpa*, and *season*. Interpret the coefficient on *season*. Is it statistically significant?

(ii) Most of the athletes who play their sport only in the fall are football players. Suppose the ability levels of football players differ systematically from those of other athletes. If ability is not adequately captured by SAT score and high school percentile, explain why the pooled OLS estimators will be biased.

(iii) Now use the data differenced across the two terms. Which variables drop out? Now test for an in-season effect.

(iv) Can you think of one or more potentially important, time-varying variables that have been omitted from the analysis?

13.14 VOTE2.RAW includes panel data on House of Representative elections in 1988 and 1990. Only winners from 1988 who are also running in 1990 appear in the sample; these are the incumbents. An unobserved effects model explaining the share of the incumbent's vote in terms of expenditures by both candidates is

$$vote_{it} = \beta_0 + \delta_0 d90_t + \beta_1 \log(inexp_{it}) + \beta_2 \log(chexp_{it}) + \beta_3 incshr_{it} + a_i + u_{it},$$

where $incshr_{it}$ is the incumbent's share of total campaign spending (in percent form). The unobserved effect a_i contains characteristics of the incumbent—such as "quality"—as well as things about the district that are constant. The incumbent's gender and party are constant over time, so these are subsumed in a_i. We are interested in the effect of campaign expenditures on election outcomes.

(i) Difference the given equation across the two years and estimate the differenced equation by OLS. Which variables are individually significant at the 5% level against a two-sided alternative?

(ii) In the equation from part (i), test for joint significance of $\Delta\log(inexp)$ and $\Delta\log(chexp)$. Report the p-value.

(iii) Reestimate the equation from part (i) using $\Delta incshr$ as the only independent variable. Interpret the coefficient on $\Delta incshr$. For example, if the incumenbent's share of spending increases by 10 percentage points, how is this predicted to affect the incumbent's share of the vote?

(iv) Redo part (iii), but now use only the pairs that have repeat challengers. [This allows us to control for characteristics of the challengers as well, which would be in a_i. Levitt (1995) conducts a much more extensive analysis.]

13.15 Use CRIME4.RAW for this exercise.

(i) Add the logs of each wage variable in the data set and estimate the model by first differencing. How does including these variables affect the coefficients on the criminal justice variables in Example 13.9?

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

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

7/14/99 7:25 PM Page 439

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

(i) Estimate the equation using first differencing. How many firms are used in the estimation? How many total observations would be used if each firm had data on all variables (in particular, *hrsemp*) for all three time periods?

(ii) Interpret the coefficient on *grant* and comment on its significance.

(iii) Is it surprising that $grant_{-1}$ is insignificant? Explain.

(iv) Do larger firms train their employees more or less, on average? How big are the differences in training?

APPENDIX 13A

Assumptions for Pooled OLS Using First Differences

In this appendix, we provide careful statements of the assumptions for the first-differencing estimator. Verification of these claims is somewhat involved, but it can be found in Wooldridge (1999, Chapter 10).

ASSUMPTION FD.1

For each i, the model is

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

where the β_j are the parameters to estimate and a_i is the unobserved effect.

ASSUMPTION FD.2

We have a random sample from the cross section.

For the next assumption, it is useful to let $\boldsymbol{X}_i$ denote the explanatory variables for all time periods for cross-sectional observation i; thus, $\boldsymbol{X}_i$ contains x_{itj}, $t = 1, \ldots, T$, $j = 1, \ldots, k$.

ASSUMPTION FD.3

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

When Assumption FD.3 holds, we sometimes say that the x_{itj} are *strictly exogenous conditional on the unobserved effect*. The idea is that, once we control for a_i, there is no correlation between the x_{isj} and the remaining error, u_{it}, for all s and t. An important implication of FD.3 is that $E(\Delta u_{it}|\boldsymbol{X}_i) = 0$, $t = 2, \ldots, T$.

7/14/99 7:25 PM Page 440

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

Under these first four assumptions, the first-difference estimators are unbiased. The key assumption is FD.3, which is strict exogeneity of the explanatory variables. Under these same assumptions, we can also show that the FD estimator is consistent with a fixed T and as $N \to \infty$ (and perhaps more generally).

ASSUMPTION FD.5
The variance of the differenced errors, conditional on all explanatory variables, is constant: $\text{Var}(\Delta u_{it}|\boldsymbol{X}_i) = \sigma^2$, $t = 2, \ldots, T$.

ASSUMPTION FD.6
For all $t \neq s$, the *differences* in the idiosyncratic errors are uncorrelated (conditional on all explanatory variables): $\text{Cov}(\Delta u_{it}, \Delta u_{is}|\boldsymbol{X}_i) = 0$, $t \neq s$.

Assumption FD.5 ensures that the differenced errors, Δu_{it}, are homoskedastic. Assumption FD.6 states that the differenced errors are serially uncorrelated, which means that the u_{it} follow a random walk across time (see Chapter 11). Under Assumptions FD.1 through FD.6, the FD estimator of the β_j is the best linear unbiased estimator (conditional on the explanatory variables).

ASSUMPTION FD.7
Conditional on $\boldsymbol{X}_i$, the Δu_{it} are independent and identically distributed normal random variables.

When we add Assumption FD.7, the FD estimators are normally distributed and the t and F statistics from pooled OLS on the differences have exact t and F distributions. Without FD.7, we can rely on the usual asymptotic approximations.