Economics Honours (Sixth Semester)

Basic Econometrics

Dummy Variable

Box 1: Measurement scale of variables

The variables that we will generally encounter fall into four broad categories: ratio scale, interval scale, ordinal scale, and nominal scale. It is important that we understand each.

Ratio Scale: For a variable X, taking two values, X_1 and X_2, the ratio X_1/X_2 and the distance $(X_2 - X_1)$ are meaningful quantities. Also, there is a natural ordering (ascending or descending) of the values along the scale. Therefore, comparisons such as $X_2 \leq X_1$ or $X_2 \geq X_1$ are meaningful. Most economic variables belong to this category. Thus, it is meaningful to ask how big this year's GDP is compared with the previous year's GDP.

Interval Scale: An interval scale variable satisfies the last two properties of the ratio scale variable but not the first. Thus, the distance between two time periods, say (2000–1995) is meaningful, but not the ratio of two time periods (2000/1995).

Ordinal Scale: A variable belongs to this category only if it satisfies the third property of the ratio scale (i.e., natural ordering). Examples are grading systems (A, B, C grades) or income class (upper, middle, lower). For these variables the ordering exists but the distances between the categories cannot be quantified. Students of economics will recall the indifference curves between two goods, each higher indifference curve indicating higher level of utility, but one cannot quantify by how much one indifference curve is higher than the others.

Nominal Scale: Variables in this category have none of the features of the ratio scale variables. Variables such as gender (male, female) and marital status (married, unmarried, divorced, separated) simply denote categories.

Source: Gujarati, D. N. Porter, D.C., Gunasekar, S. (2009), *Basic econometrics*. (Fifth ed.) McGraw-Hill Education (India).

Meaning of Dummy variables

Simple regression analysis and multiple regression analysis deal with ratio scale variables. However it must be noted that regression analysis can also be carried out with nominal scale variables. Dummy variables are nominal scale variables which are qualitative in nature like race, gender, geographical region, religion etc. Like in a survey on corporate managers we may come consider the case of male and female managers in our regression analysis. Dummy variables may indicate the

presence or absence of an attribute or classify data into mutually exclusive categories. Usually the range of dummy variables is very limited and they can take on only two quantitative values. **Dummy variables** are also called **qualitative or indicator or categorical variables**. Suppose we consider the case of corporate managers where a manager may be either male or female, then the value 0 may be used to indicate male and to indicate female we use the value 1. Like quantitative variables, dummy variables can be used in regression models. If a regression model is constructed only on the basis of dummy variables, then such models are called *Analysis of Variance (ANOVA) Model*.

Formulating Dummy Variables

We take the help of a simple example from (https://stattrek.com/multiple-regression/dummy-variables.aspx) to understand the formulation of dummy variables. In the example given below, the Table 1, shows the Test score and IQ of ten students who may be either male or female student.

Table 1: Data on Test score, IQ and Gender of ten students

Student	Test score	IQ	Gender
1	93	125	Male
2	86	120	Female
3	96	115	Male
4	81	110	Female
5	92	105	Male
6	75	100	Female
7	84	95	Male
8	77	90	Female
9	73	85	Male
10	74	80	Female

Source: https://stattrek.com/multiple-regression/dummy-variables.aspx

Here we are technically dealing with three variables on ten students, the variables IQ and Gender are used to predict the Test score of the students. Thus Test score is the dependent variable, and IQ and Gender are the two explanatory variables. Of these two explanatory variables, it is evident that IQ is a quantitative variable and Gender is a qualitative or categorical variable. Let us express the categorical variable in its dummy form. This means we indicate a male student by the number 1 or

otherwise, that is, indicate a female student as 0; then Table 2 can be reconstructed with Gender expressed as a quantitative dummy variable.

Table 2 : Reconstructing Table 1 with Gender as Dummy variable (D)

Student	Test score	IQ	D
1	93	125	1
2	86	120	0
3	96	115	1
4	81	110	0
5	92	105	1
6	75	100	0
7	84	95	1
8	77	90	0
9	73	85	1
10	74	80	0

Source: https://stattrek.com/multiple-regression/dummy-variables.aspx

The regression model can be expressed as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i, \text{ where}$$

Y_i = Test Score,

X_i = IQ,

D_i = Dummy variable on gender indicating 1 for male students and 0 for otherwise i.e. female students

β_0 = intercept,

β_1 and β_2 = slope coefficients

u_i = residual term

The Number of Dummy Variables

In our example while indicating ‘Gender’, it can assume two values i.e. male or female; however we require only one dummy variable to carry out our regression analysis. It must be noted here that in a similar manner if there are k number of categories or k different values then the total number of dummy variables required for regression analysis with dummy variables is (k-1).

Box 2: Uses in Regression Models

- Dummy variables can be used to capture changes or shifts in the intercept as evident in the regression model

$$Y = \beta_0 + \beta_1 X + \beta_2 D + u,$$

where D is 1 for one category and 0 otherwise and X is the usual quantitative explanatory variable.

- Dummy variables can be used to capture changes in the slope like in the regression model

$$Y = \beta_0 + \beta_1 X + \beta_2 XD + u,$$

where D is 1 for one category and 0 otherwise and X is the usual quantitative explanatory variable.

- Dummy variables can be used to capture changes in both intercept and slope like in the regression model

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 XD + u,$$

where D is 1 for one category and 0 otherwise and X is the usual quantitative explanatory variable.

- Dummy variables can also be used to capture differences among more than one category, such as seasons or regions:

$$Y = \beta_0 + \beta_1 X + \beta_2 D_1 + \beta_3 D_2 + \beta_4 D_3 + u,$$

where say β_0 = intercept for the first region and D_1, D_2 , D_3 refer respectively to seasons or regions 2, 3, 4.

Source: Salvatore, D., & Reagle, D. (2011). *Schaum's outline of statistics and econometrics* (2nd ed.). McGraw-Hill Education.

Box 3: Steps in Estimation

To complete a good multiple regression analysis, we want to do four things:

- Estimate regression coefficients for our regression equation.
- Assess how well the regression equation predicts test score, the dependent variable.
- Assess the extent of multicollinearity between independent variables.
- Assess the contribution of each independent variable to the prediction.

Source: https://stattrek.com/multiple-regression/dummy-variables.aspx

Box 4: Procedure of Estimation

The procedure for estimating a regression function with dummy variables is explained with the help of an example. Table 3 gives the quantity of milk (in thousands of quarts) supplied by a firm per month Q at various prices P over a 14–month period. The firm faced a strike in some of its plants during the fifth, sixth and seventh months. Run a regression of Q on P, testing only for a shift in the intercept during periods of strike and non-strike

Table3: Quantity of milk (in thousands of quarts) supplied at various prices

Month	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Q	98	100	103	105	80	87	94	113	116	118	121	123	126	128
P	0.79	0.80	0.82	0.82	0.93	0.95	0.96	0.88	0.88	0.90	0.93	0.94	0.96	0.97
D	0	0	0	0	1	1	1	0	0	0	0	0	0	0

Let the regression function be of the form $Q = \beta_0 + \beta_1 P + \beta_2 D + u$. Letting the dummy variable D = 1 during the months of strike and D = 0 otherwise, we have

Q = Quantity of milk (in thousands of quarts) supplied

P = Price

u= residual term

β_1= slope coefficient determining the influence of P on Q

β_0 = shift in intercept during period of no strike

β_2 = slope coefficient determining the influence of dummy variable D on Q, where D = 1 during the months of strike and D = 0 otherwise.

Running multiple regression analysis we get

$$\hat{Q} = -32.47 + 165.97\,P - 37.64D$$

$$(15.65) \qquad (-23.59)$$

$$R^2 = 0.98$$

Since D is statistically significant at better than the 1% level, the intercept is $\beta_0 = -32.47$ during the period of no strike and it equals to $\beta_0 + \beta_2 = -32.47-37.64 = -70.11$ during the strike period.

Taken from: Salvatore, D., & Reagle, D. (2011). *Schaum's outline of statistics and econometrics* (2nd ed.). McGraw-Hill Education.

Box 5: Dummy Variable Trap

When defining dummy variables, a common mistake is to define too many variables. If a categorical variable can take on *k* values, it is tempting to define *k* dummy variables. Resist this urge. Remember, you only need *(k – 1)* dummy variables.

A k^{th} dummy variable is redundant; it carries no new information. And it creates a severe multicollinearity problem for the analysis. Using *k* dummy variables when only *k - 1* dummy variables are required is known as the **dummy variable trap**.

Source: https://stattrek.com/multiple-regression/dummy-variables.aspx

*N.B. Explanations of all the theories have been taken from the following references:

References:

- *Dummy variables*. (n.d.). Statistics and Probability. Retrieved May 10, 2020, from https://stattrek.com/multiple-regression/dummy-variables.aspx
- Gujarati, D. N. Porter, D.C., Gunasekar, S. (2009), *Basic econometrics*. (Fifth ed.) McGraw-Hill Education (India).
- Salvatore, D., & Reagle, D. (2011). *Schaum's outline of statistics and econometrics* (2nd ed.). McGraw-Hill Education.