Chapter 7
Using Indicator Variables

Walter R. Paczkowski
Rutgers University

Chapter Contents

- 7.1 Indicator Variables
- 7.2 Applying Indicator Variables

7.1
Indicator Variables

- 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 model to predict the value of a house as a function of its characteristics:
  - size
  - Location
  - number of bedrooms
  - age

- Consider the surface 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 **indicator 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:

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

- 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:

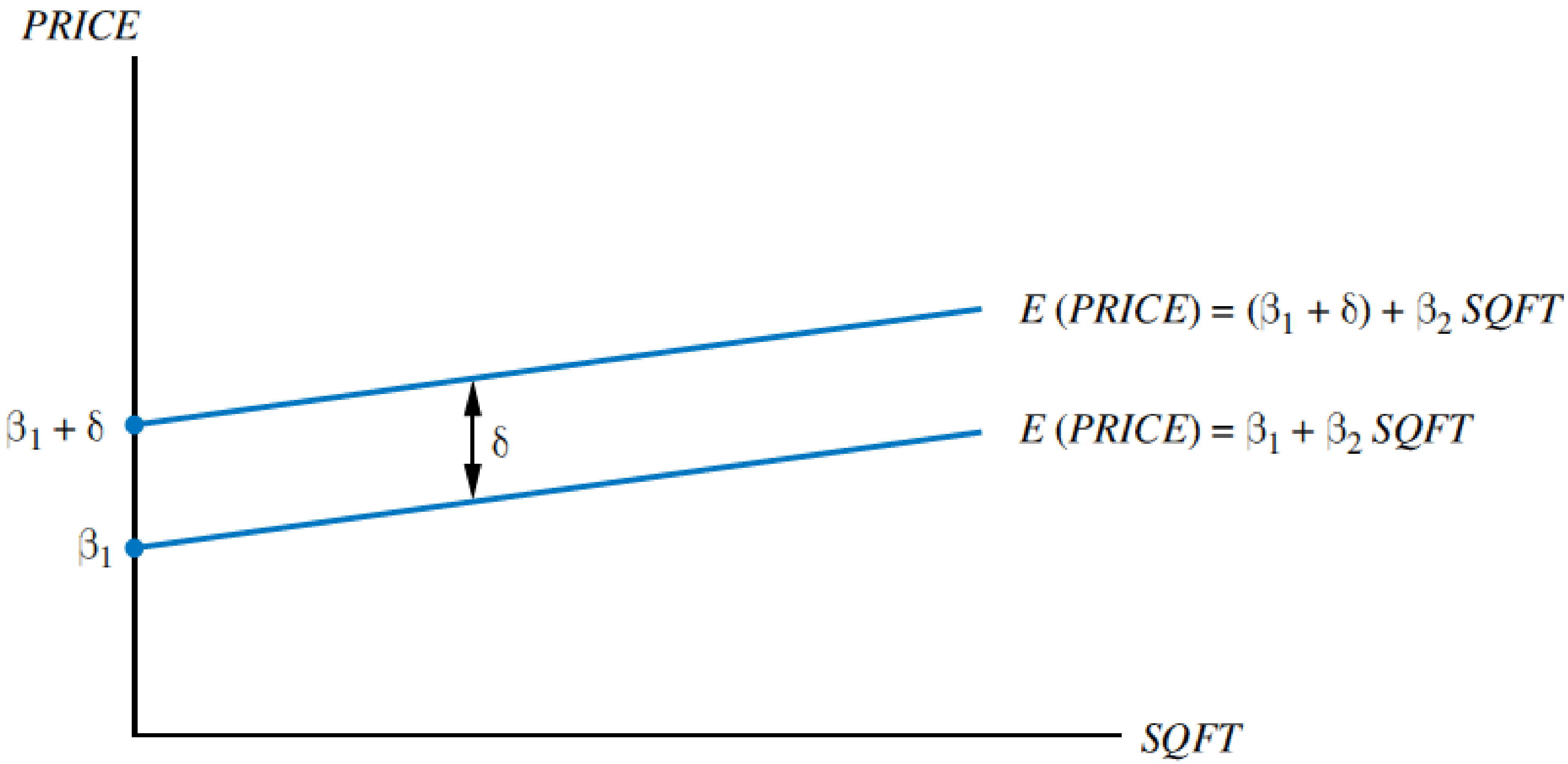
Eq. 7.4

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

- 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:

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

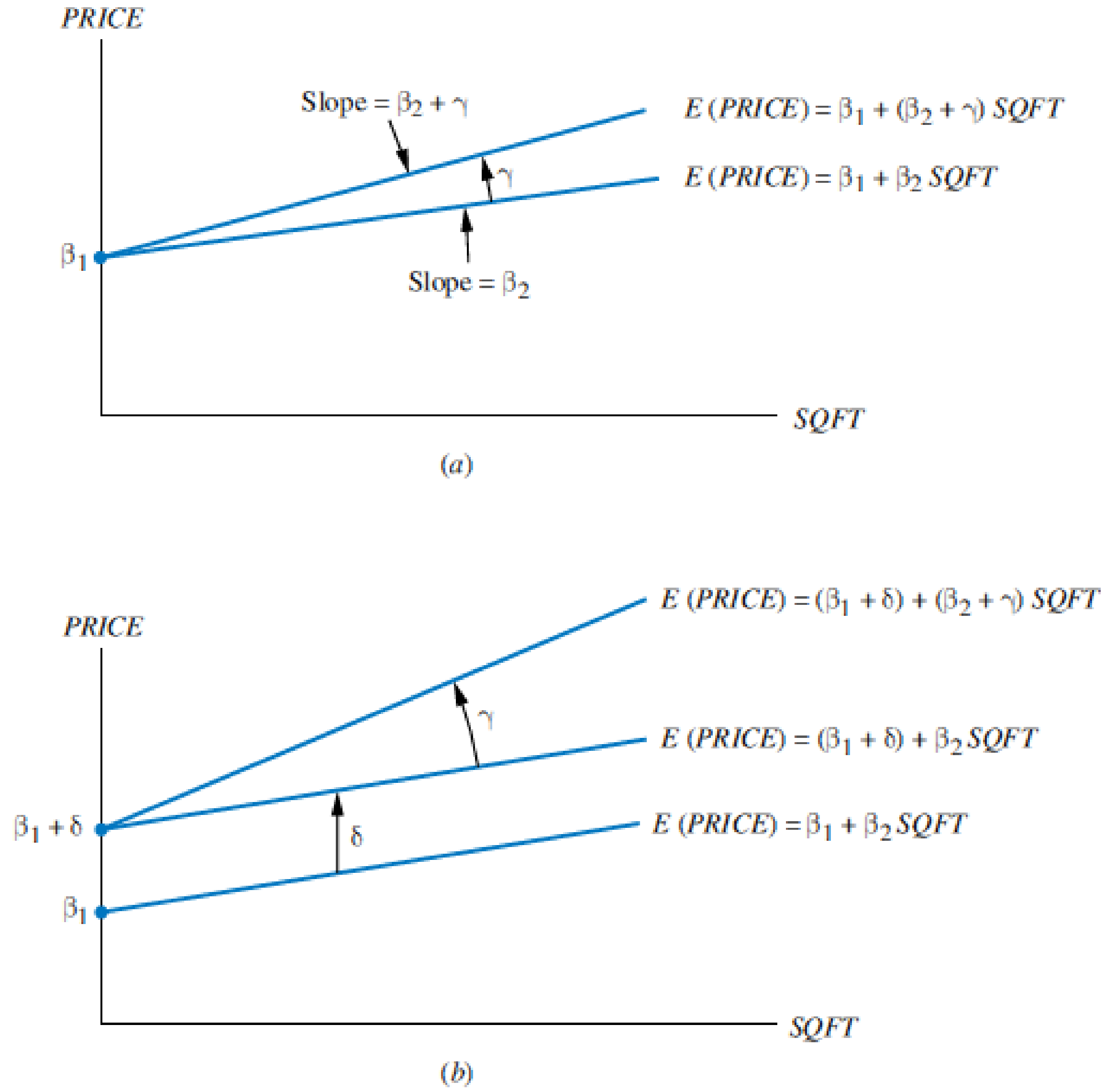
  – The new variable (*SQFT* x *D*) is the product of house size and the indicator variable
    - It is called an **interaction variable**, as it captures the interaction effect of location and size on house price
    - Alternatively, it is called a **slope-indicator variable** or a **slope dummy variable**, 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 ($SQFT \times D$) 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:

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

- 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×*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:

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

  – For a house in another area:

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

- 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:

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

  – The expected value is:

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

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:

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

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

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

- Now consider our wage equation:

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

  - "*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
    - **Chow test** is a statistical test (an *F*-test) allowing us to test the equivalence of the two regressions.

- To test this, we specify:

Eq. 7.10

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

- Now examine this version of Eq. 7.10:

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

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 $p = 0.9009$

- 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 $\delta > 0$

Key Words

Keywords

- annual indicator variables
- Chow test
- dichotomous variable
- dummy variable
- dummy variable trap
- exact collinearity
- hedonic model
- indicator variable
- interaction variable
- intercept indicator variable
- reference group
- regional indicator variable
- seasonal indicator variables
- slope-indicator variable