Econometrics Lecture 5: Limited Dependent Variable Models: Logit and Probit

R. G. Pierse

1 Introduction

In lecture 5 of last semester's course, we looked at the reasons for including dichotomous variables as explanatory variables in the regression model. Such variables can take only a limited number of possible values. They are often used as proxies for effects that cannot be quantified and are known as *dummy variables.* In the binary case only two values are possible and these can be represented by the numbers zero and one.

In this lecture we look at models where the dependent variable $\mathbf{y}$ is itself a dichotomous variable. Such models are called *limited dependent variable models*, or also *qualitative* or *catagorical* variable models. We concentrate on the binary case where y_i can take only two values. One example would be a model of success in job interviews based on observations on interview candidates. The dependent variable in this case would take the value one if the candidate was offered a job and zero if not. Various explanatory variables could be included: both continuous variables such as age and dichotomous variables such as gender or educational achievement. Other examples might include models of bank failure, or mortgage applications. In the context of survey data it is very often the case that a variable such as the purchase of a washing machine or car which might be considered as a continuous variable in the aggregate, are dichotomous when the observations are of individuals over a short time span.

2 The Linear Probability Model

The linear probability model applies the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{2.1}$$

to the case where $\mathbf{y}$ is a dichotomous variable taking the value zero or one, where we require the assumption that

$$E(\mathbf{u}) = \mathbf{0}.$$

The conditional expectation of the ith observation is given by

$$E(\mathbf{y}_i|\mathbf{x}_i) = \mathbf{x}_i'\boldsymbol{\beta}$$

where $\mathbf{x}_i$ is the $k \times 1$ vector of values of the k regressors for the ith observation, corresponding to the transpose of the ith row of matrix $\mathbf{X}$. The conditional expectation $E(\mathbf{y}_i|\mathbf{x}_i)$ in this model has to be interpreted as the probability that $y_i = 1$ given the particular value of $\mathbf{x}_i$. In practice, however, there is nothing in this model to ensure that these probabilities will lie in the admissible range $(0, 1)$.

Since y_i can only take the two values of 0 or 1, it follows that the error term u_i can only take the two values of $-\mathbf{x}_i'\boldsymbol{\beta}$ or $(1 - \mathbf{x}_i'\boldsymbol{\beta})$. From the assumption that $E(u_i) = 0$ it follows that the probabilities of these two events are given by $(1 - \mathbf{x}_i'\boldsymbol{\beta})$ and $\mathbf{x}_i'\boldsymbol{\beta}$ respectively. Thus the probability distribution of u_i, $f(u_i)$ can be represented by the table

u_i	$\Pr(u_i)$
$-\mathbf{x}_i'\boldsymbol{\beta}$	$(1 - \mathbf{x}_i'\boldsymbol{\beta})$.
$(1 - \mathbf{x}_i'\boldsymbol{\beta})$	$\mathbf{x}_i'\boldsymbol{\beta}$

It is obvious that u_i is not normally distributed. Its distribution has zero mean but *not* constant variance since

$$\begin{aligned} \operatorname{var}(u_i) &= \sum u_i^2 f(u_i) \qquad (2.2) \\ &= (-\mathbf{x}_i'\boldsymbol{\beta})^2(1 - \mathbf{x}_i'\boldsymbol{\beta}) + (1 - \mathbf{x}_i'\boldsymbol{\beta})^2\mathbf{x}_i'\boldsymbol{\beta} \\ &= \mathbf{x}_i'\boldsymbol{\beta}(1 - \mathbf{x}_i'\boldsymbol{\beta}) \end{aligned}$$

Clearly, since this variance depends on $\mathbf{x}_i$, u_i is *heteroscedastic* so that *OLS* estimation of (2.1) will not be efficient.

It is possible to devise a *feasible GLS* procedure to correct for the heteroscedastic form of (2.2). Denoting the fitted values from the *OLS* regression of (2.1) by

$$\widehat{y}_i = \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$$

then

$$\widehat{\sigma}_i^2 = \widehat{y}_i(1 - \widehat{y}_i)$$

is a consistent estimate of $\operatorname{var}(u_i)$ so that the weighted least squares regression

$$\mathbf{Ly} = \mathbf{LX}\boldsymbol{\beta} + \mathbf{Lu}$$

where

$$\mathbf{L} = \begin{bmatrix} \frac{1}{\widehat{\sigma}_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \frac{1}{\widehat{\sigma}_n} \end{bmatrix}.$$

will give *feasible GLS* estimates. However, it is theoretically possible for $\mathbf{x}_i'\widehat{\boldsymbol{\beta}}$ to take values outside the interval $(0, 1)$. In this case the estimator $\widehat{\sigma}_i^2 < 0$ so that the *GLS* procedure will fail. This illustrates the problem that the linear probability model is not really an appropriate way to estimate limited dependent variables models since it does not impose the fundamental property of the model which is that

$$E(y_i|\mathbf{x}_i) = P(y_i) \in (0, 1).$$

3 The Probit and Logit Models

Consider a different approach. Suppose that we have a regression model

$$\mathbf{y}^* = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{3.1}$$

where $\mathbf{y}^*$ is *unobserved* and is called a *latent variable.* This is related to the observed dichotomous variable $\mathbf{y}$ by

$$y_i = \begin{cases} 1, & \text{if } y_i^* > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{3.2}$$

In this model there is no reason why the error term $\mathbf{u}$ and hence the latent variable $\mathbf{y}^*$ should not be expected to have a continuous distribution. Note that the observation equation (3.2) is independent of the scale of the latent variable $\mathbf{y}^*$ in (3.1). This means that the scaling of the error process $\mathbf{u}$ can be chosen for convenience and we will assume that u_i has unit variance

$$\text{var}(u_i) = 1.$$

It follows from (3.1) and (3.2) that

$$\begin{aligned} P_i &= \Pr(y_i = 1) = \Pr(y_i^* > 0) \\ &= \Pr(u_i > -\mathbf{x}_i'\boldsymbol{\beta}) \\ &= 1 - F(-\mathbf{x}_i'\boldsymbol{\beta}) \end{aligned}$$

where $F()$ is the *cumulative distribution function* of u_i defined by

$$F(a) = \Pr(u_i \leq a) = \int_{-\infty}^{a} f(u_i)du_i.$$

If the distribution of u_i is *symmetric* then $F(a) = 1 - F(-a)$ so that

$$P_i = \Pr(y_i = 1) = F(\mathbf{x}_i'\boldsymbol{\beta}). \tag{3.3}$$

The dichotomous observations y_i will follow a *binomial distribution* with likelihood function given by

$$L(\mathbf{y}) = \prod_{y_i=1} P_i \prod_{y_i=0} (1 - P_i). \tag{3.4}$$

The exact form of P_i in (3.3) will depend on the assumption made about the distribution of the error term $\mathbf{u}$. If it is assumed that u_i follows a standard normal distribution with unit variance

$$f(u_i) = (2\pi)^{-\frac{1}{2}} \exp(-\frac{1}{2}u_i^2)$$

then

$$P_i = F(\mathbf{x}_i'\boldsymbol{\beta}) = \int_{-\infty}^{\mathbf{x}_i'\boldsymbol{\beta}} f(u_i)du_i. \tag{3.5}$$

This is known as the *probit model.*

Alternatively, if u_i is assumed to follow the *logistic distribution* defined by

$$f(u_i) = \frac{\exp(u_i)}{(1 + \exp(u_i))^2} \tag{3.6}$$

then

$$P_i = F(\mathbf{x}_i'\boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i'\boldsymbol{\beta})} = \frac{1}{1 + \exp(-\mathbf{x}_i'\boldsymbol{\beta})}. \tag{3.7}$$

The function (3.7) is known as the *logit function* and hence the model using assumption (3.6) is known as the *logit model.* The logistic distribution (3.6) has variance $\pi^2/3$ so that the estimates of $\boldsymbol{\beta}$ obtained from the logit model are conventionally rescaled by the factor $\sqrt{3}/\pi$ so as to conform with the assumption that $\text{var}(u_i) = 1$. Other scalings are possible. Amemiya (1981) recommends using the factor $1/1.6 = 0.625$ as providing a closer approximation of the logistic and standard normal distribution.

Note for the logit model that

$$\log \frac{P_i}{1 - P_i} = \mathbf{x}_i'\boldsymbol{\beta}.$$

This is known as the *log-odds ratio* and it can be seen that for this model it is linear in the variables x_i.

The cumulative normal and logistic distributions (3.5) and (3.7) are very close to each other except at the tails, so that both assumptions will give very similar results except when there are a large number of observations (and hence a lot of observations in the tails). The logit model has the advantage that its cumulative distribution function has an explicit form, whereas the integral in (3.5) must be evaluated numerically. Nevertheless, in either case, the parameter vector $\boldsymbol{\beta}$ can be obtained in a straightforward manner by numerically maximising the likelihood function (3.4) or its logarithm

$$\log L(\boldsymbol{\beta};\mathbf{y},\mathbf{X}) = \sum_{y_i=1} \log P_i + \sum_{y_i=0} \log(1-P_i). \tag{3.8}$$

For the case of the logit model

$$\frac{\partial \log P_i}{\partial \boldsymbol{\beta}} = x_i - \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta})} = (1-P_i)\mathbf{x}_i$$

and

$$\frac{\partial \log(1-P_i)}{\partial \boldsymbol{\beta}} = \frac{-\exp(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta})} = -P_i\mathbf{x}_i$$

so that the first order condition for maximising the log-likelihood function (3.8) is

$$\begin{aligned}\frac{\partial \log L}{\partial \boldsymbol{\beta}} &= \sum_{y_i=1}(1-P_i)\mathbf{x}_i - \sum_{y_i=0} P_i\mathbf{x}_i \\ &= \sum_{y_i=1}\mathbf{x}_i - \sum_{i=1}^{n} P_i\mathbf{x}_i = \mathbf{0}.\end{aligned}$$

These equations are nonlinear in $\boldsymbol{\beta}$ so that an iterative solution technique is required.

The probit and logit models can be generalised to the case of dependent variables that can take on more than two values.

3.1 Interpreting the Coefficients

Having estimated the parameters in the Logit or Probit model, we need to understand how to interpret the coefficients. For the Logit model we have

$$\log \frac{P_i}{1-P_i} = r_i = \mathbf{x}_i'\boldsymbol{\beta}$$

where r_i represents the *log-odds ratio*, so that the jth coefficient

$$\beta_j = \frac{\partial r_i}{\partial x_{ij}}$$

represents the effect of a change in the jth variable on the *log-odds ratio*. This effect is constant. The effect of a change in the jth variable on the probability P_i is given by

$$\frac{\partial P_i}{\partial x_{ij}} = \beta_j P_i(1 - P_i) \ .$$

Note that this is different for different observations $\mathbf{x}_i$.

For the Probit model we have

$$\frac{\partial P_i}{\partial x_{ij}} = \beta_j f(\mathbf{x}_i' \boldsymbol{\beta})$$

where f is the standard normal density function, which again depends on the observations. The effect of a change in the jth variable on P_i can be evaluated for different values of $\mathbf{x}_i$ or at the sample means $\overline{\mathbf{x}}$ or $\overline{P} = \overline{y}$.

3.2 Goodness of fit measures

The conventional R^2 measure of goodness of fit is problematic in limited dependent variable models where the predicted values are probabilities and the actual values are either 0 or 1. Different ways of expressing R^2 that are equivalent in the linear regression model are no longer equivalent. Several alternative goodness of fit measures have been proposed.

The conventional R^2 measure can be written as

$$R^2 = 1 - \frac{\sum (y_i - \widehat{y}_i)^2}{\sum (y_i - \overline{y}_i)^2} \ .$$

In a binary limited dependent variable model the denominator can be written as

$$\sum_{i=1}^{n} (y_i - \overline{y}_i)^2 = \sum y_i^2 - n\overline{y}^2 = n_1 - n(\frac{n_1}{n})^2 = \frac{n_1 n_2}{n}$$

where n_1 is the number of successes ($y_i = 1$) and n_2 is the number of failures ($y_i = 0$) in the sample.

The R^2 measure of *Efron* (1978) is given by

$$R^2 = 1 - \frac{n}{n_1 n_2} \sum (y_i - \widehat{y}_i)^2 \ .$$

In the standard linear regression model

$$R^2 = 1 - \left(\frac{L_r}{L_u} \right)^{2/n}$$

where L_u is the maximum of the unrestricted likelihood function and L_r is the maximum of the likelihood function where all coefficients except for the intercept have been restricted to zero. This can be used to derive an R^2 measure for the limited dependent variables model. However, in this model $L_r \leq L_u \leq 1$ so that

$$R^2 \leq 1 - L_r^{2/n}$$

and the measure needs to be rescaled to lie in the interval $[0,1]$. *Cragg and Uhler* (1970) propose the *pseudo-R^2* measure

$$\widetilde{R}^2 = \frac{L_u^{2/n} - L_r^{2/n}}{(1 - L_r^{2/n})L_u^{2/n}}.$$

Finally, *McFadden* (1974) proposes another measure based on the likelihood ratio index

$$R^2 = 1 - \frac{\log L_u}{\log L_r}.$$

4 Multi-response models

The limited dependent variable model can be extended to the situation where there are more than two possible values of the dependent variable. Two cases need to be distinguished. In the first case, the values have a natural ordering, for example owning no car, one car, or two or more cars. This gives rise to the *ordered logit* or *ordered probit* model. In the other case, there is no natural ordering of the values. An example would be the choice of country to invest in, where the possible values might be Europe, Asia, or USA. This case gives rise to the *multinomial logit* model.

4.1 Ordered response models

In the ordered response model with m categories, the regression equation

$$\mathbf{y}^* = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{4.1}$$

with *unobserved* $\mathbf{y}^*$, is related to the observed variable $\mathbf{y}$ by the conditions

$$y_i = j \quad \text{if } \gamma_{j-1} < y_i^* \leq \gamma_j, \quad j = 1, \cdots, m, \tag{4.2}$$

where the γ_j are unknown parameters, with $\gamma_0 = -\infty$ and $\gamma_m = \infty$. Standard normalisation restrictions, needed to pin down the scale, are that $\gamma_1 = 0$ and $E(u_i^2) = 1$. Assuming that u_i is *i.i.d.* standard normal gives the *ordered probit* model, whereas assuming that u_i follows the logit distribution gives rise to the *ordered logit* model.

4.2 Multinomial models

When the responses are unordered, then there is no obvious way to relate the underlying latent variable y_i^* to the observed outcome y_i. One way to impose a structure is to assume the existence of a utility function U_{ij}, giving the utility that individual i associates with alternative j. The utility function is assumed to be stochastic and is a linear function of a set of observable variables that may depend on the individual i and/or the alternative j. We can write this as

$$U_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \varepsilon_{ij}$$

where $\mathbf{x}_{ij}$ is a $k \times 1$ vector of observations on the k observable variables and ε_{ij} is a disturbance term. The probability that individual i chooses alternative j, is then given by

$$P(y_i = j) = P(U_{ij} = \max\{U_{i1}, \cdots, U_{im}\}).$$

Assuming that the disturbances ε_{ij} are independent, with a *log Weibull* distribution, it can be shown that

$$P(y_i = j) = \frac{\exp(\mathbf{x}_{ij}'\boldsymbol{\beta})}{\exp(\mathbf{x}_{i1}'\boldsymbol{\beta}) + \exp(\mathbf{x}_{i2}'\boldsymbol{\beta}) + \cdots + \exp(\mathbf{x}_{im}'\boldsymbol{\beta})} \tag{4.3}$$

which is the *multinomial logit* model. To normalise the model, it is normally assumed that $\mathbf{x}_{i1} = \mathbf{0}$. When $m = 2$, this model reduces to the standard binary logit model.

5 Censored Variables: The Tobit Model

The techniques of the logit and probit model can also be applied to a different problem: that of a variable which, though continuous, is *censored* so that it is never observed when it falls below a certain value, taken without loss of generality to be zero. It is important to note that non-positive values are supposed to exist, but are simply not observed. The observations of the explanatory variables corresponding to these unobserved negative values of the dependent variable are, however, available and form part of the estimation data set.

This model can be represented by the *latent variable* model

$$\mathbf{y}^* = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

where $\mathbf{u}$ is normally distributed with

$$\mathbf{u} \sim \mathbf{N}(\mathbf{0},\sigma^2\mathbf{I}).$$

The variable $\mathbf{y}^*$ is *unobserved* and is related to the observed variable $\mathbf{y}$ by

$$y_i = \begin{cases} \mathbf{y}_i^*, & \text{if } y_i^* > 0 \\ 0, & \text{if } y_i^* \leq 0. \end{cases} \tag{5.1}$$

This model was first analysed by Tobin (1958) and is known as the *Tobit* (Tobin's probit) model. It is also known as the *censored regression model* because some observations are not observed.

The application originally used by Tobin was that of expenditure on automobiles in the context of sample survey data of household expenditure. In a given year, some households will have bought a car so that an expenditure is observed. Other households will not have bought a car that year so that their expenditure is zero. Negative expenditure of course is never observed. On the other hand, it is not clear that this example is really a *censored* variables problem since negative expenditure is not merely *unobservable* but cannot exist because it is *logically impossible*. Despite this, there are less controversial examples of censored regression models such as measuring average earnings where only the earnings of the employed are available.

We can write down the likelihood function for the censored regression model. For the positive observations,

$$f(y_i) \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$$

so that

$$f((y_i - \mathbf{x}_i'\boldsymbol{\beta})/\sigma)$$

is standardised normal, while for the non-positive observations

$$u_i \leq -\mathbf{x}_i'\boldsymbol{\beta}$$

with

$$\begin{aligned} P(u_i &\leq -\mathbf{x}_i'\boldsymbol{\beta}) = P(u_i/\sigma \leq -\mathbf{x}_i'\boldsymbol{\beta}/\sigma) \\ &= F(-\mathbf{x}_i'\boldsymbol{\beta}/\sigma) \end{aligned}$$

where $F()$ is the cumulative distribution function of a standardised normal density.

Therefore the likelihood function can be written as

$$L(\mathbf{y}) = \prod_{y_i>0} \frac{1}{\sigma} f((y_i - \mathbf{x}_i'\boldsymbol{\beta})/\sigma) \prod_{y_i \leq 0} F(-\mathbf{x}_i'\boldsymbol{\beta}/\sigma).$$

This likelihood function has to be maximised numerically for the *ML* estimators of $\boldsymbol{\beta}$ and σ^2.

6 Truncated Variables Model

The truncated regression model deals with the case where we have no data for y_i^* or for the explanatory variables $\mathbf{x}_i$ when y_i^* is below a certain value, known as the *truncation point.* These observations are simply not sampled. *OLS* estimates will be biased.

The untruncated regression model is

$$\mathbf{y}^* = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

where $\mathbf{u}$ is normally distributed with

$$\mathbf{u} \sim \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I}).$$

However, only observations with $\mathbf{y}_i^* \geq \overline{y}$ are observed.

The area of the density function that is truncated is given by

$$\begin{aligned} P(u_i &< \overline{y} - \mathbf{x}_i'\boldsymbol{\beta}) = P(u_i/\sigma < (\overline{y} - \mathbf{x}_i'\boldsymbol{\beta})/\sigma) \\ &= F((\overline{y} - \mathbf{x}_i'\boldsymbol{\beta})/\sigma) \\ &= 1 - F((\mathbf{x}_i'\boldsymbol{\beta} - \overline{y})/\sigma) \end{aligned}$$

where $F()$ is the cumulative distribution function of a standardised normal density, and the final line follows since $F(a) = 1 - F(-a)$ for a symmetric distribution. Since the total area under any distribution should be equal to one, the density function of the truncated distribution needs to be rescaled. Thus the probability density function of the truncated sample is given by

$$g(y_i) = \frac{f((y_i - \mathbf{x}_i'\boldsymbol{\beta})/\sigma)/\sigma}{F((\mathbf{x}_i'\boldsymbol{\beta} - \overline{y})/\sigma)} \quad \text{if } y_i^* \geq \overline{y}$$

and 0 otherwise.

The likelihood function is defined by

$$\begin{aligned} L(\mathbf{y}) &= \prod_{y_i^* \geq \overline{y}} g(y_i) \\ &= \prod_{y_i^* \geq \overline{y}} \frac{1}{\sigma} f((y_i - \mathbf{x}_i'\boldsymbol{\beta})/\sigma) / \prod_{y_i^* \geq \overline{y}} F((\mathbf{x}_i'\boldsymbol{\beta} - \overline{y})/\sigma) \end{aligned}$$

Maximising this expression with respect to $\boldsymbol{\beta}$ and σ defines the maximum likelihood estimator in the truncated variables model.

References

[1] Amemiya, T. (1981), 'Qualitative response models: a survey', *Journal of Economic Literature*, 19, 483–536.

[2] Efron, B. (1978), 'Regression and ANOVA with zero-one data: measures of residual variation', *Journal of American Statistical Association*, 73,113–121.

[3] Cragg, J.G. and R. Uhler (1970), 'The demand for automobiles', *Canadian Journal of Economics*, 3, 386–406.

[4] McFadden, D. (1974), 'The measurement of urban travel demand', *Journal of Public Economics*, 3, 303–328.

[5] Tobin, J. (1958), 'Estimation of relationships for limited dependent variables', *Econometrica*, 26, 24–36.