Dummy Variables Regression Models

Multiple regression models can also include qualitative (or categorical) independent variables. Qualitative variables, unlike quantitative variables, cannot be measured on a numerical scale. Therefore, we need to code the values of the qualitative variable (called **levels**) as numbers before we can fit the model. These coded qualitative variables are called **dummy variables** since the numbers assigned to the various levels are arbitrarily selected.

The Nature of Dummy Variables

In regression analysis the dependent variable, is frequently influenced not only by ratio scale variables (e.g., income, output, prices, costs, height, temperature) but also by variables that are essentially qualitative, or nominal scale, in nature, such as sex, race, color, religion, nationality, geographical region, and party affiliation. For example, holding all other factors constant, female workers are found to earn less than their male counterparts or nonwhite workers are found to earn less than whites. This pattern may result from sex or racial discrimination, but whatever the reason, qualitative variables such as sex and race seem to influence the dependent variable and clearly should be included among the explanatory variables.

Since such variables usually indicate the presence or absence of a *"quality"* or an attribute, such as male or female, black or white, Catholic or non-Catholic, Democrat or Republican, they are essentially ***nominal scale*** variables. One way we could *"quantify"* such attributes is by constructing artificial variables that take on values of **1** or **0**, 1 indicating the *presence (or possession)* of that attribute and 0 indicating the *absence* of that attribute.

Example:

Consider a salary discrimination case where there exists a claim of gender discrimination, specifically, the claim that male executives at a large company receive higher average salaries than female executives with the same credentials and qualifications.

To test this claim, we might propose a multiple regression model for executive salaries using the gender of an executive as one of the independent variables. The dummy variable used to describe gender may be coded as follows:

$$x = \begin{cases} 1 & \text{if male} \\ 0 & \text{if female} \end{cases}$$

Variables that assume such 0 and 1 values are called dummy variables. Such variables are thus essentially a device to classify data into mutually exclusive categories such as male or female.

The advantage of using a **0 −1 coding** scheme is that the β coefficients associated with the dummy variables are easily interpreted. For example, consider the following model for executive salary Y:

$$E(Y) = \beta_0 + \beta_1 X_{1i}$$

Where:

$$X_{1i} = \begin{cases} 1 & \text{if male} \\ 0 & \text{if female} \end{cases}$$

Note that β_0 represents the mean salary for females (say, μ_f). When using a 0–1 coding convention, β_0 will always represent the mean response associated with the level of the qualitative variable assigned the value 0 (called the **base level**). The difference between the mean salary for males and the mean salary for females, $\mu_m - \mu_f$, is represented by β_1.

Therefore, with the 0–1 coding convention, β_1 will always represent the difference between the mean response for the level assigned the value 1 and the mean for the base level. Thus, for the executive salary model we have

$$\beta_0 = \mu_f \text{ (mean for based level)}$$

$$\beta_1 = \mu_m - \mu_f$$

If β_1 exceeds 0, then $\mu_m > \mu_f$ and evidence of sex discrimination at the company exists.

If a qualitative variable has k categories, introduce only **(k − 1)** dummy variables. In general, the number of dummy variables used to describe a qualitative variable will be one less than the number of levels of the qualitative variable.

A Model Relating $E(Y)$ to a Qualitative Independent Variable with Three Levels

$$E(Y) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

$$X_{1i} = \begin{cases} 1 & \text{if level A} \\ 0 & \text{if not} \end{cases} \qquad X_{2i} = \begin{cases} 1 & \text{if level B} \\ 0 & \text{if not} \end{cases} \qquad \text{Base level = Level C}$$

Where:

Interpretation of $\beta's$:

$$\beta_0 = \mu_C \text{ (mean for based level)}$$
$$\beta_1 = \mu_A - \mu_C$$
$$\beta_2 = \mu_B - \mu_C$$

Dummy variables can be incorporated in regression models just as easily as quantitative variables. As a matter of fact, a regression model may contain regressors that are all exclusively dummy, or qualitative, in nature. Such models are called **Analysis of Variance (ANOVA) models.**

Example:

Using the data regarding public school teachers' salaries by geographical region.

The data in the Excel file shows the average salary (in peso) of public school teachers in the Philippines for the year 1995. There are 51 barangay and they classified into three geographical areas: (1) North area (21 barangay in all), (2) South (17 barangay in all), and (3) West (13 barangay in all).

Consider the following model:

$$Y_i = \beta_0 + \beta_1 K_{2i} + \beta_2 K_{3i} + \varepsilon_i \qquad \textbf{(5.1)}$$

Where

Y_i = salary of public school teacher in barangay

K_{2i} = 1 if the barangay is in the North and 0 otherwise (i.e., in other areas of the country)

K_{3i} = 1 if the barangay is in the South and 0 otherwise (i.e., in other areas of the country)

Note that **(5.1)** is like any multiple regression model considered previously, except that, instead of quantitative regressors, we have only qualitative, or dummy, regressors, taking the value of 1 if the observation belongs to a particular category and 0 if it does not belong to that category or group.

Question???

What does the model **(5.1)** tell us? Assuming that the error term satisfies the usual OLS assumptions, on taking expectation of (5.1) on both sides.

The estimated model based on the data:

$$\hat{Y}_i = 26{,}158.6154 - 1{,}734.4725K_{2i} - 3{,}264.6154K_{3i} \qquad \textbf{(5.2)}$$

As these regression results show, the mean salary of teachers in the West is about Php 26,158, that of teachers in the North is lower by about Php 1,734, and that of teachers in the South is lower by about Php 3,265.

Question???

What is the actual mean salaries of two areas (North and South)?

Suppose we want to find out if the average annual salary (AAS) of public school teachers differs among the three geographical areas of the Philippines. If you take the simple arithmetic average of the average salaries of the teachers in the three areas, you will find that these averages for the three areas are as follows: Php 24,424.14 (North), Php 22,894 (South), and Php 26,158.62 (West). These numbers look different, but **are they statistically different from one another?** There are various statistical techniques to compare two or more mean values, which generally go by the name of **analysis of variance.** But the same objective can be accomplished within the framework of regression analysis.

Caution in the Use of Dummy Variables

1. If a qualitative variable has k categories, introduce only $(k-1)$ dummy variables because you will encounter **dummy variable trap**, the situation of perfect collinearity or perfect multicollinearity, if there is more than one exact relationship among the variables. This dummy variable trap happen if we have more than one qualitative variable in the model.
2. The category for which no dummy variable is assigned is known as the **base level, benchmark, control, comparison, reference,** or **omitted category.** And all comparisons are made in relation to the base level.
3. The intercept value (β_0) represents the ***mean value*** of the base level.
4. The coefficients attached to the dummy variables are known as the **differential intercept coefficients** because they tell by how much the value of the intercept that receives the value of 1 differs from the intercept coefficient of the base level.
5. If a qualitative variable has more than one category, the choice of the benchmark category is strictly up to the researcher. Sometimes the choice of the benchmark is dictated by the particular problem at hand.
6. There is a way to avoid the dummy variable trap by introducing as many dummy variables as the number of categories of that variable, ***provided we do not introduce the intercept in such a model.*** Thus, if we drop the intercept term from the model, and consider the following model,

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots \beta_{ni} X_{ni} + \varepsilon_i$$

we do not fall into the dummy variable trap, as there is no longer perfect collinearity. But make sure that when you run this regression, you ***use the non-intercept option in your regression package***.

In other words, with the intercept suppressed, and allowing a dummy variable for each category, we obtain directly the mean values of the various categories.

Question???

Which is a better method of introducing a dummy variable:

(1) introduce a dummy for each category and omit the intercept term or

(2) include the intercept term and introduce only (k − 1) dummies, where k is the number of categories of the dummy variable?

Exercise:

Modeling the shipment cost, Y, of a regional express delivery service. Suppose we want to model $E(Y)$) as a function of cargo type, where cargo type has three levels (fragile, semifragile, and durable). Costs for 15 packages of approximately the same weight and same distance shipped, but of different cargo types, are listed in table below.

Package	Cost (Y)	Cargo Type	Create a Dummy Variables	
			X_1	X_2
1	1,720	1		
2	1,110	1		
3	1,200	1		
4	1,090	1		
5	1,380	1		
6	650	2		
7	1,000	2		
8	1,150	2		
9	700	2		
10	850	2		
11	210	3		
12	130	3		
13	340	3		
14	750	3		
15	200	3		

Legend: Cargo Type: 1 - fragile, 2 - Semifragile, 3 - Durable

A. Write the estimated model relating $E(Y)$ to cargo type.

B. Interpret the estimated β coefficients in the model.

C. Conduct the F-Test for overall model utility using $\alpha = 0.05$. Determine whether there is sufficient evidence of a difference between any two of the three mean shipment costs, that is, whether cargo type is a reliable predictor of shipment cost (Y).

Prepared by:

KATRINA D. ELIZON
Department of Mathematics and Statistics
College of Science