CHAPTER 7

Using Indicator Variables

LEARNING OBJECTIVES

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

1. The difference between qualitative and quantitative economic variables.
2. How to include a 0–1 indicator variable on the right-hand side of a regression, how this affects model interpretation, and give an example.
3. How to interpret the coefficient on an indicator variable in a log-linear equation.
4. How to include a slope-indicator variable in a regression, how this affects model interpretation, and give an example.
5. How to include a product of two indicator variables in a regression, and how this affects model interpretation, giving an example.
6. How to model qualitative factors with more than two categories (similar to region of the country), and how to interpret the resulting model, giving an example.
7. The consequences of ignoring a structural change in parameters during part of the sample.
8. How to test the equivalence of two regression equations using indicator variables.
9. How to estimate and interpret a regression with an indicator dependent variable.
10. The difference between a randomized controlled experiment and a natural experiment.
11. The difference between the average treatment effect (ATE) and the average treatment effect on the treated (ATT).
12. How to use a regression discontinuity design (RDD), and explain when it is useful.

KEYWORDS

annual indicator variables
average treatment effect
Chow test
dichotomous variables
difference estimator
differences-in-differences estimator
dummy variables
dummy variable trap
exact collinearity
hedonic model
indicator variable
interaction variable
intercept indicator variable
linear probability model
log-linear model
natural experiment
quasi-experiments
reference group
regional indicator variables
regression discontinuity design
seasonal indicator variables
slope-indicator variable
treatment effect

7.1 Indicator Variables

Indicator variables, which were first introduced in Section 2.9, allow us to construct models in which some or all regression model parameters, including the intercept, change for some observations in the sample. To make matters specific, let us consider an example from real estate economics. Buyers and sellers of homes, tax assessors, real estate appraisers, and mortgage bankers are interested in predicting the current market value of a house. A common way to predict the value of a house is to use a **hedonic model**, in which the price of the house is explained as a function of its characteristics, such as its size, location, number of bedrooms, and age. The idea is to break down a good into its component pieces, and then estimate the value of each characteristic.[1]

For the present, let us assume that the size of the house, measured in square feet, $SQFT$, is the only relevant variable in determining house price, $PRICE$. Specify the regression model as

$$PRICE = \beta_1 + \beta_2 SQFT + e \tag{7.1}$$

In this model, β_2 is the value of an additional square foot of living area and β_1 is the value of the land alone.

In real estate, the three most important words are "location, location, and location." How can we take into account the effect of a property's being in a desirable neighborhood, such as one near a university, or near a golf course? Thought of this way, location is a "qualitative" characteristic of a house.

Indicator variables are used to account for qualitative factors in econometric models. They are often called **dummy**, **binary**, or **dichotomous variables** because they take just two values, usually one or zero, to indicate the presence or absence of a characteristic or to indicate whether a condition is true or false. They are also called **dummy variables**, to indicate that we are creating a numeric variable for a qualitative, nonnumeric characteristic. We use the terms *indicator variable* and *dummy variable* interchangeably. Using zero and one for the values of these variables is arbitrary, but very convenient, as we will see. Generally, we define an indicator variable D as

$$D = \begin{cases} 1 & \text{if characteristic is present} \\ 0 & \text{if characteristic is not present} \end{cases} \tag{7.2}$$

Thus, for the house price model, we can define an **indicator variable**, to account for a desirable neighborhood, as

$$D = \begin{cases} 1 & \text{if property is in the desirable neighborhood} \\ 0 & \text{if property is not in the desirable neighborhood} \end{cases}$$

Indicator variables can be used to capture changes in the model intercept, or slopes, or both. We consider these possibilities in turn.

7.1.1 Intercept Indicator Variables

The most common use of indicator variables is to modify the regression model intercept parameter. Adding the indicator variable D to the regression model, along with a new parameter δ, we obtain

$$PRICE = \beta_1 + \delta D + \beta_2 SQFT + e \tag{7.3}$$

[1] Such models have been used for many types of goods, including personal computers, automobiles and wine. This famous idea was introduced by Sherwin Rosen (1978) "Hedonic Prices and Implicit Markets," *Journal of Political Economy*, 82, 357–369. The ideas are summarized and applied to asparagus and personal computers in Ernst Berndt (1991) *The Practice of Econometrics: Classic and Contemporary*, Reading, MA: Addison-Wesley, Chapter 4.

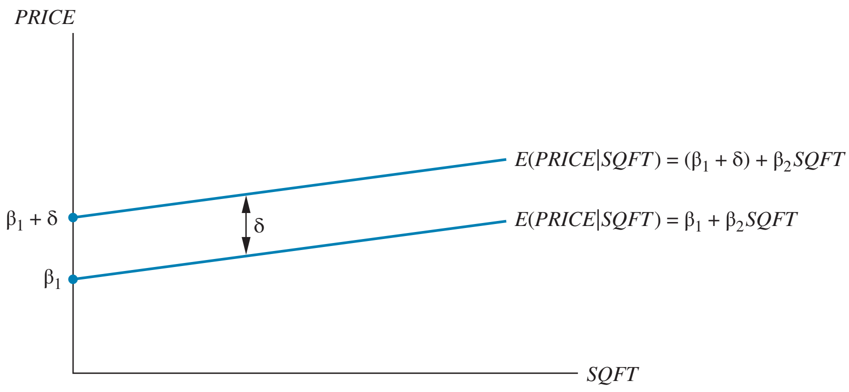

FIGURE 7.1 **An intercept indicator variable.**

The effect of the inclusion of an indicator variable D into the regression model is best seen by examining the regression function, $E(PRICE|SQFT)$, in the two locations. If the model in (7.3) is correctly specified, then $E(e|SQFT, D) = 0$ and

$$E(PRICE|SQFT) = \begin{cases} \left(\beta_1 + \delta\right) + \beta_2 SQFT & \text{when } D = 1 \\ \beta_1 + \beta_2 SQFT & \text{when } D = 0 \end{cases} \tag{7.4}$$

In the desirable neighborhood $D = 1$, and the intercept of the regression function is $\left(\beta_1 + \delta\right)$. In other areas, the regression function intercept is simply β_1. This difference is depicted in Figure 7.1, assuming that $\delta > 0$.

Adding the indicator variable D to the regression model causes a parallel shift in the relationship by the amount δ. In the context of the house price model the interpretation of the parameter δ is that it is a **location premium**, the difference in house price due to houses being located in the desirable neighborhood. An indicator variable that is incorporated into a regression model to capture a shift in the intercept as the result of some qualitative factor is called an **intercept indicator variable**, or an **intercept dummy variable**. In the house price example, we expect the price to be higher in a desirable location, and thus we anticipate that δ will be positive.

The least squares estimator's properties are not affected by the fact that one of the explanatory variables consists only of zeros and ones—D is treated as any other explanatory variable. We can construct an interval estimate for δ, or we can test the significance of its least squares estimate. Such a test is a statistical test of whether the neighborhood effect on house price is "statistically significant." If $\delta = 0$, then there is no location premium for the neighborhood in question.

Choosing the Reference Group

The convenience of the values $D = 0$ and $D = 1$ is seen in (7.4). The value $D = 0$ defines the **reference group**, or **base group**, of houses that are not in the desirable neighborhood. The expected price of these houses is simply $E(PRICE|SQFT) = \beta_1 + \beta_2 SQFT$. Using (7.3), we are comparing the house prices in the desirable neighborhood to those in the base group.

A researcher can choose whichever neighborhood is most convenient, for expository purposes, to be the reference group. For example, we can define the indicator variable LD to denote the less desirable neighborhood:

$$LD = \begin{cases} 1 & \text{if property is not in the desirable neighborhood} \\ 0 & \text{if property is in the desirable neighborhood} \end{cases}$$

This indicator variable is defined just the opposite from D, and $LD = 1 - D$. If we include LD in the model specification

$$PRICE = \beta_1 + \lambda LD + \beta_2 SQFT + e$$

then we make the reference group, $LD = 0$, the houses in the desirable neighborhood.

You may be tempted to include both D and LD in the regression model to capture the effect of each neighborhood on house prices. That is, you might consider the model

$$PRICE = \beta_1 + \delta D + \lambda LD + \beta_2 SQFT + e$$

In this model, the variables D and LD are such that $D + LD = 1$. Since the intercept variable $x_1 = 1$, we have created a model with **exact collinearity**, and as explained in Section 6.4, the least squares estimator is not defined in such cases. This error is sometimes described as falling into the **dummy variable trap**. By including only one of the indicator variables, either D or LD, the omitted variable defines the reference group, and we avoid the problem.[2]

7.1.2 Slope-Indicator Variables

Instead of assuming that the effect of location on house price causes a change in the intercept of the hedonic regression (7.1), let us assume that the change is in the slope of the relationship. We can allow for a change in a slope by including in the model an additional explanatory variable that is equal to the product of an indicator variable and a continuous variable. In our model, the slope of the relationship is the value of an additional square foot of living area. If we assume that this is one value for homes in the desirable neighborhood, and another value for homes in other neighborhoods, we can specify

$$PRICE = \beta_1 + \beta_2 SQFT + \gamma(SQFT \times D) + e \tag{7.5}$$

The new variable $(SQFT \times D)$ is the product of house size and the indicator variable, and is called an **interaction variable**, as it captures the interaction effect of location and size on house price. Alternatively, it is called a **slope-indicator variable** or a **slope dummy variable** because it allows for a change in the slope of the relationship. The slope-indicator variable takes a value equal to $SQFT$ for houses in the desirable neighborhood, when $D = 1$, and it is zero for homes in other neighborhoods. Despite its unusual nature, a slope-indicator variable is treated just like any other explanatory variable in a regression model. Examining the regression function for the two different locations best illustrates the effect of the inclusion of the slope-indicator variable into the economic model,

$$E(PRICE|SQFT,\ D) = \beta_1 + \beta_2 SQFT + \gamma(SQFT \times D) = \begin{cases} \beta_1 + \left(\beta_2 + \gamma\right) SQFT & \text{when } D = 1 \\ \beta_1 + \beta_2 SQFT & \text{when } D = 0 \end{cases}$$

In the desirable neighborhood, the price per additional square foot of a home is $\left(\beta_2 + \gamma\right)$; it is β_2 in other locations. We would anticipate $\gamma > 0$ if price per additional square foot is higher in the more desirable neighborhood. This situation is depicted in Figure 7.2a.

Another way to see the effect of including a slope-indicator variable is to use calculus. The partial derivative of expected house price with respect to size (measured in square feet), which gives the slope of the relation, is

$$\frac{\partial E(PRICE|SQFT,\ D)}{\partial SQFT} = \begin{cases} \beta_2 + \gamma & \text{when } D = 1 \\ \beta_2 & \text{when } D = 0 \end{cases}$$

If the assumptions of the regression model hold for (7.5), then the least squares estimators have their usual good properties, as discussed in Section 5.3. A test of the hypothesis that the value of an additional square foot of living area is the same in the two locations is carried out by testing the

[2] Another way to avoid the dummy variable trap is to omit the intercept from the model.

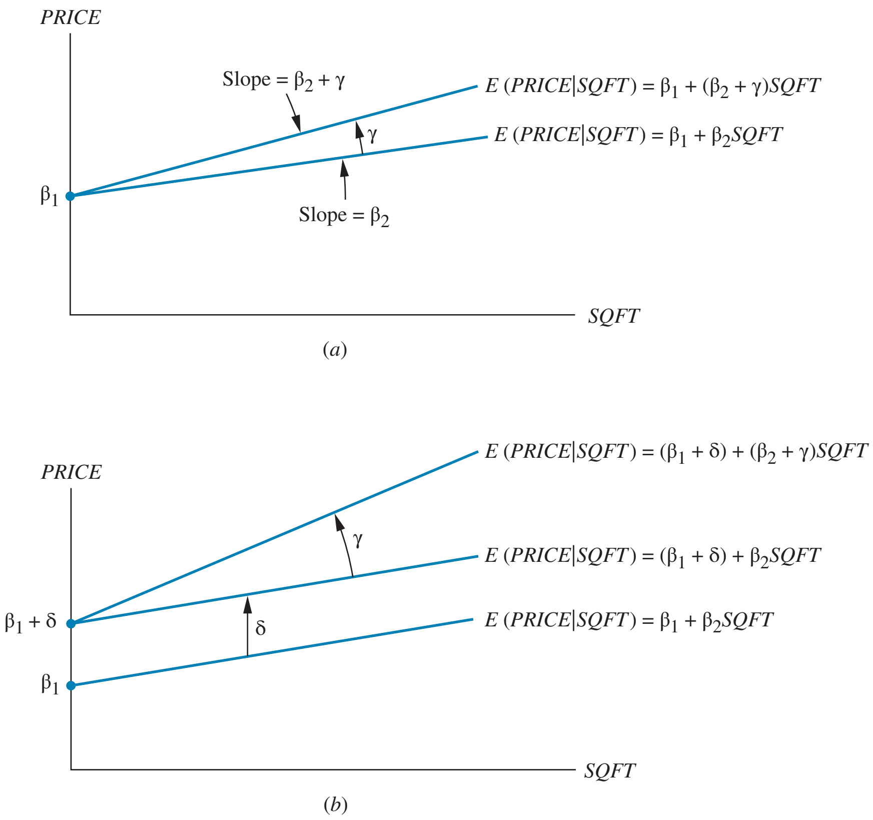

FIGURE 7.2 **(*a*) A slope-indicator variable. (*b*) Slope- and intercept-indicator variables.**

null hypothesis $H_0: \gamma = 0$ against the alternative $H_1: \gamma \neq 0$. In this case, we might test $H_0: \gamma = 0$ against $H_1: \gamma > 0$, since we expect the effect to be positive.

If we assume that house location affects *both* the intercept and the slope, then both effects can be incorporated into a single model. The resulting regression model is

$$PRICE = \beta_1 + \delta D + \beta_2 SQFT + \gamma(SQFT \times D) + e \tag{7.6}$$

In this case, the regression functions for the house prices in the two locations are

$$E(PRICE|SQFT) = \begin{cases} (\beta_1 + \delta) + (\beta_2 + \gamma)SQFT & \text{when } D = 1 \\ \beta_1 + \beta_2 SQFT & \text{when } D = 0 \end{cases}$$

In Figure 7.2b, we depict the house price relations assuming that $\delta > 0$ and $\gamma > 0$.

EXAMPLE 7.1 | The University Effect on House Prices

A real estate economist collects information on 1000 house price sales from two similar neighborhoods, one called "University Town" bordering a large state university, and another a neighborhood about three miles from the university. A few of the observations are shown in Table 7.1. The complete data file is *utown*.

House prices are given in \$1000; size ($SQFT$) is the number of hundreds of square feet of living area. For example, the first house sold for \$205,452 and has 2346 square feet of living area. Also recorded are the house AGE (in years), location ($UTOWN = 1$ for homes near the university, 0 otherwise), whether the house has a pool ($POOL = 1$ if a pool is present, 0 otherwise) and whether the house has a fireplace ($FPLACE = 1$ if a fireplace is present, 0 otherwise).

TABLE 7.1 Representative Real Estate Data Values

PRICE	*SQFT*	*AGE*	*UTOWN*	*POOL*	*FPLACE*
205.452	23.46	6	0	0	1
185.328	20.03	5	0	0	1
248.422	27.77	6	0	0	0
287.339	23.67	28	1	1	0
255.325	21.30	0	1	1	1
301.037	29.87	6	1	0	1

TABLE 7.2 House Price Equation Estimates

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	24.5000	6.1917	3.9569	0.0001
UTOWN	27.4530	8.4226	3.2594	0.0012
SQFT	7.6122	0.2452	31.0478	0.0000
SQFT × *UTOWN*	1.2994	0.3320	3.9133	0.0001
AGE	−0.1901	0.0512	−3.7123	0.0002
POOL	4.3772	1.1967	3.6577	0.0003
FPLACE	1.6492	0.9720	1.6968	0.0901
$R^2 = 0.8706$	$SSE = 230184.4$			

The economist specifies the regression equation as

$$PRICE = \beta_1 + \delta_1 UTOWN + \beta_2 SQFT + \gamma(SQFT \times UTOWN) + \beta_3 AGE + \delta_2 POOL + \delta_3 FPLACE + e \quad (7.7)$$

We anticipate that all the coefficients in this model will be positive except β_3, which is an estimate of the effect of age, or depreciation, on house price. Note that *POOL* and *FPLACE* are intercept dummy variables. By introducing these variables we are asking whether, and by how much, these features change house price. Because these variables stand alone, and are not interacted with *SQFT* or *AGE*, we are assuming that they affect the regression intercept, but not the slope. The estimated regression results are shown in Table 7.2. The goodness-of-fit statistic is $R^2 = 0.8706$, indicating that the model fits the data well. The slope-indicator variable is *SQFT* × *UTOWN*. Based on one-tail *t*-tests of significance,[3] at the $\alpha = 0.05$ level we reject zero null hypotheses for each of the parameters and accept the alternatives that they are positive, except for the coefficient on *AGE*, which we accept to be negative. In particular, based on these *t*-tests, we conclude that houses near the university have a significantly higher base price, and that their price per additional square foot is significantly higher than in the comparison neighborhood.

The estimated regression function for the houses near the university is

$$\begin{aligned}\widehat{PRICE} &= (24.5 + 27.453) + (7.6122 + 1.2994)SQFT \\ &\quad - 0.1901\,AGE + 4.3772POOL + 1.6492FPLACE \\ &= 51.953 + 8.9116SQFT - 0.1901AGE \\ &\quad + 4.3772POOL + 1.6492FPLACE\end{aligned}$$

For houses in other areas, the estimated regression function is

$$\begin{aligned}\widehat{PRICE} &= 24.5 + 7.6122SQFT - 0.1901AGE \\ &\quad + 4.3772POOL + 1.6492FPLACE\end{aligned}$$

Based on the regression results in Table 7.2, we estimate that

- The location premium for lots near the university is \$27,453.
- The change in expected price per additional square foot is \$89.12 for houses near the university and \$76.12 for houses in other areas.
- Houses depreciate \$190.10 per year.
- A pool increases the value of a home by \$4,377.20.
- A fireplace increases the value of a home by \$1,649.20.

[3] Recall that the *p*-value for a one-tail test is half of the reported two-tail *p*-value, providing that the coefficient estimate has the "correct" sign.

7.2 Applying Indicator Variables

Indicator variables can be used to ask and answer a rich variety of questions. In this section, we consider some common applications.

7.2.1 Interactions Between Qualitative Factors

We have seen how indicator variables can be used to represent qualitative factors in a regression model. Intercept indicator variables for qualitative factors are *additive*. That is, the effect of each qualitative factor is added to the regression intercept, and the effect of any indicator variable is independent of any other qualitative factor. Sometimes, however, we might question whether the effects of qualitative factors are independent.

For example, suppose we are estimating a wage equation, in which an individual's wages are explained as a function of their experience, skill, and other factors related to productivity. It is customary to include indicator variables for race and sex in such equations. If we have modeled productivity attributes well, and if wage determination is not discriminatory, then the coefficients of the race and sex indicator variables should not be significant. Including just race and sex indicator variables, however, will not capture interactions between these qualitative factors. Is there a differential in wages for black women? Separate indicator variables for being "black" and "female" will not capture this extra interaction effect. To allow for such a possibility, consider the following specification, in which for simplicity we use only education (*EDUC*) as a productivity measure:

$$\begin{aligned} WAGE = \beta_1 + \beta_2 EDUC + \delta_1 BLACK + \delta_2 FEMALE \\ + \gamma(BLACK \times FEMALE) + e \end{aligned} \tag{7.8}$$

where *BLACK* and *FEMALE* are indicator variables, and thus so is their interaction. These are intercept dummy variables because they are not interacted with any continuous explanatory variable. They have the effect of causing a parallel shift in the regression, as in Figure 7.1. When multiple dummy variables are present, and especially when there are interactions between indicator variables, it is important for proper interpretation to write out the regression function, $E(WAGE|EDUC)$, for each indicator variable combination:

$$E(WAGE|EDUC) = \begin{cases} \beta_1 + \beta_2 EDUC & WHITE - MALE \\ (\beta_1 + \delta_1) + \beta_2 EDUC & BLACK - MALE \\ (\beta_1 + \delta_2) + \beta_2 EDUC & WHITE - FEMALE \\ (\beta_1 + \delta_1 + \delta_2 + \gamma) + \beta_2 EDUC & BLACK - FEMALE \end{cases}$$

In this specification, white males are the reference group because this is the group defined when all indicator variables take the value zero, in this case $BLACK = 0$ and $FEMALE = 0$. The parameter δ_1 measures the effect of being black, relative to the reference group; the parameter δ_2 measures the effect of being female, and the parameter γ measures the effect of being black and female.

EXAMPLE 7.2 | The Effects of Race and Sex on Wage

Using CPS data (data file *cps5_small*) from 2013, we obtain the results in Table 7.3. Holding the effect of education constant, we estimate that on average black males earn \$2.07 per hour less than white males, white females earn \$4.22 less than white males, and black females earn \$5.76 less than white males. The coefficients of *EDUC* and *FEMALE* are significantly different from zero using individual *t*-tests. The coefficient of *BLACK* and the interaction effect between

BLACK and *FEMALE* are not estimated very precisely using this sample of 1200 observations, and are not statistically significant.[4]

Suppose we are asked to test the joint significance of all the qualitative factors. How do we test the hypothesis that neither a person's race nor sex affects wages? We do it by testing the joint null hypothesis $H_0: \delta_1 = 0, \delta_2 = 0, \gamma = 0$ against the alternative that at least one of the tested parameters is not zero. If the null hypothesis is true, race and sex fall out of the regression, and thus have no effect on wages.

To test this hypothesis, we use the F-test procedure that is described in Section 6.1. The test statistic for a joint hypothesis is

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)}$$

where SSE_R is the sum of squared least squares residuals from the "restricted" model in which the null hypothesis is assumed to be true, SSE_U is the sum of squared residuals from the original, "unrestricted," model, J is the number of joint hypotheses, and $N - K$ is the number of degrees of freedom in the unrestricted model. If the null hypothesis is true, then the test statistic F has an F-distribution with J numerator degrees of freedom and $N - K$ denominator degrees of freedom, $F_{(J,N-K)}$. We reject the null hypothesis if $F \geq F_c$, where F_c is the critical value, illustrated in Figure B.9, for the level of significance α. To test the $J = 3$ joint null hypotheses $H_0: \delta_1 = 0,\ \delta_2 = 0, \gamma = 0,$ we obtain the unrestricted sum of squared errors $SSE_U = 214400.9$ from the model reported in Table 7.3. The restricted sum of squares is obtained by estimating the model that assumes the null hypothesis is true, leading to the fitted model

$$\begin{array}{lcl}\widehat{WAGE} = & -10.4000 + 2.3968EDUC \\ \text{(se)} & (1.9624)\quad (0.1354)\end{array}$$

which has $SSE_R = 220062.3$. The degrees of freedom $(N - K) = (1200 - 5) = 1195$ come from the unrestricted model. The value of the F-statistic is

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \frac{(220062.3 - 214400.9)/3}{214400.9/1195}$$
$$= 10.52$$

The 1% critical value [i.e., the 99th percentile value] is $F_{(0.99,\,3,\,1195)} = 3.798$. Thus, we conclude that a worker's race and/or sex affect the wage equation.

TABLE 7.3 Wage Equation with Race and Sex

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	−9.4821	1.9580	−4.8428	0.0000
EDUC	2.4737	0.1351	18.3096	0.0000
BLACK	−2.0653	2.1616	−0.9554	0.3396
FEMALE	−4.2235	0.8249	−5.1198	0.0000
BLACK × *FEMALE*	0.5329	2.8020	0.1902	0.8492
$R^2 = 0.2277$	$SSE = 214400.9$			

7.2.2 Qualitative Factors with Several Categories

Many qualitative factors have more than two categories. An example is the variable region of the country in our wage equation. The CPS data record worker residence within one of the four regions: northeast, midwest, south, and west. Again, using just the simple wage specification for illustration, we can incorporate indicator variables into the wage equation as

$$WAGE = \beta_1 + \beta_2 EDUC + \delta_1 SOUTH + \delta_2 MIDWEST + \delta_3 WEST + e \tag{7.9}$$

[4] Estimating this model using the larger data set *cps5*, which contains 9799 observations, yields a coefficient estimate for *BLACK of* −4.3488 with a t-value of −5.81. Similarly, the coefficient of the interaction variable is 3.0873 with a $t = 3.01$. Both of these are statistically significant. Recall from Sections 2.4 and 5.3 that larger sample sizes lead to smaller standard errors and thus more precise estimation. Labor economists tend to use large data sets so that complex effects and interactions can be estimated precisely. We use the smaller data set as a text example so that results can be replicated with student versions of software.

Notice that we have not included the indicator variables for all regions. Doing so would have created a model in which exact collinearity exists. Since the regional categories are exhaustive, the sum of the **regional indicator variables** is $NORTHEAST + SOUTH + MIDWEST + WEST = 1$. Thus, the "intercept variable" $x_1 = 1$ is an exact linear combination of the region indicators. Recall, from Section 6.4, that the least squares estimator is not defined in such cases. Failure to omit one indicator variable will lead to your computer software returning a message saying that least squares estimation fails. This error is the **dummy variable trap** that we mentioned in Section 7.1.1.

The usual solution to this problem is to omit one indicator variable, which defines a **reference group**, as we shall see by examining the regression function,

$$E(WAGE|EDUC) = \begin{cases} (\beta_1 + \delta_3) + \beta_2 EDUC & WEST \\ (\beta_1 + \delta_2) + \beta_2 EDUC & MIDWEST \\ (\beta_1 + \delta_1) + \beta_2 EDUC & SOUTH \\ \beta_1 + \beta_2 EDUC & NORTHEAST \end{cases}$$

The omitted indicator variable, *NORTHEAST*, identifies the reference group for the equation, to which workers in other regions are compared. It is the group that remains when the regional indicator variables *WEST*, *MIDWEST*, and *SOUTH* are set to zero. Mathematically, it does not matter which indicator variable is omitted; the choice can be made that is most convenient for interpretation. The intercept parameter β_1 represents the base wage for a worker with no education who lives in the northeast. The parameter δ_1 measures the expected wage differential between southern workers relative to those in the northeast; δ_2 measures the expected wage differential between midwestern workers and those in the northeast.

EXAMPLE 7.3 | A Wage Equation with Regional Indicators

Using CPS data in data file *cps5_small*, let us take the specification in Table 7.3 and add the regional indicators *SOUTH*, *MIDWEST*, and *WEST*. The results are in Table 7.4. We estimate that workers in the South earn \$1.65 less per hour than workers in the Northeast, and workers in the Midwest earn \$1.94 less than workers in the Northeast, holding other factors constant. These estimates are not significantly different from zero at the 10% level.[5]

TABLE 7.4 **Wage Equation with Regional Indicator Variables**

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	−8.3708	2.1540	−3.8862	0.0001
EDUC	2.4670	0.1351	18.2603	0.0000
BLACK	−1.8777	2.1799	−0.8614	0.3892
FEMALE	−4.1861	0.8246	−5.0768	0.0000
BLACK × *FEMALE*	0.6190	2.8008	0.2210	0.8251
SOUTH	−1.6523	1.1557	−1.4297	0.1531
MIDWEST	−1.9392	1.2083	−1.6049	0.1088
WEST	−0.1452	1.2027	−0.1207	0.9039
$R^2 = 0.2308$	$SSE = 213552.1$			

[5] Using the larger CPS data file, *cps5*, the estimated regional coefficients are (t-values in parentheses): *SOUTH* −0.9405 (−2.24), *MIDWEST* −2.4299 (−5.58), and *WEST* 0.0088 (0.02).

How would we test the hypothesis that there are no regional differences? This would be a joint test of the null hypothesis that the coefficients of the regional dummies are all zero. In the context of the CPS data, $SSE_U = 213552.1$ for the wage equation in Table 7.4. Under the null hypothesis, the model in Table 7.4 reduces to that in Table 7.3 where $SSE_R = 214400.9$. This yields an F-statistic value of 1.579. The p-value for this test is 0.1926, so we fail to reject the null hypothesis that there are no regional differences in the wage equation intercept, holding other factors constant.[6]

7.2.3 Testing the Equivalence of Two Regressions

In Section 7.1.2, we introduced both intercept and slope-indicator variables into the hedonic equation for house price. The result was given in (7.6)

$$PRICE = \beta_1 + \delta D + \beta_2 SQFT + \gamma(SQFT \times D) + e$$

The regression functions for the house prices in the two locations are

$$E(PRICE|SQFT) = \begin{cases} \alpha_1 + \alpha_2 SQFT & D = 1 \\ \beta_1 + \beta_2 SQFT & D = 0 \end{cases}$$

where $\alpha_1 = \beta_1 + \delta$ and $\alpha_2 = \beta_2 + \gamma$. Figure 7.2b shows that by introducing both intercept and slope-indicator variables, we have essentially assumed that the regressions in the two neighborhoods are completely different. We could obtain the estimates for (7.6) by estimating separate regressions for each of the neighborhoods. In this section, we generalize this idea, which leads to the **Chow test**, named after econometrician Gregory Chow. The Chow test is an F-test for the equivalence of two regressions.

By including an intercept indicator variable and an interaction variable for *each* additional variable in an equation, we allow all coefficients to differ based on a qualitative factor. Consider again the wage equation in (7.8)

$$WAGE = \beta_1 + \beta_2 EDUC + \delta_1 BLACK + \delta_2 FEMALE + \gamma(BLACK \times FEMALE) + e$$

We might ask "Are there differences between the wage regressions for the south and for the rest of the country?" If there are no differences, then the data from the south and other regions can be pooled into one sample, with no allowance made for differing slope or intercept. How can we test this? We can carry out the test by creating intercept and slope-indicator variables for *every* variable in the model, and then jointly testing the significance of the indicator variable coefficients using an F-test. That is, we specify the model

$$\begin{aligned} WAGE = \beta_1 &+ \beta_2 EDUC + \delta_1 BLACK + \delta_2 FEMALE + \gamma(BLACK \times FEMALE) \\ &+ \theta_1 SOUTH + \theta_2(EDUC \times SOUTH) + \theta_3(BLACK \times SOUTH) \\ &+ \theta_4(FEMALE \times SOUTH) + \theta_5(BLACK \times FEMALE \times SOUTH) + e \end{aligned} \tag{7.10}$$

In (7.10) we have twice the number of parameters and variables than in (7.8). We have added five new variables, the *SOUTH* intercept indicator variable and interactions between *SOUTH* and the other four variables, and corresponding parameters. Estimating (7.10) is equivalent to estimating (7.8) twice—once for the southern workers and again for workers in the rest of the country.

[6] Using the larger CPS data file, *cps5*, the $F = 14.7594$ which is significant at the 1% level.

To see this, examine the regression functions. Let $\mathbf{X}$ represent (*EDUC*, *BLACK*, *FEMALE*, *SOUTH*). Then

$$E(WAGE|\mathbf{X}) = \begin{cases} \beta_1 + \beta_2 EDUC + \delta_1 BLACK + \delta_2 FEMALE & SOUTH = 0 \\ \quad + \gamma(BLACK \times FEMALE) & \\ (\beta_1 + \theta_1) + (\beta_2 + \theta_2)EDUC + (\delta_1 + \theta_3)BLACK & \\ \quad + (\delta_2 + \theta_4)FEMALE + (\gamma + \theta_5)(BLACK \times FEMALE) & SOUTH = 1 \end{cases}$$

Note that each variable has a separate coefficient for southern and nonsouthern workers.

EXAMPLE 7.4 | Testing the Equivalence of Two Regressions: The Chow Test

In column (1) of Table 7.5, we report the estimates and standard errors for the fully interacted model (7.10), using the full sample. The base model (7.8) is estimated once for workers outside the south [column (2)] and again for southern workers [column (3)]. Note that the coefficient estimates on the nonsouth data in (2) are identical to those using the full sample in (1). The standard errors differ because the estimates of the error variance, σ^2, differ. The coefficient estimates using only southern workers are obtained from the full model by adding the indicator variable interaction coefficients θ_i to the corresponding nonsouth coefficients. For example, the coefficient estimate for *BLACK* in column (3) is obtained as $\left(\hat{\delta}_1 + \hat{\theta}_3\right) = 1.1276 - 4.6204 = -3.4928$. Similarly, the coefficient on *FEMALE* in column (3) is $\left(\hat{\delta}_2 + \hat{\theta}_4\right) = -4.1520 - 0.1886 = -4.3406$.

Furthermore, note that the sum of squared residuals for the full model in column (1), but for a small rounding error, is the sum of the *SSE* from the two separate regressions

$$\begin{aligned} SSE_{full} &= SSE_{nonsouth} + SSE_{south} \\ &= 125880.0 + 87893.9 = 213773.9 \end{aligned}$$

Using this indicator variable approach, we can test for a southern regional difference. We estimate (7.10) and test the joint null hypothesis

$$H_0 : \theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta_5 = 0$$

against the alternative that at least one $\theta_i \neq 0$. This is the Chow test. If we reject this null hypothesis, we conclude that there is some difference in the wage equation in the southern

TABLE 7.5 Comparison of Fully Interacted to Separate Models

	(1) Full sample		(2) Nonsouth		(3) South	
Variable	**Coefficient**	**Std. Error**	**Coefficient**	**Std. Error**	**Coefficient**	**Std. Error**
C	−9.9991	2.3872	−9.9991	2.2273	−8.4162	3.8709
EDUC	2.5271	0.1642	2.5271	0.1532	2.3557	0.2692
BLACK	1.1276	3.5247	1.1276	3.2885	−3.4928	3.1667
FEMALE	−4.1520	0.9842	−4.1520	0.9182	−4.3406	1.7097
BLACK × *FEMALE*	−4.4540	4.4858	−4.4540	4.1852	3.6655	4.1832
SOUTH	1.5829	4.1821				
EDUC × *SOUTH*	−0.1714	0.2898				
BLACK × *SOUTH*	−4.6204	4.5071				
FEMALE × *SOUTH*	−0.1886	1.8080				
BLACK × *FEMALE* × *SOUTH*	8.1195	5.8217				
SSE	213774.0		125880.0		87893.9	
N	1200		810		390	

region relative to the rest of the country. The test can also be thought of as comparing the estimates in the nonsouth and south in columns (2) and (3) in Table 7.5.

The test ingredients are the unrestricted $SSE_U = 213774.0$ from the full model in Table 7.5 [or the sum of the SSE's from the two separate regressions], the restricted $SSE_R = 214400.9$ comes from Table 7.3. The test statistic for the $J = 5$ hypotheses is

$$F = \frac{\left(SSE_R - SSE_U\right)/J}{SSE_U/(N-K)}$$

$$= \frac{(214400.9 - 213774.0)/5}{213774.0/1190} = 0.6980$$

The denominator degrees of freedom come from the unrestricted model, $N - K = 1200 - 10$. The p-value of this test is $p = 0.6250$, and thus we fail to reject the null hypothesis that the wage regression in the South is no different from that in the rest of the country.[7]

Remark

The usual F-test of a joint hypothesis relies on the assumptions MR1–MR6 of the linear regression model. Of particular relevance for testing the equivalence of two regressions is assumption MR3, that the variance of the error term, $\text{var}\left(e_i|\mathbf{X}\right) = \sigma^2$, is the same *for all* observations. If we are considering possibly different slopes and intercepts for parts of the data, it might also be true that the error variances are different in the two parts of the data. In such a case, the usual F-test is not valid. Testing for equal variances is covered in Section 8.2, and the question of pooling in this case is covered in Section 8.4. For now, be aware that we are assuming constant error variances in the calculations above.

7.2.4 Controlling for Time

The earlier examples we have given apply indicator variables to cross-sectional data. Indicator variables are also used in regressions using time-series data, as the following examples illustrate.

Seasonal Indicators

Summer means outdoor cooking on barbeque grills. What effect might this have on the sales of charcoal briquettes, a popular fuel for grilling? To investigate, let us define a model with dependent variable y_t = the number of 20-pound bags of Royal Oak charcoal sold in week t at a supermarket. Explanatory variables would include the price of Royal Oak, the price of competitive brands (Kingsford and the store brand), the prices of complementary goods (charcoal lighter fluid, pork ribs, and sausages), and advertising (newspaper ads and coupons). While these standard demand factors are all relevant, we may also find strong seasonal effects. All other things being equal, more charcoal is sold in the warm summer months than in other seasons. Thus, we may want to include either monthly indicator variables (e.g., $AUG = 1$ if month is August, $AUG = 0$ otherwise) or **seasonal indicator variables** (in North America, $SUMMER = 1$ if month = June, July, or August; $SUMMER = 0$ otherwise) into the regression. In addition to these seasonal effects, holidays are special occasions for cookouts. In the United States, these are Memorial Day (last Monday in May), Independence Day (July 4), and Labor Day (first Monday in September). Additional sales can be expected in the week before these holidays, meaning that indicator variables for each should be included into the regression.

[7] The p-value of this test using the larger CPS data set, *cps5*, is 0.7753, so that we again fail to reject the null hypothesis.

Year Indicators

In the same spirit as seasonal indicator variables, **annual indicator variables** are used to capture year effects not otherwise measured in a model. The real estate model discussed earlier in this chapter provides an example. Real estate data are available continuously, every month, every year. Suppose we have data on house prices for a certain community covering a 10-year period. In addition to house characteristics, such as those employed in (7.7), the overall price level is affected by demand factors in the local economy, such as population change, interest rates, unemployment rate, and income growth. Economists creating "cost-of-living" or "house price" indexes for cities must include a component for housing that takes the pure price effect into account. Understanding the price index is important for tax assessors, who must reassess the market value of homes in order to compute the annual property tax. It is also important to mortgage bankers and other home lenders, who must reevaluate the value of their portfolio of loans with changing local conditions, as well as to homeowners trying to sell their houses, and to potential buyers as they attempt to agree upon a selling price.

The simplest method for capturing these price effects is to include annual indicator variables (e.g., $D99 = 1$ if year $= 1999$; $D99 = 0$ otherwise) into the hedonic regression model. An example can be found in Exercise 7.3.

Regime Effects

An economic regime is a set of structural economic conditions that exist for a certain period. The idea is that economic relations may behave one way during one regime, but may behave differently during another. Economic regimes may be associated with political regimes (conservatives in power, liberals in power), unusual economic conditions (oil embargo, recession, hyperinflation), or changes in the legal environment (tax law changes). An investment tax credit[8] was enacted in 1962 in an effort to stimulate additional investment. The law was suspended in 1966, reinstated in 1970, and eliminated in the Tax Reform Act of 1986. Thus, we might create an indicator variable

$$ITC_t = \begin{cases} 1 & \text{if } t = 1962 - 1965,\ 1970 - 1986 \\ 0 & \text{otherwise} \end{cases}$$

A macroeconomic investment equation might be

$$INV_t = \beta_1 + \delta ITC_t + \beta_2 GNP_t + \beta_3 GNP_{t-1} + e_t$$

If the tax credit was successful, then $\delta > 0$.

7.3 Log-Linear Models

In Section 4.5, we examined the log-linear model in some detail. In this section, we explore the interpretation of indicator variables in **log-linear models**. Some additional detail is provided in Appendix 7A. Let us consider the log-linear model in (7.11). We do not introduce an error term, and we take *EDUC* and *FEMALE* to be given, in order to simplify the exposition.

$$\ln(WAGE) = \beta_1 + \beta_2 EDUC + \delta FEMALE \tag{7.11}$$

What is the interpretation of the parameter δ? *FEMALE* is an intercept dummy variable, creating a parallel shift of the log-linear relationship when $FEMALE = 1$. That is,

$$\ln(WAGE) = \begin{cases} \beta_1 + \beta_2 EDUC & MALES\ \ (FEMALE = 0) \\ (\beta_1 + \delta) + \beta_2 EDUC & FEMALES\ \ (FEMALE = 1) \end{cases}$$

[8] Intriligator, Bodkin and Hsiao, *Econometric Models, Techniques and Applications*, 2nd edition, Upper Saddle River, NJ: Prentice-Hall, 1996, p. 53.

But what about the fact that the dependent variable is ln(*WAGE*)? Does that have an effect? The answer is yes—and there are two solutions.

7.3.1 A Rough Calculation

First, take the difference between ln(*WAGE*) of females and males:

$$\ln(WAGE)_{FEMALES} - \ln(WAGE)_{MALES} = \delta$$

Recall from Appendix A.1.6 and equation (A.3) that 100 times the log-difference, 100δ, is approximately the percentage difference.

EXAMPLE 7.5 | Indicator Variables in a Log-Linear Model: The Rough Approximation

Using the data file *cps5_small*, the estimated log-linear model (7.11) is

$$\begin{array}{lllll} \widehat{\ln(WAGE)} = & 1.6229 & + 0.1024EDUC & - 0.1778FEMALE \\ \text{(se)} & (0.0692) & (0.0048) & (0.0279) \end{array}$$

Thus, we would estimate that there is a 17.78% differential between male and female wages. This is quick and simple, but there is an approximation error with a difference this large.

7.3.2 An Exact Calculation

We can overcome the approximation error by doing a little algebra. The wage difference is

$$\ln(WAGE)_{FEMALES} - \ln(WAGE)_{MALES} = \ln\left(\frac{WAGE_{FEMALES}}{WAGE_{MALES}}\right) = \delta$$

using the property of logarithms that $\ln(x) - \ln(y) = \ln(x/y)$. These are natural logarithms, and the antilog is the exponential function,

$$\frac{WAGE_{FEMALES}}{WAGE_{MALES}} = e^{\delta}$$

Subtract 1 from each side (in a tricky way) to obtain

$$\frac{WAGE_{FEMALES}}{WAGE_{MALES}} - \frac{WAGE_{MALES}}{WAGE_{MALES}} = \frac{WAGE_{FEMALES} - WAGE_{MALES}}{WAGE_{MALES}} = e^{\delta} - 1$$

The percentage difference between wages of females and males is $100\left(e^{\delta} - 1\right)\%$. See Appendix 7A for a more detailed approach.

EXAMPLE 7.6 | Indicator Variables in a Log-Linear Model: An Exact Calculation

Using the data *cps5_small*, we estimate the wage differential between males and females to be

$$100\left(e^{\hat{\delta}} - 1\right)\% = 100\left(e^{-0.1778} - 1\right)\% = -16.29\%$$

The approximate standard error for this estimate is 2.34%, which is a calculation that may be provided by your software.

7.4 The Linear Probability Model

Economics is sometimes described as the "theory of choice." Many of the choices we make in life are "either—or" in nature. A few examples include the following:

- A consumer who must choose between Coke and Pepsi
- A married woman who must decide whether to enter the labor market or not
- A bank official must choose to accept a loan application or not
- A high school graduate must decide whether to attend college or not
- A member of Parliament, a Senator, or a Representative must vote for or against a piece of legislation.

To analyze and predict such outcomes using an econometric model, we represent the choice using an indicator variable, the value one if one alternative is chosen and the value zero if the other alternative is chosen. Because we are attempting to explain choice between two alternatives, the indicator variable will be the **dependent** variable rather than an independent variable in a regression model.

To begin, let us represent the variable indicating a choice as

$$y = \begin{cases} 1 & \text{if first alternative is chosen} \\ 0 & \text{if second alternative is chosen} \end{cases}$$

If we observe the choices that a random sample of individuals makes, then y is a random variable. If p is the probability that the first alternative is chosen, then $P[y = 1] = p$. The probability that the second alternative is chosen is $P[y = 0] = 1 - p$. The probability function for the binary indicator variable y is

$$f(y) = p^y(1 - p)^{1-y}, \quad y = 0, 1$$

The indicator variable y is said to follow a Bernoulli distribution. The expected value of y is $E(y) = p$, and its variance is $\text{var}(y) = p(1 - p)$.

We are interested in identifying factors that might affect the probability p using a linear regression function, or, in this context, a **linear probability model**,

$$E(y|\mathbf{X}) = p = \beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K$$

Proceeding as usual, we break the observed outcome y into a systematic portion, $E(y|\mathbf{X})$, and an unpredictable random error, e, so that the econometric model is

$$y = E(y|\mathbf{X}) + e = \beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K + e$$

One difficulty with using this model for choice behavior is that the usual error term assumptions cannot hold. The outcome y only takes two values, implying that the error term e also takes only two values, so that the usual "bell-shaped" curve describing the distribution of errors does not hold. The probability functions for y and e are

y value	e value	Probability
1	$1 - (\beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K)$	p
0	$-(\beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K)$	$1 - p$

The variance of the error term e is

$$\text{var}(e|\mathbf{X}) = p(1 - p) = (\beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K)(1 - \beta_1 - \beta_2 x_2 - \cdots - \beta_K x_K)$$

This error is not homoskedastic, so the usual formula for the variance of the least squares estimator is incorrect. A second problem associated with the linear probability model is that predicted values, $\widehat{E(y)} = \hat{p}$, can fall outside the (0, 1) interval, meaning that their interpretation as probabilities does not make sense. Despite these weaknesses, the linear probability model has the advantage of simplicity, and it has been found to provide good estimates of the marginal effects of changes in explanatory variables x_k on the choice probability p, as long as p is not too close to zero or one.[9]

EXAMPLE 7.7 | The Linear Probability Model: An Example from Marketing

A shopper is deciding between Coke and Pepsi. Define the variable *COKE*:

$$COKE = \begin{cases} 1 & \text{if Coke is chosen} \\ 0 & \text{if Pepsi is chosen} \end{cases}$$

The expected value of this variable is $E(COKE|\mathbf{X}) = p_{COKE}$ = probability that Coke is chosen given some conditioning factors. What factors might enter the choice decision? The relative price of Coke to Pepsi (*PRATIO*) is a potential factor. As the relative price of Coke rises, we should observe a reduced probability of its choice. Other factors influencing the consumer might be the presence of store displays for these products. Let *DISP_COKE* and *DISP_PEPSI* be indicator variables taking the value one if the respective store display is present and zero if it is not. We expect that the presence of a Coke display will increase the probability of a Coke purchase, and the presence of a Pepsi display will decrease the probability of a Coke purchase.

The data file *coke*[10] contains "scanner" data on 1140 individuals who purchased Coke or Pepsi. In this sample, 44.7% of the customers chose Coke. The estimated linear probability model is

$$\begin{array}{llll} \hat{p}_{COKE} = & 0.8902 - 0.4009PRATIO & + 0.0772DISP\_COKE & \\ \text{(se)} & (0.0655) \quad (0.0613) & \quad (0.0344) & \\ & -0.1657DISP\_PEPSI & & \\ & (0.0356) & & \end{array}$$

Assuming for the moment that the standard errors are reliable,[11] all the coefficients are significantly different from zero at the $\alpha = 0.05$ level. Recall that $PRATIO = 1$ if the prices of Coke and Pepsi are equal, and that $PRATIO = 1.10$ would represent a case in which Coke was 10% more expensive than Pepsi. Such an increase is estimated to reduce the probability of purchasing Coke by 0.04. A store display for Coke is estimated to increase the probability of a Coke purchase by 0.077, and a Pepsi display is estimated to reduce the probability of a Coke purchase by 0.166. The concerns about predicted probabilities falling outside (0,1) are well founded in general, but in this example only 16 of the 1140 sample observations resulted in predicted probabilities less than zero, and there were no predicted probabilities greater than one.

7.5 Treatment Effects

Consider the question "Do hospitals make people healthier?" Angrist and Pischke[12] report the results of a National Health Interview Survey that included the question "During the past 12 months, was the respondent a patient in a hospital overnight?" Also asked was "Would you say your health in general is excellent, very good, good, fair or poor?" Using the number 1 for poor health and 5 for excellent health, those *who had not* gone to the hospital had an average health score of 3.93, and those *who had been* to the hospital had an average score of 3.21. That is, individuals who had been to the hospital had poorer health than those who had not.

[9] See Chapter 16 for nonlinear models of choice, called probit and logit, which ensure that predicted probabilities fall between zero and one. These models require the use of more complex estimators and methods of inference.

[10] Obtained from the ERIM public data base, James M. Kilts Center, University of Chicago Booth School of Business. *Scanner data* is information recorded at the point of purchase by an electronic device reading a barcode.

[11] The estimates and standard errors are not terribly dissimilar from those obtained using more advanced options discussed in Chapters 8 and 16.

[12] *Mostly Harmless Econometrics: An Empiricist's Guide*, Princeton, 2009, pp. 12–13.

Books on principles of economics warn in the first chapter[13] about the faulty line of reasoning known as **post hoc**, **ergo propter hoc**, which means that one event preceding another does not necessarily make the first the cause of the second. Going to the hospital does not *cause* the poorer health status. Those who were less healthy *chose* to go to the hospital because of an illness or injury, and at the time of the survey were still less healthy than those who had not gone to the hospital. Another way to say this is embodied in the warning that "**correlation** is not the same as **causation**." We observe that those who had been in a hospital are less healthy, but observing this association does not imply that going to the hospital causes a person to be less healthy. Still another way to describe the problem we face in this example is to say that data exhibit a **selection bias** because some people chose (or **self-selected**) to go to the hospital and the others did not. When membership in the treated group is in part determined by choice, then the sample is *not* a random sample. There are systematic factors, in this case health status, contributing to the composition of the sample.

A second example of selection bias may bring the concept closer to home. Are you reading this great book because you are enrolled in an econometrics class? Is the course required, or not? If your class is an "elective," then you and your classmates are *not a random sample* from the broader student population. It is our experience that students taking econometrics as an elective have an ability level and quantitative preparation that is higher, on average, than a random sample from the university population. We also observe that a higher proportion of undergraduate students who take econometrics enroll in graduate programs in economics or related disciplines. Is this a causal relationship? In part, it certainly is, but also your abilities and future plans for graduate training may have drawn you to econometrics, so that the high success rate of our students is in part attributed to **selection bias**.

Selection bias is also an issue when asking

- "How much does an additional year of education increase the wages of married women?" The difficulty is that we are able to observe a woman's wages only if she chooses to join the labor force, and thus the observed data is not a random sample.
- "How much does participation in a job-training program increase wages?" If participation is voluntary, then we may see a greater proportion of less skilled workers taking advantage of such a program.
- "How much does a dietary supplement contribute to weight loss?" If those taking the supplement are among the severely overweight, then the results we observe may not be "typical."

In each of these cases, selection bias interferes with a straightforward examination of the data, and makes more difficult our efforts to measure a **causal effect**, or **treatment effect**.

In some situations, usually those involving the physical or medical sciences, it is clearer how we might study causal effects. For example, if we wish to measure the effect of a new type of fertilizer on rice production, we can **randomly** assign identical rice fields to be treated with a new fertilizer (the **treatment group**), with the others being treated with an existing product (the **control group**). At the end of the growing period, we compare the production on the two types of fields. The key here is that we perform a **randomized controlled experiment**. By randomly assigning subjects to treatment and control groups, we ensure that the differences we observe will result from the treatment. In medical research, the effectiveness of a new drug is measured by such experiments. Test subjects are randomly assigned to the control group, who receive a placebo drug, and the treatment group, who receive the drug being tested. By random assignment of treatment and control groups, we prevent any selection bias from occurring.

As economists, we would like to have the type of information that arises from randomized controlled experiments to study the consequences of social policy changes, such as changes in

[13] See, for example, Campbell R. McConnell and Stanley L. Brue, *Economics, Twelfth Edition*, McGraw-Hill, 1993, pp. 8–9.

laws, or changes in types and amounts of aid and training we provide the poor. The ability to perform randomized controlled experiments is limited because the subjects are people, and their economic well-being is at stake. However, there are some examples. Before we proceed, we will examine the statistical consequences of selection bias for the measurement of treatment effects.

7.5.1 The Difference Estimator

In order to understand the measurement of treatment effects, consider a simple regression model in which the explanatory variable is a dummy variable, indicating whether a particular individual is in the treatment or control group. Let y be the outcome variable, the measured characteristic the treatment is designed to effect. In the rice production example, y would be the output of rice on a particular rice field. Define the indicator variable d as

$$d_i = \begin{cases} 1 & \text{individual in treatment group} \\ 0 & \text{individual in control group} \end{cases} \tag{7.12}$$

The effect of the treatment on the outcome can be modeled as

$$y_i = \beta_1 + \beta_2 d_i + e_i, \quad i = 1, \ldots, N \tag{7.13}$$

where e_i represents the collection of other factors affecting the outcome. The regression functions for the treatment and control groups are

$$E(y_i) = \begin{cases} \beta_1 + \beta_2 & \text{if in treatment group, } d_i = 1 \\ \beta_1 & \text{if in control group, } d_i = 0 \end{cases}$$

This is the same model we used in Section 2.9 to study the effect of location on house prices. The **treatment effect** that we wish to measure is β_2. The least squares estimator of β_2 is

$$b_2 = \frac{\sum_{i=1}^{N}\left(d_i - \bar{d}\right)(y_i - \bar{y})}{\sum_{i=1}^{N}\left(d_i - \bar{d}\right)^2} = \bar{y}_1 - \bar{y}_0 \tag{7.14}$$

where $\bar{y}_1 = \sum_{i=1}^{N_1} y_i/N_1$ is the sample mean of the N_1 observations on y for the treatment group ($d = 1$) and $\bar{y}_0 = \sum_{i=1}^{N_0} y_i/N_0$ is the sample mean of the N_0 observations on y for the control group ($d = 0$). In this treatment/control framework, the estimator b_2 is called the **difference estimator** because it is the difference between the sample means of the treatment and control groups.[14]

7.5.2 Analysis of the Difference Estimator

The statistical properties of the difference estimator can be examined using the same strategy employed in Section 2.4.2. We can rewrite the difference estimator as

$$b_2 = \beta_2 + \frac{\sum_{i=1}^{N}\left(d_i - \bar{d}\right)(e_i - \bar{e})}{\sum_{i=1}^{N}\left(d_i - \bar{d}\right)^2} = \beta_2 + \left(\bar{e}_1 - \bar{e}_0\right)$$

[14] See Appendix 7B for an algebraic derivation.

In the middle equality, the factor added to β_2 has the same form as the difference estimator in (7.14), with e_i replacing y_i—hence the final equality. The difference estimator b_2 equals the true treatment effect β_2 plus the difference between the averages of the unobserved factors affecting the outcomes y for the treatment group $\left(\overline{e}_1\right)$ and for the control group $\left(\overline{e}_0\right)$. In order for the difference estimator to be unbiased, $E\left(b_2\right) = \beta_2$, it must be true that

$$E\left(\overline{e}_1 - \overline{e}_0\right) = E\left(\overline{e}_1\right) - E\left(\overline{e}_0\right) = 0$$

In words, the expected value of all the factors affecting the outcome, other than the treatment, must be **equal** for the treatment and control groups.

If we allow individuals to "self-select" into treatment and control groups, then $E\left(\overline{e}_1\right) - E\left(\overline{e}_0\right)$ is the selection bias in the estimation of the treatment effect. For example, we observed that those who had not gone to the hospital (control group) had an average health score of 3.93, and those who had been to the hospital (treatment group) had an average health score of 3.21. The estimated effect of the treatment is $(\overline{y}_1 - \overline{y}_0) = 3.21 - 3.93 = -0.72$. The estimator bias in this case arises because the preexisting health conditions for the treated group, captured by $E\left(\overline{e}_1\right)$, are poorer than the pre-existing health of the control group, captured by $E\left(\overline{e}_0\right)$, so that in this example there is a negative bias in the difference estimator.

We can anticipate that anytime some individuals **select** treatment there will be factors leading to this choice that are systematically different from those leading individuals in the control group to not select treatment, resulting in a selection bias in the difference estimator. How can we eliminate the self-selection bias? The solution is to **randomly** assign individuals to treatment and control groups, so that there are no systematic differences between the groups, except for the treatment itself. With random assignment, and the use of a large number of experiment subjects, we can be sure that $E\left(\overline{e}_1\right) = E\left(\overline{e}_0\right)$ and $E\left(b_2\right) = \beta_2$.

EXAMPLE 7.8 | An Application of Difference Estimation: Project STAR

Medical researchers use white mice to test new drugs because these mice, surprisingly, are genetically similar to humans. Mice that are bred to be identical are randomly assigned to treatment and control groups, making estimation of the treatment effect of a new drug on the mice a relatively straightforward and reproducible process. Medical research on humans is strictly regulated, and volunteers are given incentives to participate, then randomly assigned to treatment and control groups. Randomized controlled experiments in the social sciences are equally attractive from a statistician's point of view but are rare because of the difficulties in organizing and funding them. A notable example of a randomized experiment is Tennessee's Project STAR.[15]

A longitudinal experiment was conducted in Tennessee beginning in 1985 and ending in 1989. A single cohort of students was followed from kindergarten through third grade. In the experiment, children were randomly assigned within schools into three types of classes: small classes with 13–17 students, regular-sized classes with 22–25 students, and regular-sized classes with a full-time teacher aide to assist the teacher. Student scores on achievement tests were recorded, as was some information about the students, teachers, and schools. Data for the kindergarten classes is contained in the data file *star*.

Let us first compare the performance of students in small classes versus regular classes.[16]

The variable *TOTALSCORE* is the combined reading and math achievement scores and *SMALL* = 1 if the student was assigned to a small class, and zero if the student is in a regular class. In Table 7.6a and b are summary statistics for the two types of classes. First, note that on all measures **except** *TOTALSCORE* the variable means reported are very

[15] See https://dataverse.harvard.eduldataset.xhtml?persistentld=hdl: 1902.1/10766 for program description, public use data and extensive literature.

[16] Interestingly there is no significant difference in outcomes comparing a regular class to a regular class with an aide. For this example all observations for students in the third treatment group are dropped.

similar. This is because students and teachers were randomly assigned to the classes, so that there should be no patterns evident. The average value of *TOTALSCORE* in the regular classes is 918.0429 and in small classes it is 931.9419, a difference of 13.899 points. The test scores are higher in the smaller classes. The difference estimator obtain using regression will yield the same estimate, along with significance levels.

TABLE 7.6a Summary Statistics for Regular-Sized Classes

Variable	Mean	Std. Dev.	Min	Max
TOTALSCORE	918.0429	73.1380	635	1229
SMALL	0.0000	0.0000	0	0
TCHEXPER	9.0683	5.7244	0	24
BOY	0.5132	0.4999	0	1
FREELUNCH	0.4738	0.4994	0	1
WHITE_ASIAN	0.6813	0.4661	0	1
TCHWHITE	0.7980	0.4016	0	1
TCHMASTERS	0.3651	0.4816	0	1
SCHURBAN	0.3012	0.4589	0	1
SCHRURAL	0.4998	0.5001	0	1

$N = 2005$

TABLE 7.6b Summary Statistics for Small Classes

Variable	Mean	Std. Dev.	Min	Max
TOTALSCORE	931.9419	76.3586	747	1253
SMALL	1.0000	0.0000	1	1
TCHEXPER	8.9954	5.7316	0	27
BOY	0.5150	0.4999	0	1
FREELUNCH	0.4718	0.4993	0	1
WHITE_ASIAN	0.6847	0.4648	0	1
TCHWHITE	0.8625	0.3445	0	1
TCHMASTERS	0.3176	0.4657	0	1
SCHURBAN	0.3061	0.4610	0	1
SCHRURAL	0.4626	0.4987	0	1

$N = 1738$

The model of interest is

$$TOTALSCORE = \beta_1 + \beta_2 SMALL + e \tag{7.15}$$

The regression results are in column (1) of Table 7.7. The estimated "treatment effect" of putting kindergarten children into small classes is 13.899 points, the same as the difference in sample means computed above, on their achievement score total; the difference is statistically significant at the 0.01 level.

EXAMPLE 7.9 | The Difference Estimator with Additional Controls

Because of the random assignment of the students to treatment and control groups, there is no selection bias in the estimate of the treatment effect. However, if additional factors might affect the outcome variable, they can be included in the regression specification. For example, it is possible that a teacher's experience leads to greater learning and higher achievement test scores. Adding *TCHEXPER* to the base model, we obtain

$$TOTALSCORE = \beta_1 + \beta_2 SMALL + \beta_3 TCHEXPER + e \tag{7.16}$$

The least squares estimates of (7.16) are in column (2) of Table 7.7. We estimate that each additional year of teaching experience increases the test score performance by 1.156 points, which is statistically significant at the 0.01 level. This increases our understanding of the effect of small classes. The results show that the effect of small classes is the same as the effect of approximately 12 years of teaching experience.

Note that adding *TCHEXPER* to the regression changed the estimate of the effect of *SMALL* classes very little. This is exactly what we would expect if *TCHEXPER* is uncorrelated with *SMALL*. The simple correlation between *SMALL* and *TCHEXPER* is only −0.0064. Recall that omitting a variable that is uncorrelated with an included variable does not change the estimated coefficient of the included variable. Comparing the models in columns (1) and (2) of Table 7.7, the model in (1) omits the significant variable *TCHEXPER*, but there is little change in the estimate of β_2 introduced by omitting this nearly uncorrelated variable. Furthermore, we can expect, in general, to obtain an estimator with smaller standard errors if we are able to include additional controls. In (7.15), any and all factors other than small class size are included in the error term. By taking some of those factors out of the error term and including them in the regression, the variance of the error term σ^2 is reduced, which reduces estimator variance.

TABLE 7.7 **Project STAR: Kindergarten**

	(1)	(2)	(3)	(4)
C	918.0429***	907.5643***	917.0684***	908.7865***
	(1.6672)	(2.5424)	(1.4948)	(2.5323)
SMALL	13.8990***	13.9833***	15.9978***	16.0656***
	(2.4466)	(2.4373)	(2.2228)	(2.2183)
TCHEXPER		1.1555***		0.9132***
		(0.2123)		(0.2256)
SCHOOL EFFECTS	*No*	*No*	*Yes*	*Yes*
N	3743	3743	3743	3743
adj. R^2	0.008	0.016	0.221	0.225
SSE	20847551	20683680	16028908	15957534

Standard errors in parentheses
Two-tail *p*-values: $^*p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

EXAMPLE 7.10 | The Difference Estimator with Fixed Effects

It may be that assignment to treatment groups is related to one or more observable characteristics. That is, treatments are randomly assigned *given* an external factor. Prior to a medical experiment concerning weight loss, participants may fall into the "overweight" category and the "obese" category. Of those in the overweight group 30% are randomly assigned for treatment, and of the obese group 50% are randomly assigned for treatment. Given pretreatment status, the treatment is randomly assigned. If such conditioning factors are omitted and put into the error term in (7.15) or (7.16), then these factors are correlated with the treatment variable and the least squares estimator of the treatment effect is biased and inconsistent. The way to adjust to "conditional" randomization is to include the conditioning factors into the regression.

In the STAR data, another factor that we might consider affecting the outcome is the school itself. The students were randomized *within* schools (conditional randomization), but not *across* schools. Some schools may be located in wealthier school districts that can pay higher salaries, thus attracting better teachers. The students in our sample are enrolled in 79 different schools. One way to account for school effects is to include an indicator variable for each school. That is, we can introduce 78 new indicators:

$$SCHOOL\_j = \begin{cases} 1 & \text{if student is in school } j \\ 0 & \text{otherwise} \end{cases}$$

This is an "intercept" indicator variable, allowing the expected total score to differ for each school. The model including these indicator variables is

$$TOTALSCORE_i = \beta_1 + \beta_2 SMALL_i + \beta_3 TCHEXPER_i + \textstyle\sum_{j=2}^{79} \delta_j SCHOOL\_j_i + e_i \qquad (7.17)$$

The regression function for a student in school *j* is

$$E\left(TOTALSCORE_i|\mathbf{X}\right) = \begin{cases} \left(\beta_1 + \delta_j\right) + \beta_3 TCHEXPER_i & \text{student in regular class} \\ \left(\beta_1 + \delta_j + \beta_2\right) + \beta_3 TCHEXPER_i & \text{student in small class} \end{cases}$$

Here **X** represents the variables *SMALL*, *TCHEXPER*, and all the indicator variables *SCHOOL_j*. The expected score for a student in a regular class for a teacher with no experience is adjusted by the fixed amount δ_j. This **fixed effect** controls for some differences in the schools that are not accounted for by the regression model.

Columns (3) and (4) in Table 7.7 contain the estimated coefficients of interest but not the 78 indicator variable coefficients. The joint *F*-test of the hypothesis that all $\delta_j = 0$ consists of $J = 78$ hypotheses with $N - K = 3662$ degrees of freedom. The *F*-value = 14.118 is significant at the 0.001 level. We conclude that there are statistically significant individual differences among schools. The important coefficients on *SMALL* and *TCHEXPER* change a little. The estimated effect of being in a small class increases to 16.0656 achievement test points in model (4), as compared to 13.9833 points in the corresponding model (2). It appears that some effect of small classes was masked by unincorporated individual school differences. This effect is small, however, as the 95% interval estimate for the coefficient of *SMALL* [11.7165, 20.4148] in model (4) includes 13.9833. Similarly, the estimated effect of teacher experience is slightly different in the models with and without the school fixed effects.

EXAMPLE 7.11 | Linear Probability Model Check of Random Assignment

In Table 7.6a and b, we examined the summary statistics for the data sorted by whether pupils were in a regular class or a small class. Except for *TOTALSCORE*, we did not find much difference in the sample means of the variables examined. Another way to check for random assignment is to regress *SMALL* on these characteristics and check for any significant coefficients, or an overall significant relationship. If there is random assignment, we should not find any significant relationships. Because *SMALL* is an indicator variable, we use the linear probability model discussed in Section 7.4. The estimated linear probability model is

$$\begin{array}{llll} \widehat{SMALL} = 0.4665 + & 0.0014BOY + & 0.0044WHITE\_ASIAN & \\ (t) & (0.09) & (0.22) & \\ & -\ 0.0006TCHEXPER & -\ 0.0009FREELUNCH & \\ & (-0.42) & (-0.05) & \end{array}$$

First, note that none of the right-hand-side variables are statistically significant. Second, the overall F-statistic for this linear probability model is 0.06 with a $p = 0.99$. There is no evidence that students were assigned to small classes based on any of these criteria. Also, recall that the linear probability model is so named because $E(SMALL|\mathbf{X})$ is the probability of observing $SMALL = 1$ in a random draw from the population. If the coefficients of all the potential explanatory factors are zero, the estimated intercept gives the estimated probability of observing a child in a small class to be 0.4665, with 95% interval estimate [0.4171, 0.5158]. We cannot reject the null hypothesis that the intercept equals 0.5, which is what it should be if students are allocated by a "flip" of a coin. *The importance of this, again, is that by randomly assigning students to small classes we can estimate the "treatment" effect using the simple difference estimator in* (7.15). The ability to isolate the important class size effect is a powerful argument in favor of randomized controlled experiments.

7.5.3 The Differences-in-Differences Estimator

Randomized controlled experiments are somewhat rare in economics because they are expensive and involve human subjects. **Natural experiments**, also called **quasi-experiments**, rely on observing real-world conditions that approximate what would happen in a randomized controlled experiment. Treatment appears *as if* it were randomly assigned. In this section, we consider estimating treatment effects using "before and after" data.

Suppose that we observe two groups before and after a policy change, with the **treatment group** being affected by the policy, and the **control group** being unaffected by the policy. Using such data, we will examine any change that occurs to the control group and compare it to the change in the treatment group.

The analysis is explained by Figure 7.3. The outcome variable y might be an employment rate, a wage rate, a price, or so on. Before the policy change we observe the treatment group value $y = B$, and after the policy is implemented the treatment group value is $y = C$. Using only the data on the treatment group we cannot separate out the portion of the change from $y = B$ to $y = C$ that is due to the policy from the portion that is due to other factors that may affect the outcome. We say that the treatment effect is not "identified."

We can isolate the effect of the treatment by using a control group that is not affected by the policy change. Before the policy change, we observe the control group value $y = A$, and after the policy change, the control group value is $y = E$. In order to estimate the treatment effect using the four pieces of information contained in the points A, B, C, and E, we make the strong assumption that the two groups experience a **common trend**. In Figure 7.3, the dashed line $\overline{BD}$ represents what we imagine the treatment group growth would have been (the term **counterfactual** from psychology is sometimes used to describe this imagined outcome) in the absence of the policy change. The growth described by the dashed line $\overline{BD}$ is unobservable, and is obtained by assuming that the growth in the treatment group that is unrelated to the policy change is the same as the growth in the control group.

The treatment effect $\delta = \overline{CD}$ is the difference between the treatment and control values of y in the "after" period, after subtracting $\overline{DE}$, which is what the difference between the two groups would have been in the absence of the policy. Using the common growth assumption, the difference $\overline{DE}$ equals the initial difference $\overline{AB}$. Using the four observable points A, B, C, and E depicted in Figure 7.3, estimation of the treatment effect is based on data averages for the two groups in the two periods,

$$\begin{aligned}\hat{\delta} &= \left(\hat{C} - \hat{E}\right) - \left(\hat{B} - \hat{A}\right)\\ &= \left(\bar{y}_{Treatment,\ After} - \bar{y}_{Control,\ After}\right) - \left(\bar{y}_{Treatment,\ Before} - \bar{y}_{Control,\ Before}\right)\end{aligned} \qquad (7.18)$$

In (7.18), the sample means are

$$\begin{aligned}\bar{y}_{Control,\ Before} &= \hat{A} = \text{sample mean of } y \text{ for control group before policy implementation}\\ \bar{y}_{Treatment,\ Before} &= \hat{B} = \text{sample mean of } y \text{ for treatment group before policy implementation}\\ \bar{y}_{Control,\ After} &= \hat{E} = \text{sample mean of } y \text{ for control group after policy implementation}\\ \bar{y}_{Treatment,\ After} &= \hat{C} = \text{sample mean of } y \text{ for treatment group after policy implementation}\end{aligned}$$

The estimator $\hat{\delta}$ is called a **differences-in-differences** (abbreviated as D-in-D, DD, or DID) estimator of the treatment effect.

The estimator $\hat{\delta}$ can be conveniently calculated using a simple regression. Define y_{it} to be the observed outcome for individual i in period t. Let $AFTER_t$ be an indicator variable that equals one in the period after the policy change ($t = 2$) and zero in the period before the policy change ($t = 1$). Let $TREAT_i$ be a dummy variable that equals one if individual i is in the treatment group and zero if the individual is in the control group. Consider the regression model

$$y_{it} = \beta_1 + \beta_2 TREAT_i + \beta_3 AFTER_t + \delta\left(TREAT_i \times AFTER_t\right) + e_{it} \qquad (7.19)$$

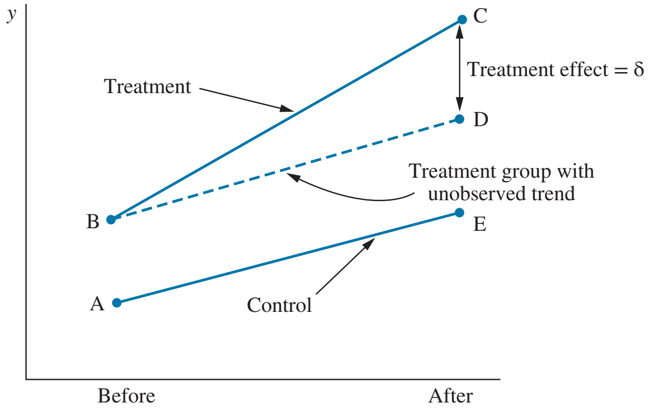

FIGURE 7.3 **Difference-in-Differences Estimation.**

The regression function is

$$E(y_{it}|\mathbf{X}) = \begin{cases} \beta_1 & TREAT = 0,\ AFTER = 0 \text{ [Control before = A]} \\ \beta_1 + \beta_2 & TREAT = 1,\ AFTER = 0 \text{ [Treatment before = B]} \\ \beta_1 + \beta_3 & TREAT = 0,\ AFTER = 1 \text{ [Control after = E]} \\ \beta_1 + \beta_2 + \beta_3 + \delta & TREAT = 1,\ AFTER = 1 \text{ [Treatment after = C]} \end{cases}$$

Here **X** contains the variables on the right-hand side of equation (7.19). In Figure 7.3, points $\text{A} = \beta_1$, $\text{B} = \beta_1 + \beta_2$, $\text{E} = \beta_1 + \beta_3$ and $\text{C} = \beta_1 + \beta_2 + \beta_3 + \delta$. Then

$$\begin{aligned} \delta &= (\text{C} - \text{E}) - (\text{B} - \text{A}) \\ &= \left[\left(\beta_1 + \beta_2 + \beta_3 + \delta\right) - \left(\beta_1 + \beta_3\right)\right] - \left[\left(\beta_1 + \beta_2\right) - \beta_1\right] \end{aligned}$$

Using this the least squares estimates b_1, b_2, b_3 and $\hat{\delta}$ from (7.19), we have

$$\begin{aligned} \hat{\delta} &= \left[\left(b_1 + b_2 + b_3 + \hat{\delta}\right) - \left(b_1 + b_3\right)\right] - \left[\left(b_1 + b_2\right) - b_1\right] \\ &= \left(\overline{y}_{Treatment,\ After} - \overline{y}_{Control,\ After}\right) - \left(\overline{y}_{Treatment,Before} - \overline{y}_{Control,Before}\right) \end{aligned}$$

EXAMPLE 7.12 | Estimating the Effect of a Minimum Wage Change: The DID Estimator

Card and Krueger (1994)[17] provide an example of a natural experiment and the **differences-in-differences estimator**. On April 1, 1992, New Jersey's minimum wage was increased from \$4.25 to \$5.05 per hour, while the minimum wage in Pennsylvania stayed at \$4.25 per hour. Card and Krueger collected data on 410 fast-food restaurants in New Jersey (the treatment group) and eastern Pennsylvania (the control group). These two groups are similar economically and close geographically, separated by only a river with multiple bridges. The "before" period is February 1992, and the "after" period is November 1992. Using these data, they estimate the effect of the "treatment," raising the New Jersey minimum wage on employment at fast-food restaurants in New Jersey. Their interesting finding, that there was no significant reduction[18] in employment, sparked a great debate and much further research.[19] In model (7.19), we will test the null and alternative hypotheses

$$H_0: \delta \geq 0 \text{ versus } H_1: \delta < 0 \tag{7.20}$$

The relevant Card and Krueger data is in the data file *njmin3*. We use the sample means of *FTE*, the number of full-time equivalent[20] employees, given in Table 7.8, to estimate the treatment effect δ using the differences-in-differences estimator.

TABLE 7.8 Full-time Equivalent Employees by State and Period

Variable	*N*	Mean	se
Pennsylvania (PA)			
Before	77	23.3312	1.3511
After	77	21.1656	0.9432
New Jersey (NJ)			
Before	321	20.4394	0.5083
After	319	21.0274	0.5203

In Pennsylvania, the control group, employment fell during the period February to November. Recall that the

[17] David Card and Alan Krueger (1994) "Minimum Wages and Employment: A Case Study of the Fast Food Industry in New Jersey and Pennsylvania," *The American Economic Review*, 84, 316–361. We thank David Card for letting us use the data.

[18] Remember that failure to reject a null hypothesis does not make it true!

[19] The issue is hotly contested and the literature extensive. See, for example, http://en.wikipedia.org/wiki/Minimum_wage, and the references listed, as a starting point.

[20] Card and Krueger calculate FTE = 0.5 × number of part time workers + number of full time workers + number of managers.

minimum wage level was changed in New Jersey, but not in Pennsylvania, so that employment levels in Pennsylvania were not affected. In New Jersey we see an increase in *FTE* in the same period. The differences-in-differences estimate of the change in employment due to the change in the minimum wage is

$$\begin{aligned}\hat{\delta} &= \left(\overline{FTE}_{NJ,After} - \overline{FTE}_{PA,After}\right) - \left(\overline{FTE}_{NJ,Before} - \overline{FTE}_{PA,Before}\right)\\ &= (21.0274 - 21.1656) - (20.4394 - 23.3312)\\ &= 2.7536 \qquad (7.21)\end{aligned}$$

We estimate that *FTE* employment increased by 2.75 employees during the period in which the New Jersey minimum wage was increased. This positive effect is contrary to what is predicted by economic theory.

Rather than compute the differences-in-differences estimate using sample means, it is easier and more general to use the regression format. In (7.19) let $y = FTE$ employment, the treatment variable is the indicator variable $NJ = 1$ if observation is from New Jersey, and zero if from Pennsylvania. The time indicator is $D = 1$ if the observation is from November and zero if it is from February. The differences-in-differences regression is then

$$FTE_{it} = \beta_1 + \beta_2 NJ_i + \beta_3 D_t + \delta\left(NJ_i \times D_t\right) + e_{it} \qquad (7.22)$$

Using the 794 complete observations in the file *njmin3*, the least squares estimates are reported in column (1) of Table 7.9. At the $\alpha = 0.05$ level of significance the rejection region for the left-tail test in (7.20) is $t \le -1.645$, so we fail to reject the null hypothesis. We cannot conclude that the increase in the minimum wage in New Jersey reduced employment at New Jersey fast-food restaurants.

As with randomized control experiments, it is interesting to see the robustness of these results. In Table 7.9 column (2), we add indicator variables for fast-food chain and whether the restaurant was company-owned rather than franchise-owned. In column (3) we add indicator variables for geographical regions within the survey area. None of these changes alter the differences-in-differences estimate, and none lead to rejection of the null hypothesis in (7.20).

TABLE 7.9 Difference-in-Differences Regressions

	(1)	(2)	(3)
C	23.3312***	25.9512***	25.3205***
	(1.072)	(1.038)	(1.211)
NJ	−2.8918*	−2.3766*	−0.9080
	(1.194)	(1.079)	(1.272)
D	−2.1656	−2.2236	−2.2119
	(1.516)	(1.368)	(1.349)
D_NJ	2.7536	2.8451	2.8149
	(1.688)	(1.523)	(1.502)
KFC		−10.4534***	−10.0580***
		(0.849)	(0.845)
ROYS		−1.6250	−1.6934*
		(0.860)	(0.859)
WENDYS		−1.0637	−1.0650
		(0.929)	(0.921)
CO_OWNED		−1.1685	−0.7163
		(0.716)	(0.719)
SOUTHJ			−3.7018***
			(0.780)
CENTRALJ			0.0079
			(0.897)
PA1			0.9239
			(1.385)
N	794	794	794
R^2	0.007	0.196	0.221
adj. R^2	0.004	0.189	0.211

Standard errors in parentheses
Two-tail *p*-values: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

EXAMPLE 7.13 | Estimating the Effect of a Minimum Wage Change: Using Panel Data

In the previous section's differences-in-differences analysis, we did not exploit one very important feature of Card and Krueger's data—namely, that the *same* fast-food restaurants were observed on two occasions. We have "before" and "after" data on 384 of the 410 restaurants. These are called **paired data** observations, or **repeat data** observations, or **panel data** observations. In Chapter 1 we introduced the notion of a **panel** of data—we observe the same individual-level units over several periods. The Card and Krueger data includes $T = 2$ observations on $N = 384$ individual restaurants among the 410 restaurants surveyed. The remaining 26 restaurants had missing data on *FTE* either in the "before" or "after" period. There are powerful advantages to using panel data, some of which we will describe here. See Chapter 15 for a much more extensive discussion.

Using panel data, we can control for **unobserved individual-specific characteristics**. There are characteristics of the restaurants that we do not observe. Some restaurants will have preferred locations, some may have superior managers, and so on. These unobserved individual specific characteristics are included in the error term of the regression (7.22). Let c_i denote any unobserved characteristics of individual restaurant i that do not change over time. Adding c_i to (7.22), we have

$$FTE_{it} = \beta_1 + \beta_2 NJ_i + \beta_3 D_t + \delta(NJ_i \times D_t) + c_i + e_{it} \quad (7.23)$$

Whatever c_i might be, it contaminates this regression model. A solution is at hand *if* we have a panel of data. If we have $T = 2$ repeat observations, we can *eliminate* c_i by analyzing the changes in FTE from period one to period two. Recall that $D_t = 0$ in period one, so $D_1 = 0$; and $D_t = 1$ in period two, so $D_2 = 1$. Subtract the observation for $t = 1$ from that for $t = 2$

$$\begin{array}{c} FTE_{i2} = \beta_1 + \beta_2 NJ_i + \beta_3 1 + \delta(NJ_i \times 1) + c_i + e_{i2} \\ \underline{-(FTE_{i1} = \beta_1 + \beta_2 NJ_i + \beta_3 0 + \delta(NJ_i \times 0) + c_i + e_{i1})} \\ \Delta FTE_i = \beta_3 + \delta NJ_i + \Delta e_i \end{array}$$

where $\Delta FTE_i = FTE_{i2} - FTE_{i1}$ and $\Delta e_i = e_{i2} - e_{i1}$. Using the **differenced data**, the regression model of interest becomes

$$\Delta FTE_i = \beta_3 + \delta NJ_i + \Delta e_i \quad (7.24)$$

Observe that the contaminating factor c_i has dropped out! Whatever those unobservable features might have been, they are now gone. The intercept β_1 and the coefficient β_2 have also dropped out, with the parameter β_3 becoming the new intercept. The most important parameter, δ, measuring the treatment effect is the coefficient of the indicator variable NJ_i, which identifies the treatment (New Jersey) and control group (Pennsylvania) observations.

The estimated model (7.24) is

$$\begin{array}{llll} \widehat{\Delta FTE} & = -2.2833 & + \ 2.7500 NJ & R^2 = 0.0146 \\ \text{(se)} & \quad (1.036) & \quad (1.154) & \end{array}$$

The estimate of the treatment effect $\hat{\delta} = 2.75$ using the differenced data, which accounts for any unobserved individual differences, is very close to the differences-in-differences estimate. Once again we fail to conclude that the minimum wage increase has reduced employment in these New Jersey fast-food restaurants.

7.6 Treatment Effects and Causal Modeling

In Section 7.5, we provided the basics of treatment effect models. In this section, we present extensions and enhancements using the framework of **potential outcomes**, sometimes called the **Rubin Causal Model (RCM)**, in recognition of Donald B. Rubin who formulated this approach.[21]

7.6.1 The Nature of Causal Effects

Economists are interested in causal relationships between variables. **Causality**, or causation, means that a change in one variable is the direct consequence of a change in another variable. For example, if you receive an hourly wage rate, then increasing your work hours (the cause) will lead to an increase in your income (the effect). Another example is from the standard supply and demand model for a normal good. If consumer incomes rise (the cause), demand increases, and there is a subsequent increase in the market price and quantities bought and sold (the effect).

A cause must precede, or be contemporaneous with, the effect. The confusion between correlation and causation is widespread, and correlation does not imply causation. We observe many associations between variables that are not causal. The correlation between the divorce rate in

[21] The literature in this area has grown dramatically in recent years, and continues to grow. In this section we draw heavily on a survey by Guido W. Imbens and Jeffrey M. Wooldridge (2009) "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47(1), 5–86, Jeffrey M. Wooldridge (2010) *Econometric Analysis of Cross Section and Panel Data, Second Edition*, MIT Press, Chapter 21; and Joshua D. Angrist and Jörn-Steffen Pischke (2009) *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press. These references are advanced. See also Joshua D. Angrist and Jörn-Steffen Pischke (2015) *Mastering Metrics: The Path from Cause to Effect*, Princeton University Press.

the state of Maine and the U.S. per capita consumption of margarine is 0.9926[22] over the period 2000–2009. We doubt that this high correlation is a causal relationship. Not all confusions, or spurious correlations, are amusing and harmless. There is a concern among some parents about the relationship between childhood vaccinations and subsequent negative health outcomes, such as autism. Despite intense study by the U.S. Centers for Disease Control and Prevention (CDC), finding no causal relationship, there has been a movement among parents to not have some vaccinations for their children, resulting in concern by health officials that some childhood diseases will make a widespread comeback.

7.6.2 Treatment Effect Models

Treatment effect models seek to estimate a causal effect. Let the treatment, which might be an individual receiving a new drug, or some additional job training, be denoted as $d_i = 1$, whereas not receiving the treatment is $d_i = 0$. The outcome of interest might be a cholesterol level if the treatment is a new drug. If the treatment is job training, the outcome might be a worker's performance on completing a particular task. For each individual there are two possible, or potential, outcomes, y_{1i} if an individual receives treatment $\left(d_i = 1\right)$, and y_{0i} if the individual does not receive treatment $\left(d_i = 0\right)$. We would like to know the **causal effect** $y_{1i} - y_{0i}$, the difference in the outcome for individual i if they receive the treatment versus if they do not. An advantage of the potential outcomes framework is that it forces us to recognize that the treatment effect varies across individuals—it is individual specific. The difficulty is that we never observe both y_{1i} and y_{0i}. We only observe one or the other. The outcome we observe is

$$y_i = \begin{cases} y_{1i} & \text{if } d_i = 1 \\ y_{0i} & \text{if } d_i = 0 \end{cases} \tag{7.25}$$

Written another way, what we observe is

$$y_i = y_{1i}d_i + y_{0i}\left(1 - d_i\right) = y_{0i} + \left(y_{1i} - y_{0i}\right) d_i \tag{7.26}$$

Instead of being able to estimate $y_{1i} - y_{0i}$ for each individual, what we are able to estimate is the population **average treatment effect (ATE)**, $\tau_{ATE} = E\left(y_{1i} - y_{0i}\right)$. To see this, express the difference between the conditional expectation of y_i, the outcome we actually observe, for those who receive treatment, $\left(d_i = 1\right)$, and those who do not, $\left(d_i = 0\right)$;

$$E\left(y_i|d_i = 1\right) - E\left(y_i|d_i = 0\right) = E\left(y_{1i}|d_i = 1\right) - E\left(y_{0i}|d_i = 0\right) \tag{7.27}$$

In a randomized, controlled experiment, individuals are randomly selected from the population and then randomly assigned to a group receiving the treatment (the **treatment group**), for whom $\left(d_i = 1\right)$, or to a group not receiving the treatment (the **control group**), for whom $(d_i = 0)$. In this way the treatment, d_i, is statistically independent of the potential outcomes y_{1i} and y_{0i} so that

$$\begin{aligned} E\left(y_i|d_i = 1\right) - E\left(y_i|d_i = 0\right) &= E\left(y_{1i}|d_i = 1\right) - E\left(y_{0i}|d_i = 0\right) \\ &= E\left(y_{1i}\right) - E\left(y_{0i}\right) = E\left(y_{1i} - y_{0i}\right) \\ &= \tau_{ATE} \end{aligned} \tag{7.28}$$

From the first line to the second we use the fact that if two random variables, say X and Y, are statistically independent,[23] then $E(Y|X = x) = E(Y)$. To see that this is true, suppose X and Y are discrete random variables. Then

$$E(Y) = \sum yP(Y = y) \text{ and } E(Y|X = x) = \sum yP(Y = y|X = x)$$

[22] http://www.tylervigen.com/

[23] We revert to the notation from the probability primer here, with upper case Y and X being random variables and lower case y and x being values of the random variables.

If X and Y are statistically independent, then

$$P(Y = y|X = x) = P(Y = y)$$

so that

$$E(Y|X = x) = \sum yP(Y = y|X = x) = \sum yP(Y = y) = E(Y)$$

If we randomly choose population members and randomly assign them to the treatment and control groups, then treatment, d_i, is statistically independent of the potential outcomes of the experiment. An unbiased estimator of $E(y_i|d_i = 1)$ is the sample mean of the N_1 outcomes for the treatment group, $\bar{y}_1 = \sum_{i=1}^{N_1} y_{1i}/N_1$. An unbiased estimator of $E(y_i|d_i = 0)$ is the sample mean of the N_0 outcomes for the control group, $\bar{y}_0 = \sum_{i=1}^{N_0} y_{0i}/N_0$. An unbiased estimator of the population average treatment effect is $\hat{\tau}_{ATE} = \bar{y}_1 - \bar{y}_0$. This is the **difference estimator** in equation (7.14). That is, we can obtain the estimator of the average treatment effect from the simple regression $y_i = \alpha + \tau_{ATE}d_i + e_i$ using all $N = N_0 + N_1$ observations.

7.6.3 Decomposing the Treatment Effect

Using equation (7.27) $\left[E(y_i|d_i = 1) - E(y_i|d_i = 0) = E(y_{1i}|d_i = 1) - E(y_{0i}|d_i = 0)\right]$, we can gain additional insight into the simple regression $y_i = \alpha + \tau_{ATE}d_i + e_i$. Add and subtract $E(y_{0i}|d_i = 1)$ to the right-hand side, and rearrange to obtain

$$\begin{aligned} E(y_i|d_i = 1) - E(y_i|d_i = 0) = {} & \left[E(y_{1i}|d_i = 1) - E(y_{0i}|d_i = 1)\right] \\ & + \left[E(y_{0i}|d_i = 1) - E(y_{0i}|d_i = 0)\right] \end{aligned} \tag{7.29}$$

The left-hand side is the difference in average outcomes for the treatment group $(d_i = 1)$ and the control group $(d_i = 0)$. The difference $\left[E(y_{1i}|d_i = 1) - E(y_{0i}|d_i = 1)\right]$ is average difference in potential outcomes for those who received the treatment, or as called in this literature, the **average treatment effect on the treated (ATT)**, which we denote by τ_{ATT}. The second term $E(y_{0i}|d_i = 1) - E(y_{0i}|d_i = 0)$ is the average potential outcome for those in the treatment group should they not receive treatment minus the average outcome for those in the control group. If individuals are truly randomly assigned to treatment and control groups $E(y_{0i}|d_i = 1) - E(y_{0i}|d_i = 0)$ will be zero, meaning that there are no differences between the expected potential outcomes for the treatment and control groups if they had remained untreated. In this case, the treatment effect $\tau_{ATE} = E(y_i|d_i = 1) - E(y_i|d_i = 0)$ equals $\tau_{ATT} = E(y_{1i}|d_i = 1) - E(y_{0i}|d_i = 1)$, the average treatment effect on the treated.

In equation (7.29), if the second term in brackets is not zero, or $E(y_{0i}|d_i = 1) - E(y_{0i}|d_i = 0) \neq 0$, then there is **selection bias**. It means that individuals are not randomly assigned to the treatment and control groups because the average of the potential outcomes if untreated, y_{0i}, in the treatment and control groups are different. If the treatment is receiving a new drug, there is selection bias if (i) a screener looks at a randomly chosen person and thinks "This person looks sickly and could use this drug, so I'll assign him to the treatment group;" or (ii) a person thinks the treatment might be good for him, and manages to be added to the treatment group. Either way, there is a difference in the average untreated health y_{0i} of the treatment and control groups. The term $E(y_{0i}|d_i = 1) - E(y_{0i}|d_i = 0)$ is called **selection bias** for this reason. Random assignment of individuals to treatment and control groups eliminates selection bias. If there is selection bias, then the difference estimator $\hat{\tau}_{ATE} = \bar{y}_1 - \bar{y}_0$ is not an unbiased estimator of the average treatment effect, and the average treatment effect is not the average treatment effect on the treated.

To summarize, in a randomized experiment the treatment indicator d_i is statistically independent of the potential outcomes y_{0i} and y_{1i}. We do not observe both potential outcomes but rather $y_i = y_{0i} + (y_{1i} - y_{0i})d_i$. If treatment d_i is statistically independent of the potential outcomes, then

$$\tau_{ATE} = \tau_{ATT} = E(y_i|d_i = 1) - E(y_i|d_i = 0) \tag{7.30}$$

and an unbiased estimator is

$$\hat{\tau}_{ATE} = \hat{\tau}_{ATT} = \bar{y}_1 - \bar{y}_0 \tag{7.31}$$

The equality $\tau_{ATE} = \tau_{ATT}$ actually holds under a weaker assumption than statistical independence. From (7.29)

$$\tau_{ATE} = \tau_{ATT} + E(y_{0i}|d_i = 1) - E(y_{0i}|d_i = 0) \tag{7.32}$$

The selection bias term $E(y_{0i}|d_i = 1) - E(y_{0i}|d_i = 0) = 0$ if $E(y_{0i}|d_i = 1) = E(y_{i0})$ and $E(y_{0i}|d_i = 0) = E(y_{i0})$. This is called the **conditional independence assumption (CIA)**, or **conditional mean independence**. While this is a less stringent condition than statistical independence between the treatment and the potential outcomes, it is still strong. It suggests that being in the treatment or control group is unrelated to the average outcome for the untreated.

7.6.4 Introducing Control Variables

A **control variable**, x_i, is not the object of interest in a study. It is included in the model to hold constant factors that, if neglected, would lead to selection bias. See Section 6.3.4. In treatment effect models, control variables are introduced in order to allow unbiased estimation of the treatment effect when the potential outcomes, y_{0i} and y_{1i}, might be correlated with the treatment variable, d_i. Ideally, by conditioning on a control variable x_i the treatment becomes "as good as" randomized, allowing us to estimate the average causal or treatment effect. We consider only a single control variable to simplify our presentation. The methods discussed as follows carry over to the case with multiple control variables. The key is an extension of the **conditional independence assumption**,[24]

$$E(y_{0i}|d_i, x_i) = E(y_{0i}|x_i) \quad \text{and} \quad E(y_{1i}|d_i, x_i) = E(y_{1i}|x_i) \tag{7.33}$$

Once we condition on the control variables, then the expected potential outcomes do not depend the treatment. In a sense, having good control variables *is as good as* having a randomized controlled experiment. Good control variables have the feature of being "predetermined" in the sense that they are fixed, and given, at the time the treatment is assigned. Enough control variables should be added so that the conditional independence assumption holds. Avoid "bad control" variables that might be outcomes of the treatment.

When potential outcomes depend on x_i, then the average treatment effect depends on x_i, and is

$$\tau_{ATE}(x_i) = E(y_{1i}|d_i, x_i) - E(y_{0i}|d_i, x_i) = E(y_{1i}|x_i) - E(y_{0i}|x_i)$$

Assuming a linear regression structure for the expectations, and recalling that the observed outcome is $y_i = y_{0i} + (y_{1i} - y_{0i})\, d_i$, let

$$E(y_i|x_i, d_i = 0) = E(y_{i0}|x_i, d_i = 0) = E(y_{i0}|x_i) = \alpha_0 + \beta_0 x_i \tag{7.34a}$$

$$E(y_i|x_i, d_i = 1) = E(y_{i1}|x_i, d_i = 1) = E(y_{i1}|x_i) = \alpha_1 + \beta_1 x_i \tag{7.34b}$$

The treatment effect is the difference between equations (7.34b) and (7.34a), or

$$\tau_{ATE}(x_i) = (\alpha_1 + \beta_1 x_i) - (\alpha_0 + \beta_0 x_i) = (\alpha_1 - \alpha_0) - (\beta_1 - \beta_0)x_i \tag{7.35}$$

Because $\tau_{ATE}(x_i)$ depends on x_i, the average treatment effect will be obtained by "averaging" over the population distribution of x_i. Recall from the probability primer that a "population average"

[24] This assumption has been called *unconfoundedness* and also *ignorability*. The literature on causal modeling spans several disciplines, and the terminology can be quite different in each. The following development follows Woodridge (2010, 919–920).

is an expected value. So we define the average treatment effect as $\tau_{ATE} = E_x\left[\tau_{ATE}(x_i)\right]$ where the subscript x on the expectation operator means that we are treating x as random.

In practice, we can estimate the regression functions separately on the treatment and control groups:

1. Obtain $\hat{\alpha}_0 + \hat{\beta}_0 x_i$ from a regression of y_i on x_i for the control group, $\left(d_i = 0\right)$
2. Obtain $\hat{\alpha}_1 + \hat{\beta}_1 x_i$ from a regression of y_i on x_i for the treatment group, $\left(d_i = 1\right)$

Then

$$\hat{\tau}_{ATE}(x_i) = \hat{\alpha}_1 + \hat{\beta}_1 x_i - \left(\hat{\alpha}_0 + \hat{\beta}_0 x_i\right) = \left(\hat{\alpha}_1 - \hat{\alpha}_0\right) + \left(\hat{\beta}_1 - \hat{\beta}_0\right)x_i \tag{7.36}$$

Averaging the estimated value across the sample values gives

$$\begin{aligned}\hat{\tau}_{ATE} &= N^{-1}\sum_{i=1}^{N}\hat{\tau}_{ATE}(x_i) = N^{-1}\sum_{i=1}^{N}\left[\left(\hat{\alpha}_1 - \hat{\alpha}_0\right) + \left(\hat{\beta}_1 - \hat{\beta}_0\right)x_i\right] \\ &= \left(\hat{\alpha}_1 - \hat{\alpha}_0\right) + \left(\hat{\beta}_1 - \hat{\beta}_0\right)\left(N^{-1}\sum_{i=1}^{N}x_i\right) \\ &= \left(\hat{\alpha}_1 - \hat{\alpha}_0\right) + \left(\hat{\beta}_1 - \hat{\beta}_0\right)\overline{x} \end{aligned} \tag{7.37}$$

Using slope and intercept indicator variables, we can estimate the average treatment effect in a pooled regression, and calculate a standard error for the estimate $\hat{\tau}_{ATE}$. The pooled regression is

$$y_i = \alpha + \theta d_i + \beta x_i + \gamma\left(d_i x_i\right) + e_i \tag{7.38}$$

The regression functions for the treatment and control groups are

$$E\left(y_i | d_i, x_i\right) = \begin{cases} \alpha + \beta x_i & \text{if } d_i = 0 \\ (\alpha + \theta) + (\beta + \gamma)x_i & \text{if } d_i = 1 \end{cases} \tag{7.39}$$

In terms of the separate regression coefficients

$$\alpha = \alpha_0, \quad \beta = \beta_0, \quad \alpha + \theta = \alpha_1, \quad \text{and} \quad \beta + \gamma = \beta_1 \tag{7.40}$$

It follows that from the pooled regression (7.38) the estimates $\hat{\theta} = \hat{\alpha}_1 - \hat{\alpha}_0$ and $\hat{\gamma} = \hat{\beta}_1 - \hat{\beta}_0$. The relation of these estimates to $\hat{\tau}_{ATE}$ is

$$\hat{\theta} = \hat{\tau}_{ATE} - \overline{x}\left(\hat{\beta}_1 - \hat{\beta}_0\right) = \hat{\tau}_{ATE} - \overline{x}\hat{\gamma}$$

or

$$\hat{\tau}_{ATE} = \hat{\theta} + \overline{x}\hat{\gamma}$$

We can modify the pooled regression so that τ_{ATE} appears in the pooled regression. In the pooled regression (7.38) add and subtract the term $\gamma\left(d_i\overline{x}\right)$

$$\begin{aligned} y_i &= \alpha + \theta d_i + \beta x_i + \gamma\left(d_i x_i\right) + \left[\gamma d_i\overline{x} - \gamma d_i\overline{x}\right] + e_i \\ &= \alpha + \left(\theta + \gamma\overline{x}\right)d_i + \beta x_i + \gamma\left[d_i\left(x_i - \overline{x}\right)\right] + e_i \\ &= \alpha + \tau_{ATE}d_i + \beta x_i + \gamma\left(d_i\tilde{x}_i\right) + e_i \end{aligned} \tag{7.41}$$

Now the population average treatment effect τ_{ATE} is a parameter in the pooled regression. The term $\tilde{x}_i = \left(x_i - \overline{x}\right)$ is notation for deviations about the mean. By using least squares regression, we obtain $\hat{\tau}_{ATE}$. Your software will also report a standard error $\text{se}\left(\hat{\tau}_{ATE}\right)$.[25]

The average treatment effect in the population, $\tau_{ATE} = E\left(y_{1i} - y_{0i}\right)$, may not be the parameter of interest in some applications. By slightly modifying the pooled regression, we can obtain the

[25] Wooldridge (2010, p. 919) notes that the usual estimator of the standard error is not quite valid in this case because it ignores the additional variability added by including the sample mean in $\tilde{x}_i = \left(x_i - \overline{x}\right)$. One alternative to the usual standard error is to use the **bootstrap** standard error, discussed in Appendix 5B.5.

average treatment effect of a subpopulation. For example, how large is the average treatment effect on those who actually received treatment? The **average treatment effect on the treated**, τ_{ATT}, where the subscript ATT denotes the target group, is obtained by estimating the pooled regression

$$y_i = \alpha + \tau_{ATT} d_i + \beta x_i + \gamma\left(d_i \tilde{x}_{i1}\right) + e_i \tag{7.42}$$

where $\tilde{x}_{i1} = \left(x_i - \bar{x}_1\right)$ and $\bar{x}_1 = N_1^{-1}\sum_{i=1}^{N_1} x_i$ for the treatment group, where $d_i = 1$.

Similarly, we can restrict measurement of the treatment effect to other subpopulations of interest. For example, if we are considering the effects of a job training program, we may not want to include the extremely wealthy. We could specify the population of interest to be those with incomes in the lowest 25% of society. Denote this restricted group of interest by R and let $\tau_{ATE,\,R}$ be the average treatment effect on this group. Let $\tilde{x}_{iR} = \left(x_i - \bar{x}_R\right)$, where $\bar{x}_R = N_R^{-1}\sum_{i \in R} x_i$, with $i \in R$ indicating that we are restricting the sum to those individuals i falling in the target group, R, and N_R is the number of individuals in the sample satisfying the condition. Then we can estimate $\tau_{ATE,R}$ from the pooled regression

$$y_i = \alpha + \tau_{ATE,R} d_i + \beta x_i + \gamma\left(d_i \tilde{x}_{iR}\right) + e_i \tag{7.43}$$

7.6.5 The Overlap Assumption

The so-called **overlap assumption** must hold, in addition to the conditional independence assumption in equation (7.33). The overlap assumption says that for each value of x_i it must be possible to see an individual in the treatment and control groups, or $0 < P\left(d_i = 1|x_i\right) < 1$ and $0 < P\left(d_i = 0|x_i\right) = 1 - P\left(d_i = 1|x_i\right) < 1$. A rule of thumb is to compute the normalized difference

$$\frac{\bar{x}_1 - \bar{x}_0}{\left(s_1^2 + s_0^2\right)^{1/2}} \tag{7.44}$$

where s_1^2 and s_0^2 are the sample variances of the explanatory variable x for the treatment and control groups. If the normalized difference is greater in absolute value than 0.25,[26] then there is cause for concern. If the overlap assumption fails, then redefining the population of interest may be required. To see the impact of the difference of means, $\bar{x}_1 - \bar{x}_0$, on the average treatment effect, let $f_0 = N_0/N$ and $f_1 = N_1/N$ be the fractions of observations in the control and treatment groups, respectively. In Appendix 7C, we show that

$$\hat{\tau}_{ATE} = \left(\bar{y}_1 - \bar{y}_0\right) - \left(f_0\hat{\beta}_1 + f_1\hat{\beta}_0\right)\left(\bar{x}_1 - \bar{x}_0\right)$$

If the difference in the sample means of the treatment and control groups is large, the estimated slopes from the regressions in (7.34), $\hat{\beta}_1$ and $\hat{\beta}_0$, have a larger influence in the estimate $\hat{\tau}_{ATE}$ of the average treatment effect.

7.6.6 Regression Discontinuity Designs

Regression discontinuity (**RD**) **designs**[27] arise when the separation into treatment and control groups follows a deterministic rule, such as "Students receiving 75% or higher on the midterm exam will receive an award." How the award affects future academic outcomes might be the question of interest. The key insight about the RD designs is that that students receiving "close to

[26] Wooldridge (2010, p. 917)

[27] In this section we draw heavily on a survey by David S. Lee and Thomas Lemieux (2010) "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, 48(1), 5-86, Jeffrey M. Wooldridge (2010) *Econometric Analysis of Cross Section and Panel Data, Second Edition*, MIT Press, Chapter 21 and Joshua D. Angrist and Jörn-Steffen Pischke (2009) *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, Chapter 6. These references are advanced. See also Joshua D. Angrist and Jörn-Steffen Pischke (2015) *Mastering Metrics: The Path from Cause to Effect*, Princeton University Press, Chapter 4.

75%" are likely very similar in most regards (a condition that can be checked) so that those just below the cutoff point are a good comparison group for those just above the cutoff. Using individuals close to the cutoff is "just as good as" a random assignment, for the purpose of estimating a treatment effect.

Suppose that x_i is the single variable determining whether an individual is assigned to the treatment group or control group. In this literature, x_i is called the **forcing variable**. The treatment indicator variable $d_i = 1$ if $x_i \geq c$, where c is a preassigned cutoff value and $d_i = 0$ if $x_i < c$. This is said to be a **sharp regression discontinuity design** because the treatment is definitely given if the forcing variable crosses the threshold. The observed outcome is $y_i = (1 - d_i)y_{0i} + d_i y_{1i}$, where y_{0i} is the potential outcome for individual i when not receiving treatment and y_{1i} is the potential outcome for individual i when receiving the treatment. For the sharp RD design, the conditional independence assumption in equation (7.33)

$$E(y_{0i}|d_i, x_i) = E(y_{0i}|x_i) \quad \text{and} \quad E(y_{1i}|d_i, x_i) = E(y_{1i}|x_i)$$

is automatically satisfied because the treatment is completely determined by the forcing variable, x_i. Interestingly, the overlap assumption fails completely. For a given value of x_i, we cannot hope to observe individuals in both treatment and control groups. Rather than trying to estimate a population average treatment effect, in the RD design we estimate the treatment effect "at the cutoff,"

$$\tau_c = E(y_{1i} - y_{0i}|x_i = c) = E(y_{1i}|x_i = c) - E(y_{0i}|x_i = c) \tag{7.45}$$

One required assumption is "continuity." That is, $E(y_{1i}|x_i)$ and $E(y_{0i}|x_i)$ must meet smoothly at $x_i = c$ except for a "jump." The jump is the treatment effect at the cutoff, τ_c.

A picture is worth a thousand words, especially with RD designs, so let us look at a graph. Suppose we give a 100 point midterm exam (the forcing variable x) and award a new laptop computer to students receiving a score of 75 (the cutoff value c) or over. The outcome we measure is student performance, y, on a 400 point final exam.

In Figure 7.4, based on simulated data, we see that at midterm score 75 there is a jump in the final exam score. That jump is what we seek to measure. The RDD idea is that students receiving just under and just over 75 are basically very similar, so that if we compare them it is just as good as randomly assigning treatment. Another way to picture the outcomes is to divide the forcing

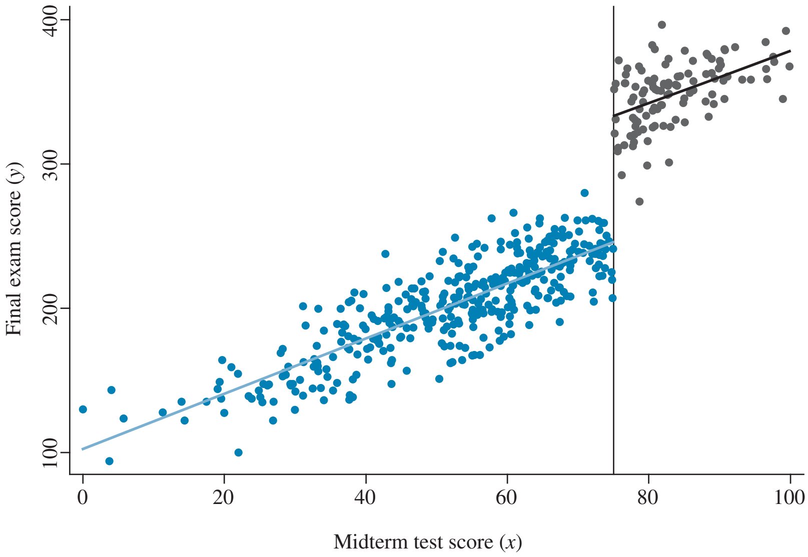

FIGURE 7.4 **Regression Discontinuity Design.**

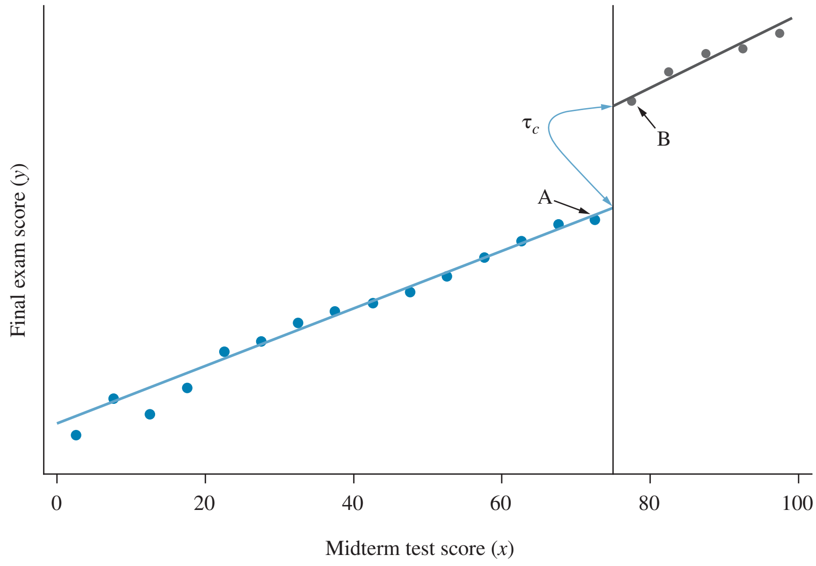

FIGURE 7.5 **Conditional Means Graph.**

variable (x) into intervals, or bins, and calculate and plot the mean, or median, of the outcome variable (y). Figure 7.5 is based on five point bins.

The difference between the mean scores of the two groups (A and B) just to either side of the cutoff is an estimate of the treatment effect at the cutoff, in this case $\hat{\tau}_c = B - A = 326.7 - 243.6 = 83.1$. We estimate that for students near the cutoff, getting a 75 or higher on the midterm, and thus receiving a new computer, had scores on the final exam that were 83.1 points higher than those who were also near the cutoff, but not receiving the prize, all other things being equal. This estimator is reasonable and intuitive. The difficulty is that students in the 70–75 range of test scores may not be as similar as we would like to students with test scores 75–80. If we make the bin widths smaller and smaller, then the groups to either side of the cutoff become more and more similar, but the number of observations in each bin gets smaller and smaller, reducing the reliability of this estimator of the treatment effect.[28]

Instead, let us use all the observations and use regression analysis to estimate the treatment effect at the cutoff, τ_c. Estimate the regression functions separately on the two groups, using as explanatory variable $x_i - c$:

1. Obtain $\hat{\alpha}_0 + \hat{\beta}_0(x_i - c)$ from a regression of y_i on $x_i - c$ for individuals below the cutoff, $(x_i < c)$.
2. Obtain $\hat{\alpha}_1 + \hat{\beta}_1(x_i - c)$ from a regression of y_i on $x_i - c$ for individuals above the cutoff, $(x_i \geq c)$.

The estimate of τ_c is $\hat{\tau}_c = \hat{\alpha}_1 - \hat{\alpha}_0$. Equivalently, we can use a pooled regression with an indicator variable. Define $d_i = 1$ if $x_i \geq c$, and $d_i = 0$ if $x_i < c$. Then the equivalent pooled regression is

$$y_i = \alpha + \tau_c d_i + \beta(x_i - c) + \gamma[d_i(x_i - c)] + e_i \tag{7.46}$$

There are some additional considerations when using RD designs. First, using the full range of the data may not be a good idea. The goal is to estimate the regression "jump" at the cutoff value

[28] Selecting bin width is an important issue in RDD analysis. See Lee and Lemieux (2010, pp. 307–314).

$x_i = c$. With sufficient observations, we can make the estimate "local" by only using data within a certain distance h of the cutoff. That is, use observations for which $c - h \le x_i \le c + h$. Checking the robustness of findings to various choices of h is a good idea.

Second, it is important to build into the regression sufficient flexibility to capture a nonlinear relationship. For example, if the true relationship between the outcome y and the test score x is nonlinear, then using linear relationships in the RDD can give a biased estimator of the treatment effect. In Figure 7.6, we illustrate a situation when there is no "jump" in the underlying relationship but using RDD with an assumed linear fit makes there appear be a positive treatment effect at $x_i = c$.

For this reason, researchers often use additional powers of $(x_i - c)$ in the regression relation, such as $(x_i - c)^2$, $(x_i - c)^3$, and $(x_i - c)^4$. If we use up to the third power, the pooled regression becomes

$$y_i = \alpha + \tau_c d_i + \sum_{q=1}^{3} \beta_q (x_i - c)^q + \sum_{p=1}^{3} \gamma_p \left[d_i (x_i - c)^p\right] + e_i \tag{7.47}$$

For the data in Figure 7.6, the estimated treatment effect from (7.47), $\hat{\tau}_c$, is not statistically different from zero, with a $t = 1.11$ and a p-value of 0.268. Alternatively, the recognition of a "nonjump" could be detected by using local observations for which $c - h \le x_i \le c + h$.

Third, it is possible that variables other than the forcing variable, say z_i, may influence the outcome. These can be added to the RDD model in equation (7.47).

Fourth, the illustration we have provided assumes that those with test scores at 75 or above are given a new computer whether they want one or not. We could instead offer those with test scores 75 and above a heavily discounted price on a new computer before the final exam. Some will elect to purchase the new machine using the discount and others will not. Some with test scores below 75 could, of course, also buy new computers. These issues lead to what is known as a **fuzzy regression discontinuity design**. The key in this case is that there is a "jump" in the **probability of treatment** (receiving a new computer before the final exam) at $x_i = c$. In this case, we must use an estimation alternative to least squares called **instrumental variables estimation**. This topic is considered in Chapter 10.

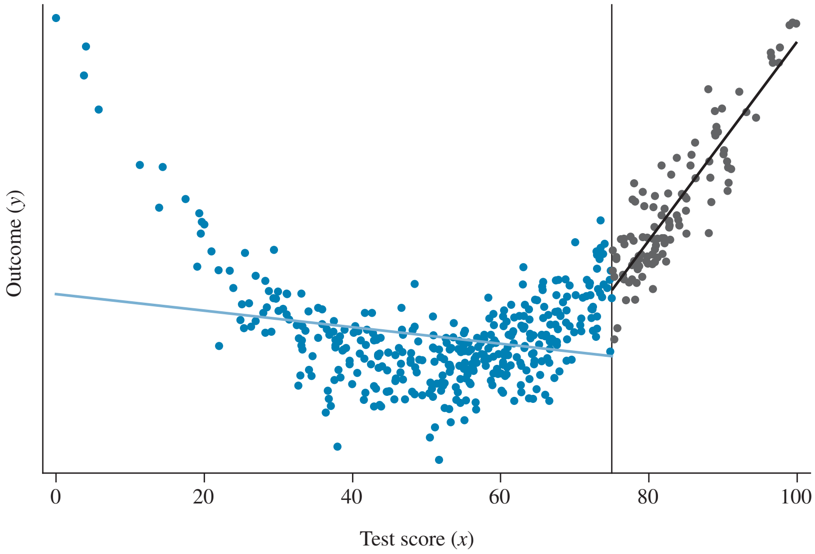

FIGURE 7.6 **RDD bias.**

7.7 Exercises

7.7.1 Problems

7.1 Suppose we are able to collect a random sample of data on economics majors at a large university. Further suppose that, for those entering the workforce, we observe their employment status and salary 5 years after graduation. Let $SAL = \$$ salary for those employed, $GPA =$ grade point average on a 4.0 scale during their undergraduate program, with $METRICS = 1$ if student took econometrics, $METRICS = 0$ otherwise.

a. Consider the regression model $SAL = \beta_1 + \beta_2 GPA + \beta_3 METRICS + e$. Should we consider this a causal model, or a predictive model? Explain your reasoning.

b. Assuming β_2 and β_3 are positive, draw a sketch of $E(SAL|GPA, METRICS) = \beta_1 + \beta_2 GPA + \beta_3 METRICS$.

c. Define a dummy variable $FEMALE = 1$, if the student is female; 0 otherwise. Modify the regression model to be $SAL = \beta_1 + \beta_2 GPA + \beta_3 METRICS + \delta_1 FEMALE + e$. What is the expected salary of a male who has not taken econometrics? What is the expected salary of a female who has taken econometrics?

d. Consider the regression model

$$\begin{aligned} SAL = \beta_1 + \beta_2 GPA + \beta_3 METRICS + \delta_1 FEMALE \\ + \delta_2 (FEMALE \times METRICS) + e \end{aligned} \tag{XR7.1.1}$$

What is the expected salary of a male who has not taken econometrics? What is the expected salary of a female who has taken econometrics?

e. In the equation (XR7.1.1), assume that $\delta_1 < 0$ and $\delta_2 < 0$. Sketch $E(SAL|GPA, METRICS, FEMALE)$ versus GPA for (i) males not taking econometrics, (ii) males taking econometrics, (iii) females not taking econometrics, and (iv) females taking econometrics.

f. In equation (XR7.1.1), what are the null and alternative hypotheses, in terms of model parameters, for testing that econometrics training does not affect the average salary of economics majors? In order to use the test statistic in equation (6.4), what regression must you estimate in addition to (XR7.1.1)? What is the distribution of the test statistic if the null hypothesis is true assuming $N = 300$? What is the rejection region for a 5% test?

7.2 In September of 1998, a local TV station contacted an econometrician to analyze some data for them. They were going to do a Halloween story on the legend of full moons affecting behavior in strange ways. They collected data from a local hospital on emergency room cases for the period from January 1, 1998 until mid-August. There were 229 observations. During this time, there were eight full moons and seven new moons (a related myth concerns new moons) and three holidays (New Year's day, Memorial Day, and Easter). If there is a full-moon effect, then hospital administrators will adjust numbers of emergency room doctors and nurses, and local police may change the number of officers on duty. Let T be a time trend ($T = 1, 2, 3, \ldots, 229$). Let the indicator variables $HOLIDAY = 1$ if the day is a holiday, $= 0$ otherwise; $FRIDAY = 1$ if the day is a Friday, $= 0$ otherwise; $SATURDAY = 1$ if the day is a Saturday, $= 0$ otherwise; $FULLMOON = 1$ if there is a full moon, $= 0$ otherwise; $NEWMOON = 1$ if there is a new moon, $= 0$ otherwise. Consider the model

$$\begin{aligned} CASES = \beta_1 + \beta_2 T + \delta_1 HOLIDAY + \delta_2 FRIDAY + \delta_3 SATURDAY \\ + \theta_1 FULLMOON + \theta_2 NEWMOON + e \end{aligned} \tag{XR7.2.1}$$

a. What is the expected number of emergency room cases for day $T = 100$, which was a Friday with neither a full or new moon?

b. What is the expected number of emergency room cases for day $T = 185$, which was a holiday Saturday?

c. In terms of the model parameters, what are the null and alternative hypotheses for testing that neither a full moon nor a new moon have any effect on the number of emergency room cases? What is the test statistic? What is the distribution of the test statistic if the null hypothesis is true? What is the rejection region for a 5% test?

d. The sum of squared residuals from the regression in (XR7.2.1) is 27109. If full moon and new moon are omitted from the model the sum of squared residuals is 27424. Carry out the test in (c). What is your conclusion?

e. Using the model in equation (XR7.2.1), the estimated coefficient of *SATURDAY* is 10.59 with standard error 2.12, and the estimated coefficient for *FRIDAY* is 6.91, with standard error 2.11. The estimated covariance between the coefficient estimators is 0.75. Should the hospitals prepare for significantly more emergency room patients on Saturday than Friday? State the relevant null and alternative hypotheses in terms of the model parameters. What is the test statistic? What is the distribution of the test statistic if the null hypothesis is true? What is the rejection region for a test at the 10% level? Carry out the test and state your conclusion?

7.3 One of the key problems regarding housing prices in a region concerns construction of "price indexes." That is, holding other factors constant, have prices increased, decreased or stayed relatively constant in a particular area? As an illustration, consider a regression model for house prices (in $1000s) on home sales from 1991 to 1996 in Stockton, CA, including as explanatory variables the size of the house (*SQFT*, in 100s of square feet), the age of the house (*AGE*) and annual indicator variables, such as $D92 = 1$ if the year is 1992 and 0 otherwise.

$$\begin{aligned} PRICE = {} & \beta_1 + \beta_2 SQFT + \beta_3 AGE + \delta_1 D92 + \delta_2 D93 + \delta_3 D94 + \delta_4 D95 \\ & + \delta_5 D96 + e \end{aligned} \tag{XR7.3.1}$$

An alternative model employs a "trend" variable $YEAR = 0, 1, \ldots, 5$ for the years 1991–1996.

$$PRICE = \beta_1 + \beta_2 SQFT + \beta_3 AGE + \tau YEAR + e \tag{XR7.3.2}$$

a. What is the expected selling price of a 10-year-old house with 2000 square feet of living space in each of the years 1991–1996 using equation (XR7.3.1)?

b. What is the expected selling price of a 10-year-old house with 2000 square feet of living space in each of the years 1991–1996 using equation (XR7.3.2)?

c. In order to choose between the models in (XR7.3.1) and (XR7.3.2), we propose a hypothesis test. What set of parameter constraints, or restrictions, would result in equation (XR7.3.1) equaling (XR7.3.2)? The sum of squared residuals from (XR7.3.1) is 2385745 and from (XR7.3.2) is 2387476. What is the test statistic for testing the restrictions that would make the two models equivalent? What is the distribution of the test statistic if the null hypotheses are true? What is the rejection region for a test at the 5% level? If the sample size is $N = 4682$, what do you conclude?

d. Using the model in (XR7.3.1) the estimated coefficients of the indicator variables for 1992 and 1994, and their standard errors, are −4.393 (1.271) and −13.174 (1.211), respectively. The estimated covariance between these two coefficient estimators is 0.87825. Test the null hypothesis that $\delta_3 = 3\delta_1$ against the alternative that $\delta_3 \neq 3\delta_1$ if $N = 4682$, at the 5% level.

e. The estimated value of τ in equation (XR7.3.2) is –4.12. What is the estimated difference in the expected house price for a 10-year-old house with 2000 square feet of living space in 1992 and 1994. Using information in (d), how does this compare to the result using (XR7.3.1)?

7.4 Angrist and Pischke[29] report estimation results of log-earnings equations using a large sample of college graduates. The predictors of interest (there are others included in their model) are the indicator variable *PRIVATE* (=1 if the individual attended a private college or university, = 0 if the individual attended a public college or university) and *SAT*/100, the individual's SAT score divided by 100. In the estimated regression equations, the dependent variable is ln(*EARNINGS*) and they include an intercept. The coefficient estimates, with standard errors in brackets, for two regressions that they estimate, are as follows.

$$0.212[0.060]PRIVATE \tag{XR7.4.1}$$

$$0.152\,[0.057]PRIVATE + 0.051[0.008](SAT/100) \tag{XR7.4.2}$$

a. In each model, what is the approximate effect on earnings of attending a private university rather than a public university?

[29] Joshua D. Angrist and Jörn-Steffen Pischke (2015) *Mastering Metrics: The Path from Cause to Effect*, Princeton University Press, p. 66.

b. In the second model, what is the predicted effect on earnings of a 100-point increase in SAT score?

c. The estimated coefficient of *PRIVATE* is smaller in the second model than in the first model. Use the concept of "omitted variables bias" to explain this result.

d. What should happen to the estimated coefficients in equation (XR7.4.2) if parental income is included as an explanatory variable? Explain.

7.5 In 1985, the state of Tennessee carried out a statewide experiment with primary school students. Teachers and students were randomly assigned to be in a regular-sized class or a small class. The outcome of interest is a student's score on a math achievement test (*MATHSCORE*). Let $SMALL = 1$ if the student is in a small class and $SMALL = 0$ otherwise. The other variable of interest is the number of years of teacher experience, *TCHEXPER*.

a. Write down the econometric specification of the linear regression model explaining *MATHSCORE* as a function of *SMALL* and *TCHEXPER*. Use β_1, β_2, and β_3 as the model parameters. In this model, what is the expected math score for a child in a regular-sized class with a teacher having 10 years of experience? What is the expected math score for a child in a small class with a teacher having 10 years of experience?

b. Let $BOY = 1$ if the child is male and $BOY = 0$ if the child is female. Modify the model in part (a) to include the variables *BOY* and $BOY \times SMALL$, with parameters θ_1 and θ_2. Using this model

i. What is the expected math score for a boy in a small class with a teacher having 10 years of experience?

ii. What is the expected math score for a girl in a regular-sized class with a teacher having 10 years of experience?

iii. What is the null hypothesis, written in terms of the model parameters, that the sex of the child has no effect on expected math score? What is the alternative hypothesis? What is the test statistic for the null hypothesis and what is its distribution if the null hypothesis is true? What is the test rejection region for a 5% test when $N = 1200$?

iv. It is conjectured that boys may benefit from small classes more than girls. What null and alternative hypothesis would you test to examine this conjecture? [*Hint*: Let the conjecture be the alternative hypothesis.]

7.6 In 1985, the state of Tennessee carried out a statewide experiment with primary school students. Teachers and students were randomly assigned to be in a regular-sized class or a small class. The outcome of interest is a student's score on a math achievement test (*MATHSCORE*). Let $SMALL = 1$ if the student is in a small class and $SMALL = 0$ otherwise. The other variable of interest is the number of years of teacher experience, *TCHEXPER*. Let $BOY = 1$ if the child is male and $BOY = 0$ if the child is female.

a. Write down the econometric specification of the linear regression model explaining *MATHSCORE* as a function of *SMALL*, *TCHEXPER*, *BOY* and $BOY \times TCHEXPER$, with parameters $\beta_1, \beta_2, \ldots$.

i. What is the expected math score for a boy in a small class with a teacher having 10 years of experience?

ii. What is the expected math score for a girl in a regular-sized class with a teacher having 10 years of experience?

iii. What is the *change* in the expected math score for a boy in a small class with a teacher having 11 years of experience rather than 10?

iv. What is the *change* in the expected math score for a boy in a small class with a teacher having 13 years of experience rather than 12?

v. State, in terms of the model parameters, the null hypothesis that the marginal effect of teacher experience on expected math score does not differ between boys and girls, against the alternative that boys benefit more from additional teacher experience. What test statistic would you use to carry out this test? What is the distribution of the test statistic assuming then null hypothesis is true, if $N = 1200$? What is the rejection region for a 5% test?

b. Modify the model in part (a) to include $SMALL \times BOY$.

i. What is the expected math score for a boy in a small class with a teacher having 10 years of experience?

ii. What is the expected math score for a girl in a regular-sized class with a teacher having 10 years of experience?

iii. What is the expected math score for a boy? What is it for a girl?

iv. State, in terms of the part (b) model parameters, the null hypothesis that the expected math score does not differ between boys and girls, against the alternative that there is a difference in expected math score for boys and girls. What test statistic would you use to carry out this test? What is the distribution of the test statistic assuming the null hypothesis is true, if $N = 1200$? What is the rejection region for a 5% test?

7.7 Can monetary policy reduce the impact of a severe recession? A **natural experiment** is provided by the State of Mississippi. In December of 1930, there were a series of bank failures in the southern United States. The central portion of Mississippi falls into two Federal Reserve Districts: the sixth (Atlanta Fed) and the eighth (St. Louis Fed). The Atlanta Fed offered "easy money" to banks while the St. Louis Fed did not. On July 1, 1930 (just before the crisis), there were 105 State Charter banks in Mississippi in the sixth district and 154 banks in the eighth district. On July 1, 1931 (just after the crisis), there were 96 banks remaining in the sixth district and 126 in the eighth district. These data values are from Table 1, Gary Richardson and William Troost (2009) "Monetary Intervention Mitigated Banking Panics during the Great Depression: Quasi-Experimental Evidence from a Federal Reserve District Border, 1929–1933," *Journal of Political Economy*, 117(6), 1031–1073.

a. Let the eighth district be the control group and the sixth district be the treatment group. Construct a figure similar to Figure 7.3 using the four observations rather than sample means. Identify the treatment effect on the figure.

b. How many banks did each district lose during the crisis? Calculate the magnitude of the treatment effect using (7.18) with these four observations, rather than sample means.

c. Suppose we have data on these two districts for 1929–1934, so $N = 12$. Let $AFTER_t = 1$ for years after 1930, and let $AFTER_t = 0$ for years 1929 and 1930. Let $TREAT_i = 1$ for the sixth district and let $TREAT_i = 0$ for banks in the eighth district. Let $BANKS_{it}$ be the number of banks in each district in each year. Angrist and Pischke (2015, p. 188) report the estimated equation

$$\begin{aligned}\widehat{BANKS_{it}} &= 167 - \underset{(8.8)}{2.9}TREAT_i - \underset{(7.6)}{49}AFTER_t + \underset{(10.7)}{20.5}\left(TREAT_i \times AFTER_t\right)\\ &\text{(se)}\end{aligned}$$

Compare the estimated treatment effect from this equation to the calculation in (b). Is the estimated treatment effect significant, at the 5% level?

7.8 Using $N = 2005$ observations, we examine the relationship between food expenditures away from home per person in the past month as a function of household monthly income, the highest level of education of a household member, and region of the country. The full equation of interest is

$$\begin{aligned}\ln(FOODAWAY) = \beta_1 + \beta_2\ln(INCOME) + \delta_1 COLLEGE + \delta_2 ADVANCED\\ + \theta_1 MIDWEST + \theta_2 SOUTH + \theta_3 WEST + e\end{aligned}$$

where $COLLEGE = 1$ if the highest education of a household member is a college degree, $ADVANCED = 1$ if the highest education of a household member is an advanced degree (such as a Master's or Ph.D.). The regional indicators equal one if the household lives in that region and are zero otherwise.

a. The estimated value of β_2 is 0.427 with a standard error of 0.035. Construct and interpret a 95% interval estimate.

b. The estimated value of δ_2 is 0.270 with a standard error of 0.0544. Construct and interpret a 95% interval estimate using the rough calculation in Section 7.3.1.

c. Use the exact calculation discussed in Section 7.3.1 to estimate the predicted effect on food expenditure per person away from home for a household having a member with an advanced degree.

d. What is the null hypothesis, in terms of the model parameters, that the highest level of education achieved by a household member does not matter? What is the test statistic for this hypothesis? What is the 5% rejection region? The sum of squared residuals from the full model is 1586 and *SSE* from the model omitting the education variables is 1609. Can we conclude that the education variables are important predictors of food expenditures away from home?

e. In the full model, the reported t-value for $COLLEGE$ is 0.34. What can we conclude from that? [*Hint*: What is the reference group?]

f. The estimated value of θ_2 is 0.088. What is the estimated expected value of ln(*FOODAWAY*) for a household with \$10,000 per month income, with a member with an advanced degree, and who live in the south? Calculate the natural and corrected predictors of expenditure on food away from home per member for this household. [*Hint*: A relevant piece of information is in part (b).]

7.9 Suppose we wish to estimate a model of household expenditures on alcohol (*ALC*, in dollars per month) as a function of household income (*INCOME*, \$100's per month), and some other demographic variables.

a. Let $KIDS = 0, 1, 2, \ldots$ be the number children in the household. Is *KIDS* a qualitative or quantitative variable? Interpret the coefficient of *KIDS* in the model

$$ALC = \beta_1 + \beta_2 INCOME + \delta KIDS + e \qquad \text{(XR7.9.1)}$$

What is the marginal impact of the second child? What is the marginal impact of the fourth child?

b. Let $ONEKID = 1$ if there is one child, and zero otherwise. Let $TWOKIDS = 1$ if there are two children, and zero otherwise. Let $MANY = 1$ if there are three or more children, and zero otherwise. Consider the model

$$ALC = \beta_1 + \beta_2 INCOME + \delta_1 ONEKID + \delta_2 TWOKIDS + \delta_3 MANY + e \qquad \text{(XR7.9.2)}$$

Compare the interpretation of this model to that in part (a). Is the impact of an additional child the same as in the model in (a)? What is the impact of the first child on expected household expenditure on alcohol? What is the impact of having a fourth child on the expected household expenditure on alcohol?

c. Is there a set of parameter restrictions, or constraints, that we can impose on equation (XR7.9.2) to make it equivalent to equation (XR7.9.1)?

7.10 Suppose we wish to estimate a model of household expenditures on alcohol (*ALC*, in dollars per month) as a function of household income (*INCOME*, \$100's per month), and some other demographic variables.

a. Let $RELIGIOUS = 0, 1, 2, 3,$ or 4 if the household considers itself not religious, a little religious, moderately religious, very religious, or extremely religious, respectively. Is *RELIGIOUS* a quantitative or qualitative variable? Explain your choice.

b. Consider the model

$$ALC = \beta_1 + \beta_2 INCOME + \beta_3 RELIGIOUS + e$$

What is the expected household expenditure on alcohol for a household that considers itself not religious? What is the expected household expenditure for a household that considers itself a little religious? What is the expected household expenditure for a household that considers itself moderately religious?

c. If we test the hypothesis $\beta_3 = 0$ in model (b), what behavioral assumption are we testing? What is the expected household expenditure on alcohol if the hypothesis is true?

d. Let $LITTLE = 1$ if the household considers itself a little religious, and zero otherwise. Similarly define the indicator variables *MODERATELY*, *VERY*, and *EXTREMELY*. Consider the model

$$ALC = \gamma_0 + \gamma_1 INCOME + \gamma_2 LITTLE + \gamma_3 MODERATELY + \gamma_4 VERY + \gamma_5 EXTREMELY + e$$

What is the expected household expenditure for a household that considers itself not religious? What is the expected household expenditure for a household that considers itself a little religious? What is the expected household expenditure for a household that considers itself moderately religious? Very religious? Extremely religious?

e. If we impose the restrictions $\gamma_3 = 2\gamma_2$, $\gamma_4 = 3\gamma_2$, $\gamma_5 = 4\gamma_2$ on the model in part (d), how does the restricted model compare to the model in (b)?

7.11 Consider the log-linear regression model $\ln(y) = \beta_1 + \beta_2 x + \delta_1 D + \delta_2 (x \times D) + e$. If the regression errors are normally distributed $N(0, \sigma^2)$, then

$$E(y|x, D) = \exp\left(\beta_1 + \beta_2 x + \delta_1 D + \delta_2 (x \times D)\right) \exp\left(\sigma^2/2\right) \qquad \text{(XR7.11.1)}$$

a. Use Derivative Rule 7 to show that

$$\frac{\partial E(y|x,D)}{\partial x} = \exp\left(\beta_1 + \beta_2 x + \delta_1 D + \delta_2 (x \times D)\right) \exp\left(\sigma^2/2\right)\left(\beta_2 + \delta_2 D\right) \quad \text{(XR7.11.2)}$$

b. Divide both sides of the result in (a) by $E(y|x,D)$ to show that

$$\frac{\partial E(y|x,D)}{\partial x}\frac{1}{E(y|x,D)} = \frac{\partial E(y|x,D)/E(y|x,D)}{\partial x} = \left(\beta_2 + \delta_2 D\right) \quad \text{(XR7.11.3)}$$

c. Multiply both sides of the equation in (b) by 100 to obtain

$$100\frac{\partial E(y|x,D)/E(y|x,D)}{\partial x} = \%\Delta E(y|x,D) = 100\left(\beta_2 + \delta_2 D\right) \quad \text{(XR7.11.4)}$$

This is the marginal effect, the percentage change, in $E(y|x,D)$ given a unit change in x in the log-linear model.

d. A fitted log-linear model for house price, where $SQFT(x)$ is the house's living area (100s of square feet) and $UTOWN(D)$ is an indicator variable with $UTOWN = 1$ for houses near a university, and zero otherwise, is

$$\widehat{\ln(PRICE)} = 4.456 + 0.362SQFT + 0.336UTOWN - 0.00349(SQFT \times UTOWN)$$

Use equation (XR7.11.4) to calculate the marginal effect of $SQFT$ on house price, for a house with $UTOWN = 1$ and for a house with $UTOWN = 0$.

e. Let b_2 and d_2 be the least squares estimators of β_2 and δ_2 in equation (XR7.11.4). Write down the formula for the standard error of the estimated value $100\left(b_2 + d_2 D\right)$, for a given D.

f. Multiply both sides in (XR7.11.3) by x, and by 100/100, and rearrange to obtain

$$\frac{\partial E(y|x,D)/E(y|x,D)}{\partial x}x = \frac{100\partial E(y|x,D)/E(y|x,D)}{100\partial x/x} = \left(\beta_2 + \delta_2 D\right)x \quad \text{(XR7.11.5)}$$

Interpreting $100\partial x/x$ as the percentage change in x, we find that the elasticity of expected price with respect to a percentage change in x is $\left(\beta_2 + \delta_2 D\right)x$.

g. Apply the result in equation (XR7.11.5) to calculate the elasticities of expected house price with respect to a change in price for a house of 2500 square feet, when $UTOWN = 1$ and when $UTOWN = 0$.

h. Let b_2 and d_2 be the least squares estimators of β_2 and δ_2 in equation (XR7.11.5). Write down the formula for the standard error of the estimated value $\left(b_2 + d_2 D\right)x$, given D and x.

7.12 Consider the log-linear regression model $\ln(y) = \beta_1 + \beta_2 x + \delta_1 D + \delta_2 (x \times D) + e$. If the regression errors are normally distributed $N\left(0, \sigma^2\right)$, then $E(y|x,D)$ is given in equation (XR7.11.1).

a. Find $E(y|x, D = 1)$ and $E(y|x, D = 0)$.

b. Show that

$$\frac{100\left[E(y|x,D=1) - E(y|x,D=0)\right]}{E(y|x,D=0)} = 100\left[\exp\left(\delta_1 + \delta_2 x\right) - 1\right] \quad \text{(XR7.12.1)}$$

This is the percentage change in the expected value of y, given x, when the indicator variable changes from $D = 0$ to $D = 1$.

c. Given the log-linear model, the value of $\ln(y)$ when $D = 0$ is $\ln(y|D = 0, x) = \beta_1 + \beta_2 x + e$, and when $D = 1$ we have $\ln(y|D = 1, x) = \left(\beta_1 + \delta_1\right) + \left(\beta_2 + \delta_2\right)x + e$. Subtract $\ln(y|D = 0, x)$ from $\ln(y|D = 1, x)$, and multiply by 100, to obtain

$$100\left[\ln(y|D=1,x) - \ln(y|D=0,x)\right] \simeq \%\Delta(y|x) = 100\left(\delta_1 + \delta_2 x\right) \quad \text{(XR7.12.2)}$$

d. A fitted log-linear model for house price, where $SQFT(x)$ is the house's living area (100s of square feet) and $UTOWN(D)$ is an indicator variable with $UTOWN = 1$ for houses near a university, and zero otherwise, is

$$\widehat{\ln(PRICE)} = 4.456 + 0.362SQFT + 0.336UTOWN - 0.00349(SQFT \times UTOWN)$$

Calculate the percentage change in the expected value of $PRICE$ for a house of 2500 square feet using (XR7.12.1). Also calculate the approximate value in (XR7.12.2).

e. If d_1 and d_2 are the least squares estimators of δ_1 and δ_2 in equation (XR7.12.2), write down the formula for the standard error of $100(d_1 + d_2x)$, given x.

f. Let $\lambda = 100\left[\exp(\delta_1 + \delta_2 x) - 1\right]$ and $\hat{\lambda} = 100\left[\exp\left(d_1 + d_2 x\right) - 1\right]$. Use Derivative Rule 7, in Appendix A.3.1, to show that $\partial\lambda/\partial\delta_1 = 100\exp\left(\delta_1 + \delta_2 x\right)$ and $\partial\lambda/\partial\delta_2 = 100\exp(\delta_1 + \delta_2 x)x$. The "delta method" for finding the variance of a nonlinear function, such as $\hat{\lambda}$, is discussed in Section 5.7.4 and also Appendix 5B.5. Using the delta method, write out the expression for standard error of $\hat{\lambda}$.

7.13 Many cities in California have passed Inclusionary Zoning policies (also known as below-market housing mandates) as an attempt to make housing more affordable. These policies require developers to sell some units below the market price on a percentage of the new homes built. For example, in a development of 10 new homes each with market value $850,000, the developer may have to sell 5 of the units at $180,000. Means and Stringham (2012)[30] examine the effects of such policies on house prices and number of housing units available using 1990 and 2000 census data on 311 California cities.

a. Let *LNPRICE* be the log of average home price, and let *LNUNITS* be the log of the number of housing units. Using only the data for 2000, we compare the sample means of *LNPRICE* and *LNUNITS* for cities with an Inclusionary Zoning policy, *IZLAW* = 1, to those without the policy, *IZLAW* = 0. The following table displays the sample means of *LNPRICE* and *LNUNITS*.

2000	*IZLAW* = 1	*IZLAW* = 0
$\overline{LNPRICE}$	12.8914	12.2851
$\overline{LNUNITS}$	9.9950	9.5449

Based on these estimates, what is the percentage difference in prices and number of units for cities with and without the law? Use the approximation $100\left[\ln(y_1) - \ln\left(y_0\right)\right]$ for the percentage difference between y_0 and y_1. Does the law appear to achieve its purpose?

b. Using the data for 1990, we compare the sample means of *LNPRICE* and *LNUNITS* for cities with an Inclusionary Zoning policy, *IZLAW* = 1, to those without the policy, *IZLAW* = 0. The following table displays the sample means of *LNPRICE* and *LNUNITS*.

1990	*IZLAW* = 1	*IZLAW* = 0
$\overline{LNPRICE}$	12.3383	12.0646
$\overline{LNUNITS}$	9.8992	9.4176

Use the existence of an Inclusionary Zoning policy as a "treatment." Consider those cities that did not pass such a law, *IZLAW* = 0, the "control" group. Draw a figure similar to Figure 7.3 comparing treatment and control groups for *LNPRICE*, and determine the "treatment effect." Are your conclusions about the effect of the policy the same as in (a)?

c. Draw a figure similar to Figure 7.3 comparing treatment and control groups for *LNUNITS*, and determine the "treatment effect." Are your conclusions about the effect of the policy the same as in (a)?

7.14 Consider a model explaining the weekly sales (*SALES* = 100's cans sold) of a popular brand (the "target" brand) of canned tuna as a function of its price (*PRICE* = average price in cents), the average prices of two competitors (*PRICE*2, *PRICE*3, also in cents). Also included is an indicator variable *DISP* = 1 if there is a store display but **no** newspaper ad during the week for the target brand, and 0 otherwise. The indicator variable *DISPAD* = 1 if there is a store display during the week for the target

[30] Tom Means and Edward P. Stringham (2012) "Unintended or Intended consequences? The effect of below-market housing mandates on housing markets in California," *Journal of Public Finance and Public Choice*, p. 39–64. The authors wish to thank Tom Means for providing the data and insights into this exercise.

brand **and** newspaper ads, 0 otherwise. The estimated log-linear model is

$$\begin{array}{lllll} \widehat{\ln(SALES)} = 2.077 - 0.0375PRICE + 0.0115PRICE2 + 0.0129PRICE3 + 0.424DISP \\ \text{(se)} \quad (0.646) \quad (0.00577) \quad (0.00449) \quad (0.00605) \quad (0.105) \\ + 1.431DISPAD \qquad R^2 = 0.84 \qquad N = 52 \\ (0.156) \end{array}$$

a. Discuss and interpret the coefficients of the price variables.

b. Are the signs and relative magnitudes of the advertising variables consistent with economic logic? Provide both the "rough" and "exact" calculations for the effects of *DISP* and *DISPAD* from Sections 7.3.1 and 7.3.2.

c. Test the significance of the advertising variables using a two-tail test, at the 1% level of significance. What do you conclude?

d. The *F*-test statistic value for the joint significance of the two advertising variables is 42.0. What can we conclude about the significance of advertising? If you were going to use the form of the *F*-statistic in equation (6.4), what additional regression would you need to run?

e. Label the parameters in the equation $\beta_1, \beta_2, \ldots$ If the null hypothesis is $H_0 : \beta_6 \leq \beta_5$, state the alternative hypothesis. Why is the test of this null hypothesis and alternative hypothesis interesting? Carry out the test at the 1% level of significance, given that the calculated *t*-value is 6.86. What do you conclude?

7.15 Mortgage lenders are interested in determining borrower and loan characteristics that may lead to delinquency or foreclosure. We estimate a regression model using 1000 observations and the following variables. The dependent variable of interest is *MISSED*, an indicator variable = 1 if the borrower missed at least three payments (90+ days late), but 0 otherwise. Explanatory variables are *RATE* = initial interest rate of the mortgage; *AMOUNT* = dollar value of mortgage (in \$100, 000); and *ARM* = 1 if mortgage has an adjustable rate, and = 0 if mortgage has a fixed rate. The estimated equation is

$$\begin{array}{l} \widehat{MISSED} = -0.348 + 0.0452RATE + 0.0732AMOUNT + 0.0834ARM \\ \text{(se)} \qquad\qquad\qquad (0.00841) \qquad (0.0144) \qquad (0.0326) \end{array}$$

a. Interpret the signs and significance of each of the coefficients.

b. Two borrowers who did not miss a payment had loans with the following characteristics: $(RATE = 8.2,\ AMOUNT = 1.912,\ ARM = 1)$ and $(RATE = 9.1, AMOUNT = 8.6665, ARM = 1)$. For each of these borrowers, predict the probability that they will miss a payment.

c. Two borrowers who did miss a payment had loans with the following characteristics: $(RATE = 12.0, AMOUNT = 0.71, ARM = 0)$ and $(RATE = 6.45, AMOUNT = 8.5, ARM = 1)$. For each of these borrowers, predict the probability that they will miss a payment.

d. For a borrower seeking an adjustable rate mortgage, with an initial interest rate of 6.0, above what loan amount would you predict a missed payment with probability 0.51?

7.7.2 Computer Exercises

7.16 In this exercise, we examine the hours of market work by married women as a function of their education and number of children. Use data file *cps5mw_small* for this exercise. The data file *cps5mw* contains more observations.

a. Estimate the linear regression model

$$HRSWORK = \beta_1 + \beta_2 WAGE + \beta_3 EDUC + \beta_4 NCHILD + e \qquad \text{(XR7.16.1)}$$

Interpret the coefficient of *NCHILD*. Estimate the expected hours worked by a married woman whose wage is \$20 per hour, who has 16 years of education, and who has no children. Do the same calculation for a woman with one child, two children, and three children. How much does the expected number of hours change with each additional child?

b. Define the indicator variables *POSTGRAD* = 1 if *EDUC* > 16, 0 otherwise; *COLLEGE* = 1 if *EDUC* = 16, 0 otherwise; and *SOMECOLLEGE* if 12 < *EDUC* < 16. Estimate the *HRSWORK* equation (XR7.16.1) replacing *EDUC* by these three indicator variables. Interpret the coefficients of the education indicator variables. Estimate the expected hours worked by a married woman whose wage is \$20 per hour, who has 12 years of education, and who has no children. Do the same calculation for a woman with $EDUC = 13, 14, 15, 16$, and 17. Is the marginal effect of education constant?

c. Define indicator variables $ONEKID = 1$ if $NCHILD = 1$, 0 otherwise; $TWOKIDS = 1$ if $NCHILD = 2$, 0 otherwise; and $MOREKIDS = 1$ if $NCHILD > 2$, 0 otherwise. Estimate the *HRSWORK* equation (XR7.16.1) but replace *NCHILD* by these three indicator variables. Interpret the estimated coefficients of the three indicator variables. Estimate the expected hours worked by a married woman with 16 years of education, whose wage is \$20 per hour with no children, one child, two children, and more than two children. Compare and contrast these estimates to those in (a).

d. Estimate the model (XR7.16.1) replacing *EDUC* with the three indicator variables in (b) and replacing *NCHILD* with the three indicator variables in (c). Compare and contrast this model to the models in (a)–(c).

e. Define the indicator variable $EDUC12 = 1$ if $EDUC = 12$, 0 otherwise. Define indicator variables *EDUC12*, *EDUC13*, *EDUC14*, *EDUC16* similarly. In this sample, there are no women with 15 years of education. Define $EDUC18 = 1$ if $EDUC > 16$, 0 otherwise. Estimate the *HRSWORK* equation (XR7.16.1) replacing *NCHILD* by the three indicator variables and *EDUC* by the five new indicator variables. Have any essential conclusions changed by using this specification?

f. Which of the specifications in (a)–(e) has the highest R^2? The highest adjusted-R^2, the smallest *SCHWARZ* criterion (SC or BIC) value? Which model do you prefer taking into account economic, econometric, and fit aspects?

7.17 Does a mother's smoking affect the birthweight of her child? Using data in the file *bweight_small* taken from Cattaneo (2010),[31] we explore this question. The file *bweight* contains more observations.

a. Calculate the sample means of *BWEIGHT* for mothers who smoke ($MBSMOKE = 1$) and those who do not smoke ($MBSMOKE = 0$). Use the t-test of the equality of population means given in Appendix C.7.2, Case 1, to test whether the mean birthweight for smoking and nonsmoking mothers is the same. Use the 5% level of significance.

b. Estimate the regression $BWEIGHT = \beta_1 + \beta_2 MBSMOKE + e$. Interpret the coefficient of *MBSMOKE*. Can we interpret the coefficient as the "average treatment effect" of smoking? Test the null hypothesis that $\beta_2 \geq 0$ against $\beta_2 < 0$ at the 5% level of significance.

c. Add to the model in (b) control variables *MMARRIED*, *MAGE*, *PRENATAL1*, and *FBABY*. Are any of these variables significant predictors of an infant's birthweight? Which signs of the significant coefficients are consistent with your expectations? Does the estimate of the coefficient of *MBSMOKE* change much?

d. Estimate the regression of *BWEIGHT* on *MMARRIED*, *MAGE*, *PRENATAL1*, and *FBABY* for mothers who smoke ($MBSMOKE = 1$) and those who do not smoke ($MBSMOKE = 0$). Carry out a Chow test of the equivalence of these two regressions at the 5% level.

e. Use equation (7.37) to obtain the estimate of the average treatment effect using the results from (d). Compare this estimate of the average treatment effect to the estimates in (b) and (c).

7.18 Does a mother's smoking affect the birthweight of her child? Using the data file *bweight_small*, we explore this question. The file *bweight* contains more observations.

a. Estimate the regression model represented by equation (7.38) for *BWEIGHT*. Include as explanatory variables *MMARRIED*, *MAGE*, *PRENATAL1*, and *FBABY*, along with *MBSMOKE* and interactions between *MBSMOKE* and the other variables. Use equation (7.40), and the discussion below equation (7.40), to estimate the average treatment effect.

b. Use equation (7.41) to estimate the average treatment effect of mother smoking on infant birthweight, and construct a 95% interval estimate for τ_{ATE}.

c. Calculate the normalized difference equation (7.44) for each of the variables *MMARRIED*, *MAGE*, *PRENATAL1*, and *FBABY*. Are any of the normalized differences bigger than the rule of thumb threshold of 0.25?

d. Use equation (7.42) to estimate the average treatment effect on the treated, τ_{ATT}. How much does it differ from your estimate of the population average treatment effect?

[31] Efficient semiparametric estimation of multi-valued treatment effects under ignorability, *Journal of Econometrics*, 155, 138–154. The authors would like to thank Matias Cattaneo for providing the data. The dataset is used in *Stata Treatment-Effects Reference Manual, Release 14* for examples as well.

e. Use equation (7.43) to estimate the average treatment effect on the population of mothers who are Hispanic ($MHISP = 1$). How does it compare to the estimated population average treatment effect?

f. Use equation (7.43) to estimate the average treatment effect on the population of mothers who are white ($MWHITE = 1$). How does this compare to the population average treatment effect estimate?

7.19 Does a mother's smoking affect the birthweight of her child? Using the data file *bweight_small* we explore this question. The file *bweight* contains more observations. The variable *MSMOKE* is the number of cigarettes smoked daily during pregnancy. Nonsmokers ($MBSMOKE = 0$) smoke zero daily. Among smokers ($MBSMOKE = 1$), the variable $MSMOKE = 1$ if 1–5 cigarettes are smoked daily; $MSMOKE = 2$ if 6–10 cigarettes are smoked daily; and $MSMOKE = 3$ if 11 or more cigarettes are smoked daily.

a. Estimate a regression model for *BWEIGHT*. Include as explanatory variables *MMARRIED*, *MAGE*, *PRENATAL1*, and *FBABY*, along with *MSMOKE*. Interpret the estimated coefficient of *MSMOKE*.

b. From *MSMOKE* create three indicator variables, $SMOKE2 = 1$ if a mother smokes 1–5 cigarettes per day, 0 otherwise; $SMOKE3 = 1$ if a mother smokes 6–10 cigarettes per day, 0 otherwise; $SMOKE4 = 1$ if a mother smokes 11 or more cigarettes per day, 0 otherwise. Estimate a regression model for *BWEIGHT*. Include as explanatory variables *MMARRIED*, *MAGE*, *PRENATAL1*, and *FBABY*, along with *SMOKE2*, *SMOKE3*, and *SMOKE4*. Interpret the estimated coefficients of *SMOKE2*, *SMOKE3*, and *SMOKE4*. Does smoking 1–5 cigarettes per day have a statistically significant negative effect on infant birthweight?

c. Using the results in (b), test the null hypothesis that smoking 11 or more cigarettes per day reduces birthweight by no more than smoking 6–10 cigarettes per day, against the alternative that smoking 11 or more cigarettes per day reduces birthweight by more than smoking 6–10 cigarettes per day.

d. Using the results in (b), test the null hypothesis that smoking 11 or more cigarettes per day reduces birthweight by no more than smoking 1–5 cigarettes per day, against the alternative that smoking 11 or more cigarettes per day reduces birthweight by more than smoking 1–5 cigarettes per day.

e. Estimate a regression model for *BWEIGHT*. Include as explanatory variables *MMARRIED*, *MAGE*, *PRENATAL1*, and *FBABY*. Estimate the model separately for $MSMOKE = 0$, 1, 2, and 3. Using each model, estimate the expected birthweight of a child of a married woman who is 25 years old whose first prenatal visit was in the first trimester and who had already given birth to at least one child. What do you observe?

f. Estimate the linear probability model with dependent variable *LBWEIGHT* as a function of explanatory variables *MMARRIED*, *MAGE*, *PRENATAL1*, and *FBABY*, along with *MSMOKE*. Predict the probability of a low-birthweight infant for $MSMOKE = 0$, 1, 2, and 3 of a married woman who is 25 years old whose first prenatal visit was in the first trimester and who had already given birth to at least one child. What do you observe?

7.20 In this exercise, we will explore some of the factors predicting costs at American universities using the data file *poolcoll2* and observations outside the great recession. Let $TC =$ the real ($2008) total cost per student, $FTUG =$ number of full-time undergraduate students, $FTGRAD =$ number of full-time graduate students, $FTEF =$ full-time faculty per 100 students, $CF =$ number of contract faculty per 100 students, $FTENAP =$ full-time nonacademic professionals per 100 students.

a. Estimate the regression of $\ln(TC)$ on the remaining variables. What are the predicted effects of additional undergraduate students and graduate students on total cost per student?

b. What are the predicted effects of additional full-time faculty, contract faculty, and nonacademic professionals on total cost per student?

c. Add the indicator variable *PRIVATE* to the model. Do you predict higher or lower total cost per student at private universities? Is this a statistically significant factor in predicting total cost per student?

d. Add to the model not only *PRIVATE* but also $PRIVATE \times FTEF$. Are these variables individually and jointly significant at the 5% level?

e. Add to the model not only *PRIVATE* but also *PRIVATE* times all the other variables. Test the joint significance of *PRIVATE* and *PRIVATE* times all the other variables using an F-test. What do you conclude about the model in (a) that does not distinguish between private and public universities?

f. Estimate the model in (a) twice, once for private universities and once for public universities. Call the sum of squared residuals for the private universities $SSE1$, and the sum of squared residuals for the public universities $SSE0$. Compare $SSE1 + SSE0$ to the sum of squared residuals in part (e).

7.21 In this exercise, we explore some of the factors predicting costs at American public universities using the data file *pubcoll*. Let TC = the real ($2008) total cost per student, $FTUG$ = number of full-time undergraduate students, $FTGRAD$ = number of full-time graduate students, $FTEF$ = full-time faculty per 100 students, CF = number of contract faculty per 100 students, and $FTENAP$ = full-time nonacademic professionals per 100 students.

- **a.** Estimate the regression of $\ln(TC)$ on the remaining variables. What are the predicted effects of additional undergraduate students and graduate students on total cost per student?
- **b.** What are the predicted effects of additional full-time faculty, contract faculty, and nonacademic professionals on total cost per student?
- **c.** Add indicator variables for the years 1989, 1991, 1999, 2005, 2008, 2010, and 2011. Are these variables jointly and individually significant? Using your favorite site for macroeconomic data, plot the quarterly percentage change in the real U.S. GDP from January 1987 to January 1993. Does this help explain the signs and significance of any of the indicator variable coefficients?
- **d.** The variable $CRASH = 1$ during 2008, 2010, and 2011. Add to the model in (c) interactions between $CRASH$ and each of the variables $FTEF$, CF, and $FTENAP$. Are these variables individually significant at the 5% level? Are they jointly significant?
- **e.** Add to the model in (d) interactions between $CRASH$ and each of the variables $FTUG$ and $FTGRAD$. Considering all the interaction variables, which are significant at the 5% level? Test the joint significance of all the interaction variables at the 5% level.

7.22 In this exercise, we explore some of the factors predicting costs at American public universities using the data file *pubcoll*. Let TC = the real ($2008) total cost per student, $FTUG$ = number of full-time undergraduate students, $FTGRAD$ = number of full-time graduate students, $FTEF$ = full-time faculty per 100 students, CF = number of contract faculty per 100 students, and $FTENAP$ = full-time nonacademic professionals per 100 students. Use only the data for years prior to 2008. Include in the model year indicator variables $D1989$, $D1991$, $D1999$, and $D2005$.

- **a.** Estimate the regression of $\ln(TC)$ on the remaining variables. What are the predicted effects of additional undergraduate students and graduate students on total cost per student?
- **b.** What are the predicted effects of additional full-time faculty, contract faculty, and nonacademic professionals on total cost per student?
- **c.** Using the estimates from part (a), compute the normal and corrected predictors of total cost using 2005 data for University of Arizona (unitid = 104179), Indiana University-Bloomington (unitid 151351), and The University of Texas at Austin (unitid = 228778). Compare the predicted values to the reported TC for 2005. Which schools had actual total cost TC higher than predicted?
- **d.** Add an indicator variable for each different university except the first, which is the reference group. Test the joint significance of these indicator variables at the 5% level of significance using the F-test given in equation (6.4). Are there individual differences among the universities?
- **e.** Using the estimates from part (d), compute the normal and corrected predictors of total cost using 2005 data for University of Arizona (unitid = 104179), Indiana University-Bloomington (unitid 151351), and The University of Texas at Austin (unitid = 228778). Compare the predicted values to the reported TC for 2005. Which schools had actual total cost TC higher than predicted?

7.23 In the STAR experiment (Section 7.5.3), children were randomly assigned within schools into three types of classes: small classes with 13–17 students, regular-sized classes with 22–25 students, and regular-sized classes with a full-time teacher aide to assist the teacher. Student scores on achievement tests were recorded as well as some information about the students, teachers, and schools. Data for the kindergarten classes is contained in the data file *star5_small2*.

- **a.** Calculate the average of $MATHSCORE$ for (i) students in regular-sized classrooms with full-time teachers but no aide; (ii) students in regular-sized classrooms with full-time teachers and an aide; and (iii) students in small classrooms. What do you observe about test scores in these three types of learning environments?
- **b.** Estimate the regression model $MATHSCORE_i = \beta_1 + \beta_2 SMALL_i + \beta_3 AIDE_i + e_i$, where $AIDE$ is an indicator variable equaling 1 for classes taught by a teacher and an aide, and 0 otherwise. What is the relation of the estimated coefficients from this regression to the sample means in part (a)? Test the statistical significance of β_3 at the 5% level.
- **c.** To the regression in (b) add the additional explanatory variable $TCHEXPER$. Is this variable statistically significant? Does its addition to the model affect the estimates of β_2 and β_3? Construct a 95% interval estimate of expected math score for a student in a small class with a teacher having

10 years of experience. Construct a 95% interval estimate of expected math score for a student in a class with an aide and having a teacher with 10 years of experience. Calculate the least squares residuals from this model, calling them *EHAT*. This variable will be used in the next part.

d. To the regression in (c), add the additional indicator variable *FREELUNCH*. Students from lower income households receive a free lunch at school. Is this variable statistically significant? Does its addition to the model affect the estimates of β_2 and β_3? What explains the sign of *FREELUNCH*? Calculate the sample average of *EHAT*, from part (c), for students receiving a free lunch, and for students who do not receive a free lunch. Are the residual averages consistent with the regression that includes *FREELUNCH*?

e. To the model in (d), add interaction variables between *FREELUNCH* and *SMALL*, *AIDE* and *TCHEXPER*. Are any of these individually significant? Test the joint significance of these three interaction variables at the 5% level. What do you conclude?

f. Carry out a Chow test for the equivalence of the regression $MATHSCORE_i = \beta_1 + \beta_2 SMALL_i + \beta_3 AIDE_i + \beta_4 TCHEXPER + e_i$ for students who receive a free lunch and those who do not receive a free lunch. How does this test result compare to the test result in part (e)?

7.24 Many cities in California have passed Inclusionary Zoning policies (also known as below-market housing mandates) as an attempt to make housing more affordable. These policies require developers to sell some units below the market price on a percentage of the new homes built. For example, in a development of 10 new homes each with market value $850,000, the developer may have to sell 5 of the units at $180,000. Means and Stringham (2012), and exercise 7.13, examine the effects of such policies on house prices and number of housing units available using 1990 and 2000 census data on California cities. Use the data file *means* for the following exercises.

a. Use *LNPRICE* and *LNUNITS* as dependent variables in difference-in-difference regressions, with explanatory variables *D*, the indicator variable for year 2000; *IZLAW*, and the interaction of *D* and *IZLAW*. Is the estimate of the treatment effect statistically significant, and of the anticipated sign?

b. To the regressions in (a) add the control variable *LMEDHHINC*. Interpret the estimate of the new variable, including its sign and significance. How does this addition affect the estimates of the treatment effect?

c. To the regressions in (b) add the variables 100(*EDUCATTAIN*), 100(*PROPPOVERTY*), and *LPOP*. Interpret the estimates of these new variables, including their signs and significance. How do these additions affect the estimates of the treatment effect?

d. Consider the differences-in-differences regression for *LNPRICE*

$$\ln\left(PRICE_{it}\right) = \beta_1 + \beta_2 IZLAW_i + \beta_3 D_t + \delta\left(IZLAW_i \times D_t\right) + \theta CITY_i + e_{it}$$

In this model, $CITY_i$ represents some unobservable characteristic of each city that stays constant over time. Write this model for the year 2000 $\left(D_t = 1\right)$. Write this model for the year 1990 $\left(D_t = 0\right)$. Subtract the expression for 1990 from the expression for 2000. The dependent variable is

$$DLNPRICE_i = \left[\ln\left(PRICE_{i,2000}\right) - \ln\left(PRICE_{i,2000}\right)\right] \simeq \%\Delta PRICE_i/100$$

which is the decimal equivalent of the percentage change in price for city *i*. What parameters and variables remain on the right-hand side after the subtraction?

e. Regress $DLNPRICE_i$ against $IZLAW_i$ and compare the result to the *LNPRICE* regression in part (a).

7.25 Professor Ray C. Fair's voting model was introduced in Exercise 2.23. He builds models that explain and predict the U.S. presidential elections. See his website at http://fairmodel.econ.yale.edu/vote2016/index2.htm and see in particular his paper entitled "Presidential and Congressional Vote-Share Equations: November 2014 Update." The basic premise of the model is that the Democratic party's share of the two-party [Democratic and Republican] popular vote is affected by a number of factors relating to the economy, and variables relating to the politics, such as how long the incumbent party has been in power, and whether the President is running for reelection. Data for 1916–2016 are in the data file *fair5*. The dependent variable is *VOTE* = percentage share of the popular vote won by the Democratic party. In addition to *GROWTH* and *INFLAT*, the explanatory variables include the following:

INCUMB = 1 if there is a Democratic incumbent at the time of the election and −1 if there is a Republican incumbent.

GOODNEWS = (number of quarters in the first 15 quarters of the administration in which the growth rate of real per capita GDP is greater than 3.2% at an annual rate except for 1920, 1944, and 1948, where the values are zero) × *INCUMB*.

DPER = 1 if the incumbent is running for election and 0 otherwise.

DUR = 0 if the Democratic party has been in power for one term, 1[−1] if the Democratic [Republican] party has been in power for two consecutive terms, 1.25[−1.25] if the Democratic [Republican] party has been in power for three consecutive terms, 1.50 for four consecutive terms, and so on.

WAR = 1 for the elections of 1920, 1944, and 1948 and 0 otherwise.

a. Consider the regression model

$$VOTE = \beta_1 + \beta_2 GROWTH + \beta_3 INFLAT + \beta_4 GOODNEWS + \beta_5 DPER$$
$$+ \beta_6 DUR + \beta_7 INCUMB + \beta_8 WAR + e$$

Discuss the anticipated effects of the dummy variable *DPER*.

b. The variable *INCUMB* is somewhat different than dummy variables we have considered. Write out the regression function $E(VOTE)$ when there is a Democratic incumbent. Write out the regression function $E(VOTE)$ when there is a Republican incumbent. Recall that the signs of *GOODNEWS*, *GROWTH*, and *INFLAT* depend on *INCUMB*. Discuss the effects of this specification.

c. Use the data for the period 1916–2012 to estimate the proposed model. Discuss the estimation results. Are the signs as expected? Are the estimates statistically significant? How well does the model fit the data?

d. Use the regression result from part (c) to predict the value of *VOTE* for the 2016 election using the actual values of the explanatory variables.

e. Use the regression result from part (c) to construct a 95% prediction interval for the value of *VOTE* for the 2016 election using the actual values of the explanatory variables.

f. Use the data for the period 1916–2012 to estimate the proposed model. In election year 2016, *INCUMB* = 1, *DPER* = 0, *DUR* = 1, and *WAR* = 0. Using *GROWTH* = 2.16, *INFLAT* = 1.37, and *GOODNEWS* = 3, predict the vote in favor of the Democratic party candidate in 2016.

g. Using the results in (f), predict the vote in favor of the Democratic party in 2016 if *GOODNEWS* = 3, *GROWTH* = 2.16, and *INFLAT* = 0.

h. Using the results in (f), predict the vote in favor of the Democratic party in 2016 if *GOODNEWS* = 3, *GROWTH* = 4.0, and *INFLAT* = 0.

7.26 The data file *br2* contains data on 1080 house sales in Baton Rouge, Louisiana, during July and August 2005. The variables are: *PRICE* ($), *SQFT* (total square feet), *BEDROOMS* (number), *BATHS* (number), *AGE* (years), *OWNER* (= 1 if occupied by owner; 0 if vacant or rented), *TRADITIONAL* (= 1 if traditional style; 0 if other style), *FIREPLACE* (= 1 if present), *WATERFRONT* (= 1 if on waterfront).

a. Compute the data summary statistics and comment. In particular, construct a histogram of *PRICE*. What do you observe?

b. Estimate a regression model explaining ln(*PRICE*/1000) as a function of the remaining variables. Divide the variable *SQFT* by 100 prior to estimation. Comment on how well the model fits the data. Discuss the signs and statistical significance of the estimated coefficients. Are the signs what you expect? Give an exact interpretation of the coefficient of *WATERFRONT.*

c. Create a variable that is the product of *WATERFRONT* and *TRADITIONAL*. Add this variable to the model and reestimate. What is the effect of adding this variable? Interpret the coefficient of this interaction variable and discuss its sign and statistical significance.

d. It is arguable that the traditional style homes may have a different regression function from the diverse set of nontraditional styles. Carry out a Chow test of the equivalence of the regression models for traditional versus nontraditional styles. What do you conclude?

e. Predict the value of a traditional style house with 2500 square feet of area, that is 20 years old, which is owner occupied at the time of sale, with a fireplace, but no pool, and not on the waterfront.

7.27 The three most important words in real estate are "location, location, location!" We explore this question using 500, single-family home sales in Baton Rouge, LA from 2009 to 2013 in the data file *collegetown*. See *collegetown.def* for variable definitions.

a. Estimate the log-log model $\ln(PRICE) = \beta_1 + \beta_2 \ln(SQFT) + \delta_1 CLOSE + e$. Interpret the estimated coefficients of ln(*SQFT*) and *CLOSE*. Is the location variable *CLOSE* statistically significant at the 5% level?

b. Estimate the log-log model $\ln(PRICE) = \beta_1 + \beta_2 \ln(SQFT) + \delta_2[CLOSE \times \ln(SQFT)] + e$. Interpret the estimated coefficients of ln(*SQFT*) and $[CLOSE \times \ln(SQFT)]$. Is the location variable $[CLOSE \times \ln(SQFT)]$ statistically significant at the 5% level?

c. Estimate the log-log model

$$\ln(PRICE) = \beta_1 + \beta_2 \ln(SQFT) + \delta_1 CLOSE + \delta_2[CLOSE \times \ln(SQFT)] + e$$

Are the location variables *CLOSE* and $[CLOSE \times \ln(SQFT)]$ individually and jointly statistically significant at the 5% level?

d. Using the model in (c), predict the prices of two houses with 2500 square feet, one close to the university and another that is not close. Use the corrected predictor.

e. Add *FIREPLACE*, *TWOSTORY*, and *OCCUPIED* to the model in (c). How do these features affect the price of a house?

f. Carry out a Chow test for the log-log model, comparing houses that are close to the university to those that are not close, using explanatory variables ln(*SQFT*), *FIREPLACE*, *TWOSTORY*, and *OCCUPIED*. What is the *p*-value of the test?

7.28 How much of an incumbency advantage do winners in U.S. House elections enjoy? This is the topic of a paper by David S. Lee (2008) "Randomized experiments from nonrandom selection in U.S. House elections," *Journal of Econometrics*, 142(2), 675–697. Lee uses a regression discontinuity approach to estimate the effect. There are 435 Congressional districts in the United States and elections are held every 2 years. Representatives serve a term of 2 years. We employ a subset of Lee's data. The data file *rddhouse_small* has 1200 observations. See the *rddhouse_small.def* for data details. The data file *rddhouse* is larger. The forcing variable is *SHARE*, which is the Democratic share of the votes in a election in year *t* minus 0.50, so that *SHARE* is the Democratic margin of victory. The outcome of interest is the Democratic share of the vote in the next election, *SHARENEXT*.

a. Create a scatter plot with *SHARE* on the horizontal axis and *SHARENEXT* on the vertical axis. Does there appear to be positive relationship, an inverse relationship, or no relationship?

b. The dummy variable $D = 1$ if $SHARE > 0$ and $D = 0$ if $SHARE < 0$. Estimate the regression model with *SHARENEXT* as dependent variable, and *SHARE*, *D,* and $SHARE \times D$ as explanatory variables. Interpret the magnitudes, signs, and significance of the coefficients of *D* and $SHARE \times D$. Graph the fitted value from this regression against *SHARE*.

c. The variable *BIN* is the center of an interval of width 0.005, starting at −0.25. There are 100 bins between −0.25 and 0.25. Define a "narrow" win or loss as being an election where the margin of victory, or loss, is within the interval −0.005 to 0.005. Calculate the sample means of *SHARENEXT* when $BIN = -0.0025$ and when $BIN = 0.0025$. Is the difference in means an estimate of the value of incumbency? Explain how.

d. Treat the two groups created in (c) as two populations. Carry out a test of the difference between the two population means using the test in Appendix C.7.2, Case 1. Using a two-tail test and the 5% level of significance, do we reject the equality of the two population means, or not?

e. The variables *SHARE2*, *SHARE3*, and *SHARE4* are *SHARE* raised to the second, third, and fourth power, respectively. Estimate the regression model with *SHARENEXT* as dependent variable, with explanatory variables *SHARE* and its powers, *D* and *D* times *SHARE* and its powers. Interpret the magnitudes, signs, and significance of the coefficients of *D*, and *D* times *SHARE*.

f. Graph the fitted value from the regression in (e) against *SHARE*. Is the fitted line similar to the one in (b)?

g. Estimate the regression with *SHARENEXT* as dependent variable with explanatory variables *SHARE* and its powers, for the observations when $D = 0$. Reestimate the regression for the observations when $D = 1$. Compare these results to those in (e).

h. The variable *BIN* in part (c) was created using the equation $BIN = SHARE - \text{mod}(SHARE, 0.005) + 0.0025$, where "mod" is the "modulus operator," a common software function. In particular, $\text{mod}(x, y) = x - y \times \text{floor}(x/y)$ where the operator "floor" rounds the argument down to the next integer. Explain how this operator works in this application to create "bins" of width 0.005.

7.29 How much of an incumbency advantage do winners in U.S. Senate elections enjoy? This issue is examined by Matias D. Cattaneo, Brigham R. Frandsen and Rocío Titiunik (2015) "Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate, *Journal of Causal Inference*, 3(1): 1–24.[32] As they describe (p. 11): "Term length in the U.S. Senate is 6 years and there are 100 seats. These Senate seats are divided into three classes of

[32] Also in "Robust Data-Driven Inference in the Regression-Discontinuity Design," by Sebastian Calonico, Matias D. Cattaneo and Rocio Titiunik, *Stata Journal* 14(4): 909–946, 4th Quarter 2014.

roughly equal size (Class I, Class II, and Class III), and every 2 years only the seats in one class are up for election. As a result, the terms are staggered: In every general election, which occurs every 2 years, only one-third of Senate seats are up for election. Each state elects two senators in different classes to serve a 6-year term in popular statewide elections. Since its two senators belong to different classes, each state has Senate elections separated by alternating 2-year and 4-year intervals." We employ a subset of their data, contained in the file *rddsenate*. See *rddsenate.def* for data details. The forcing variable is *MARGIN*, which is the Democratic share of the votes in an election in year t minus 50: it is the Democratic margin of victory. The outcome of interest is the Democratic share of the vote in the next election for that Senate seat, *VOTE*.

- **a.** Create a scatter plot with *MARGIN* on the horizontal axis and *VOTE* on the vertical axis. Does there appear to be a positive relationship, an inverse relationship, or no relationship?
- **b.** The dummy variable $D = 1$ if $MARGIN > 0$ and $D = 0$ if $MARGIN < 0$. Estimate the regression model with *VOTE* as dependent variable, and *MARGIN*, *D*, and $MARGIN \times D$ as explanatory variables. Interpret the magnitudes, signs, and significance of the coefficients of *D* and $MARGIN \times D$. Graph the fitted value from this regression against *MARGIN*.
- **c.** The variable *BIN* is the center of an interval of width 5, starting at −97.5 and ending at 102.5. Define a "narrow" win or loss as being an election where the margin of victory, or loss, is within the interval −2.5 to 2.5. Calculate the sample means of *VOTE* when $BIN = -2.5$ and when $BIN = 2.5$. Is the difference in means an estimate of the value of incumbency? Explain how.
- **d.** Treat the two groups created in (c) as two populations. Carry out a test of the difference between the two population means using the test in Appendix C.7.2, Case 1: Using a two-tail test and the 5% level of significance, do we reject the equality of the two population means, or not?
- **e.** The variables *MARGIN2*, *MARGIN3*, and *MARGIN4* are *MARGIN* raised to the second, third, and fourth powers, respectively. Estimate the regression model with *VOTE* as dependent variable, with explanatory variables *MARGIN* and its powers, *D* and *D* times *MARGIN* and its powers. Interpret the magnitudes, signs, and significance of the coefficients of *D* and *D* times *MARGIN*.
- **f.** Graph the fitted value from the regression in (e) against *MARGIN*. Is the fitted line similar to the one in (b)?
- **g.** How would the results of (e) compare to the regression with *VOTE* as dependent variable with explanatory variables *MARGIN* and its powers, for the observations when $D = 0$. What if the regression was estimated for the observations when $D = 1$?

7.30 What effect does having public health insurance have on the number of doctor visits a person has during a year? Using 1988 data, *rwm88_small*, from Germany we will explore this question. The data file *rwm88* contains more observations. The data were used by Regina T. Riphahn, Achim Wambach, and Andreas Million, "Incentive Effects in the Demand for Health Care: A Bivariate Panel Count Data Estimation," *Journal of Applied Econometrics*, Vol. 18, No. 4, 2003, pp. 387–405.

- **a.** Construct a histogram of *DOCVIS*. How many doctor visits do most patients in the survey have during the year? What are the mean and median number of doctor visits? What is the 90th percentile?
- **b.** Test the null hypothesis that the population mean number of doctor visits for those with public insurance is the same as those who do not have public insurance. Use the 5% level of significance and a one-tail test.
- **c.** Estimate the regression model with dependent variable *DOCVIS* and explanatory variables *FEMALE*, *HHKIDS*, *MARRIED*, *SELF*, *EDUC2*, *HHNINC2*. Comment on the signs and significance of these predictor variables.
- **d.** Estimate the regression model with dependent variable *DOCVIS* and explanatory variables *FEMALE*, *HHKIDS*, *MARRIED*, *SELF*, *EDUC2*, *HHNINC2* separately for those with public insurance and those who do not have public insurance. Use equation (7.37) to obtain the estimate of the average treatment effect of public insurance.
- **e.** Estimate the regression model with dependent variable *DOCVIS* and the explanatory variables *FEMALE*, *HHKIDS*, *MARRIED*, *SELF*, *EDUC2*, *HHNINC2* in "deviation from the mean" form. That is, for each variable x create the variable $\tilde{x} = x - \bar{x}$, where $\bar{x}$ is the sample mean. Compare these results to those in (c).
- **f.** Estimate the regression model with dependent variable *DOCVIS* and the explanatory variables *FEMALE*, *HHKIDS*, *MARRIED*, *SELF*, *EDUC2*, *HHNINC2*, along with *PUBLIC* and *PUBLIC* times each of the variables in deviation about the mean form. What is the estimated average treatment effect? Is it statistically significant at the 5% level?

Appendix 7A Details of Log-Linear Model Interpretation

You may have noticed that in Section 7.3, while discussing the interpretation of the log-linear model, we omitted the error term, and we did not discuss the regression function $E(WAGE|\mathbf{x})$. To do so, we make use of the properties of the log-normal distribution in Appendix B.3.9 and discussed in Problem 7.11. There we noted that for the log-linear model $\ln(y) = \beta_1 + \beta_2 x + e$, if the error term $e \sim N(0, \sigma^2)$, then the expected value of y is

$$E(y|\mathbf{x}) = \exp\left(\beta_1 + \beta_2 x + \sigma^2/2\right) = \exp\left(\beta_1 + \beta_2 x\right) \times \exp\left(\sigma^2/2\right)$$

Starting from this equation, we can explore the interpretation of dummy variables and interaction terms.

Let D be a dummy variable. Adding this to our log-linear model, we have $\ln(y) = \beta_1 + \beta_2 x + \delta D + e$ and

$$E(y|\mathbf{x}) = \exp\left(\beta_1 + \beta_2 x + \delta D\right) \times \exp\left(\sigma^2/2\right)$$

If we let $E(y_1|\mathbf{x})$ and $E(y_0|\mathbf{x})$ denote the cases when $D = 1$ and $D = 0$, respectively, then we can compute their percentage difference as

$$\begin{aligned}
\%\Delta E(y|\mathbf{x}) &= 100\left[\frac{E(y_1|\mathbf{x}) - E(y_0|\mathbf{x})}{E(y_0|\mathbf{x})}\right]\%, \\
&= 100\left[\frac{\exp(\beta_1 + \beta_2 x + \delta) \times \exp(\sigma^2/2) - \exp(\beta_1 + \beta_2 x) \times \exp(\sigma^2/2)}{\exp(\beta_1 + \beta_2 x) \times \exp(\sigma^2/2)}\right]\% \\
&= 100\left[\frac{\exp(\beta_1 + \beta_2 x)\exp(\delta) - \exp(\beta_1 + \beta_2 x)}{\exp(\beta_1 + \beta_2 x)}\right]\% = 100\left[\exp(\delta) - 1\right]\%
\end{aligned}$$

The interpretation of dummy variables in log-linear models carries over to the regression function. The percentage difference in the *expected* value of y is $100[\exp(\delta) - 1]\%$.

Appendix 7B Derivation of the Differences-in-Differences Estimator

To verify the expression for the differences-in-differences estimator in (7.14), note that the numerator can be expressed as

$$\begin{aligned}
\sum_{i=1}^{N}\left(d_i - \bar{d}\right)(y_i - \bar{y}) &= \sum_{i=1}^{N} d_i(y_i - \bar{y}) - \bar{d}\sum_{i=1}^{N}(y_i - \bar{y}) \\
&= \sum_{i=1}^{N} d_i(y_i - \bar{y}) \quad \left[\text{using } \sum_{i=1}^{N}(y_i - \bar{y}) = 0\right] \\
&= \sum_{i=1}^{N} d_i y_i - \bar{y}\sum_{i=1}^{N} d_i \\
&= N_1\bar{y}_1 - N_1\bar{y} \\
&= N_1\bar{y}_1 - N_1\left(N_1\bar{y}_1 + N_0\bar{y}_0\right)/N \\
&= \frac{N_0 N_1}{N}(\bar{y}_1 - \bar{y}_0) \quad \left[\text{using } N = N_1 + N_0\right]
\end{aligned}$$

The denominator of b_2 is

$$\begin{aligned}
\sum_{i=1}^{N}\left(d_i-\overline{d}\right)^2 &= \sum_{i=1}^{N} d_i^2 - 2\overline{d}\sum_{i=1}^{N} d_i + \sum_{i=1}^{N}\overline{d}^2 \\
&= \sum_{i=1}^{N} d_i - 2\overline{d}N_1 + N\overline{d}^2 \quad \left[\text{using } d_i^2 = d_i \text{ and } \sum_{i=1}^{N} d_i = N_1\right] \\
&= N_1 - 2\frac{N_1}{N}N_1 + N\left(\frac{N_1}{N}\right)^2 \\
&= \frac{N_0N_1}{N} \quad \left[\text{using } N = N_0 + N_1\right]
\end{aligned}$$

Combining the expressions for numerator and denominator, we obtain the result for the difference estimator in (7.14).

Appendix 7C The Overlap Assumption: Details

To see the impact of the difference of means, $\overline{x}_1 - \overline{x}_0$, on the average treatment effect we begin with the separate regressions on the control and treatment groups used to compute the average treatment effect in Section 7.6.4, $\hat{\alpha}_0 + \hat{\beta}_0 x_i$ and $\hat{\alpha}_1 + \hat{\beta}_1 x_i$. Using the property of least squares fitted lines, the estimated intercepts are

$$\hat{\alpha}_0 = \overline{y}_0 - \hat{\beta}_0\overline{x}_0 \text{ and } \hat{\alpha}_1 = \overline{y}_1 - \hat{\beta}_0\overline{x}_1$$

We can express the sample mean of the control variable as

$$\begin{aligned}
\overline{x} &= N^{-1}\sum_{i=1}^{N} x_i = N^{-1}\left[\sum_{i=1}^{N_0} x_i + \sum_{i=N_0+1}^{N} x_i\right] = N^{-1}\left[N_0\overline{x}_0 + N_1\overline{x}_1\right] \\
&= \frac{N_0\overline{x}_0}{N} + \frac{N_1\overline{x}_1}{N} = f_0\overline{x}_0 + f_1\overline{x}_1
\end{aligned}$$

The control variable sample mean $\overline{x}$ is a weighted average of $\overline{x}_0$ and $\overline{x}_1$, where the weight f_0 is the fraction of the observations in the control group and f_1 is the fraction of observations in the treatment group. Then

$$\begin{aligned}
\hat{\tau}_{ATE} &= \left(\hat{\alpha}_1 - \hat{\alpha}_0\right) + \left(\hat{\beta}_1 - \hat{\beta}_0\right)\overline{x} \\
&= \left[\left(\overline{y}_1 - \hat{\beta}_1\overline{x}_1\right) - \left(\overline{y}_0 - \hat{\beta}_0\overline{x}_0\right)\right] + \left(\hat{\beta}_1 - \hat{\beta}_0\right)\left(f_0\overline{x}_0 + f_1\overline{x}_1\right) \\
&= \left(\overline{y}_1 - \overline{y}_0\right) - \hat{\beta}_1\overline{x}_1 + \hat{\beta}_0\overline{x}_0 + f_0\hat{\beta}_1\overline{x}_0 + f_1\hat{\beta}_1\overline{x}_1 - f_0\hat{\beta}_0\overline{x}_0 - f_1\hat{\beta}_0\overline{x}_1 \\
&= \left(\overline{y}_1 - \overline{y}_0\right) + \left(f_1\hat{\beta}_1\overline{x}_1 - \hat{\beta}_1\overline{x}_1\right) - \left(f_0\hat{\beta}_0\overline{x}_0 - \hat{\beta}_0\overline{x}_0\right) + f_0\hat{\beta}_1\overline{x}_0 - f_1\hat{\beta}_0\overline{x}_1 \\
&= \left(\overline{y}_1 - \overline{y}_0\right) + \left(f_1 - 1\right)\hat{\beta}_1\overline{x}_1 - \left(f_0 - 1\right)\hat{\beta}_0\overline{x}_0 + f_0\hat{\beta}_1\overline{x}_0 - f_1\hat{\beta}_0\overline{x}_1
\end{aligned}$$

But

$$f_1 - 1 = \frac{N_1 - \left(N_0 + N_1\right)}{N_0 + N_1} = -\frac{N_0}{N_0 + N_1} = -f_0$$

and

$$f_0 - 1 = \frac{N_0 - \left(N_0 + N_1\right)}{N_0 + N_1} = -\frac{N_1}{N_0 + N_1} = -f_1$$

Therefore,

$$\begin{aligned}
\hat{\tau}_{ATE} &= \left(\overline{y}_1 - \overline{y}_0\right) - f_0\hat{\beta}_1\overline{x}_1 + f_1\hat{\beta}_0\overline{x}_0 + f_0\hat{\beta}_1\overline{x}_0 - f_1\hat{\beta}_0\overline{x}_1 \\
&= \left(\overline{y}_1 - \overline{y}_0\right) + \left(f_0\hat{\beta}_1 + f_1\hat{\beta}_0\right)\overline{x}_0 - \left(f_0\hat{\beta}_1 + f_1\hat{\beta}_0\right)\overline{x}_1 \\
&= \left(\overline{y}_1 - \overline{y}_0\right) - \left(f_0\hat{\beta}_1 + f_1\hat{\beta}_0\right)\left(\overline{x}_1 - \overline{x}_0\right)
\end{aligned}$$