[Academic Script]

Regression Analysis for Dummy Dependent Variable

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 – 2 Regression Analysis for Dummy Dependent Variable

Academic Script

1. Introduction

Hello friend's greetings of the day. Today we are learning a special category of problems corresponding to Dummy Variable. Hence Dummy Variables occurs as dependent variable. As for example if a person holds a credit card or debit card depends upon his income, his work experience etc. So the answer is either yes or no. Only two category so it is a dummy dependent variable. Our usual models will not be apply here due to the special nature of this problem. So here we will study what is known as linear problem model, Logit Model, Probit Model etc. So let us now listen to the lectures.

In all our studies so far pertaining to regression analysis, the dependent variable (or regressand) is a quantitative variable, whereas the explanatory variables can be quantitative or qualitative or a mixture of both quantitative and qualitative. Since dependent variable is under the effect of all other variables which are independent (or exogenous), it is also called **response variable**.

In practical applications, we do not always have a dependent variable which is quantitative in nature. That is, the dependent or response variable can be qualitative in nature.

To understand this let us take two illustrations as under

Illustration (1)

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

Here Y = Labour force participation

Y_i = 1 If a person is in the labour force

= 0 Otherwise

X = Average wage rate

D_i = 1 If the person is male

= 0 Otherwise

U = Disturbance term

Here response variable Y is a dummy dependent variable with two categories, X is quantitative variable and D_i is dummy independent variable.

Illustration (2)

$Yi = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \delta D_i + U_i$

Here, Y means whether a person is having Debit Card or not.

Yi = 1 If the person has a debit card

= 0 Otherwise

X_1 = Monthly income of the person

X_2 = Years of service of the person

D_i = 1 If the person is educated

= 0 Otherwise

U = Disturbance term

Here also we have one dummy dependent variable, two independent explanatory variables and one independent dummy variable.

We want to study now the case where the dependent variable is a qualitative variable. Such study is also known as study of **qualitative response regression models**.

Looking to the nature of the problem, it is evident that its study needs a separate treatment compared to our earlier studies.

If there are only two categories of dummy dependent variable, (i.e. yes or no, or 0 and 1) it is called **binary** or **dichotomous** variable. There can also be more categories viz. person belonging to political party A or party B or Part C – which is called **trichotomous** and in this way for many categories it is called **Polychotomous (multiple category)**.

We want to study here **<u>binary response regression</u>** models. There are certain models under this study.

(1) Linear Probability Model (LPM)

(2) Logit Model

(3) Probit Model

(4) Tobit Model etc.

2. Linear Probability Model (LPM)

We want to discuss here two variables model with Y as dummy dependent variable with two categories and one explanatory independent variable.

Some more illustrations for binary response variable can be given as under

(i) Y = 1 If a certain drug is effective

= 0 If it is not effective

(ii) Y = 1 If a family has disability insurance

= 0 If the family does not have insurance

(iii) Y = 1 If a company wants to declare bonus shares etc.

= 0 If not

Let us consider here the following model.

$Y_i = \beta_1 + \beta_2 X_i + Ui$ . . . . (1)

(i = 1, 2, ... n)

Here Y_i = 1 If the family owns a house

= 0 If it does not

X = Family annual income

U = Disturbance term

Model described above is called **<u>Linear Probability Model</u> (LPM)**.

We define $P_i = P (Y_i = 1)$ →Prob. that family owns a house

(Prob. of Success →Prob. that event will occur)

then $1 - P_i = P(Y_i = 0)$ →Prob. that family does not own a house (Prob. of Failure →Prob. that event will not occur)

Here variable Y_i has the following probability distribution.

Y_i	Probability
0	$1 - P_i$
1	P_i
	Sum = 1

We assume that $E(U_i) = 0$

Then from (1), $E(Y_i / X_i) = \beta_1 + \beta_2 X_i$ . . . . (2)

Now $E(Y_i) = 0.(1 - P_i) + 1.P_i = P_i$ . . . . (3)

Hence from (2) and (3),

$E(Y_i / X_i) = \beta_1 + \beta_2 X_i = P_i$ . . . . (4)

Since $0 \leq P_i \leq 1$

We get $0 \leq E(Y_i / X_i) \leq 1$ . . . . (5)

2. Estimation in LPM

While estimating LPM, there are certain problems which can be summarised briefly as under

(1) Non-normality of the disturbance term

From the model $Y_i = \beta_1 + \beta_2 X_i + U_i$

$U_i = Y_i - \beta_1 - \beta_2 X_i$

Hence When $Y_i = 1$, $U_i = 1 - \beta_1 - \beta_2 X_i$

and when $Y_i = 0$, $U_i = - \beta_1 - \beta_2 X_i$

Thus U_i takes only two values as above. Due to this it is incorrect to assume normality for U_i. However for large samples, normality assumption can be considered to be valid. U_i

In fact, for estimation purpose, normality assumption is not needed. So OLS estimators will be unbiased, and in large

samples, the inference of LPM will follow the usual OLS procedures under the normality assumptions.

(2) Heteroscadastic Variable of U_i

For LPM, we assume that $E(U_i) = 0$ and also $E(U_i U_j) = 0$ $(i \neq j)$ for all

$(i, j = 1, 2, \ldots n)$

(i.e. there is no serial correlation between the disturbances).

We obtain probability distribution of U_i

Y_i	U_i	Probability
0	$-\beta_1 - \beta_2 X_i$	$1 - P_i$
1	$1 - \beta_1 - \beta_2 X_i$	P_i
		Sum = 1

Since $E(U_i) = 0$, $V(U_i) = E(U_i^2)$

$$= (-\beta_1 - \beta_2 X_i)^2 (1 - P_i) + (1 - \beta_1 - \beta_2 X_i) P_i$$

Put $P_i = \beta_1 + \beta_2 X_i$ then simplification gives

$V(Ui) = (\beta_1 + \beta_2 X_i)(1 - \beta_1 - \beta_2 X_i)$

$$= P_i (1 - P_i) \quad \ldots . (6)$$

$$= E(Yi / X_i)[1 - E(Y_i / X_i)] \quad \ldots . (7)$$

Thus there is heteroscadasticity.

In the presence of heteroscadasticity OLS estimators are unbiased but not efficient. This can be dealt with by the usual remedial measures to overcome the problem of heteroscadasticity.

Method

Let $\sqrt{W_i} = \sqrt{Pi(1-Pi)} \quad \ldots . (8)$

then we divide both the sides of (1) by $\sqrt{W_i}$ so that

$$\frac{Y_i}{\sqrt{Wi}} = \frac{\beta_1}{\sqrt{Wi}} + \beta_2 \left(\frac{X_i}{\sqrt{Wi}}\right) + \left(\frac{U_i}{\sqrt{Wi}}\right) \quad \ldots . (9)$$

It can be easily checked that even though original model has heteroscadasticity, the transformed model has homoscadasticity. We have to estimate W_i to operate on the model.

Since $E(Y_i / X_i) = P_i$ and $W_i = P_i (1 - P_i)$

We get $\hat{W}_i = \hat{P}_i (1 - \hat{P}_i) = E(\hat{Y}_i / Xi) [1 - E(\hat{Y}_i / X_i)]$

$= \hat{Y}_i \quad \ldots . (10)$

We have two stage procedure to operate on the model

Stage – I. As usual run the regression of Yi on Xi and find OLS estimates and hence find $\hat{Y}_i$.

Hence $\hat{W}_i = \hat{Y}_i(1 - \hat{Y}_i) \quad \ldots . (11)$

Stage – II. Run the regression of the transformed model given in (9) above and find the estimators.

(3) Non-fulfillment of the condition $(0 \leq E(Y_i / X_i) \leq 1)$

Theoretically, we expect that $(0 \leq E(Y_i / X_i) \leq 1)$

However, while computing Yi may not follow such a rule. This becomes a crucial problem with LPM.

There are two ways for finding out whether $(0 \leq Y_i \leq 1)$ or not.

(I) Estimate LPM by usual OLS method and find Y_i. Judge about the values obtained for Y_i. If some values are negative, take them as zero. Similarly, if some values are greater than 1, take them as 1. (This is the usual practical approach for dealing with the problem).

(II) Some special methods are to be devised to check this point to ascertain that $(0 \leq Y_i \leq 1)$.

(Note:

(1) To overcome this difficulty, further studies for Logit and Probit models can be done.

(2) It is not proper to conclude about the correctness of the LPM model based upon the value of R^2. In most of the cases, R^2 lies

between 0.2 and 0.6. It is only when the scatter plot shows a cluster of points that appear very near in a line. In such cases R^2 will be in excess of 0.8.

Hence Use of R^2 as a summary statistic should be avoided with qualitative dependent variables.)

3. **LOGIT Model:-**

Now we consider further modification to LPM model so that the drawbacks or limitations of the model can be overcome by considering some new way of presentation. This is achieved by what is called LOGIT Model.

We shall continue our earlier illustration with the model.

$Y_i = \beta_1 + \beta_2 X_i + U_i$ . . . . (1) (i = 1, 2, ... n)

Where X = Family Income

Y_i = 1 If the family owns a house

= 0 Otherwise

As done in LPM, we assume E (U_i) = 0 for all i = 1, 2, . . . n

So that E (Y_i / X_i) = $\beta_1 + \beta_2 Xi$ . . . . (2)

We set Pi = Prob.that the family owns a house

$\therefore$ P_i = E (Y_i = 1/ X_i) = $\beta_1 + \beta_2 X_i$ . . . . (3)

Let us define $P_i = \frac{1}{1+\exp(-\beta_1-\beta_2 Xi)}$ . . . . (4)

If we write Z_i $= \beta_1 + \beta_2 X_i$ . . . . (5)

then $P_i = \frac{1}{1+\exp(-Zi)}$

Hence $P_i = \frac{e^{zi}}{1+e^{zi}}$ . . . . (6)

Equation (6) is called c.d.f. of logistic distribution (that is logistic distribution function).

Note that as Z_i ranges from $-\infty$ to $+\infty$, Pi ranges between 0 and 1, and Pi is non-linearly related to Z_i (and hence to Xi). This

partially solves the problem found in LPM. However we have created an estimation problem because Pi is non-linear not only in Xi but also in the Beta co-efficients. This means that we can not use OLS estimation procedure to estimate the parameters.

To solve this problem let us consider the following ratio.

$$\frac{\text{Prob. that family owns a house}}{\text{Prob. that family does not own a house}} = \frac{Pi}{1-P_i}$$

$$= \frac{1+e^{zi}}{1+e^{-zi}} = e^{zi} \qquad \ldots . (7)$$

This ratio is called <u>Odds ratio</u> (Ratio of prob. of success to prob. of failure)

(e.g. If Pi = 0.6, Odd ratio = 1.5, which means that odds are 3 to 2 in favour of the family owning a house)

Let us take logarithm to natural base for this ratio and denote it by L_i.

then $L_i = L_n \left(\frac{Pi}{1-Pi}\right) = Z_i = \beta_1 + \beta_2 X_i \qquad \ldots . (8)$

This last equation $L_i = \beta_1 + \beta_2 Xi$ shows that the log of the odds ratio is not only linear in X but it is also linear in parameters.

L is called LOGIT and hence we have what is called LOGIT Model given by

$$Li = Ln\left(\frac{P_i}{(1-P_i)}\right) = Zi = \beta_1 + \beta_2 X_i + U_i. \ldots . (9)$$

<u>Some features of LOGIT Model</u>

(1) As P goes from 0 to 1 (i.e. as Z varies from - ∞ to + ∞), the Logit L goes from - ∞ to + ∞. Thus even though probabilities lie between 0 and 1, the logits are not so bounded.

(2) LPM assumes that P_i is linearly related to X_i, the logit model assumes that log of odds ratio is linearly related to X_i.

(3) β_2 is the slope indicating change in L for a unit change in X. This means how is the log of odds ratio in the favour of owning a house as income changes by a unit.

β_1 is the intercept term for LOGIT model.

(4) L is linear in X but the probability by themselves are not.

(5) We have included a single explanatory independent Variable X, but we can add many such explanatory variables as regressors and odds ratio and hence L can be defined accordingly.

3. LOGIT Model -How to estimate Logit Model?

Case – I. Data at the individual level :-

If we have data on the individual families, we can not use equation (8) above due to the nature of the data.

(e.g. $Y_i = 1 \rightarrow P_i = 1 \rightarrow L_i = \text{Ln}\left(\frac{1}{0}\right)$, If family owns a house

$Yi = 0 \rightarrow P_i = 0 \rightarrow L_i = \text{Ln}\left(\frac{0}{1}\right)$, if family does not own a house.

Both of these are meaningless.

Here we should use Maximum Likelihood Estimator (MLE). This is achieved by using ML method. First we write likelihood equation, then Logarithm of likelihood equation Lin the parameters β_1 and β_2. Necessary condition is $\frac{\partial L}{\partial \beta_i} = 0$ and sufficient condition is $\frac{\partial^2 L}{\partial \beta_i^2} < 0$. (i = 1, 2)

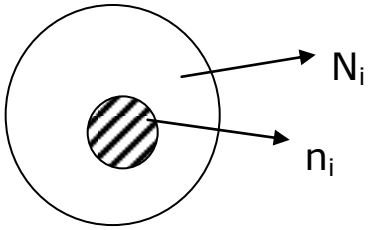
We get very complicated form and we can solve it by adopting usual methods for obtaining MLE. Some software packages like MICROFIT, EVIEWS, LIMDEP, SHZAM, PC-GIVE, STATA and

MINITAB have built in routines to estimate the logit model at the individual level. We can use them to solve the problem.

Case (II) Grouped or Replicated Data

There can be data for several families (for our problem stated earlier) which are grouped or replicated according to the number of families owning a house at each income level.

Thus corresponding to each income level X_i, there are Ni families and n_i among them are those who own their house.



Hence we can computer the relative frequency which measures the probability (using frequency definition of probability).

Hence for our above stated LOGIT model.

$$Ln = L_n\left(\frac{Pi}{1-Pi}\right) = Z_i = \beta_1 + \beta_2 X_i + U_i \qquad \ldots . (9)$$

We can have $P_i = \frac{n_i}{N_i}$ which estimates P_i.

Thus $$\hat{P}_i = p_i = \frac{n_i}{N_i} \qquad \ldots . (10)$$

Hence $$L_i = = L_n\left(\frac{\hat{P}i}{1-\hat{P}i}\right) = Z_i = \beta_1 + \beta_2 X_i + U_i \qquad \ldots . (11)$$

We have to run the regression based upon (11) and find estimators of β_1 and β_2.

Hence $$\hat{L}_i = L_n\left(\frac{\hat{P}i}{1-\hat{P}i}\right) = \hat{\beta}_1 + \hat{\beta}_2 Xi \qquad \ldots . (12)$$

Where $\hat{\beta}_1$ and $\hat{\beta}_2$ are the estimates of β_1 and β_2.

Can we run OLS regression? Answer is no. Why? The reason is that there is heteroscadasticity of disturbances. Let us examine this first before running regression.

Under normality assumptions, if N_i is fairly large, $U_i \sim N$ $\left[0, \frac{1}{N_i P_i (1-P_i)}\right]$

Thus U_i follows normal distribution with mean zero and variance V (U_i) = $\left[\frac{1}{N_i P_i (1-P_i)}\right]$

This shows that there is heteroscadasticity.

We can write $\hat{\sigma}^2 = \left[\frac{1}{N_i \hat{P}_i (1-\hat{P}_i)}\right] = \frac{1}{\sqrt{Wi}}$

Where $W_i = N_i \hat{P}_i (1-\hat{P}_i) = N_i p_i (1-p_i)$

We can divide Logit model by $\frac{1}{\sqrt{W_i}}$ so that

$\sqrt{W_i} L_i = \beta_1 \sqrt{W_i} + \beta_2 (\sqrt{W_i}) X_i + (\sqrt{W_i}) U_i$

Thus $L_i^* = \beta_1^* + \beta_2^* X_i + v_i^*$

Where $L_i^* = \sqrt{W_i} L_i$, $\beta_1^* = \beta_1 \sqrt{W_i}$

$\beta_2^* = \beta_2 (\sqrt{W_i})$, $v_i^* = \sqrt{W_i} . U_i$

Check that E (v_i^*) = 0, V (v_i^*) = σ^2

Hence the transformed model has homoscadasticity even though originally there is heteroscadasticity. We follow the estimation procedure as under

(1) For each income level X_i, computer the prob. of owning a house as $p_i = \hat{P}_i = \frac{n_i}{N_i}$.

(2) For each X_i, obtain the Logit as $\hat{L}_i = \text{Ln} \left[\frac{\hat{P}_i}{(1-\hat{P}_i)}\right]$

(3) Obtain $W_i = Ni\ \hat{P}_i\ (1 - \hat{P}_i)$

(4) Find $L^*_i = \sqrt{Wi}.Li$, $\beta_1^* = \beta_1\sqrt{Wi}$

$\beta_2^* = \beta_2\sqrt{Wi}$

(5) Run the regression for the transformed data

$L^*_i = \beta^*_1 + \beta_2^* Xi + U^*_i$ and obtain OLSE of β^*_1 and β^*_2.

(Note that (i) this procedure is valid if the sample is of reasonably large size.)

(ii) we can obtain confidence intervals and do testing problems as usual.

(iii) We can get inference about odds ratio for interpretation purpose.)

4. Probit Model

In LOGIT model we have used cumulative logistic function. This is not the only c.d.f. that one can use. In some applications, the normal c.d.f. is found to be useful. The estimating model that emerges from the normal c.d.f. is popularly known as PROBIT Model. It is also known as NORMIT Model.

To understand about this model, we consider further our house ownership example whether i^{th} family own a house or not, depends upon utility index Ii (also called Latent variable).

Thus larger the value of I_i greater is the prob. of owning a house. We express this index as

$I_i = \beta_1 + \beta_2 X_i$ . . . . (1)

Where X_i = Income of the i^{th} family. As before Y = 1, if the family owns a house and Y = 0 otherwise. We assume that there is a threshold level of the index denoted by I_i^* such that if $I_i > I_i^*$, the family will own a house, otherwise it will not. I_i as well as I_i^* are not directly observable but with the help of normal distribution assumption we can deal with the problem.

Under normality assumption;

$P_i = P(Y_i = 1 / X) = P(I_i^* \leq I_i)$

$= P(Z_i \leq \beta_1 + \beta_2 X_i)$

$= F(\beta_1 + \beta_2 X_i)$ . . . . (2)

Z_i is standard normal variate, and F_i is its c.d.f.

$= F(I_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{I_i} \exp(\frac{-z^2}{2}) dz$

$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\beta_1+\beta_2 X_i} \exp(\frac{-z^2}{2}) dz$ . . . . (3)

To obtain information on I_i, the utility index as well as on β_1 and β_2, we take inverse of relation given in (2)

Thus $I_i = F^{-1}(I_i) = F^{-1}(P_i) = \beta_1 + \beta_2 X_i$ . . . (4)

Here F^{-1} is the inverse of the normal p.d.f.

We can deal with this problem by using tables of normal distribution c.d.f. once we obtain I_i, then for given X_i, we can run the regression of equation (4) using OLS method to estimate Beta Co-efficients.

This model seems to be mathematically difficult. There are sophisticated statistical packages which are available to solve the problem.

It may be noted that both Logit and Probit models give similar results, and also both are equally popular for application purposes.

(Note : (1) An extension of PROBIT Model is TOBIT Model which is developed by James Tobin. It uses censored data. Hence it is also called Limited Dependent Variable Regression Model.

(2) Further studies on these lines are very interesting – e.g. Poisson Regression Model, Ordinal Logit and Probit models, Multinomial Logit and Probit models, Duration models etc.)