Chapter 7

QUALITATIVE REGRESSION MODEL (LOGIT, PROBIT, TOBIT)

Shivaswamy G. P., K. N. Singh and Anuja A. R.

INTRODUCTION

Regression analysis is a statistical method of studying functional relationship between a dependent or response variable and one or more independent or explanatory variables. In classical linear regression models (CLRM), response variable is implicitly assumed as quantitative and explanatory variables can be quantitative or qualitative. Parameter estimation in the CLRM is based on important assumptions such as linearity of model in parameters, though response and explanatory variables may or may not be linearly related; explanatory variables are independent of the error terms; independent and identically distributed error terms with zero mean and constant variance; and equal reliability of observations. However, when the response variable is qualitative, these basic assumptions of CLRM may not hold. For instance, one wants to study the adoption of high yielding varieties of a crop which is a binary response variable. It takes only two values: 1 if the variety is adopted and 0 if it is not. There are several instances, where the response variable is binary. Suppose one wants to study the determinants of access to institutional credit such as age, gender, education, social status, land holding etc. Whether a farmer has access to institutional credit is a binary variable taking values 0 or 1, 0 meaning no access and 1 meaning access to institutional credit. Similarly, other examples can be crop diversification status as a function of various quantitative or qualitative independent variables. In such cases, error terms are not normality distributed and the variance is not constant thereby violating the homoscedasticity assumption. The statistical models preferred for the analysis of such a binary response are logit and probit models as these models do not make assumptions on the distribution of explanatory variables.

LOGISTIC REGRESSION

Logistic regression analysis is used when the dependent variable is qualitative and normality assumption is not satisfied. Cox (1958) developed this model. Logistic regression is appropriate when the dependent and independent variables are non-linearly related. Logit is transformation of logistic regression to make it linear.

In case of logit transformation, binary variable (Koutsoyiannis, 2001) of adoption of a rice variety, the decision to adopt or not to adopt by i^{th} individual depends on the latent

variable Y_i^* which in turn depends on explanatory variables such as age, education, farm size, access to institutional credit, irrigation facility and training.

$$Y_i^* = BX + u_i \ldots \quad (1)$$

where, B vector of parameters, X, vector of explanatory variables; and μ_i error term of i[th] individual

It is assumed that $Y_i = \begin{cases} 0 \; if \; Y_i^* \leq 0 \\ 1 \; if \; Y_i^* > 0 \end{cases}$

That is if individual's utility index exceeds Y_i^* farmer will adopt a variety, but if it is less than Y_i^*then variety is not adopted.

In case of probability of adoption of variety (Y=1), the logistic regression function can be expressed as

$$P_i = Pr(Y_i = 1 | X = x) = \frac{1}{(1 + e^{-Z_i})} \ldots \quad (2)$$

where, $Z_i = BX + u_i$

The probability that variety is not adopted (Y=0) is given by

$$(1 - P_i) = \frac{1}{(1 + e^{Z_i})}$$

where, as Z_i ranges from - ∞ to + ∞, P_i ranges between 0 and 1. And, the model is nonlinear both in response variables X and parameters Bs.

Further, to make the logistic regression function linear in the parameters, we take the ratio of probability that farmer adopts a variety to probability that he is not;

$$\frac{P_i}{1 - P_i} = \frac{\frac{1}{(1 + e^{-Z_i})}}{\frac{1}{(1 + e^{Z_i})}}$$

$$\frac{P_i}{1 - P_i} = e^{Z_i} \quad \ldots \quad (3)$$

where, $P_i/(1 - P_i)$ is known as the odds ratio in favor of adoption of a variety i.e. the ratio of probability that a farmer adopts a variety to probability that he does not adopt.

Equation can be transformed by taking natural logarithm as follows

$$\ln\left(\frac{P_i}{1 - P_i}\right) = Z_i \ldots \quad (4)$$

Log of the odds ratio is known as the logit which is nothing but a linear transformation of the logistic regression model.

Estimation of logit model

Usual OLS method cannot be used to estimate the logit model despite its linearity properties due to problem of undefined expressions. Rather, Maximum likelihood estimation method is used for estimation.

Variables used for the logit analysis of determinants of variety adoption example are as follows

Adoption=1 for adopters and 0 for non-adopters

Age in years

Education=1 if educated; 0 otherwise Credit=1 if there is access to institutional credit; 0 other wise

Irrigation=1 if there is access to irrigation; 0 for non-access

Training=1 if undergone training; 0 otherwise

Table 1: Sample data set for logit and probit analysis

Observations	Adoption	Age	Education	Farm size	Credit	Irrigation	Training
1	0	70	1	9.01	0	0	0
2	0	30	0	2.136	1	1	0
3	1	40	1	1.12	1	1	0
4	1	60	0	1.003	0	1	0
5	0	30	1	2.61	1	0	1
6	0	60	1	1.23	0	1	1
7	1	45	1	2.434	1	0	0
.							
.							
.							
1763	1	52	1	1.705	1	0	0

Table 2 provides the results of logit model for the adoption of variety example, which are obtained by *STATA* using the command *logit.*

Table 2: Logit estimates of adoption of a rice variety

Particulars	Coefficient	Standard error	Z statistic	Prob>Z
Age	0.002	0.003	0.06	0.951
Education	-0.070	0.102	-0.07	0.945
Farm size	-0.014	0.015	-0.92	0.36

Particulars	Coefficient	Standard error	Z statistic	Prob>Z
Institutional credit	0.489	0.098	4.98	0
Irrigation access	0.299	0.097	3.07	0.002
Training	0.096	0.285	0.34	0.735
Constant	-0.351	0.213	-1.64	0.1
Number of observations	1763			
McFadden R^2	0.014			

The results of logit model show that access to institutional credit and irrigation are statistically significant at 1 percent level of significance. It is interpreted as access to institutional credit increases the average logit value by 0.489. Access to irrigation is also interpreted similarly. Other variables such as age, education and training are statistically insignificant meaning they do not have visible impact on adoption of a variety.

In case of CLRM, R^2 indicates the goodness of fit showing the proportion of variation in the dependent variable explained by the independent variables in the model. But, in case of binary regression models, R^2 is not meaningful for which McFadden R^2 or pseudo R^2 is discussed in the literature. The value of McFadden R^2 ranges between 0 and 1. In our example its value is 0.014. It should be noted that in qualitative regression models, the expected sign of the regression coefficients and their statistical significance are more important than the goodness of fit measures.

We can express the logit coefficients in terms of odds ratio (Table 3) by using following *STATA* command *logit adoption age education farmsize credit irrigation access training.*

Table 3: Odds ratio for adoption versus non-adoption

	Odds ratio	Standard error	Z statistic	Prob>Z
Age	1.000	0.004	0.060	0.951
Education	0.993	0.102	-0.070	0.945
Farmsize	0.986	0.015	-0.920	0.360
Institutional credit	1.632	0.161	4.980	0.000
Irrigation	1.349	0.131	3.070	0.002
Training	1.101	0.315	0.340	0.735
Constant	0.704	0.151	-1.640	0.100
Number of observations	1763			
McFadden R^2	0.014			

The odds ratios are obtained by taking the exponential of logit coefficients given in Table 2. The interpretation of the odds ratio depends on whether its value is greater than 1 or less than 1. Odds ratios of greater than 1 indicate the increased chance of adoption as compared to non-adoption. On the other hand, odds ratio of less than 1 indicates the decreased chance of adoption. Odds ratio of 1 suggests that chances of adopting and not adopting are even. In our example, two variables institutional credit and irrigation have the odds ratios of greater than 1 meaning increased chance of adopting a variety as against non-adoption.

Estimation of marginal effects

Marginal effects depict the marginal impact of one unit change in the explanatory variable on the probability of adoption of a variety. It is a way depicting the model estimates in terms of probabilities which helps in interpreting in terms of magnitude. Note that instead of computing the marginal effect for each independent variable on the probability of adoption, it is computed for the average values of variables. It is to be noted that for quantitative response variables marginal effect is the derivative $(\frac{dy}{dx})$ of dependent variable (y) with respect of independent variable (x) that is rate of change of y with respect to x. However for qualitative independent variable which takes the discrete values 0 and 1 as in our example, marginal effect is estimated for the discrete change in the qualitative variable from 0 to 1.

Table 4: Marginal effects of logit model

Particulars	Marginal effects	Standard error	Z statistic	Prob>Z	Mean value
Age	0	0.001	0.060	0.951	50.558
Education	-0.001	0.026	-0.070	0.945	0.647
Farmsize	-0.003	0.004	-0.920	0.360	2.771
Institutional credit	0.121	0.024	5.030	0.000	0.450
Irrigation	0.074	0.024	3.080	0.002	0.573
Training	0.024	0.071	0.340	0.735	0.029
Number of observations	1763				
McFadden R^2	0.014				

The interpretation of marginal effects in our example is that unit change in age and farm size does not have statistically significant impact on the rate of change in the probability of adoption. For qualitative variable, an access to institutional credit has significant positive impact in the probability of adoption of variety by about 0.121. Similarly access to irrigation increases the probability of adoption by about 0.074.

PROBIT MODEL

Like logit, probit model is used when the response variable is qualitative. Error terms in the probit model follow normal distribution.

For arriving at the probit model, equation 1 can be translated into,

$$Pr(Y_i = 1|X = x) = Y_i^* > 0$$

$$= Pr(u_i > -B'X)$$

$$= Pr\left(\frac{u_i}{\sigma} > \frac{-B'X}{\sigma}\right) \quad \ldots \quad (5)$$

$$Pr(Y_i = 1|X = x) = \Phi\left(\frac{-B'X}{\sigma}\right)$$

Estimation of probit model

Probit model is estimated based on the maximum likelihood function which finds coefficients that maximize the probability of $Y_i = 1$.

Using *STATA* command *probit,* ML estimates of the probit model for adoption of variety are given in Table 5 (Spermann, 2008).

Table 5: Probit estimates of adoption of variety

	Coefficient	Standard error	Z statistic	Prob>Z
Age	0.000	0.002	0.070	0.943
Education	-0.004	0.064	-0.060	0.951
Farm size	-0.009	0.010	-0.930	0.354
Institutional credit	0.306	0.061	4.990	0.000
Irrigation	0.187	0.061	3.070	0.002
Training	0.061	0.178	0.340	0.730
Constant	-0.221	0.134	-1.650	0.098
Number of observations	1763			
McFadden R^2	0.014			

Although coefficients of logit and probit models are different, the interpretation of coefficients is similar. Institutional credit and irrigation are statistically significant at 1 percent level of significance. It should be noted that only sign of the logit and probit models are interpreted but not the magnitude. The coefficients of logit and probit models are different and can be comparable after multiplying probit coefficients by about 1.81. However, marginal effects of probit and logit models are similar (Table

6). Logit and probit functions are almost similar with both s shaped curves. The main difference between the logit and probit models is that the logistic distribution has slightly flatter tails. Therefore, there is no compelling reason for choosing one model over another (Halloran, 2018).

Table 6: Marginal effects of probit model

	Marginal effects	Standard error	Z statistic	Prob>Z	Mean value
Age	0	0.001	0.060	0.951	50.558
Education	-0.001	0.026	-0.070	0.945	0.647
Farmsize	-0.003	0.004	-0.920	0.360	2.771
Institutional credit	0.121	0.024	5.030	0.000	0.450
Irrigation	0.074	0.024	3.080	0.002	0.573
Training	0.024	0.071	0.340	0.735	0.029
Number of observations	1763				
McFadden R^2	0.014				

TOBIT MODEL

Tobit model, also called as censored regression model or limited dependent variable regression model, was proposed by Tobin in 1958. A censored sample is a sample in which information on dependent variable is available for only some observations in a sample. If we use OLS on censured data set, estimates obtained will be inconsistent meaning coefficients will not necessarily approach the true population parameters as sample size increases (Gujarati, 2003). In such cases, Tobit model is used for analyzing censored sample.

The model can be expressed as

$$Y = X\beta + u \quad \text{If } \beta'X + u > 0;$$

$$= 0 \text{ Otherwise}$$

Such that the residual, $u \sim N(0, \sigma^2)$

where *Y*, (nx1) vector of dependent variable; β (kX1) vector of unknown parameters; and X vector of exogenous variables.

The model can be estimated using maximum likelihood method or Heckman two step procedure.

The application of the model can be explained with the help of labour economics example. Using the data set which contains information on both working and non-

working married women, suppose we want to estimate the extent of participation of working women in labour force. Here the dependent variable (extent of participation in hours per year) is continuous and lower limit of dependent variable is zero which means non-participation in labour market. Tobit model can be applied here as it employs all information collected for both working women (where dependent variable is more than zero) and non-working women (dependent variable -zero).

In another example, suppose we want to estimate the amount of money spent by an individual on meat in relation to socio economic variables. Now there are two groups of consumers. One set $m_{1,}$ about whom we have information on the independent variables (age, education, income etc.) as well as the dependent variable (the amount of money spent on meat items) and another set m2 about whom we have information only on the independent variables but not on dependent variable. Here we cannot neglect observations on second group as the OLS estimates of parameters using only first group of observations will be biased and inconsistent. In this case we can use Tobit model for a better estimation of parameters.

REFERENCES

Cox, D. R. (1958), The regression analysis of binary sequences (with discussion). *Journal of the Royal Statistical Society,* B, 20: 215-242

Gujarati, D. N. (2003), Basic Econometrics. 4th Edition, New York: Mc Graw Hill Publications

Halloran, S. (2018), Logit and probit models, accessed in September 2019 http://www.columbia.edu/~so33/SusDev/Lecture_9.pdf

Koutsoyiannis, A. (2001), Theory of Econometrics, (2nd ed.), New York: Palgrave Macmillan Limited

Spermann, A. (2009), The probit model, accessed in September 2019 https://www.empiwifo.uni-freiburg.de/lehre-teaching-1/summer-term-09/materials-microeconometrics/probit_7-5 09.pdf.