Chapter 7
Using Indicator Variables

Chapter Contents

7.1 Indicator Variables 1 of 3

- Indicator variables allow us to construct models in which some or all regression model parameters, including the intercept, change for some observations in the sample
- Consider a **hedonic model** to predict the value of a house as a function of its characteristics:
  - Size, location, number of bedrooms, age
- Consider the square footage at first: (7.1) $PRICE = \beta_1 + \beta_2 SQFT + e$
- β_2 is the value of an additional square foot of living area and β_1 is the value of the land alone

7.1 Indicator Variables 2 of 3

- How do we account for location, which is a qualitative variable?
- 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, non-numeric characteristic
- We use the terms indicator variable and dummy variable interchangeably

7.1 Indicator Variables 3 of 3

- Generally, we define an indicator variable D as:

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

- So, to account for location, a qualitative variable, we would have:

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

7.1.1 Intercept Indicator Variables 1 of 4

- Adding our indicator variable to our model:
  - (7.3) $PRICE = \beta_1 + \delta D + \beta_2 SQFT + e$
- If our model is correctly specified, then:
  - (7.4) $E(PRICE|SQFT) \begin{cases} (\beta_1 + \delta) + \beta_2 SQFT \ when \ D = 1 \\ \beta_1 + \beta_2 SQFT \ when \ D = 1 \end{cases}$

7.1.1 Intercept Indicator Variables 2 of 4

- Adding the indicator variable D to the regression model causes a parallel shift in the relationship by the amount δ

- An indicator variable like *D* 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

- The least squares estimator's properties are not affected by the fact that one of the explanatory variables consists only of zeros and ones

7.1.1 Intercept Indicator Variables 3 of 4

- *D* is treated as any other explanatory variable
- We can construct an interval estimate for *D*, or we can test the significance of its least squares estimate
- The value $D = 0$ defines the **reference group**, or **base group**
- We could pick any base
- For example: $$LD = \begin{cases} 1 & \text{if property is not in the desirable neighborhood} \\ 0 & \text{if property is in the desirable neighborhood} \end{cases}$$

7.1.1 Intercept Indicator Variables 4 of 4

- Then our model would be: $PRICE = \beta_1 + \lambda LD + \beta_2 SQFT + e$
- Suppose we included both D and LD: $PRICE = \beta_1 + \delta D + \lambda LD + \beta_2 SQFT + e$
- 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**
- We have fallen into the **dummy variable trap**.
- By including only one of the indicator variables the omitted variable defines the reference group and we avoid the problem

7.1.2 Slope-Indicator Variables 1 of 3

- Suppose we specify our model as:
  - (7.5) $PRICE = \beta_1 + \beta_2 SQFT + \gamma\left(SQFT \times D\right) + e$
- The new variable (*SQFT* x *D*) is the product of house size and the indicator variable
- It 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

7.1.2 Slope-Indicator Variables 2 of 3

- Now we can write:
  - $E(PRICE|SQFT, D) = \beta_1 + \beta_2 SQFT + \gamma(SQFT\ x\ D)$

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

- The slope can be expressed as:

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

Figure 7.2

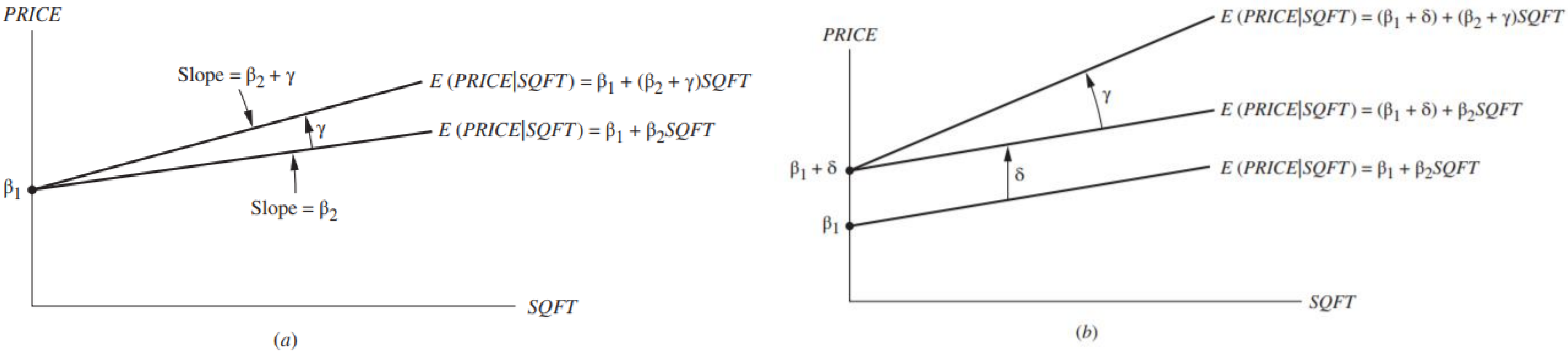

FIGURE 7.2 (*a*) A slope-indicator variable. (*b*) Slope- and intercept-indicator variables.

7.1.2 Slope-Indicator Variables 3 of 3

- Assume that house location affects both the intercept and the slope, then both effects can be incorporated into a single model:
  - (7.6) $PRICE = \beta_1 + \delta D + \beta_2 SQFT + \gamma\left(SQFT \times D\right) + e$
- The variable (*SQFTD*) is the product of house size and the indicator variable, and is called an **interaction variable**
- Alternatively, it is called a **slope-indicator variable** or a **slope dummy variable**

Now we can see that: $E(PRICE|SQFT) = \begin{cases} (\beta_1 + \delta) + (\beta_2 + \gamma)SQFT \text{ when } D = 1 \\ \beta_1 + \beta_2 SQFT \text{ when } D = 0 \end{cases}$

7.2 Applying Indicator Variables

We can apply indicator variables to a number of problems

7.2.1 Interactions Between Qualitative Factors

- Consider the wage equation:

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

- The expected value is:

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

7.2.2 Qualitative Factors with Several Categories 1 of 2

- Consider including regions in the wage equation:
  - (7.9) $WAGE = \beta_1 + \beta_2 EDUC + \delta_1 SOUTH + \delta_2 MIDWEST + \delta_3 WEST + e$
- the sum of the regional indicator variables is *NORTHEAST* + *SOUTH* + *MIDWEST* + *WEST* = 1
- Thus, the “intercept variable” x1 = 1 is an exact linear combination of the region indicators

7.2.2 Qualitative Factors with Several Categories 2 of 2

- Failure to omit one indicator variable will lead to the dummy variable trap

- Omitting one indicator variable defines a reference group so our equation is:

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

- The omitted indicator variable, *NORTHEAST*, identifies the reference

7.2.3 Testing the Equivalence of Two Regressions 1 of 6

- Suppose we have:

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

- And for two locations:

- $\text{E(WAGE|EDUC)} = \begin{cases} \propto_1 + \propto_2 SQFT & D = 1 \\ \beta_1 + \beta_2 SQFT & D = 0 \end{cases}$

- where $\alpha_1 = \beta_1 + \delta$ and $\alpha_2 = \beta_2 + \gamma$

7.2.3 Testing the Equivalence of Two Regressions 2 of 6

- 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
  - The **Chow test** is an *F*-test for the equivalence of two regressions

7.2.3 Testing the Equivalence of Two Regressions 3 of 6

- Now consider our wage equation:

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

- "*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

7.2.3 Testing the Equivalence of Two Regressions 4 of 6

- To test this, we specify:

- (7.10)

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

7.2.3 Testing the Equivalence of Two Regressions 5 of 6

- Now examine this version of (7.10):

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

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

7.2.3 Testing the Equivalence of Two Regressions 6 of 6

- 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, $\text{var}(e_i) = \sigma^2$, 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

7.2.4 Controlling for Time 1 of 2

- Indicator variables are also used in regressions using time-series data
- We may want to include an effect for different seasons of the year
- In the same spirit as seasonal indicator variables, annual indicator variables are used to capture year effects not otherwise measured in a model
- An economic regime is a set of structural economic conditions that exist for a certain period

7.2.4 Controlling for Time 2 of 2

- The idea is that economic relations may behave one way during one regime, but may behave differently during another
- An example of a regime effect: the investment tax credit:

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

- The model is then: $INV_t = \beta_1 + \delta ITC_t + \beta_2 GNP_t + \beta_3 GNP_{t-1} + e_t$
  - If the tax credit was successful, then $\delta > 0$

7.3 Log-Linear Models 1 of 2

- Consider the wage equation in log-linear form:
- (7.11) $\ln(WAGE) = \beta_1 + \beta_2 EDUC + \delta FEMALE$
- What is the interpretation of δ?
- FEMALE is an intercept dummy variable, creating a parallel shift of the log-linear relationship when FEMALE = 1

7.3 Log-Linear Models 2 of 2

- Expanding our model, we have:

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

- But what about the fact that the dependent variable is ln(WAGE)? Does that have an effect?

- Yes and there are two solutions

7.3.1 A Rough Calculation

- Let's first write the difference between females and males:

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

- This is approximately the percentage difference
- The estimated model is:
- $\ln(\widehat{WAGE}) = 1.6229 + 0.1024EDUC - 0.1778FEMALE$
  - $(se) \quad = (0.0692) \ (0.0048) \qquad (0.0279)$
- We estimate that there is a 24.32% differential between male and female wages

7.3.2 An Exact Calculation 1 of 2

- For a better calculation, 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,

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

7.3.2 An Exact Calculation 2 of 2

- Subtracting 1 from both sides:

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

- The percentage difference between wages of females and males is $100(e^{\delta} - 1)\%$

- We estimate the wage differential between males and females to be:

- $100(e^{\delta} - 1)\% = 100(e^{-0.2432} - 1)\% = -21.59\%$

7.4 The Linear Probability Model 1 of 6

- Many of the choices we make are “either-or” in nature:
  - 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

7.4 The Linear Probability Model

- Because we are trying to explain choice, the indicator variable is the dependent variable
- Let us represent the variable indicating a choice is a choice problem as:

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

- The probability that the first alternative is chosen is $P[\,y = 1] = p$
- The probability that the second alternative is chosen is $P[\,y = 0] = 1 - p$

7.4 The Linear Probability Model 3 of 6

- 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 $\text{var}(y) = p(1 - p)$

7.4 The Linear Probability Model

- A **linear probability model** is:
- $E(y|\boldsymbol{X}) = p = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k$
- An econometric model is:
- $E(y|\boldsymbol{X}) + e = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$

- The probability functions for y and e are:

y value	e value	Probability
1	$1-(\beta_1+\beta_2x_2+\cdots+\beta_Kx_K)$	p
0	$-(\beta_1+\beta_2x_2+\cdots+\beta_Kx_K)$	$1-p$

- The variance of the error term is

- $\text{var}(e|\boldsymbol{X}) = p(1-p) = (\beta_1+\beta_2x_2+\cdots+\beta_kx_k)((1-\beta_1-\beta_2x_2-\cdots-\beta_kx_k))$

7.4 The Linear Probability Model 6 of 6

- The error term is not homoskedastic
- The predicted values, $\widehat{E(y)} = \hat{p}$, can fall outside the (0, 1) interval
  - Any interpretation as probabilities would not make sense
- Despite these weaknesses, the linear probability model has the advantage of simplicity

Example 7.7 The linear probability model: an example from marketing

- A shopper must chose between Coke and Pepsi:
  - Define *COKE* as: $COKE = \begin{cases} 1 & \text{if Coke is chosen} \\ 0 & \text{if Pepsi is chosen} \end{cases}$
- The estimated equation is:
- $\hat{p}COKE = 0.8902 - 0.4009\text{PRATIO} + 0.0772\text{DISP\_COKE} - 0.1657\text{DISP\_PEPSI}$

  (se) = (0.0655) (0.0613) (0.0344) (0.0356)

7.5 Treatment Effects 1 of 3

- Avoid the faulty line of reasoning known as **post hoc, ergo propter hoc**
- One event’s preceding another does not necessarily make the first the cause of the second
- Another way to say this is embodied in the warning that “correlation is not the same as causation”
- 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

7.5 Treatment Effects 2 of 3

- Selection bias is also an issue when asking:
  - "How much does an additional year of education increase the wages of married women?"
  - "How much does participation in a job-training program increase wages?"
  - "How much does a dietary supplement contribute to weight loss?"
- Selection bias interferes with a straightforward examination of the data, and makes more difficult our efforts to measure a causal effect, or treatment effect

7.5 Treatment Effects 3 of 3

- We would like to randomly assign items to a **treatment group**, with others being treated as a **control group**
  - We could then compare the two groups
  - The key is a **randomized controlled experiment**
- The ability to perform randomized controlled experiments in economics is limited because the subjects are people, and their economic well-being is at stake

7.5.1 The Difference Estimator 1 of 2

- Define the indicator variable d as:

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

  - The model is then

$$(7.13) \quad y_i = \beta_1 + \beta_2 d_i + e_i, \quad i = 1, \ldots, N$$

- And the regression functions are: $E(y_i) = \begin{cases} \beta_1 + \beta_2 & \text{if in treatment group, } d_i = 1 \\ \beta_1 & \text{if in control group, } d_i = 0 \end{cases}$

7.5.1 The Difference Estimator 2 of 2

- A The least squares estimator for β_2, the **treatment effect**, is:

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

- with: $\bar{y}_1 = \sum_{i=1}^{N_1} y_i / N_1, \bar{y}_0 = \sum_{i=1}^{N_0} y_i / N_0$

- The estimator b_2 is called the **difference estimator**, because it is the difference between the sample means of the treatment and control groups

7.5.2 Analysis of the Difference Estimator 1 of 2

- The difference estimator can be rewritten as:

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

- To be unbiased, we must have: $E\left(\bar{e}_1 - \bar{e}_0\right) = E\left(\bar{e}_1\right) - E\left(\bar{e}_0\right) = 0$

- If we allow individuals to "self-select" into treatment and control groups, then:

7.5.2 Analysis of the Difference Estimator 2 of 2

$$E\left(\bar{e}_1\right) - E\left(\bar{e}_0\right)$$

- Is the selection bias in the estimation of the treatment effect
- We can eliminate the self-selection bias is we randomly assign individuals to treatment and control groups, so that there are no systematic differences between the groups, except for the treatment itself

7.5.3 The Differences-in-Differences Estimator 1 of 5

- Randomized controlled experiments are 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

7.5.3 The Differences-in-Differences Estimator 2 of 5

- Estimation of the treatment effect is based on data averages for the two groups in the two periods:

$$\hat{\delta} = \left(\hat{C} - \hat{E}\right) - \left(B - A\right)$$

$$= \left(\bar{y}_{Treatment,After} - \bar{y}_{Control,After}\right) - \left(\bar{y}_{Treatment,Before} - \bar{y}_{Control,Before}\right)$$

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

7.5.3 The Differences-in-Differences Estimator 3 of 5

- The sample means are:

$$\bar{y}_{Control,Before} = \hat{A} = \text{mean for control group before policy}$$

$$\bar{y}_{Treatment,Before} = \hat{B} = \text{mean for treatment group before policy}$$

$$\bar{y}_{Control,After} = \hat{E} = \text{mean for control group after policy}$$

$$\bar{y}_{Treatment,After} = \hat{C} = \text{mean for treatment group after policy}$$

7.5.3 The Differences-in-Differences Estimator 4 of 5

- Consider the regression model:

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

- The regression function is:

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

7.5.3 The Differences-in-Differences Estimator 5 of 5

- Using the points in the figure:

$$\delta = (C - E) - (B - A) = \left[(\beta_1 + \beta_2 + \beta_3 + \delta) - (\beta_1 + \beta_3) \right] - \left[(\beta_1 + \beta_2) - \beta_1 \right]$$

- Using the least squares estimates, we have:

$$\begin{aligned} \hat{\delta} &= \left[\left(b_1 + b_2 + \mathrm{b}_3 + \delta \right) - \left(b_1 + \mathrm{b}_3 \right) \right] - \left[\left(b_1 + b_2 \right) - b_1 \right] \\ &= \left(\bar{y}_{Treatment,After} - \bar{y}_{Conrol,After} \right) - \left(\bar{y}_{Treatment,Before} - \bar{y}_{Conrol,Before} \right) \end{aligned}$$

7.6 Treatment Effects and Causal Modeling

- This section presents extensions and enhancements
  - Using the framework of **potential outcomes**
  - Sometimes called the **Rubin Causal Model (RCM)**

7.6.1 The Nature of Causal Effects

- 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)
- 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

7.6.2 Treatment Effect Models 1 of 2

- Treatment effect models seek to estimate a causal effect
- Let the treatment be denoted as $d_i = 1$, and not receiving treatment $d_i = 0$
- 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
- The outcome we observe is $\begin{cases} y_{1i} & if \quad d_i = 1 \\ y_{0i} & if \quad d_i = 0 \end{cases}$

7.6.2 Treatment Effect Models 2 of 2

- 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):** $\tau_{ATE} = E(y_{1i} - y_{0i})$
- In an experiment the **treatment group is** $(d_i = 1)$ and the **control group is** $(d_i = 0)$
- d_i, is statistically independent of the potential outcomes y_{1i} and y_{0i} so that

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

- An unbiased estimator of the population average treatment effect is $\tau_{ATE} = y_1 - y_0$. This is the difference estimator in equation (7.14)

7.6.3 Decomposing the Treatment Effect 1 of 2

- (7.29) $E(y_i | d_i = 1) - E(y_i | d_i = 0) = [E(y_{1i} | d_i = 1) - E(y_{0i} | d_i = 1)] + [E(y_{0i} | d_i = 1) - E(y_{0i} | d_i = 0)]$
- The left-hand side is the difference in average outcomes for the treatment group ($d_i = 1$) and the control group ($d_i = 0$)
- $[E(y_{1i} | d_i = 1) - E(y_{0i} | d_i = 1)]$ is the **average treatment effect on the treated (ATT)** denoted as τ_{ATT}
- (7.30) $\tau_{ATT} = E(y_{1i} | d_i = 1) - E(y_{0i} | d_i = 1) = \tau_{ATE}$ when d_i is statistically independent of the potential outcomes

7.6.3 Decomposing the Treatment Effect 2 of 2

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

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

- 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**

7.6.4 Introducing Control Variables

- 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
- 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
- The key is an extension of the **conditional independence assumption**
- The **average treatment effect on the treated**, τ_{ATT}, where the subscript ATT denotes the target group, is obtained by estimating the pooled regression

7.6.5 The Overlap Assumption

- The so-called overlap assumption must hold, in addition to the conditional independence assumption in the equation

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

- 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
- If the difference in the sample means of the treatment and control groups is large, $\hat{\beta}_1$ and $\hat{\beta}_0$, have a larger influence in the estimate $\hat{\tau}_{ATT}$ of the average treatment effect.

7.6.6 Regression Discontinuity Designs

- Regression discontinuity (RD) designs 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. Those just below the cut-off point are a good comparison to those just above it
- x_i is the single variable determining whether an individual is assigned to the treatment group or control group (**forcing variable**)

Figure 7.4 Regression Discontinuity Design

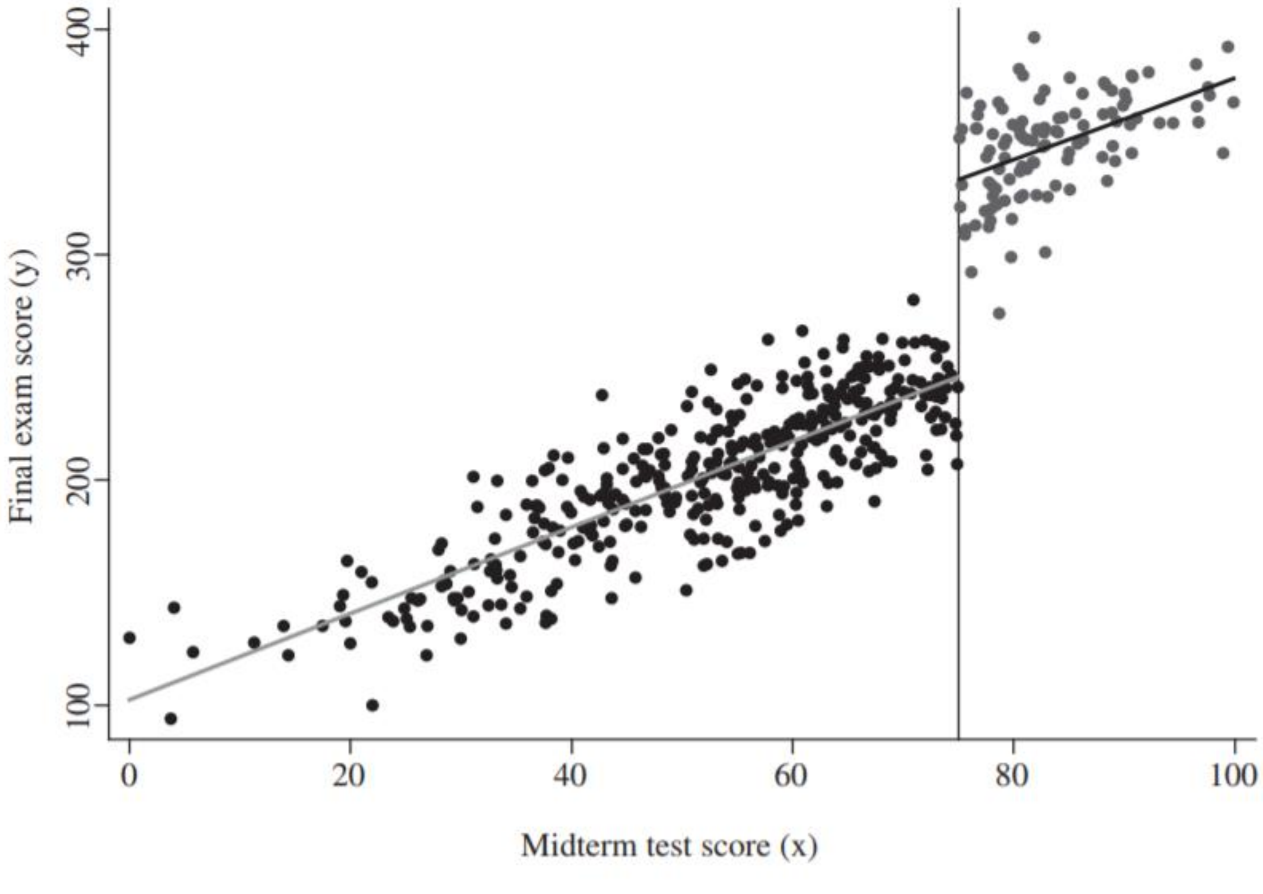

FIGURE 7.4 **Regression Discontinuity Design**

Figure 7.6 RDD bias

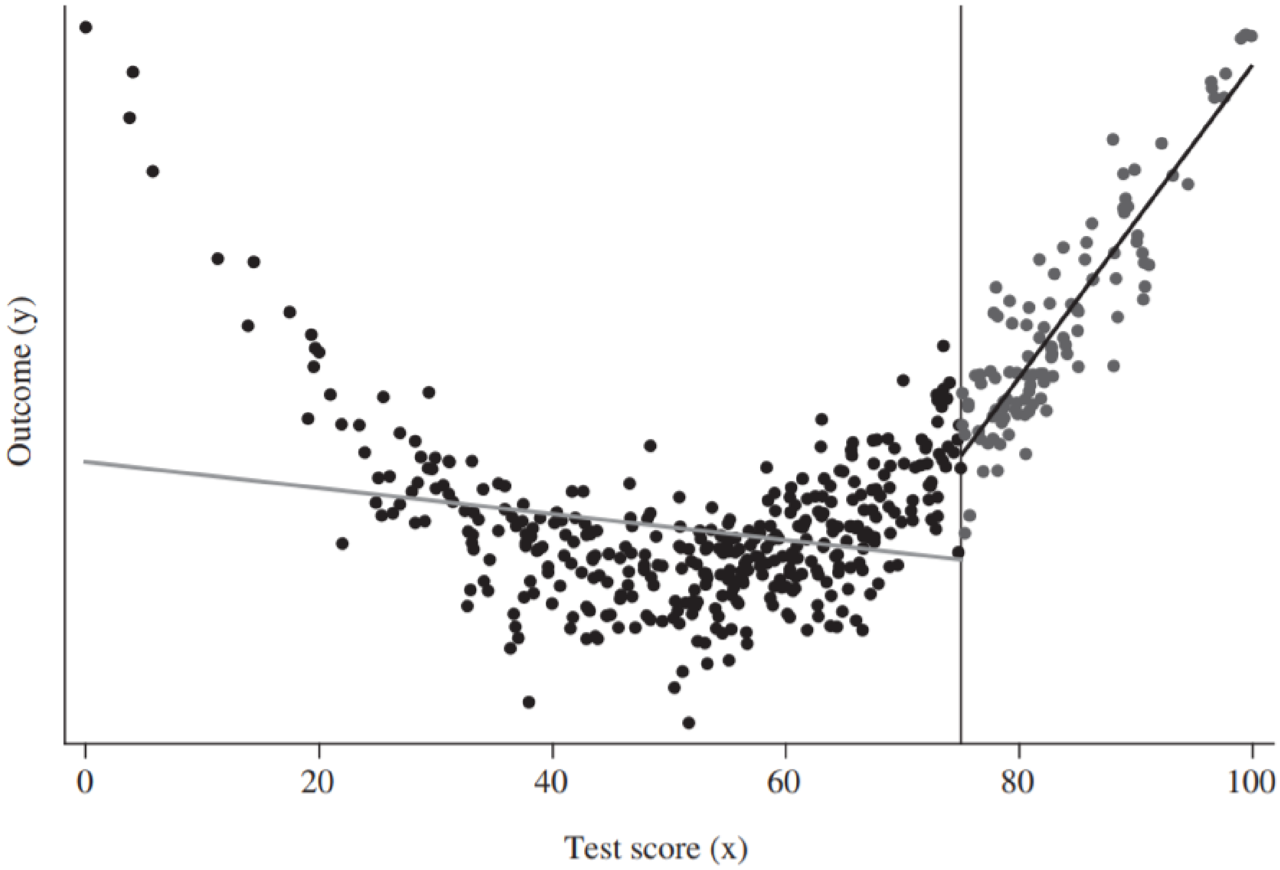

FIGURE 7.6 RDD bias

Key Words

- 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

Copyright