**Limited Dependent Variable**

Econometric Methods, ECON 370

We had previously discussed the possibility of running regressions even when the dependent variable is dichotomous in nature (binary dependent variable, which we can think of as probability measures). However, the caveat we noted is that there is nothing to guarantee that the predicted dependent variable would fall between 0 and 1, which is the support by definition of a probability measure.

# 1 Logit & Probit Models for Binary Response

As noted, the key complaints against the Linear Probability Model (LPM) is that,

1. Predicted dependent variable may not be within the support.
2. Partial Effects are constant for all explanatory variables.

In the binary response model, the principle concern is with the response probability,

$$\Pr(y = 1|x) = \Pr(y = 1|x_1, x_2, ..., x_k) \tag{1}$$

Suppose what we are examining is the probability of high school graduation after a new compulsory education policy, then $y$ could be 1 if the child graduates, and 0 otherwise. While $x$ would include family characteristics, and include an indicator or dummy variable for whether the child lived through the new law or otherwise.

To eliminate the limitations of the LPM, we will assume the following,

$$\Pr(y = 1|x) = F(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k) \tag{2}$$

where $F(.)$ is a function such that $F : x \longmapsto [0, 1]$, $\forall x \in \mathbb{R}$. There are various function suggested for $F(.)$, of which we will be discussing the two most popular.

Firstly, the **Logit Model** is based on the assumption that $F(.)$ follows a logistic (cumulative) distribution,

$$F(x) = \frac{exp(x)}{1 + exp(x)} = \Lambda(x) \tag{3}$$

The other is the **Probit Model** assumes that the function $F(.)$ follows a normal (cumulative) distribution,

$$F(x) = \Phi(x) = \int_{-\infty}^{x} \phi(z)dz \tag{4}$$

where $\phi(z)$ is the normal density function,

$$\phi(z) = \frac{exp(-\frac{z^2}{2})}{\sqrt{2\pi}} \tag{5}$$

The logit and probit models can be derived from an **latent variable model**. Let $y^*$ be an unobserved or latent variable determined by,

$$y^* = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon \tag{6}$$

The idea here is that the observed variable, $y$, will take on a value of 1 if $y^*$ is greater than 0 ($\mathbb{I}(y^* > 0)$), and 0 otherwise, where $\mathbb{I}(.)$ is an indicator function, and takes on the value 1 if the term in brackets is true. In order to estimate this function, we will still assume that the expected value of the error terms given the independent variables is 0, i.e. that there are uncorrelated. The distribution of the error term is dependent on the underlying assumption made about $F(.)$ of course (note that both Logistic and Normal distribution functions are symmetric about 0). Given the assumptions on the distribution functions, and the specification for the latent variables, we can derive the response probabilities then,

$$\begin{aligned} \Pr(y = 1|x) &= \Pr(y^* > 0|x) = \Pr(\epsilon > -\beta_0 - \beta_1 x_1 - \beta_2 x_2 - ... - \beta_k x_k|x) \\ &= 1 - F(-\beta_0 - \beta_1 x_1 - \beta_2 x_2 - ... - \beta_k x_k) \\ &= F(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k) \end{aligned}$$

Note that the last equality is exactly what we want, as in equation (2).

Because $y^*$ typically does not have a measure that is easily interpretable, when examining the effect of the independent variable, we examine it in relation to the effect it has on $\Pr(y = 1|x)$. To find the partial effect of say variable $x_j$, just note that,

$$\frac{\partial F(x\beta)}{\partial x_j} = f(x\beta)\beta_j \tag{7}$$

if $x_j$ is a continuous variable. $f(.)$ is the density function, which if we assume a logistic or normal distribution function, $F(.)$ is strictly increasing, and $f(.) > 0\ \forall x$, and therefore the partial effect always has the sign of the coefficient, $\beta_j$.

An interesting fact to note is that the relative effects of any two continuous independent variable does not depend on $x$ since the ratio of the partial effects between two independent variables $x_j$ and $x_k$ is just $\frac{\beta_j}{\beta_k}$.

If $x_j$ were instead a dichotomous or binary variable, then the partial effect is,

$$F(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{j-1} x_{j-1} + \beta_j + \beta_{j+1} x_{j+1} + ... + \beta_k x_k)$$
$$-F(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{j-1} x_{j-1} + \beta_{j+1} x_{j+1} + ... + \beta_k x_k)$$

and as before, the effect will be dependent on the values of $x$, which will affect your estimate of the partial effect. Most statistical programs calculates this at the mean. However, when the researcher has substantial number of dummy variables, it is more accurate to calculate this partial effects for each group or category, at means for the continuous variables for observations which belong to the category.

The principal difficulty is due to the scale factor, i.e. the fact that the partial effect is dependent on the independent variables, which consequently makes the interpretation of partial effect difficult. Some commonly used is to take the average of the partial effects estimated for each observation. Consequently, this is commonly called the **average partial effect**. For continuous variables, this involves calculating,

$$n^{-1} \sum_{i=1}^{n} f(x_i \widehat{\beta}) \widehat{\beta}_j$$

Here, even if the independent variables include dummy variables, by averaging the partial effect over the entire sample, you would have implicitly weighted the partial effect by the proportion of the observations you observed them in their respective categories. The average partial effect for dummy variables are similar in the sense that they are still averages across the sample, but we have to account for the discrete change (note that the formula here works both for dummy, and discrete variables, such as year of education, which very often would occur as an independent variable, particularly in wage regressions). The formula for dummy variables would be,

$$n^{-1} \sum_{i=1}^{n} \Big( G(x_{-j} \widehat{\beta}_{-j} + x_j \widehat{\beta}_j) - G(x_{-j} \widehat{\beta}_{-j}) \Big)$$

while that for discrete variables if it increases by the usual unit measure,

$$n^{-1} \sum_{i=1}^{n} \Big( G(x_{-j} \widehat{\beta}_{-j} + (x_{j+1} + 1) \widehat{\beta}_j) - G(x_{-j} \widehat{\beta}_{-j} + x_{j+1} \widehat{\beta}_j) \Big)$$

where the subscript $-j$, implies all variables excluding that indexed by $j$.

There is no restriction as to what kind of functional form for the independent variables that the researcher can use, that is you can use log values and quadratic values of the independent variables in the regression. Of course, the partial effects has to be calculated for interpretation. The care in interpretation is especially crucial when examining interaction terms.

## 1.1 Maximum Likelihood Estimation

Since the regression equation is non-linear in both parameter and variables, OLS and WLS are not applicable. Of course, we can always use Non-Linear Least Squares (which is not part of the curriculum), it is far easier to use Maximum Likelihood Estimation.

Suppose we have on hand $n$ observations for all the variables. Then the probability of observing any outcome is just,

$$L_i(y_i|x_i;\beta) = (F(x_i\beta))_i^y \left(1 - F(x_i\beta)\right)^{1-y_i} \tag{8}$$

We call the above the likelihood function. Consequently, the log likelihood function for observation $i$ is just,

$$l_i(\beta) = y_i \log(F(x_i\beta)) + (1 - y_i) \log(1 - F(x_i\beta)) \tag{9}$$

Since the logistic and normal distributions are increasing and convex, the log transformation ensures that the problem is well behaved. Then the log Likelihood function for the entire sample is just

$$l(\beta) = \sum_{i=1}^{n} y_i \log(F(x_i\beta)) + \sum_{i=1}^{n} (1 - y_i) \log(1 - F(x_i\beta)) = \sum_{i=1}^{n} l_i(\beta) \tag{10}$$

Under general assumptions, the estimates of the coefficients using MLE are consistent, asymptotically normal and efficient. The asymptotic variance is as follows,

$$p\lim var(\widehat{\beta}) = \left( \sum_{i=1}^{n} \frac{[f(x_i\widehat{\beta})]^2 x_i' x_i}{F(x_i\widehat{\beta})(1 - F(x_i\widehat{\beta}))} \right)^{-1} \tag{11}$$

Given the variance, the test of significance of the estimates are the usual $t$ test.

## 1.2 Testing Multiple Hypotheses

There are three ways of testing multiple restrictions in Probit and Logit models, namely

1. **Lagrange Multiplier or Score Test**
2. **Wald Test**
3. **Likelihood Ratio Test**

The most commonly used test, and most easily calculated test is the last, Likelihood Ratio Test. It is used when we wish to test our exclusion restrictions, i.e. whether we should or should not exclude a variable or set of variables. The idea is a simple one, since what we are maximizing is the log likelihood function, and since as variables are excluded from the regression relationship, the objective function falls. The question then is whether we have a significant fall in the log likelihood function value. The likelihood ratio statistic is just twice the difference in the log likelihood functions:

$$LR = 2(L_{ur} - L_r)$$

where $L_{ur}$ and $L_r$ are the log likelihoods for the unrestricted and restricted models. It must be noted that since the distribution functions within the log are strictly $\in [0, 1]$, this implies that taking log would yield negative numbers. Nonetheless, there is no correction needed, you simply take the difference between the two numbers still. The $LR$ statistics follows a $\chi^2_q$ distribution asymptotically, where $q$ is the degrees of freedom, which also correspond with the number of exclusion restrictions. Finally note that since $|L_{ur}| \leq L_r$, the $LR$ statistic is always necessarily a positive number.

A possible goodness of fit measure in Probit and Logit models is the **percent correctly predicted** measure. Because it is typically not reported by most statistical software, to create it would need additional calculations. The idea is to compare for how many of the predicted probabilities correspond with the observed action of the observation. Let the predicted probability be $F(x_i\beta)$ for observation 1. You would then have to create a new variable $\widetilde{y}$ such that it takes on the value of 1 if $F(x_i\beta) \geq 0.5$, and 0 otherwise. Next create the difference between the observed action, $y$ and the predicted $\widetilde{y}$. If the prediction is correct, this difference is 0, else not. By counting the number of zeros and divided the count by the total number of observations then gives you the measure of goodness of fit you want here.

Another common goodness of fit which is reported is that proposed by McFadden(1974),

$$1 - \frac{L_{ur}}{L_o}$$

where as before $L_{ur}$ is the log likelihood value you obtain from the regression, and $L_o$ is that generated by a regression (either Probit or Logit. But note that this goodness of fit measure is also used and reported often by statistical packages for any regression that involves MLE) with only the intercept. The key idea here is that $|L_{ur}| \leq L_o$ so that the fraction above is always between 0 and 1. So that the proposed goodness of fit would be close to 0 is the regression has no explanatory power, and if good, would be close to 1. This is called a **Pseudo** $R^2$. There are other versions of the Pseudo $R^2$, one of which is also touched on in your text, page 590. For the similar concerns such as heteroskedasticity and endogeneity it is recommended that you read Wooldridge's reference book "Econometric Analysis of Cross Section and Panel Data".