

Chapter 7

Specification: Choosing a Functional Form

Lecture 2

Masao Ogaki

Department of Economics,
Ohio State University

November 13, 2007

Using Dummy Variables

- ▶ numerical or quantitative explanatory variables
- ▶ *qualitative* explanatory variables.

Examples: education (college graduates, high school graduates, and non-high school graduates, etc.) sex, race, religion, and nationality are qualitative variables.

These qualitative variables are represented by *dummy variables*. A dummy variable takes on the value of 0 or 1, depending on a qualitative attribute such as gender.

Consider the following example, in which the only explanatory variable in the two-variable regression is a dummy variable:

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i \quad (1)$$

Y_i : the annual starting salary

$$D_i = \begin{cases} 1 & \text{if college graduate} \\ 0 & \text{otherwise (i.e., noncollege graduate)} \end{cases}$$

We obtain from Regression (1) the following for the mean starting salary of noncollege graduates:

$$E(Y_i|D_i = 0) = \beta_0 + \beta_1 \times 0 = \beta_0 \quad (2)$$

For the mean starting salary of college graduates:

$$E(Y_i|D_i = 1) = \beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1 \quad (3)$$

Imagine that the OLS results for (1) for 18 individuals are

$$\hat{Y}_i = 23.03 + 2.56 X_i$$

$t =$

$N = 18, R^2 = 0.753$

Exercise: Fill in the missing numbers. What is the point estimate for the mean salary of noncollege graduates and estimates? What is the point estimate for the mean salary of college graduates? Do you reject the hypothesis $H_0 : \beta_1 = 0$ with a two-tailed test? Which alternative hypothesis is appropriate for a one-tailed test? Do you reject the hypothesis with the appropriate one-tailed test?

Consider the following example:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \epsilon_i \quad (4)$$

Y_i =: the annual salary of a high school teacher

X_i =: years of teaching experience

$$D_i = \begin{cases} 1 & \text{if male} \\ 0 & \text{otherwise (i.e., female)} \end{cases}$$

The meaning of Regression (4) is as follows.
Mean salary of a female college teacher:

$$E(Y_i|X_i, D_i = 0) = \beta_0 + \beta_2 X_i$$

Mean salary of a male college teacher:

$$E(Y_i|X_i, D_i = 1) = (\beta_0 + \beta_1) + \beta_2 X_i$$

Imagine that the OLS results for the data are

$$\hat{Y}_i = 19.7 + 1.459 D_i + 3.458 X_{2i}$$

(0.678) (\quad)
 $t =$ 4.64
 $N = 17, R^2 =$ 0.768

Exercises: Fill in the missing numbers. Interpret the slope coefficient for X_i . What is the mean salary of female college teachers with 5 years of experience? What is the mean salary of male high school teachers with 1 year of experience? Test the hypothesis that the average salaries of the male teachers and female teachers with same years of experience are the same at the 5 % significance level.

It should be noted that we had two categories (college graduates and noncollege graduates in the first example; female and male teachers in the second example) in the two examples examined above. We only used one dummy variable to distinguish the two categories in each example. This is because our regression model contains an intercept term. Consider a model with two dummy variables

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 X_i + \epsilon_i \quad (5)$$

where Y and X are as defined before for (4) and where

$$D_{1i} = \begin{cases} 1 & \text{if male} \\ 0 & \text{otherwise} \end{cases}$$

$$D_{2i} = \begin{cases} 1 & \text{if female} \\ 0 & \text{otherwise} \end{cases}$$

Model (5) suffers from perfect multicollinearity. To see this, note that (5) can be written as

$$Y_i = \beta_0 \times 1 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 X_i + \epsilon_i$$

and that

$$D_{1i} + D_{2i} = 1,$$

because a teacher is either male or female. In general, the number of dummies should be one less than the number of categories of the variable.

Three Categories

Consider annual starting salary (Y), High School GPA (X), and education. For education, we have three categories: non-high school graduates, high school graduates, and college graduates.

Exercises: When we have three categories, how many dummy variables should we use?. Write down a regression equation.