[Academic Script]

Regression Analysis for Qualitative Variables

Subject:	Business Economics
Course:	B. A. (Hons.), 5th Semester, Undergraduate
Paper No. & Title:	Paper – 531 Elective Paper Q1 – Advanced Econometrics
Unit No. & Title:	Unit – 4 Binary Data and Limited Dependent Variable Models
Lecture No. & Title:	Lecture – 1 Regression Analysis for Qualitative Variables

Academic Script

1. Introduction

Hello friends nice meeting you. Today we are discussing specialized techniques known as Dummy Variables techniques. Which are used for dealing with qualitative data. Mostly we have quantitative data, but many times we also have qualitative data. As for example sex, religion etc. for that we have to adopt separate methods. So all qualitative variables are to be converted in to quantitative by use of dummy variables. There are very nice applications for this techniques and we want to study this techniques in detail in this lecture. So let us start our lecture.

While dealing with analysis of data collected, we find that the concerned variables are quantitative or qualitative.

e.g.

(1) Data connected with output, price, income, expenditure, cost, height, weight, blood pressure etc. are representing quantitative variables. This means that they are measured by means of some scaling units.

(2) Data connected with sex, race, religion, marital status, educational status, caste structure, colours, strikes, wars etc. are representing qualitative variables. They are not measured by means of some scaling units.

The question is how we can deal with the qualitative nature of these variables?

By some approach we may introduce them so that they may be converted from qualitative into quantitative nature. Here we use what is called Dummy Variable or Categorical Variable or Dichotonomous Variables.

We want to study in this lecture about regression analysis concerning such dummy variables. Let us consider some illustrations to begin with.

(1) <u>One dummy independent variable</u>

Let $Y_i = \alpha + \beta D_i + U_i$

Y_i = Annual income of a person

D_i = 1 If the person is male

= 0 if female

Here D is the dummy variable representing sex which is a qualitative variable.

(2) <u>Two dummy independent variables</u>

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

Y_i = Annual Income

D_{1i} = 1 For Male

= 0 For Female

D_{2i} = 1 For Married Person

= 0 For Unmarried Person

Here sex and marital status are represented by means of two dummy variables D_1 and D_2.

(3) <u>One explanatory variable and one dummy independent variable</u>

$Y_i = \alpha + \beta_1 X_i + \beta_2 D_i + U_i$

Y_i =Annual Income

X_i =Years of Service

D_i = 1 For Male

= 0 For Female

This is a mixture of explanatory variable i.e. service experience with qualitative variable sex.

(Note: <u>In model (1)</u>

$E(Y_i|D_i) = \alpha + \beta D_i$

So that $E(Y_i|D_i = 1) = \alpha + \beta =$ Average Annual Income of male

$E(Y_i|D_i = 0) = \alpha =$Average Annual Income of female

<u>In Model (2)</u>

$E(Y_i|D_{1i}, D_{2i}) = \beta_1 + \beta_2 D_{1i} + \beta_3 D_{2i}$

<u>In Model (3)</u>

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

From the above presentation it may be clear that if we do not use dummy variables then we have to run separate regressions for male and female etc. Thus use of dummy variables simplifies the method by considering only one regression.

Dummy variables occur in practice in three ways

(1) Dummy independent (explanatory) Variables

(2) Dummy dependent Variables

(3) Mixture of both dummy dependent and independent variables

At present we want to consider only the first category.

(2) <u>Some essential features of the dummy variables</u>

(1) If a qualitative variable has m categories (or classes) then include (m – 1) dummy variables. If this is not observed, there is multicollinearity in the model. Such a situation is called <u>dummy variables trap</u>.

e.g. For the qualitative variable sex, if we define

$D_{1i} = 1$ For male

$= 0$ Otherwise

$D_{2i} = 1$ For Female

$= 0$ Otherwise

Then if we have a sample 3 males and 2 females (with income and experience as other variables) we have the following matrix.

$$\begin{array}{cccccc} & & & D_1 & D_2 & X \\ M & Y_1 & 1 & 1 & 0 & X_1 \end{array}$$

M	Y_2	1	1	0	X_2
F	Y_3	1	0	1	X_3
M	Y_4	1	1	0	X_4
F	Y_5	1	0	1	X_5

Here $D_2 = 1 - D_1$

Which shows that there is multicollinearity which is dummy variables trap. Here only one dummy variables should be used.

(2)In general, dummy variable can take any value $z = a + b\,D$ $(b \neq 0)$

So that if $D = 0$, $z = a$ and if $D = d$, $z = a + b$

Thus (0, 1) can be replaced by (a, a + b) and its choice is purely arbitrary.

The value 0 is called Base or Control or omitted category.

(3)The co-efficient α_1 attached to the dummy variable D_1 is called Differential Intercept Co-efficient (DIC). The co-efficient attached with the explanatory variable (other than dummy) is called differential slope coefficient.

(4)A regression model which contains explanatory variables which are exclusively dummy variables is called Analysis of Variance (ANOVA) model.

A regression model which contains a mixture of both dummy explanatory variable as well as some explanatory variables is called Analysis of Covariance (ANCOVA) model.

2. Some applications of dummy variables techniques

(3.1) Combining two regressions

This is also called pooling of cross section and time series data.

Suppose that we consider bivariate regression for the data in two ranges of periods. First sample contains n_1 observations, second one contains n_2 observations. If we combine these two samples, we have ($n_1 + n_2 = n$) observations.

We may think to run two separate regressions for both of them and also third one for the entire period. Instead of that we can run only one regression by using dummy variables technique.

Sample (I) $Y_t = \alpha_1 + \beta X_t + U_{1t} \quad (t = 1, 2, \ldots, n_1)$ → (1)

Sample (II) $Y_t = \alpha_2 + \beta X_t + U_{2t} \quad (t = 1, 2, \ldots, n_2)$ → (2)

Sample (III) $Y_t = \alpha_1 Z_{1t} + \alpha_2 Z_{2t} + \beta X_t + U_t$

$(t = 1, 2, \ldots, n_1 + 1, .. n_1 + n_2 = n)$ → (3)

Here assumption is that slope does not change and also homoscadasticity in the respective regressions.

In the combined model,

Z_{1t} = 1 if $t = 1, 2, \ldots, n_1$

= 0 if $t = n_1 + 1, .. n_1 + n_2$

Z_{2t} = 1 if $t = n_1 + 1, .. n_1 + n_2$

= 0 if $t = 1, 2, \ldots, n_1$

Z_{1t} and Z_{2t} are the dummy variables.

We can find OLSE of $\hat{\beta}$ for the first two models which are

$$\hat{\beta}_{(1)} = \frac{\sum_1^{n_1} (Y_t - \bar{Y}_1)(X_t - \bar{X}_1)}{\sum (X_t - \bar{X}_1)^2} \quad \rightarrow (4)$$

$$\hat{\beta}_{(2)} = \frac{\sum_{n_1+1}^{n_2} (Y_t - \bar{Y}_2)(X_t - \bar{X}_2)}{\sum_{n_1+1}^{n_2} (X_t - \bar{X}_2)^2} \quad \rightarrow (5)$$

With $V\left(\hat{\beta}_{(1)}\right) = \frac{\sigma^2}{S_1^2}$ →(6)

$$V\left(\hat{\beta}_{(2)}\right) = \frac{\sigma^2}{S_2^2} \quad \rightarrow (7)$$

Instead of running these two separate regressions, suppose we run the regression in (3) using dummy variables method, then we can write the normal equations as usual and by solving them we can obtain the estimator of β given by

$$\hat{\beta}_* = \frac{S_1^2 \hat{\beta}_{(1)} + S_2^2 \hat{\beta}_{(2)}}{S_1^2 + S_2^2} \quad \rightarrow (8)$$

And $V(\hat{\beta}_*) = \frac{\sigma^2}{S_1^2 + S_2^2}$ →(9)

Comparing these results with the above two separate regressions, we find that $\hat{\beta}_*$ is unbiased estimator for β and $V(\hat{\beta}_*)$ < $V(\hat{\beta}_{(i)})$ (i = 1, 2)

This shows that $\hat{\beta}_*$ is more efficient than $\hat{\beta}_{(1)}$ and $\hat{\beta}_{(2)}$.

Which means that by using dummy variables technique not only that we can solve the problem of running separate regression, but also we can obtain estimator which is unbiased and more precise (i.e. more efficient).

(3.2) Interaction effects using Dummy Variables

This is an interesting application for using dummy variables method.

Let us consider the following model

$Y_t = \alpha_1 + \alpha_2 D_t + \beta_1 X_t + \beta_2 (D_t X_t) + U_i \rightarrow$ (1)

Where Y = Literary rate

X = Per capita Net State Domestic product

We use two dummy variables D_2 and D_3, where D_2 is for Gender.

and D_3 is for residential status.

D_2 = 1 if Male

= 0 if Female

D_3 = 1 if Urban

= 0 if Rural

Here $D_t X_3$ is the combined term (product of D_t and X_3) showing interaction.

You will notice that we would have used different regressions for urban males, urban females, dummy variables. In addition, here the model presented in (1) above also gives the interaction effect considering gender with residential status. (i.e. males from urban area have more literacy rate or not etc.).

The above model given in (1) above is an example of additive model.

We can also have a multiplicative model as shown below.

$Yt = \alpha_1 + \alpha_2 D_{2t} + \alpha_3 D_{3t} + \alpha_4 (D_{2t}. D_{3t}) + \beta X_t + U_t$ →(2)

The above extension given in (2) is also called Dummy Interaction due to the product term $D_{2t}. D_{3t}$. We can think of such dummy interaction when two or more qualitative variables are included in the model.

(3.3) Deseasonalisation of time series data

In time series analysis, we have seen different methods to separate the seasonal component. Here dummy variables technique can be useful to deseasonlise the given time series.

We collect time series data on the basis of months and quarters in a given year. Suppose that the data are collected on quarterly basis for n years. Thus we have 4n observations. Seasonal effect is a very important part under temporal factors, hence every time series is to be deseasonalised.

Let y_{ij} = Value of Y in the i^{th} year and j^{th} quarter

(i = 1, 2, . . n_i, j = 1, 2, 3, 4)

We write

$$\underline{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ \vdots \\ y_{n1} \\ y_{n2} \\ y_{n3} \\ y_{n4} \end{bmatrix}_{:4n\times 1} \qquad \underline{b} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix}_{:4\times 1}$$

then we have the model

$\underline{y} = D \cdot \underline{b} + \underline{y}^{\alpha}$

Where $D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 1 \end{bmatrix}_{:4n\times 4}$

Here $\underline{y}^{\alpha} =: 4n \times 1$ is the disturbance term

OLSE of $\underline{b}$ is given by

$\hat{\underline{b}} = (D'D)^{-1}D'\underline{y}$

With $V(\hat{\underline{b}}) = (D'D)^{-1}\sigma^2$

We have $E\left(\underline{y}^{\alpha}\underline{y}^{\alpha'}\right) = \sigma^2 I_{4n}$

(Note that in the above presentation. sample matrix D is composed for 4 dummy variables defined by D_{it} = 1 if t occur in

ith quarter i = 1, 2, 3, 4

= 0 otherwise)

$\hat{\underline{y}}^{\alpha} = \underline{y} - \hat{\underline{y}} = \underline{y} - D\hat{\underline{b}} = M\underline{y}$

Where $M = [I - D(D'D)^{-1}D']$ which is Symmetric Idempotent Matrix with MD = 0.

$\hat{\underline{y}}^{\alpha}$ is the deseasonalised series.

<u>Extension of the model</u>

In the above presentation, the deseasonalised series $\hat{\underline{y}}^{\alpha}$ has a drawback that the components of $\hat{\underline{y}}^{\alpha}$ add to zero. The real deseasonalised vector is $\hat{\underline{y}}^{\alpha} + \bar{y}$ where

$\bar{y} = \Sigma\Sigma\, y_{ij}/4n$

In fact in a time series data, there are other components of trend, cyclic fluctuations etc. Thus the above model is to be extended further. We write the model as

We write the model as

$$\underline{y} = (P,D)\begin{pmatrix} \underline{a} \\ \underline{b} \end{pmatrix} + \underline{\hat{y}}^{\alpha}$$

Here P is chosen suitably

$$P = \begin{bmatrix} 1 & 1^2 & \dots\dots 1^p \\ 2 & 2^2 & \dots\dots 2^p \\ 3 & 3^2 & \dots\dots 3^p \\ . & . & \dots\dots\dots \\ 4n & 4n^2 & \dots 4n^p \end{bmatrix} : 4n \times p$$

Where the elements of P are the powers of time.

We may write the model as

$$\underline{y} = (P,D)\underline{C} + \underline{y}^{\alpha} = X\underline{C} + \underline{y}^{\alpha}$$

Where $\underline{C} = \begin{pmatrix} \underline{a} \\ \underline{b} \end{pmatrix}$ $and\, X = (P,D)$

under usual conditions of homoscadasticity, we use OLS method to find $\underline{\hat{C}}$ which is given by

$$\underline{\hat{C}} = \begin{pmatrix} \underline{\hat{a}} \\ \underline{\hat{b}} \end{pmatrix} = (X'X)^{-1}X'\underline{y}$$

$$X'X = \begin{pmatrix} P' \\ D' \end{pmatrix}(P,D) = \begin{pmatrix} P'P & P'D \\ D'P & D'D \end{pmatrix}$$

$$X'\underline{y} = \begin{pmatrix} P' \\ D' \end{pmatrix}\underline{y} = \begin{pmatrix} P'\underline{y} \\ D'\underline{y} \end{pmatrix}$$

Here $\underline{\hat{C}} = \begin{pmatrix} \underline{\hat{a}} \\ \underline{\hat{b}} \end{pmatrix} = \begin{bmatrix} P'P & P'D \\ D'P & D'D \end{bmatrix}^{-1} \cdot \begin{pmatrix} P'\underline{y} \\ D'\underline{y} \end{pmatrix}$

We are interested to find the estimate of $\underline{\hat{b}}$ which is given by $\underline{\hat{b}} = (D'ND)^{-1}D'N\underline{y}$

Where $N = [I - P(P'P)^{-1}P']$

Then $\underline{\hat{y}}^{\alpha} = \underline{y}^{ds} = \underline{y} - D\,\underline{\hat{b}}$

$$= \underline{y} - D(D'ND)^{-1}D'N\underline{y}$$

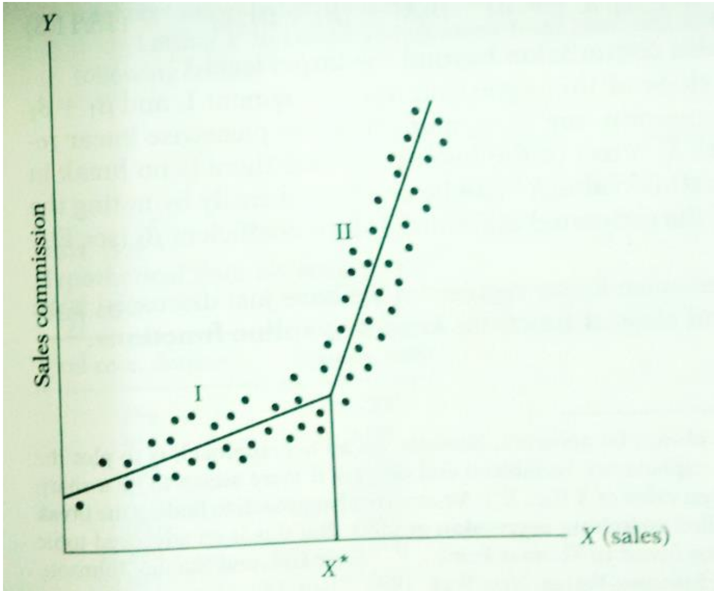
$$= [I - D(D'ND)^{-1}D'N]\underline{y}$$

Hence $\underline{y}^{ds} = T \cdot \underline{y}$

Where $T = [I - D(D'ND)^{-1}D'N]$ is an important matrix (but it is not symmetric), and TD=0. $\underline{y}^{ds}$ Series is the deseasonalised series.

3. Piecewise Regression

A very interesting application in dummy variables is what is called Piecewise regression. Suppose that a company gives remuneration to its sales employees (sales persons) as per their performance. This sales commission has a specific behavior that up to certain stage there is some amounts given as commissions, but after that stage the structure of commission stages. Thus we have two situations: Commission up to a given stage and commission after the given stage. This stage is called target or threshold. In both the cases we observe linearity, but these two linear functions are not the same. This is represented by what is known as Piecewise Regression. (Note that besides the sales, some other factors can also be effective but they are assumed to be represented by the stochastic disturbance term). This behavior is expressed clearly by the following diagram.



Here X^* is the target or threshold values for sales. Up to X^*, the situation is labeled as (I) and beyond X^*, the situation is labeled as (II). Not that both of them have linearity but under label (II) situation, the slope of line is steeper, thus indicating higher commission after when you increase sales as compared to threshold value.

Both the above linear relationships in situations (I) and (II) are represented by a single equation as under

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

Where Y_i= Sales commission

X_i = Volume of sales by the sales person

X^* = Target or threshold value of sales

D_i is the dummy variable such that

D_i = 1 if X_i> X^*

D_i = 0 if X_i < X^*

Assuming $E(U_i) = 0$ as usual we get

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

= Mean sales commission upto the target X_i

And

$E(Y_i | D_i = 1, X_i, X^*) = \alpha_1 - \beta_2 X^* + (\beta_1 + \beta_2) X_i$

= Mean sales commission beyond the target value X^*.

β_1 is the slope of the line in Segment I and $(\beta_1 + \beta_2)$ is the slope of the line in Segment II. Here advantage is that instead of two regressions segment wise, we can have only one regression dealing with the situation.

(It may be noted here that this illustrates what is known as Spline function. Further extensions on these lead to what is known as piecewise polynomials of order k and so on).

4. **Dummy variables in Semilogarithmic regression**:

When we consider Log-lin model concerning quantitative variables, the co-efficient attached to explanatory variables give the elasticity measure. We want to know what happens if we use dummy variables model.

Let us write our model as

$Ln(Y_i) = \beta_1 + \beta_2 D_i + U_i$ →(1)

Where Y_i = Literacy rate in percentage

D_i = 1 if male

= 0 if female

Ln is logarithm to the natural base

U_i is the disturbance term.

Under the assumption E(U_i) = 0, we get E [$Ln(Y_i)/D_i$= 1] = $\beta_1 + \beta_2$→(2)

and E [$Ln(Y_i)/D_i$ = 0] =β_1 →(3)

β_1 is the mean log literary rate and the slope co-efficient gives the difference in the mean log literacy rate of males and females. If we take antilog of β_1, what we obtain is not the mean but the median literacy rate of females. Similarly if we take antilog of ($\beta_1 + \beta_2$) we obtain the median literary rate of males.

5. **Further studies**

(1) What happens if we have dummy dependent variable? We shall study it separately with its relevant further extensions in the next lecture.

(2) There are further studies related with

(i) Random or varying parametric models

(ii) Switching regression models

(iii) Disequilibrium models etc.

(Note: we end up this discussion here looking to the limitations of our syllabus.)