Subject: Quantitative techniques for Economics
Course code: ECON6002
Topic: Dummy variable regression models
Ph.D. Economics (1st. Semester)

Dr. Kailash Chandra Pradhan

Mahatma Gandhi Central University,
Department of Economics

Dummy variable models

- In the regression analysis variable is frequently influenced not only by ratio scale variable (For example, the variables like income, output, prices, costs, temperature, hight etc.) but also by qualitative or nominal scale such as sex, race, color, religion etc.
- The qualitative variables do not have any natural scale of measurement. Such variables usually indicate the presence or absence of a "quality" or an attribute like male or female, employed or unemployed, graduate or non-graduate, yes or no, acceptance or rejection.
- One way to quantify such attributes is by constructing artificial variables that take on value of 1 or 0, 1 indicating the presence (or possession) of that attribute and 0 indicating the absence of that attribute.
- For example, 1 may indicate that a person is a female and 0 may indicate a male.

Dummy variable models

- Such variables classify the data into mutually exclusive categories such as male or female.
- In the regression model, the qualitative and quantitative variables may also occur together, i.e., some variables may be qualitative and others are quantitative.
- If all independent or explanatory variables are quantitative, then the model is called a regression model.
- If all independent or explanatory variables are qualitative, then the model is called an analysis of variance model (ANOVA).
- If all independent or explanatory variables are admixture of quantitative and qualitative, then the model is called a analysis of covariance model (ANCOVA).

ANOVA Models

- **Example 1:** On average salary (in dollars) of public school teachers in 50 states and the District of Columbia for the year 1985. These 51 areas are classified into three geographical regions: (1) Northeast and North Central (21 states in all), (2) South (17 states in all), and (3) West (13 states in all). Suppose we want to find out if the average annual salary (AAS) of public school teachers differs among the three geographical regions of the country.
- The ANOVA model can be as written as:

$$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_{3i} D_{3i} + u_i$$

where Y_i = *(average) salary of public school teacher in state i*

D_{2i} = *1 if the state is in the Northeast or North Central*

= *0 otherwise (i.e., in other regions of the country)*

D_{3i} = *1 if the state is in the South*

= *0 otherwise (i.e., in other regions of the country)*

In the above model, 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.

ANOVA Models cntd...

Interpretation

- Mean salary of public school teachers in the Northeast and North Central:

  $E(Y_i | D_{2i} = 1, D_{3i} = 0) = \beta_1 + \beta_2$

- Mean salary of public school teachers in the South:

  $E(Y_i | D_{2i} = 0, D_{3i} = 1) = \beta_1 + \beta_3$

- Mean salary of public school teachers in the West:

  $E(Y_i | D_{2i} = 0, D_3 i = 0) = \beta_1$

Empirical Results

- Using the example 1, suppose we obtain the following results:

  $\hat{Y}_i = 26{,}158.62 - 1734.473D_{2i} - 3264.615D_{3i}$

  se = (1128.523) (1435.953) (1499.615)

  t = (23.1759) (−1.2078) (−2.1776)

  (0.0000)* (0.2330)* (0.0349)* $R^2 = 0.0901$

  where * indicates the *p values*.

ANOVA Models cntd…

Interpretation of empirical results:

- The results show from the regression, that the mean salary of teachers in the West is about $26,158, that of teachers in the Northeast and North Central is lower by about $1734, and that of teachers in the South is lower by about $3265.
- The actual mean salaries in the last two regions can be easily obtained by adding these differential salaries to the mean salary of teachers in the West
- We can find that the mean salaries in the latter two regions are about (= $26,158- $1734= $24,424) and (= $26,158- $3265 = $22,894).

Caution in the Use of Dummy Variables

First

- In previous Example have three regions, we used only two dummy variables, D_2 *and* D_3.
- *We have not used three dummies to distinguish* the three regions? Suppose we do that and write the model as:

  $Y_i = \alpha + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i$

  where D_{1i} *takes a value of 1 for states in the West and 0 otherwise.*
- Now, we have a dummy variable for each of the three geographical regions.
- If run the regression using the above model, the computer will "refuse" to run the regression. The reason is that we have a dummy variable for each category or group and also an intercept, and in this case, we get **perfect collinearity,** that is, exact linear relationships among the variables.
- If we do not follow this rule, we will fall into what is called the **dummy variable trap**
- **If a qualitative variable has m categories, introduce only (m − 1) dummy variables.**

Caution in the Use of Dummy Variables

Second

- The category for which no dummy variable is assigned is known as the **base, benchmark, control, comparison, reference, or omitted category.** All comparisons are made in relation to the benchmark category.

Third

- The intercept value (β_1) represents the mean value of the benchmark category. In Example 1, the benchmark category is the Western region. Hence, the intercept value of about 26,159 represents the mean salary of teachers in the Western states.

Four

- We already have studied about the dummy variable trap. There is a way to circumvent this 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.
- If we drop the intercept term from example 1, and we can consider the following model,

$$Y_i = \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i$$

Caution in the Use of Dummy Variables

- In this model, we do not fall into the dummy variable trap, as there is no longer perfect collinearity. But we have to make sure that when we run this regression, without intercept.
- Here we interpret regression like

  β_1 = *mean salary of teachers in the West*

  β_2 = *mean salary of teachers in the Northeast and North Central.*

  β_3 = *mean salary of teachers in the South.*

Five

- Most researchers find the equation with an intercept more convenient because it allows them to address more easily the questions in which they usually have the most interest, namely, whether or not the categorization makes a difference, and if so, by how much.
- If the categorization does make a difference, by how much is measured directly by the dummy variable coefficient estimates.

ANOVA models with two qualitative variables

- The regression model can written as:
- $\hat{Y}_i = 8.8148 + 1.0997D_{2i} - 1.6729D_{3i}$

  se = (0.4015) (0.4642) (0.4854)

  t = (21.9528) (2.3688) (−3.4462)

  (0.0000)* (0.0182)* (0.0006)* $R^2 = 0.0322$

* denotes the *p values*

where Y = hourly wage ($)

D2 = married status, 1 = married, 0 = otherwise

D3 = region of residence; 1 = South, 0 = otherwise

- In this example we have two qualitative regressors, each with two categories. Hence we have assigned a single dummy variable for each category

ANOVA models with two qualitative variables

Interpretation

- It is unmarried, non-South residence is a base category.
- In other words, unmarried persons who do not live in the South are the omitted category. Therefore, all comparisons are made in relation to this group.
- The mean hourly wage in this benchmark is about $8.81. Compared with this, the average hourly wage of those who are married is higher by about $1.10, for an actual average wage of $9.91 (= 8.81 + 1.10).
- By contrast, for those who live in the South, the average hourly wage is lower by about $1.67, for an actual average hourly wage of $7.14 (= 8.81 - 1.67).

ANCOVA Models

- We can extend the previous model by adding the variables of Expenditure on public schools by local authorities, as public education is primarily a local and state question. The model can be written as:

  $Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 X_i + u_i$

  where Y_i = *average annual salary of public school teachers in state ($)*

  X_i = *spending on public school per pupil ($)*

  D_{2i} = *1, if the state is in the Northeast or North Central*

  = *0, otherwise*

  D_{3i} = *1, if the state is in the South*

  = *0, otherwise*

- With two qualitative regressors, we added a quantitative variable, X, which is called as ANCOVA models is also known as a **covariate.**

ANCOVA Models cntd…

Empirical Model

- $\hat{Y}_i = 13{,}269.11 - 1673.514D_{2i} - 1144.157D_{3i} + 3.2889X_i$

  se = (1395.056) (801.1703) (861.1182) (0.3176)

  t = *(9.5115)** *(−2.0889)** *(−1.3286)*** *(10.3539)** $R^2 = 0.7266$

  *where * indicates p values less than 5 percent, and ** indicates p values greater than 5 percent.*

- The results suggest, *ceteris paribus:* as public expenditure goes up by a dollar, on average, a public school teacher's salary goes up by about $3.29.
- Controlling for spending on education, we now see that the differential intercept coefficient is significant for the Northeast and North-Central region, but not for the South. These results are different from previous results.

Interaction effects using dummy variables

1. Suppose, the interaction of dummy variable model is given by

$Y_i = \alpha_1 + \alpha_2 D_{2i} + \beta_1 X_i + \beta_2 (D_{2i} * X_i) + u_i$

where *Y = literacy rate (per cent)*

X= per capita household income (Rs)

D_2 = *1 if Urban, 0 otherwise*

$D_{2i} * X$=*Interaction of urban household dummy with per capita household income*

α_2 = *differential effect of being a resident of Urban*

β_1 = *differential effect of per capita household income*

β_2 = *differential effect of being a urban area with per capita household income*

Interaction effects using dummy variables cntd…

2. Suppose, the interaction of dummy variable model is given by

$$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 (D_{2i} * D_{3i}) + \beta X_i + u_i$$

where Y = *hourly wage in Rs.*

X = *education (years of schooling)*

D_2 = *1 if female, 0 otherwise*

D_3 = *1 if Urban, 0 otherwise*

α_2 = *differential effect of being a female*

α_3 = *differential effect of being a urban household*

α_4 = *differential effect of being a female urban household*

Piecewise linear regression

Suppose the model is as follows

$Y_i = \alpha_1 + \beta_1 X_i + \beta_2 (X_i - X^*) D_i + u_i$

where Y_i = sales commission, X_i = volume of sales generated by the sales person, X^* = threshold value of sales also known as a **knot** (known in advance)

$D = 1$ if $X_i > X^*$

$= 0$ if $X_i < X^*$

Assuming $E(u_i) = 0$,

- $E(Y_i \mid D_i = 0, X_i, X^*) = \alpha_1 + \beta_1 X_i$

  which gives the mean sales commission up to the target level X^* and
- $E(Y_i \mid D_i = 1, X_i, X^*) = \alpha_1 - \beta_2 X^* + (\beta_1 + \beta_2) X_i$

  which gives the mean sales commission beyond the target level X^*.

Piecewise linear regression cntd…

- Thus, β_1 gives the slope of the regression line in segment I, and $\beta_1 + \beta_2$ gives the slope of the regression line in segment II of the piecewise linear regression.
- A test of the hypothesis that there is no break in the regression at the threshold value X*. This can be conducted by testing the statistical significance of the estimated differential slope coefficient β_2
- Incidentally, the piecewise linear regression is an example of a more general class of functions known as **spline functions.**

The Interpretation of Dummy Variables in Semilogarithmic Regressions

- The Dummy Variables in Semilogarithmic Regressions model is as follows.

  $\ln Y_i = \beta_1 + \beta_2 D_i + u_i$

  where Y = hourly wage rate (Rs.) and D = 1 for female and 0 for male.
- Assuming $E(u_i) = 0$, we obtain Wage function for male workers:

  $E(\ln Y_i \mid D_i = 0) = \beta_1$
- Wage function for female workers:

  $E(\ln Y_i \mid D_i = 1) = \beta_1 + \beta_2$
- The intercept β_1 gives the mean log hourly earnings and the "slope" coefficient gives the difference in the mean log hourly earnings of male and females. But, this is very difficult to interpret.
- We can take the antilog of β_1, what we obtain is not the mean hourly wages of male workers, but their **median wages.** We know that mean, median, and mode are the three measures of central tendency of a random variable. And if we take the antilog of $(\beta_1 + \beta_2)$, we obtain the median hourly wages of female workers.

Thank you..