Chapter 7
Using Indicator Variables

Walter R. Paczkowski
Rutgers University

Chapter Contents

- 7.1 Indicator Variables
- 7.2 Applying Indicator Variables
- 7.3 Log-linear Models
- 7.4 The Linear Probability Model
- 7.5 Treatment Effects

- Economists develop and evaluate theories about economic behavior
  - Hypothesis testing procedures are used to test these theories
  - Theories economists develop sometimes provide **nonsample information** that can be used along with the sample information to estimate the parameters of a regression model
  - A procedure that combines these two types of information is called **restricted least squares**

6.1

Joint Hypothesis Testing

- 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:

Eq. 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

- 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

- Generally, we define an indicator variable D as:

Eq. 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

- Adding our indicator variable to our model:

Eq. 7.3

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

  – If our model is correctly specified, then:

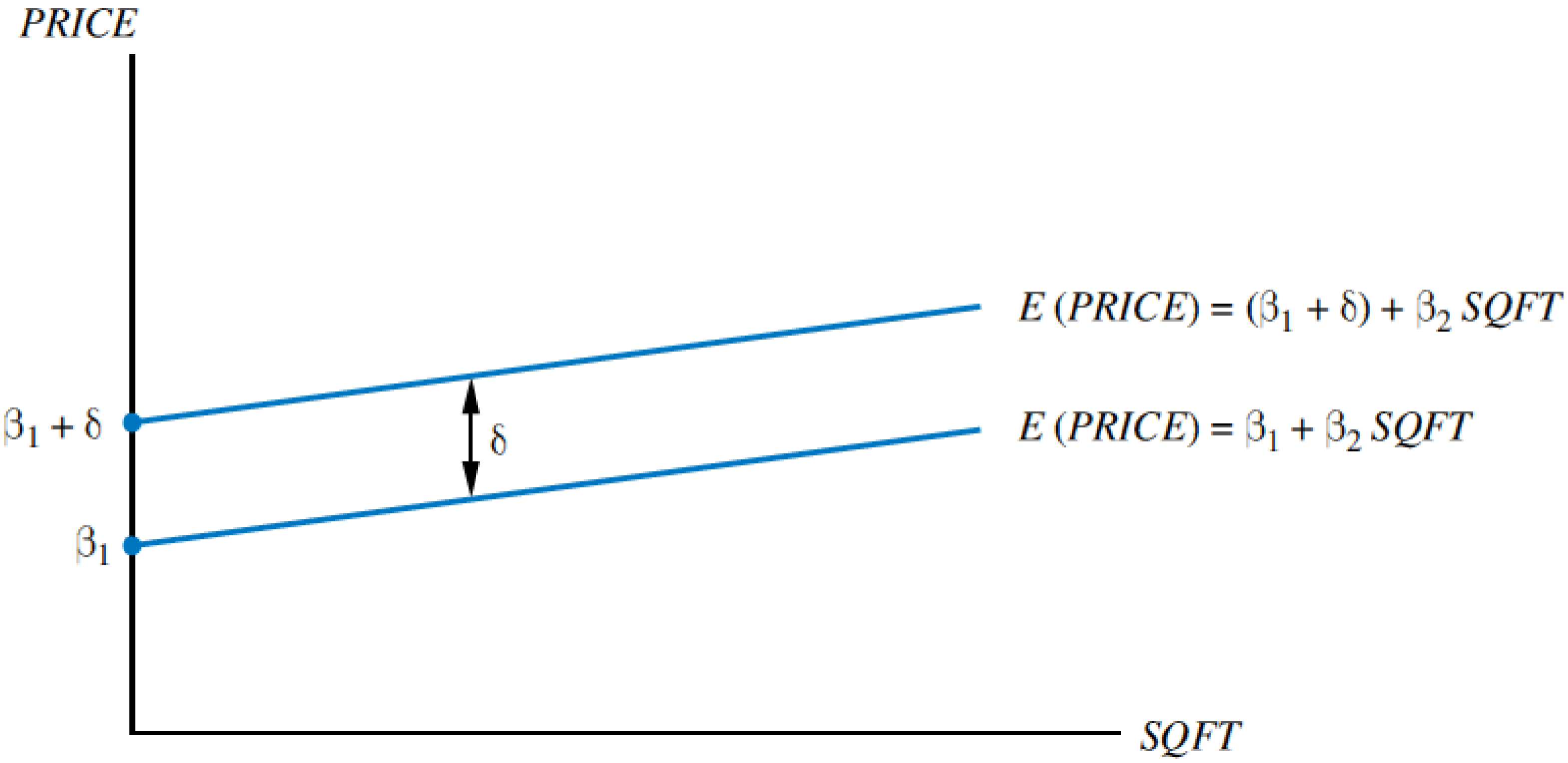
Eq. 7.4

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

- Adding an indicator variable 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
  - *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

FIGURE 7.1 An intercept indicator variable



- The value $D = 0$ defines the **reference group**, or **base group**
  - We could pick any base
  - For example:

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

- 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

Eq. 7.5

- Suppose we specify our model as:

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

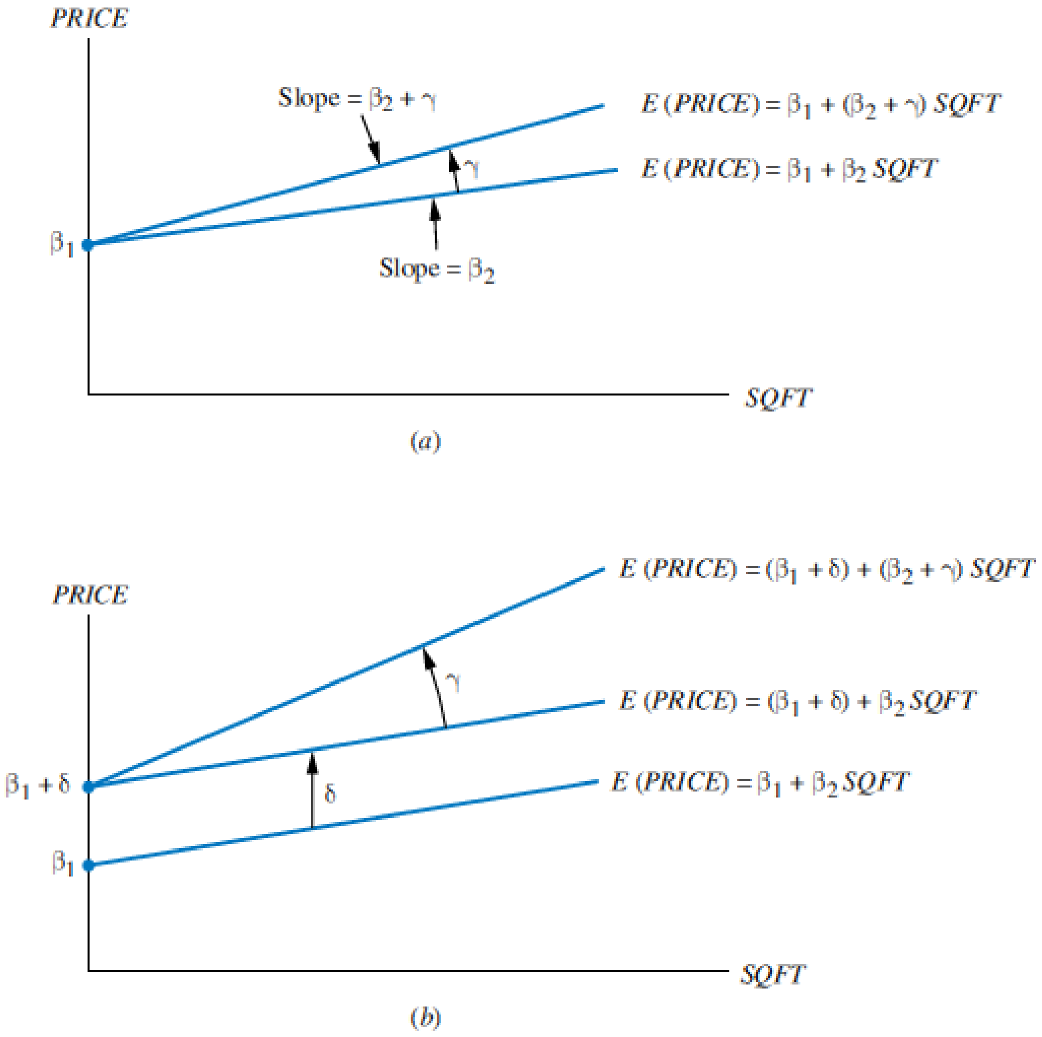
  - The new variable (*SQFT* x *D*) is the product of house size and the indicator variable
    - It is called an **interaction variable**, as it captures the interaction effect of location and size on house price
    - Alternatively, it is called a **slope-indicator variable** or a **slope dummy variable**, because it allows for a change in the slope of the relationship

- Now we can write:

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

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

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



- The slope can be expressed as:

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

- Assume that house location affects both the intercept and the slope, then both effects can be incorporated into a single model:

Eq. 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:

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

- Suppose an economist specifies a regression equation for house prices as:

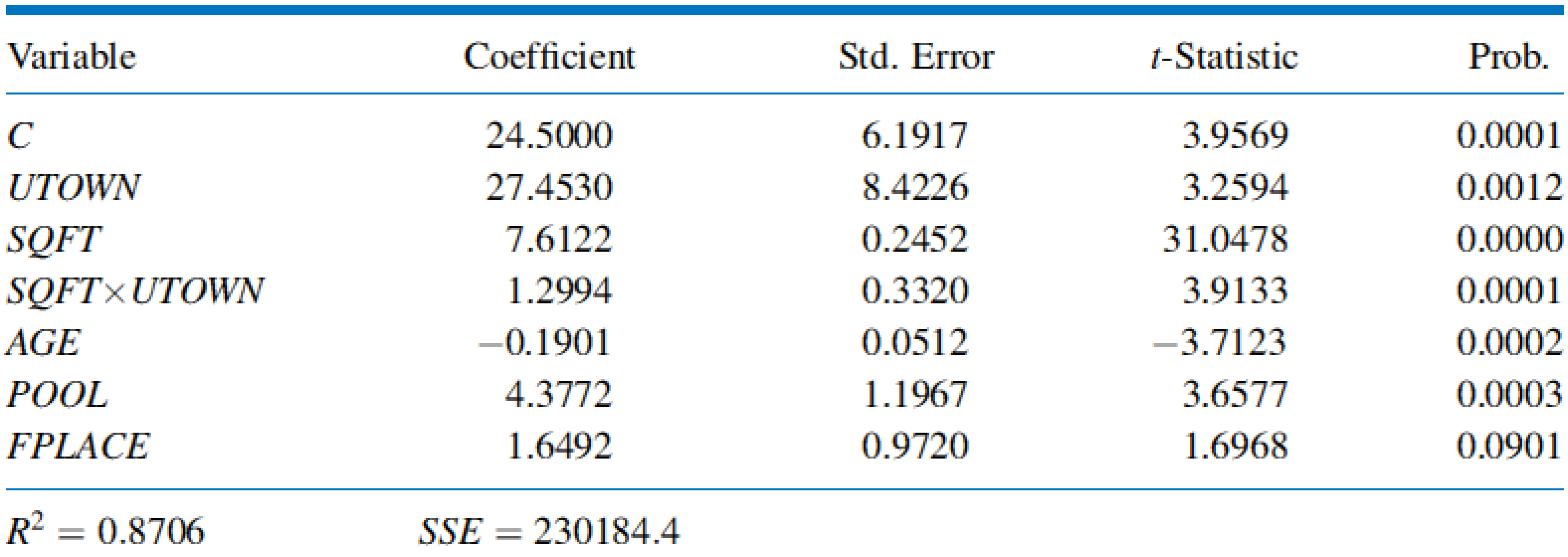
Eq. 7.7

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

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
$SQFT \times UTOWN$	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 estimated regression equation is for a house near the university is:

$$\hat{PRICE} = (24.5 + 27.453) + (7.6122 + 1.2994)SQFT +$$
$$-0.1901AGE + 4.3772POOL + 1.6492FPLACE$$
$$= 51.953 + 8.9116SQFT - 0.1901AGE$$
$$+4.3772POOL + 1.6492FPLACE$$

  - For a house in another area:

$$\hat{PRICE} = 24.5 + 7.6122SQFT - 0.1901AGE +$$
$$4.3772POOL + 1.6492FPLACE$$

- We therefore 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

7.2
Applying Indicator Variables

- We can apply indicator variables to a number of problems

Eq. 7.8

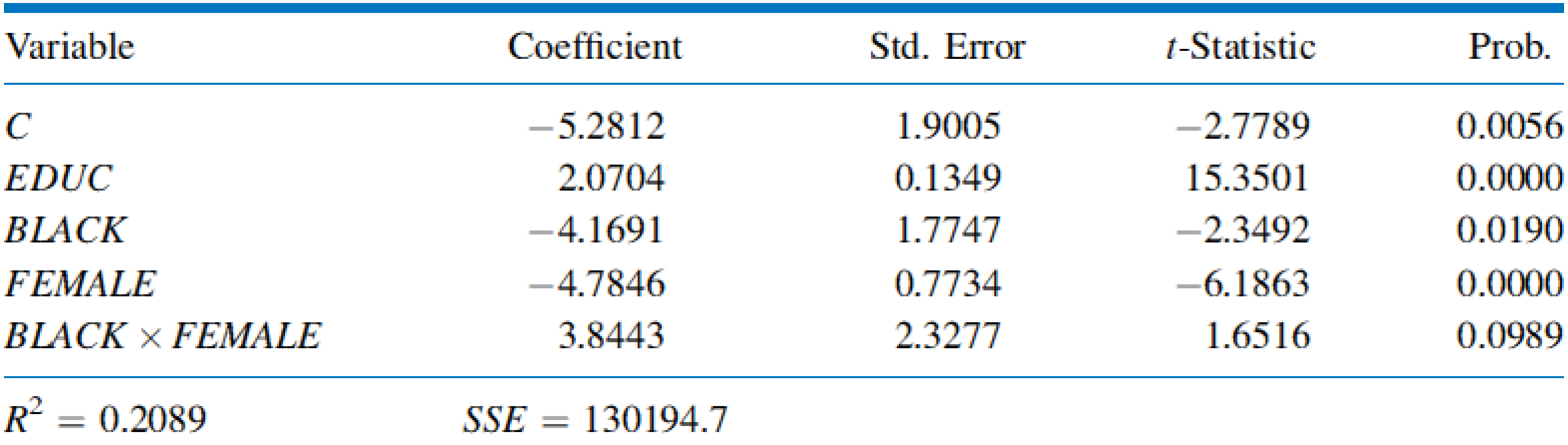
- Consider the wage equation:

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

  - The expected value is:

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

Table 7.3 Wage Equation with Race and Gender



Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	−5.2812	1.9005	−2.7789	0.0056
$EDUC$	2.0704	0.1349	15.3501	0.0000
$BLACK$	−4.1691	1.7747	−2.3492	0.0190
$FEMALE$	−4.7846	0.7734	−6.1863	0.0000
$BLACK \times FEMALE$	3.8443	2.3277	1.6516	0.0989

$R^2 = 0.2089$ $SSE = 130194.7$

- Recall that the test statistic for a joint hypothesis is:

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

- To test the $J = 3$ joint null hypotheses H_0: $\delta_1 = 0$, $\delta_2 = 0$, $\gamma = 0$, we use $SSE_U = 130194.7$ from Table 7.3
  - The SSE_R comes from fitting the model:

$$\widehat{WAGE} = -6.7103 + 1.9803EDUC$$
$$(se) \qquad (1.9142) \; (0.1361)$$

for which $SSE_R = 135771.1$

- Therefore:

$$F = \frac{\left(SSE_R - SSE_U\right)/J}{SSE_U/\left(N-K\right)} = \frac{\left(135771.1 - 130194.7\right)/3}{130194.7/995} = 14.21$$

  - The 1% critical value (i.e., the 99th percentile value) is $F_{(0.99,3,995)}$ = 3.80.
    - Thus, we conclude that race and/or gender affect the wage equation.

- Consider including regions in the wage equation:

Eq. 7.9

$$WAGE = \beta_1 + \beta_2 EDUC + \delta_1 SOUTH + \delta_2 MIDWEST + \delta_3 WEST + e$$

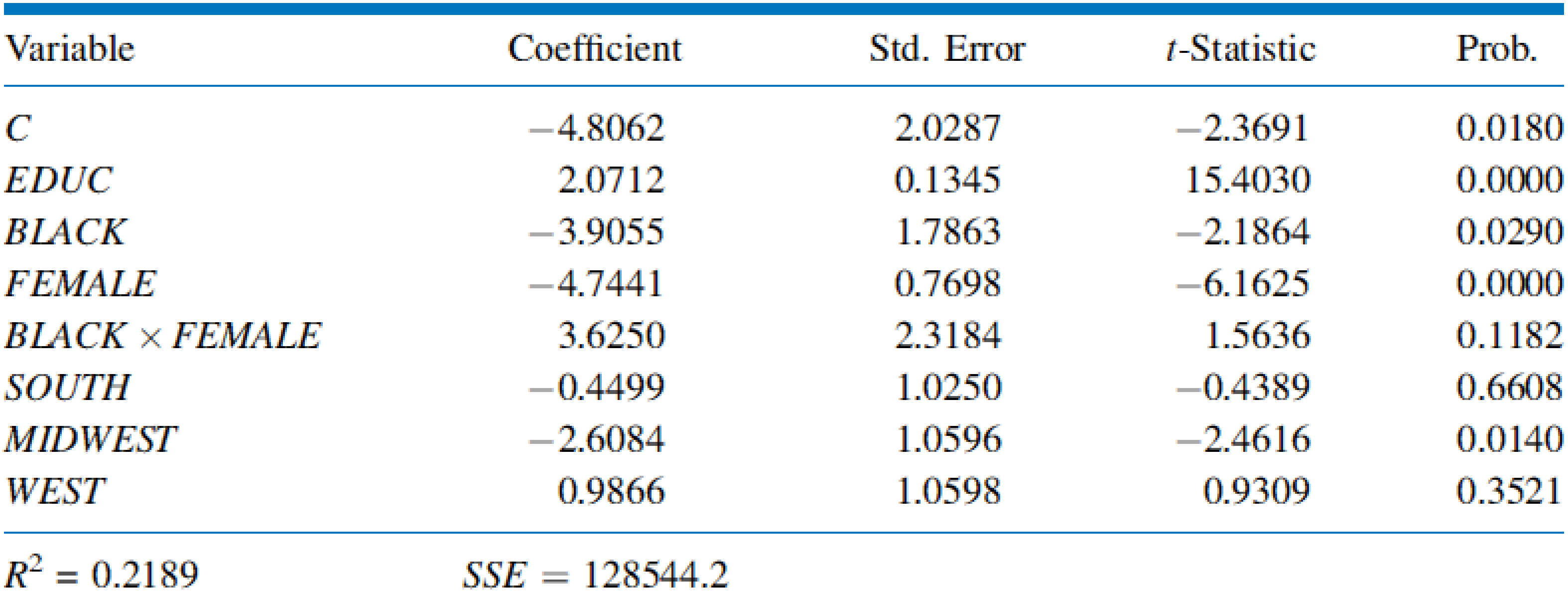
- – Since the regional categories are exhaustive, the sum of the regional indicator variables is $NORTHEAST + SOUTH + MIDWEST + WEST = 1$
- – Failure to omit one indicator variable will lead to the dummy variable trap

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

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

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

Table 7.4 Wage Equation with Regional Indicator Variables



Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	−4.8062	2.0287	−2.3691	0.0180
$EDUC$	2.0712	0.1345	15.4030	0.0000
$BLACK$	−3.9055	1.7863	−2.1864	0.0290
$FEMALE$	−4.7441	0.7698	−6.1625	0.0000
$BLACK \times FEMALE$	3.6250	2.3184	1.5636	0.1182
$SOUTH$	−0.4499	1.0250	−0.4389	0.6608
$MIDWEST$	−2.6084	1.0596	−2.4616	0.0140
$WEST$	0.9866	1.0598	0.9309	0.3521

$R^2 = 0.2189$ $SSE = 128544.2$

- Suppose we have:

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

and for two locations:

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

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

- 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 Eq. 7.6 by estimating separate regressions for each of the neighborhoods
  - The **Chow test** is an *F*-test for the equivalence of two regressions

- 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

- To test this, we specify:

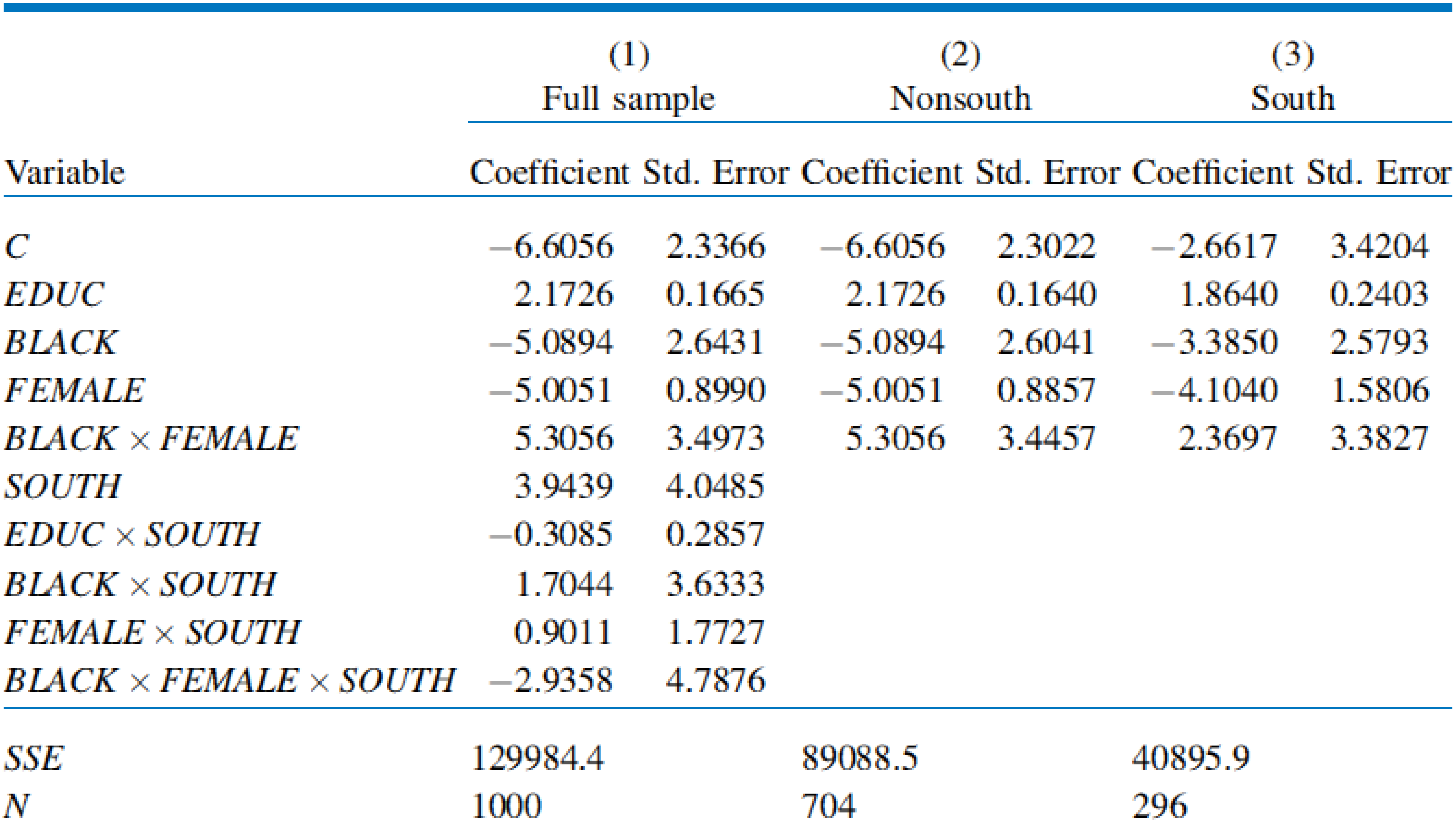
Eq. 7.10

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

- Now examine this version of Eq. 7.10:

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

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	−6.6056	2.3366	−6.6056	2.3022	−2.6617	3.4204
EDUC	2.1726	0.1665	2.1726	0.1640	1.8640	0.2403
BLACK	−5.0894	2.6431	−5.0894	2.6041	−3.3850	2.5793
FEMALE	−5.0051	0.8990	−5.0051	0.8857	−4.1040	1.5806
BLACK × *FEMALE*	5.3056	3.4973	5.3056	3.4457	2.3697	3.3827
SOUTH	3.9439	4.0485				
EDUC × *SOUTH*	−0.3085	0.2857				
BLACK × *SOUTH*	1.7044	3.6333				
FEMALE × *SOUTH*	0.9011	1.7727				
BLACK × *FEMALE* × *SOUTH*	−2.9358	4.7876				
SSE	129984.4		89088.5		40895.9	
N	1000		704		296	

- From the table, we note that:

$$\begin{aligned} SSE_{full} &= SSE_{nonsouth} + SSE_{south} \\ &= 89088.5 + 40895.9 \\ &= 129984.4 \end{aligned}$$

- We can test for a southern regional difference.
  - We estimate Eq. 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

- The *F*-statistic is:

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

$$= \frac{\left(130194.7 - 129984.4\right)/5}{129984.4/990}$$

$$= 0.3203$$

  – The 10% critical value is $F_c = 1.85$, and thus we fail to reject the hypothesis that the wage equation is the same in the southern region and the remainder of the country at the 10% level of significance

    • The *p*-value of this test is $p = 0.9009$

- 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, $\text{var}(e_\text{i}) = \sigma^2$, 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.

- 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
  - 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:

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

  – 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

- Consider the wage equation in log-linear form:

Eq. 7.11

$$\ln(WAGE) = \beta_1 + \beta_2 EDUC + \delta FEMALE$$

  - What is the interpretation of δ?

- Expanding our model, we have:

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

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

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

  – This is approximately the percentage difference

- The estimated model is:

$$\widehat{\ln}\left(WAGE\right)=1.6539+0.0962EDUC-0.2432FEMALE$$
$$\left(se\right)\qquad\left(0.0844\right)\ \left(0.0060\right)\qquad\qquad\left(0.0327\right)$$

  – We estimate that there is a 24.32% differential between male and female wages

- For a better calculation, the wage difference is:

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

  – But, by the property of logs:

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

- 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

- 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

- 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 $P[\,y = 0] = 1 - p$

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

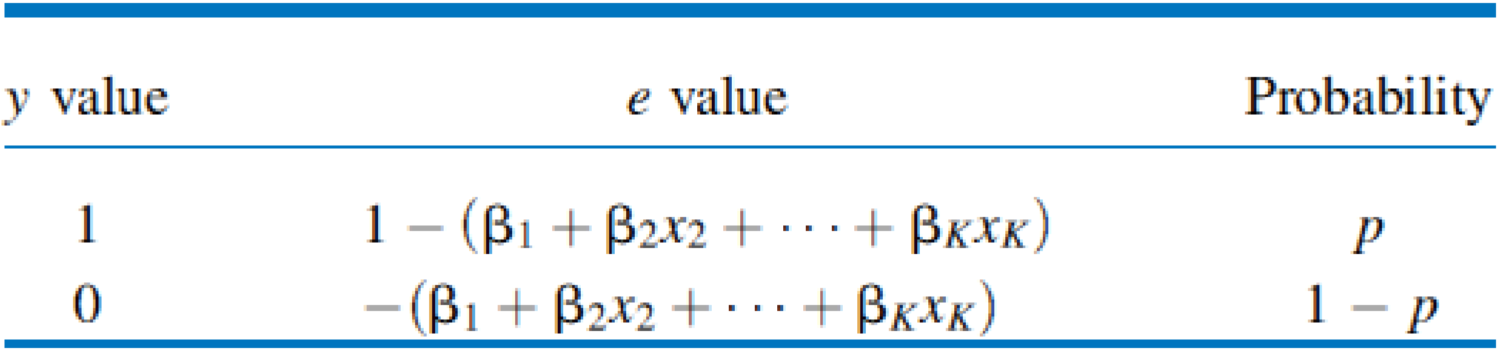
- A **linear probability model** is:

$$E\left(y\right)=p=\beta_1+\beta_2x_2+\mathrm{L}\ \ +\beta_{Ks}x_K$$

  – An econometric model is:

$$y=E\left(y\right)+e=\beta_1+\beta_2x_2+\mathrm{L}\ \ +\beta_{Ks}x_K+e$$

- 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:

$$\operatorname{var}(e) = p(1 - p)$$
$$= (\beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K)(1 - \beta_1 - \beta_2 x_2 - \cdots - \beta_K x_K)$$

  – The error term is not homoskedastic

- The predicted values, $\hat{E}(y) = \hat{p}$, can fall outside the (0, 1) interval
  - Any interpretation would not make sense

- 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:

$$\begin{array}{lllll} \widehat{E}\left(COKE\right) = \hat{p}_{COKE} = 0.8902 & - \; 0.4009PRATIO & + \; 0.0772DISP\_COKE & - \; 0.1657DISP\_PEPSI \\ \quad \left(se\right) \qquad\qquad \left(0.0655\right) & \left(0.0613\right) & \left(0.0344\right) & \left(0.0356\right) \end{array}$$

7.5
Treatment Effects

- 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
    - When membership in the treated group is in part determined by choice, then the sample is not a random sample

- 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

- 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

- Define the indicator variable d as:

Eq. 7.12

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

  - The model is then:

Eq. 7.13

$$y_i = \beta_1 + \beta_2 d_i + e_i, \quad i = 1, \mathrm{K}, N$$

  - And the regression functions are:

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

- The least squares estimator for β$_2$, the **treatment effect**, is:

Eq. 7.14

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

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

- 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:

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

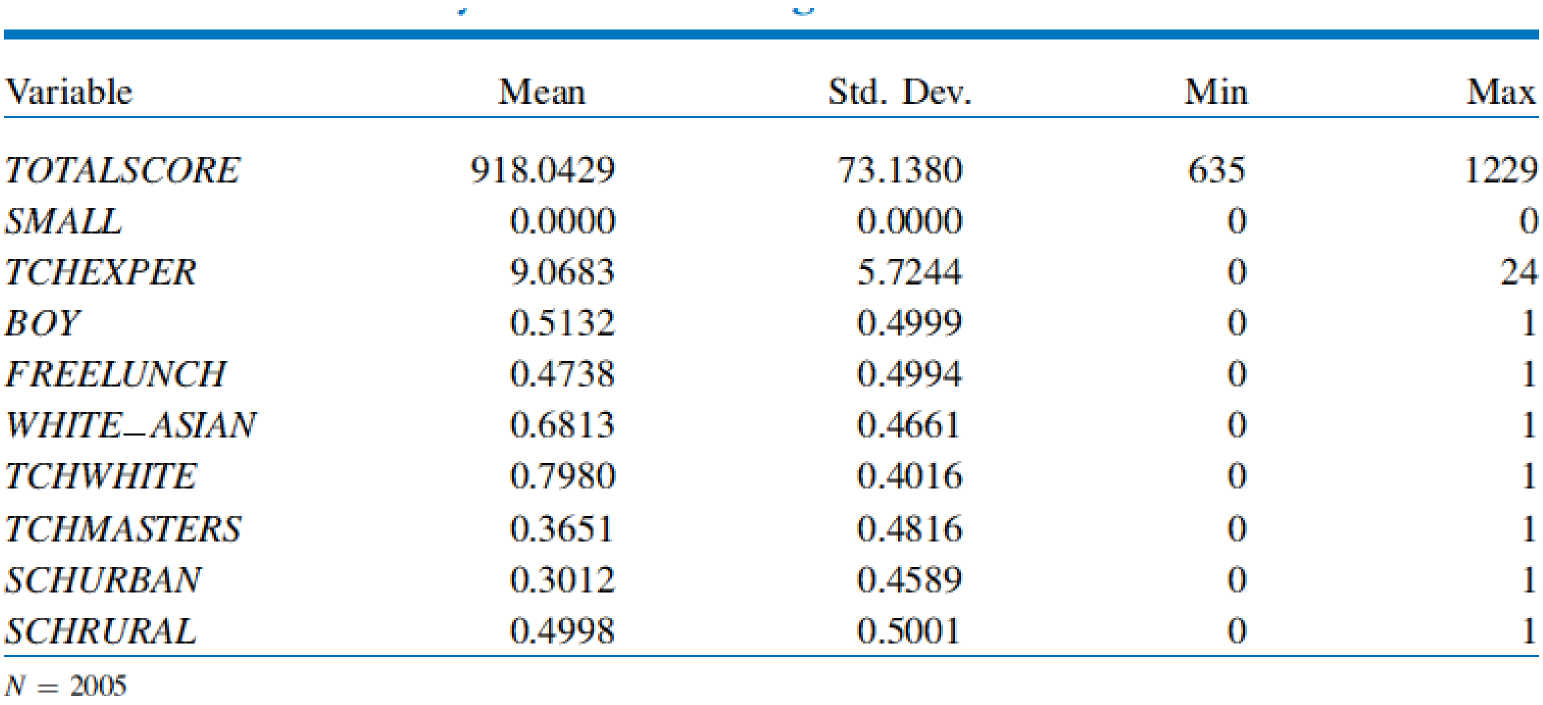
- If we allow individuals to "self-select" into treatment and control groups, then:

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

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

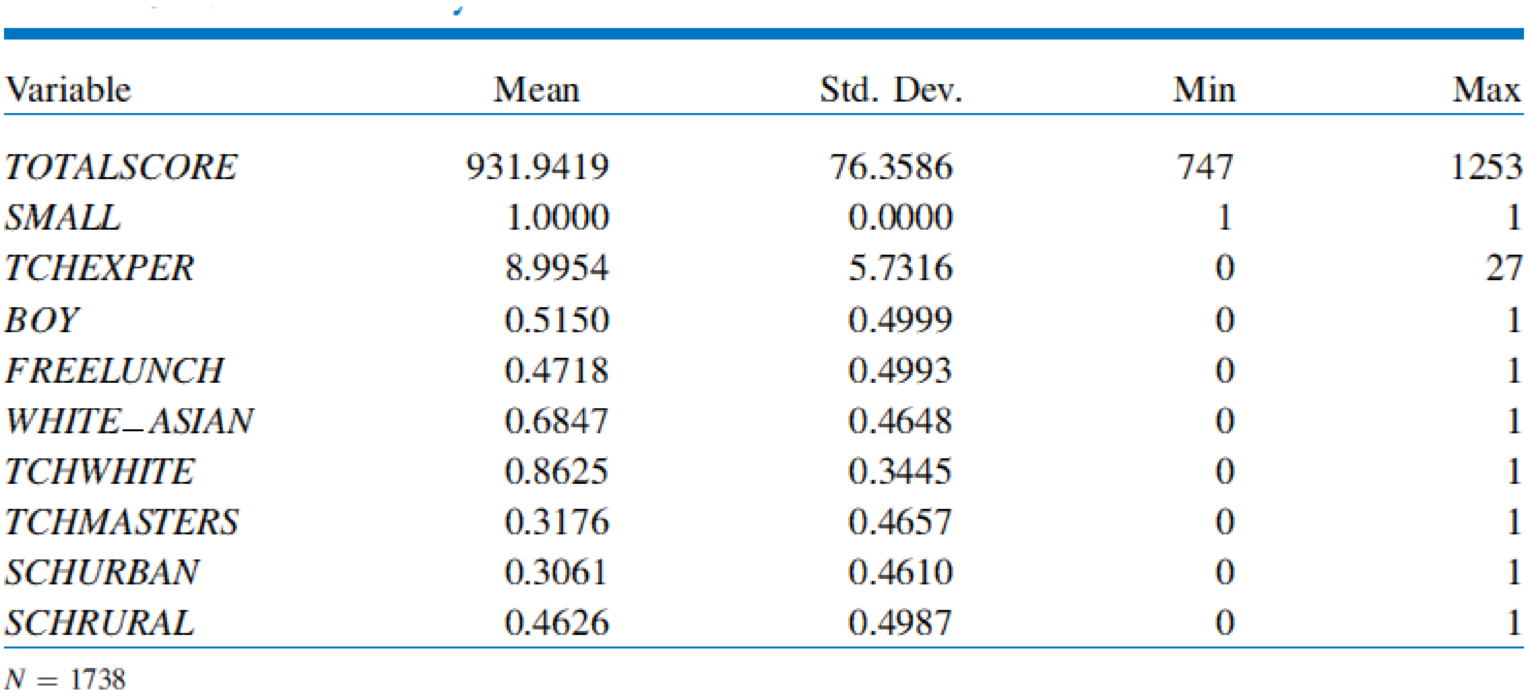
Table 7.6a Summary Statistics for Regular-Sized Classes



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

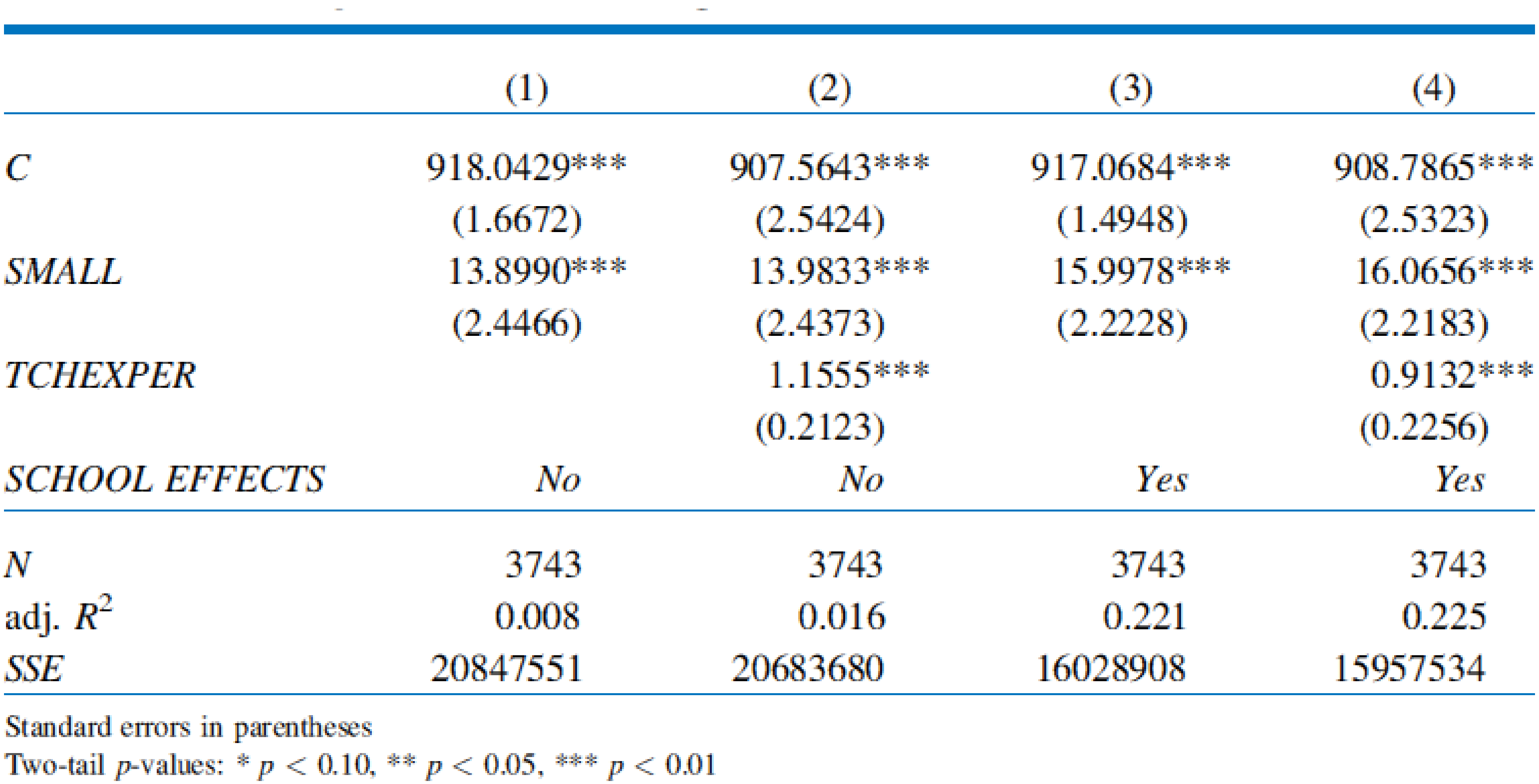
$N = 1738$

- The model of interest is:

Eq. 7.15

$$TOTALSCORE = \beta_1 + \beta_2 SMALL + e$$

Table 7.7 Project STAR: Kindergarden



	(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: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

- Adding *TCHEXPER* to the base model we obtain:

Eq. 7.16

$$TOTALSCORE = \beta_1 + \beta_2 SMALL + \beta_3 TCHEXPER + e$$

- 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:

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

- The model is now:

Eq. 7.17

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

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

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

7.5.4b
Linear
Probability
Model Check of
Random
Assignment

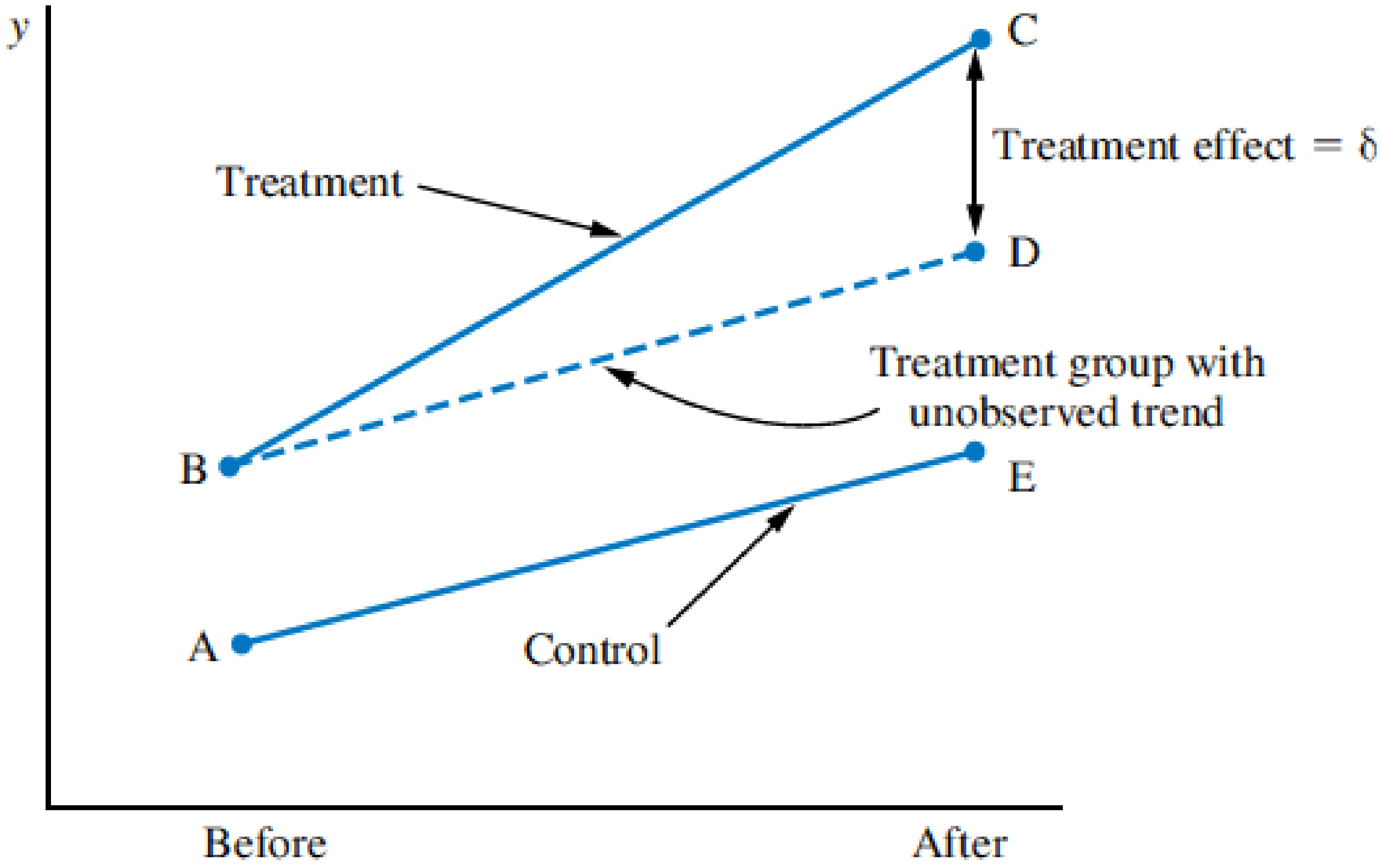
- Another way to check for random assignment is to regress *SMALL* on these characteristics and check for any significant coefficients, or an overall significant relationship
  - If there is random assignment, we should not find any significant relationships
  - Because *SMALL* is an indicator variable, we use the linear probability model

- The estimated linear probability model is:

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

- 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

FIGURE 7.3 Difference-in-Differences Estimation



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

Eq. 7.18

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

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

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

- The sample means are:

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

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

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

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

- Consider the regression model:

Eq. 7.19

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

- The regression function is:

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

- Using the points in the figure:

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

  – Using the least squares estimates, we have:

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

- We will test the null and alternative hypotheses:

Eq. 7.20

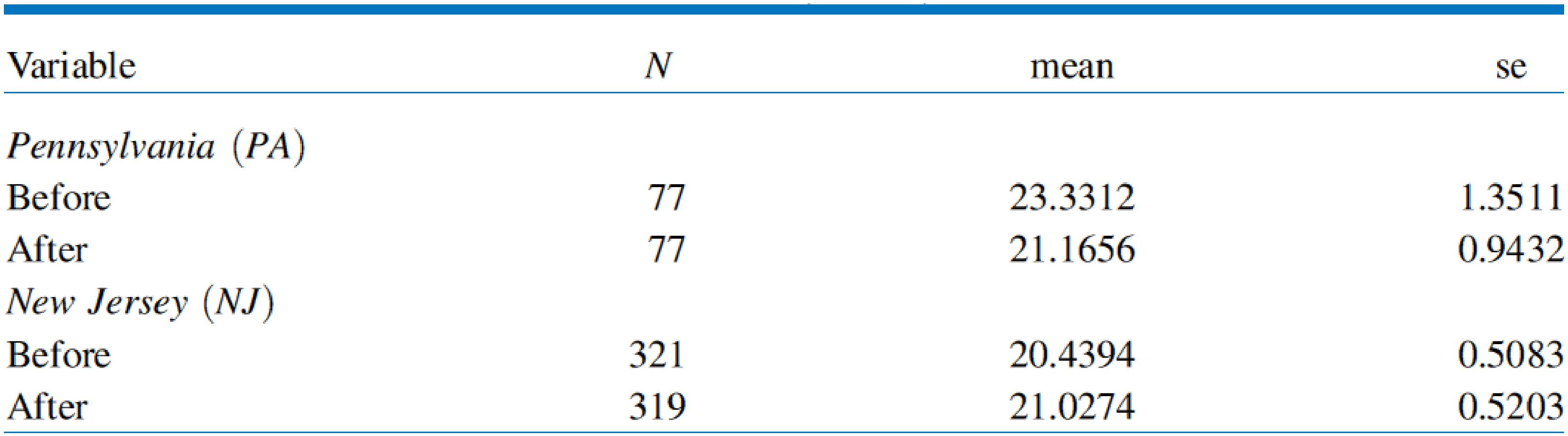
$$H_0 : \delta \geq 0 \text{ versus } H_1 : \delta < 0$$

– The differences-in-differences estimate of the change in employment due to the change in the minimum wage is:

Eq. 7.21

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

Table 7.8 Full-time Equivalent Employees by State and Period



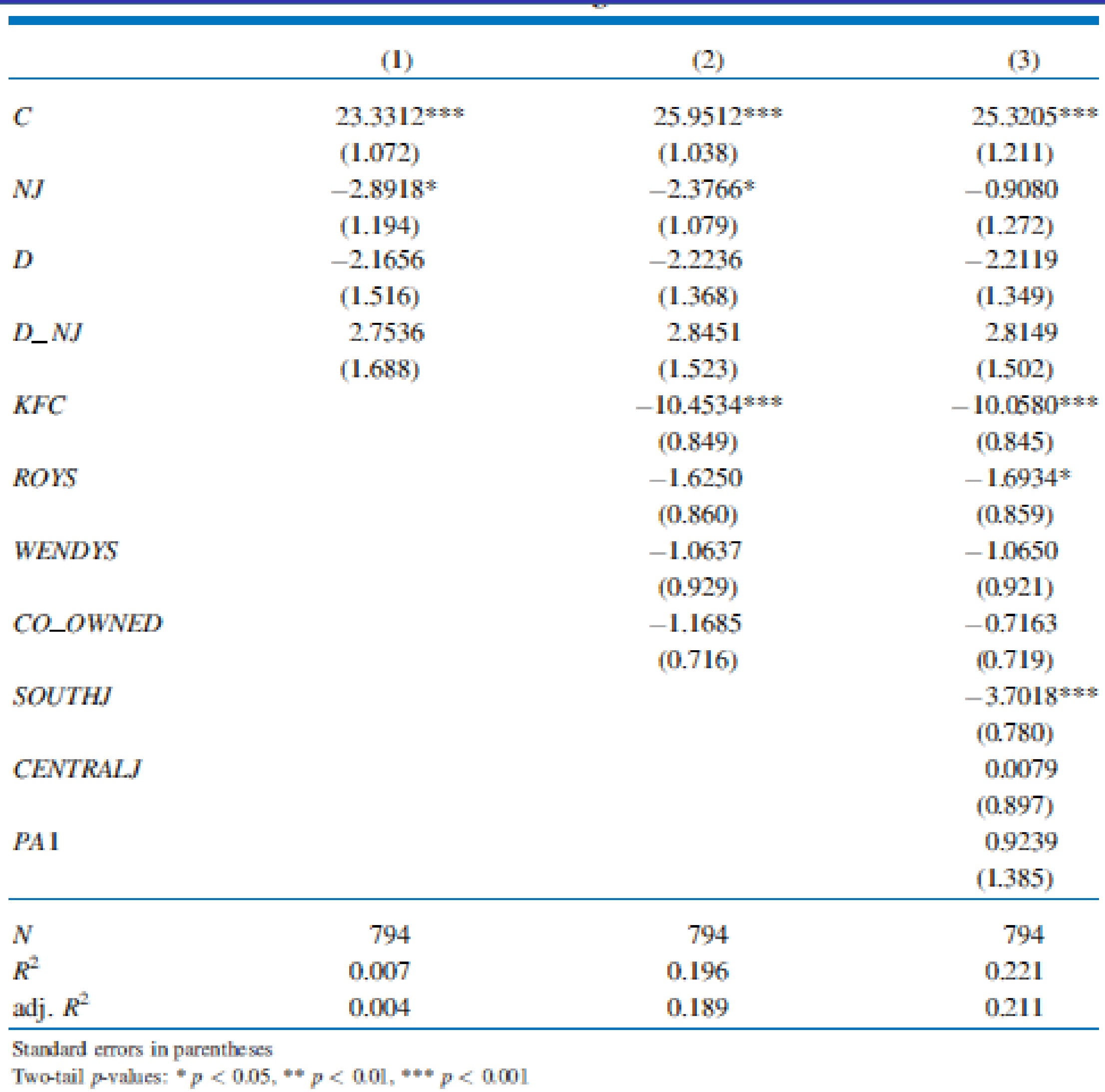
Variable	N	mean	se
Pennsylvania (PA)			
Before	77	23.3312	1.3511
After	77	21.1656	0.9432
New Jersey (NJ)			
Before	321	20.4394	0.5083
After	319	21.0274	0.5203

- Rather than compute the differences-in-differences estimate using sample means, it is easier and more general to use the regression format
  - The differences-in-differences regression is:

Eq. 7.22

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

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: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

- In our differences-in-differences analysis, we did not exploit one very important feature of the data - namely, that the same fast food restaurants were observed on two occasions
  - We have “before” and “after” data
  - These are called **paired data** observations, or **repeat data** observations, or **panel data** observations

- We previously introduced the notion of a **panel** of data – we observe the same individual-level units over several periods
  - Using panel data we can control for unobserved individual-specific characteristics

- Let c_i denote any unobserved characteristics of individual restaurant *i* that do not change over time:

Eq. 7.23

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

- Subtract the observation for t = 1 from that for t = 2:

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

where:

$$\Delta FTE_i = FTE_{i2} - FTE_{i1}$$

$$\Delta e_i = e_{i2} - e_{i1}$$

- Using the differenced data, the regression model of interest becomes:

Eq. 7.24

$$\Delta FTE_i = \beta_3 + \delta NJ_i + \Delta e_i$$

- The estimated model is:

$$\widehat{\Delta FTE} = -2.2833 + 2.7500 NJ \qquad R^2 = 0.0146$$
$$(se) \qquad (1.036) \quad (1.154)$$

  - 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
  - We fail to conclude that the minimum wage increase has reduced employment in these New Jersey fast food restaurants

Key Words

Keywords

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

Appendices

- 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:

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

- 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:

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

- We can compute the percentage difference as:

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

  - 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(δ) -1]%

- To verify Eq. 7.14, note that the numerator is:

$$\sum_{i=1}^{N}\left(d_i-\bar{d}\right)\left(y_i-\bar{y}\right)=\sum_{i=1}^{N}d_i\left(y_i-\bar{y}\right)-\bar{d}\sum_{i=1}^{N}\left(y_i-\bar{y}\right)$$

$$=\sum_{i=1}^{N}d_i\left(y_i-\bar{y}\right)\qquad\left[\text{using }\sum_{i=1}^{N}\left(y_i-\bar{y}\right)=0\right]$$

$$=\sum_{i=1}^{N}d_iy_i-\bar{y}\sum_{i=1}^{N}d_i$$

$$=N_1\bar{y}_1-N_1\bar{y}$$

$$=N_1\bar{y}_1-N_1\left(N_1\bar{y}_1+N_0\bar{y}_0\right)/N$$

$$=\frac{N_0N_1}{N}\left(\bar{y}_1-\bar{y}_0\right)\qquad\left[\text{using }\ N=N_1+N_0\right]$$

7B
Derivation of the Differences-in-Differences Estimator

- The denominator is:

$$\sum_{i=1}^{N}\left(d_i-\bar{d}\right)^2=\sum_{i=1}^{N}d_i^2-2\bar{d}\sum_{i=1}^{N}d_i+\sum_{i=1}^{N}\bar{d}^{\,2}$$

$$=\sum_{i=1}^{N}d_i-2\bar{d}N_1+N\bar{d}^{\,2}\quad\left[\text{using } d_i^2=d_i \text{ and } \sum_{i=1}^{N}d_i=N_1\right]$$

$$=N_1-2\frac{N_1}{N}N_1+N\left(\frac{N_1}{N}\right)^2$$

$$=\frac{N_0N_1}{N}\qquad\left[\text{using } N=N_0+N_1\right]$$

  - Combining the two expression gives us Eq. 7.14