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 rumor 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 relative to the price of more distant houses.

For illustration, we define a house to be near the incinerator if it is within three miles. [In Computer Exercise C3, 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:

$$rprice = \gamma_0 + \gamma_1 nearinc + u, \qquad [13.3]$$

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

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

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 and far from the incinerator are 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

$$\widehat{rprice} = 82{,}517.23 - 18{,}824.37\ nearinc$$
$$(2{,}653.79) \quad (4{,}744.59) \qquad [13.5]$$
$$n = 179,\ R^2 = .082.$$

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* is

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

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

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

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

$$rprice = \beta_0 + \delta_0 y81 + \beta_1 nearinc + \delta_1 y81 \cdot nearinc + u, \qquad [13.7]$$

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) shows 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 $y81 \cdot nearinc$: 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. 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 $\approx .057$).

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 homes

TABLE 13.2 Effects of Incinerator Location on Housing Prices

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

selling near the incinerator in 1981 might have been systematically different than those selling near the incinerator in 1978; if so, it can be important to control for such characteristics. Second, even if the relevant house characteristics did not change, 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 quadratic. 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

$$\widehat{\log(price)} = 11.29 + .457\, y81 - .340\, nearinc - .063\, y81 \cdot nearinc \qquad \textbf{[13.9]}$$
$$(.31)\ (.045)\qquad (.055)\qquad (.083)$$
$$n = 321,\ R^2 = .409.$$

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 used in 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 a true experiment, in which treatment and control groups are randomly and explicitly chosen, the control and treatment groups in natural experiments arise from the particular policy change. To control for systematic differences between the control and treatment groups, we need two years of data, one before the policy change 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 *C* the control group and *T* the treatment group, letting *dT* equal unity for those in the treatment group *T*, and zero otherwise. Then, letting *d2* denote a dummy variable for the second (post–policy change) time period, the equation of interest is

$$y = \beta_0 + \delta_0 d2 + \beta_1 dT + \delta_1 d2 \cdot dT + other\ factors, \qquad [13.10]$$

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:

$$\hat{\delta}_1 = (\bar{y}_{2,T} - \bar{y}_{2,C}) - (\bar{y}_{1,T} - \bar{y}_{1,C}), \qquad [13.11]$$

where the bar denotes average, the first subscript denotes the year, and the second subscript denotes the group.

The general difference-in-differences setup is shown in Table 13.3. Table 13.3 suggests that the parameter δ_1, sometimes called the **average treatment effect** (because it measures the effect of the "treatment" or policy on the average outcome of y), can be estimated in two ways: (1) Compute the differences in averages between the treatment and control groups in each time period, and then difference the results over time; this is just as

TABLE 13.3 Illustration of the Difference-in-Differences Estimator

	Before	After	After – Before
Control	β_0	$\beta_0 + \delta_0$	δ_0
Treatment	$\beta_0 + \beta_1$	$\beta_0 + \delta_0 + \beta_1 + \delta_1$	$\delta_0 + \delta_1$
Treatment − Control	β_1	$\beta_1 + \delta_1$	δ_1