[Academic Script]

Dummy Variables

Subject:	Business Economics
Course:	B.A., 3rd Semester, Undergraduate
Paper No. & Title:	Paper – 304 Basic Econometrics
Unit No. & Title:	Unit - 5 Dummy Variables
Lecture No. & Title:	Lecture – 1 Dummy Variables

Academic Script

1. Introduction

Consider the regression equation $Y= a+b_1X_{1+} b_2X_2+U$

Where, Y= Income of a person.

X_1= Age of a person.

X_2= Length of service.

Here dependent variable Y and independent variables X_1 and X_2 all are quantitative variables and U is disturbance term, But sometime qualitative variables may be used as independent variables in the regression model.

If we add third independent variables X_3 for gender of a person then here X_3(gender) is a qualitative variable.

We may rewrite the above model as

$Y= a+b_1X_1+ b_2X_2+ b_3X_3+U.$

Then question may arise here how to use qualitative variables like: religion, education, marital status etc. in the regression model as independent variables?

This is done by using dummy variables.

2. Dummy Variables

A dummy variable is a qualitative variable that can be divided into two or more categories. When qualitative variable divided into two categories, e.g. gender can be divided as male or female, it can be used as an indication variable which assumes the values 0 and 1 we may take 0 for female and 1 for male.

Sometime qualitative variable is divided into 3 or more categories like:

(i) Sales region: east, west, north and south.

(ii) Method of payment: cash, cheque and credit card.

(iii) Season: summer, winter, monsoon.

For such type of qualitative variables, dummy variables are used in the regression model.

3. How to use dummy variable?

To avoid inter relationship among the categories of dummy variable, the thumb rule.

If the qualitative variable X has K categories then K-1 dummy variable are used for the qualitative variable.

e.g.: **(i)** For gender of a person, there are two categories Male and Female.

We use only one dummy variable D for gender, defined as follow:

$$D_1 = \begin{cases} 1, & if\ i^{th}\ person\ is\ male \\ 0, & if\ i^{th}\ person\ is\ female \end{cases}$$

(ii) For seasons, there are three categories viz: Summer, Winter and monsoon for seasons, defined as

$$D_2 = \begin{cases} 1, & if\ i^{th}\ season\ is\ 'Summer' \\ 0, & otherwise \end{cases}$$

$$D_3 = \begin{cases} 1, & if\ i^{th}\ season\ is\ 'Winter' \\ 0, & otherwise \end{cases}$$

Here if $D_2=0$ and $D_3=0$ that means the season is monsoon. We may use the above qualitative variables in the regression model to estimate sales of a person based on gender and season as,

$Y= a+b_1D_1+ b_2D_2+ b_3D_3+U.$

Here Y denotes the sales of a person.

Thus dummy variables are used in regression models to analyse and estimate difference among groups; and they are used as other explanatory variables in the multiple regression model.

i.e. Dummy variable are used with quantitative variable as explanatory variables in the regression model.

If we consider the regression model to estimate sales of a person based on the age of a person and the season.

We use the model as $Y= a+b_1X_1+ b_2X_2+ b_3X_3 +U$

Where Y= sales of a person.

X_1= age of a person.

$$D_2 = \begin{cases} 1, & if\ i^{th}\ season\ is\ 'summer' \\ 0, & otherwise \end{cases}$$

$$D_3 = \begin{cases} 1, & if\ i^{th}\ season\ is\ 'winter' \\ 0, & otherwise \end{cases}$$

Here average of values X gives average age of persons while average values of dummy variable D_2 gives the proportion of season 'summer' in the given data.

(a) Reference group

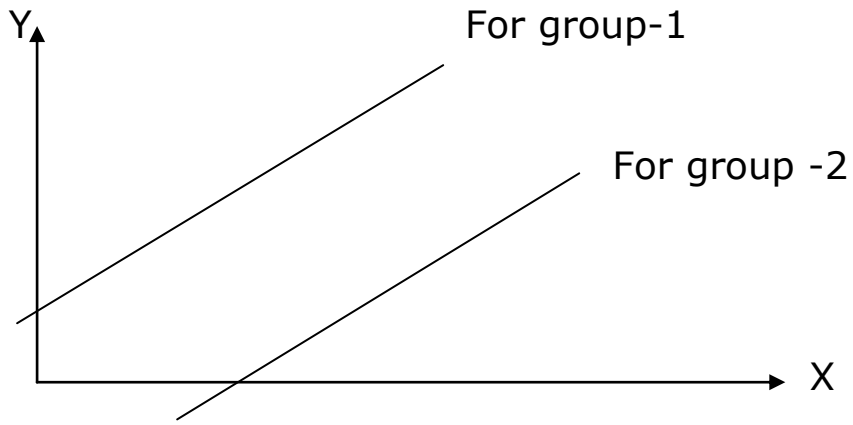
One of the two groups in a definition of a dummy variable is called the 'included' and the other is called 'excluded' group. The group identified with a value of 1 in the definition of the dummy variable is called 'included group'. The excluded group carries a value 0 and is also known as the control group.

For instance, if D=1 for male and D=0 for female, control group and the included group is male which is called reference group also.

(b) Common slope

Suppose that the effect of X on Y is the same for both groups and that regardless of the level of X, there is systematic

difference between the two groups. Graphically the solution is depicted by two parallel lines with different intercepts as shown in the following figure:



(c) Interactions

An interaction between two qualitative variables (dummy variable) or between a qualitative variable (dummy variable) and quantitative variable allows the analyst to estimate difference in the slope among groups.

For instance, in the salary equation if we include the product of the two variables sex (dummy variable) and experience as 'sex * experience' the coefficient of the variable would indicate whether there is any difference between the additional salary that males and females can obtain with an additional year of experience.

When the interaction of two qualitative (dummy) variables is used then a third dummy variable is created for the product of the two dummy variable D_3 as,

$$D_3 = \begin{cases} 1, & \text{if gender} = \text{`male (1) and Marital status} = \text{married (1)} \\ 0, & \textit{otherwise} \end{cases}$$

Example: suppose we have 50 information technology firms which are categories (i) for equipment & software and (ii) other types say telecom & electronics. The data include R & D budget and net income of the companies.

The fitted equation without category of the companies may be like,

R & D = 1987 + 0.3471 (income) ..(1)

But if the dummy variable is used for the categories of the firm as,

$$D = \begin{cases} 1, & \text{if the firm is for equipment } and\ software \\ 0, & \text{if the firm is for telecom } electronics \end{cases}$$

Then regression equation including dummy variable may be obtained as,

R & D = 1271 + 0.3162 (income) + 994.62D ..(2)

Putting D=1, we get the regression equation for equipment & software firms as,

Y = 2265.62 + 0.3162(income) ..(3)

And the equation for telecom & electronics firm as,

Y = 1271 + 0.3162 (income) ..(4)

Here both the regression line (3) and (4) have same slope (0.3162) but different intercepts 2265.62 and 1271 respectively.

The equation (1) is cancelling out the group difference and thus report misleading intercept. The difference between two groups of firms looks substantial, which can be verified by the difference of intercepts of equations (3) & (4). (difference = coefficient of dummy variable in equation (2))

Note: **(i)** if we code the dummy variable reversely,

i.e.

$$D = \begin{cases} 1, & \text{if the firm is for equipment and } software \\ 0, & \text{if the firm is for telecom } and\ electronics \end{cases}$$

It gives equivalent results.

(ii) Numerical variables are interval or ratio scale variables whose values are directly comparable, e.g. 60 kg is twice as much as 30 kg or height 62'' is 3 inches more than the height 59''. But for dummy variables the numbers are used to indicate the category. e.g. 0 for primary level education, 2 for higher secondary level education & 3 for graduate or more level of education. Such numbers do not have intrinsic meaning of their own. Here 3 plus 2 or 3 minus 2 do not mean anything. Dummy variables are created to trick the regression algorithm into correctly analysing attributes variables.

4. Construction of regression with dummies

Now, consider a situation where three (or more) groups are needed to consider the categories of a qualitative variable.

Example: let us consider a data on sales of a company in there different region made by 20 salesmen of ages between 19 to 30 years.

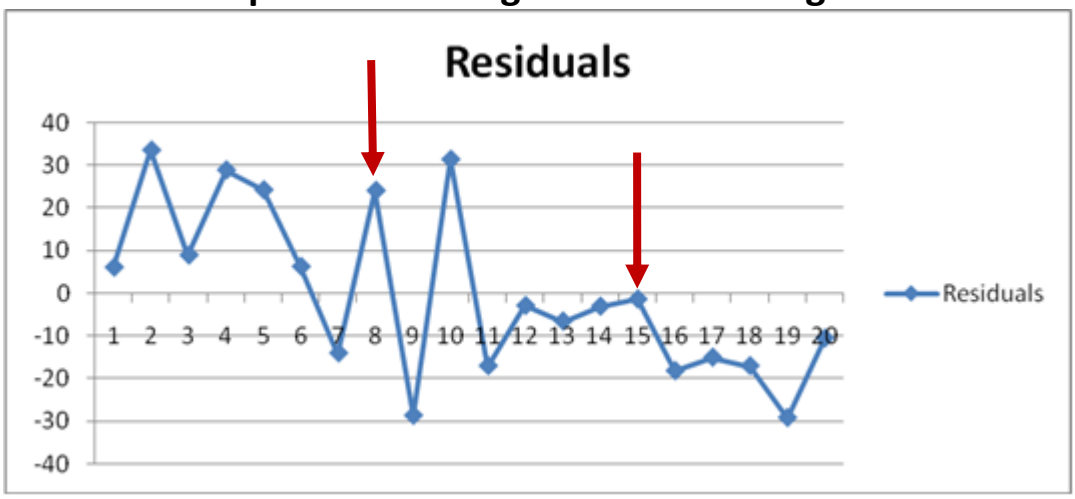
Sales Executive	Gender	Age	Region	Sales
1	1	25	1	50
2	1	22	1	75
3	1	20	2	11
4	1	27	2	77
5	1	28	3	45
6	1	24	1	52
7	1	24	2	26
8	1	23	3	24
9	2	24	3	28
10	2	30	3	31
11	2	19	2	36
12	2	24	1	72
13	2	26	1	69
14	2	26	1	51
15	2	21	2	34
16	2	24	2	40
17	2	29	3	18
18	2	27	3	35
19	2	24	1	29
20	2	25	1	68

Solution:

(i) The regression equation of sales on age without dummy variable can be obtained from the data as,

Sales= 23.039 + 0.834(Age) (R^2=0.014)

Residuals predicted using the variable " Age"



Red arrows denote the partition according to regions

(ii) Regression of sales on age & regions (without dummy variable techniques)

Here we use two explanatory variables age and regions (region are labelled as 1, 2, 3) then fitted equation is:

Sale = 16.550 + 2.396(Age) – 16.806 (region)

= 16.550 + b_1 (Age) – b_2 (region)

Here for given age (fixed) the difference in sales between region 2 and region 1 is given by,

$(b_0 + 2\ b_2) - (b_0 + b_2) = b_2$

And the difference between region 3 and 2 is,

$(b_0 + 3\ b_2) - (b_0 + 2b_2) = b_2$

i.e. the difference in sales between the region 2 and 1 is same as that between the region 3 and 2.This is not necessarily true. Therefore this method of use of qualitative variable region is not proper.

(iii) Let us use two dummy variables for the three types of region defined as,

$$D_{r1} = \begin{cases} 1, & for\ region\ 1 \\ 0, & otherwise \end{cases}$$

$$D_{r2} = \begin{cases} 1, & for\ region\ 2 \\ 0, & otherwise \end{cases}$$

and the fitted equation is obtained as,

Sales = -36.731 + 2.493(Age) + 33.901 D_{r1} + 17.970 D_{r2} (R^2=0.479)

This equation does not impose any restriction on sales for the three regions.

(iv) Let us use now all the three dummy variables, ignoring intercept in the regression model. The fitted equation becomes,

Sales = 2.493(Age) – 2.830 (D_{r1}) - 18.761 (D_{r2}) – 36.731 (D_{r3}) (R^2=0.913)

Note: **1)** Here we have not included intercept in the model i.e. fit the regression through the origin. In such situation R^2 measures the proportion of the variability in the dependent variable about the origin explained by regression. This cannot be compared to R^2 for models which include an intercept.

2) When we include intercept in such model (including all the 3 dummy variables for region) we will be caught in a so called 'dummy variable trap''. This creates multicollinearity and the regression equation becomes unsolvable.

5. Dummies for testing for the presences of seasonal trends

Many economics time series data based on quarterly or monthly data exhibit seasonal patterns (regular oscillatory movements) e.g. sales of woollen clothes at winter and other seasons, demand of soft drinks during summer and other seasons etc. Often it is advisable to remove the seasonal component from a time series, So that one can concentrate on trend of the data. This process of removing the seasonal component from a time series is known as deaseonalization and the time series thus obtained is called deseasonalized time series.

There are several methods of daseasonalizing a time series. One of them is the method of dummy variables.

Consider a quarterly data on the sales of refrigerators over the sample period 1978 to 1985. For four quarter (For data see page 313, table 9.3, Book: Basics of Econometrics by D. Gujarati) the arrangement of the data is as follows:

Year	Quarter			
	I	II	III	IV
1978	1317	1615	1662	1295
1979	1271	1555	1639	1238
1980	1277	1258	1417	1185
1981	1196	1410	1417	919
1982	943	1175	1269	973
1983	1102	1344	1641	1225
1984	1429	1699	1749	1117
1985	1242	1684	1764	1328

Here we consider the following model:

$$Y= \beta_0+ \beta_1 D_{2+} \beta_2 D_3+ \beta_3 D_4+U$$

Where D_2=1 for 2nd quarter, 0 otherwise

D_3=1 for 3rd quarter, 0 otherwise

D_4=1 for 4th quarter, 0 otherwise.

Here we have assumes first quarter as the reference quarter assigned dummier to the second; third and fourth quarters.

On fitting of the above model to the above sales data, one can obtained the regression equation,

$$Y=1222.1250 + 245.3750\ D_2 + 347.6250\ D_3 - 62.1250 D_4 \quad ..(1)$$

t: $(20.3720)^*$ $(2.8922)^*$ $(4.0974)^*$ (-0.7322)

$R^2 = 0.5318$

Where * indicates p values less than 0.05.

(a) Interpretation and test for seasonal trends

Since we have considered the first quarter as the reference (benchmark) quarter, the coefficients attached to the various dummies showing by how much the average value of Y in the

quarter that receives a dummy values of 1 differs from that of the benchmark (first) quarter.

i.e. the coefficients on the seasonal dummies will give the seasonal increases as decreases in the average value of Y relative to the base season.

By adding various differential intercept values to the benchmark (first quarter) average value of 1222.125 we will get the average value for the various quarters.

Here the average value of Y for the fourth quarter is not significant(p-value is grater then 0.05) at 5% level of significance from the average value for the first quarter, as the dummy coefficient for the fourth quarters are statistically different from the average value for the first quarter is not statistically significant, While other quarter have significant dummy coefficient (p-value < 0.05). i.e. the average values of Y for second and third quarter are statistically different from the average value for the first quarter.

(b) How to get deseasonalized time series?

Using the fitted regression equation (1), we get estimates of the values of Y_t say $\hat{Y}_t$ for each observation and subtracting them from the actual values of Y. i.e. calculation

$Y_t - \hat{Y}_t$, which are nothing but the residuals from the regression. Such deseasonalized time series contains trend, cycle and random components.

Deseasonalized time series

Year	Quarter			
	I	II	III	IV
1978	94.875	147.500	92.250	135.000
1979	48.875	87.500	69.250	78.000
1980	54.875	-209.500	-152.750	25.000
1981	-26.125	-57.500	-152.750	-241.000
1982	-279.125	-292.500	-300.750	-187.000
1983	-120.125	-123.500	71.250	65.000
1984	206.875	231.500	179.250	-43.000
1985	19.875	216.500	194.250	168.000

6. Summary

When qualitative variables are used in the regression model as an explanatory (independent) variable, a dummy variable is defined for such qualitative variable and this dummy variable is used in the regression model. We have discussed the technique of creating the dummy variable based on the category of a qualitative variable. We may use dummy variables for the qualitative variables like: gender, Education status, region for job, profit-loss category etc. We may use both qualitative and quantitative variables as independent variables in the regression model. Here we assume dependent variable as quantitative variable. When dependent variable is qualitative then logistic regression model are used. Dummy variables have many applications in the different fields. In economics important application of dummy variable is to test the presence of seasonal trends in the given time series and make it desisonalized.