Limited Dependent Variables

Studenmund talks about regression models for a dummy (0/1) dependent variable. There are many other ways in which the range of dependent variables can be limited. Even if you do not acquire a deep understanding of these methods, you should be aware of their existence so that they can be in your toolbox to address appropriate classes of research problems.

What we will do in class is to describe the setting in which these methods are appropriate (so that you can recognize them when you need to) and briefly discuss how to implement them in Stata, including computing marginal effects of various kinds. We won't derive the likelihood function or analyze the underlying probability distributions involved.

The pages that follow are a brief appendix to Chapter 11 of Stock and Watson's text. We will talk about these methods more in class. The situations and methods are summarized in the table below:

Values of *Y*	**Techniques for estimation**
0 or 1	Linear probability model, logit, or probit
Positive, but often 0	Tobit (censored regression model)
Positive, but often 0, with no data for 0s	Heckit (truncated regression model)
0, 1, 2, 3, …	Poisson or negative binomial regression
Ordinal variable	Ordered probit or logit
Multiple discrete choices	Multinomial logit or probit

Standard Errors for Predicted Probabilities

For simplicity, consider the case of a single regressor in the probit model. Then the predicted probability at a fixed value of that regressor, x, is $\hat{p}(x) = \Phi(\hat{\beta}_0^{MLE} + \hat{\beta}_1^{MLE}x)$, where $\hat{\beta}_0^{MLE}$ and $\hat{\beta}_1^{MLE}$ are the MLEs of the two probit coefficients. Because this predicted probability depends on the estimators $\hat{\beta}_0^{MLE}$ and $\hat{\beta}_1^{MLE}$, and because those estimators have a sampling distribution, the predicted probability will also have a sampling distribution.

The variance of the sampling distribution of $\hat{p}(x)$ is calculated by approximating the function $\Phi(\hat{\beta}_0^{MLE} + \hat{\beta}_1^{MLE}x)$, a nonlinear function of $\hat{\beta}_0^{MLE}$ and $\hat{\beta}_1^{MLE}$, by a linear function of $\hat{\beta}_0^{MLE}$ and $\hat{\beta}_1^{MLE}$. Specifically, let

$$\hat{p}(x) = \Phi(\hat{\beta}_0^{MLE} + \hat{\beta}_1^{MLE}x) \cong c + a_0(\hat{\beta}_0^{MLE} - \beta_0) + a_1(\hat{\beta}_1^{MLE} - \beta_1) \qquad (11.19)$$

where the constant c and factors a_0 and a_1 depend on x and are obtained from calculus. [Equation (11.19) is a first-order Taylor series expansion; $c = \Phi(\beta_0 + \beta_1 x)$; and a_0 and a_1 are the partial derivatives, $a_0 = \partial\Phi(\beta_0 + \beta_1 x)/\partial\beta_0|_{\hat{\beta}_0^{MLE}, \hat{\beta}_1^{MLE}}$ and $a_1 = \partial\Phi(\beta_0 + \beta_1 x)/\partial\beta_1|_{\hat{\beta}_0^{MLE}, \hat{\beta}_1^{MLE}}$.] The variance of $\hat{p}(x)$ now can be calculated using the approximation in Equation (11.19) and the expression for the variance of the sum of two random variables in Equation (2.31):

$$\begin{aligned} \text{var}[\hat{p}(x)] &\cong \text{var}[c + a_0(\hat{\beta}_0^{MLE} - \beta_0) + a_1(\hat{\beta}_1^{MLE} - \beta_1)] \\ &= a_0^2\,\text{var}(\hat{\beta}_0^{MLE}) + a_1^2\,\text{var}(\hat{\beta}_1^{MLE}) + 2a_0a_1\text{cov}(\hat{\beta}_0^{MLE}, \hat{\beta}_1^{MLE}). \end{aligned} \qquad (11.20)$$

Using Equation (11.20), the standard error of $\hat{p}(x)$ can be calculated using estimates of the variances and covariance of the MLE's.

APPENDIX

11.3 Other Limited Dependent Variable Models

This appendix surveys some models for limited dependent variables, other than binary variables, found in econometric applications. In most cases the OLS estimators of the parameters of limited dependent variable models are inconsistent, and estimation is routinely done using maximum likelihood. There are several advanced references available to the reader interested in further details; see, for example, Ruud (2000) and Wooldridge (2002).

Censored and Truncated Regression Models

Suppose that you have cross-sectional data on car purchases by individuals in a given year. Car buyers have positive expenditures, which can reasonably be treated as continuous

random variables, but nonbuyers spent \$0. Thus the distribution of car expenditures is a combination of a discrete distribution (at zero) and a continuous distribution.

Nobel laureate James Tobin developed a useful model for a dependent variable with a partly continuous and partly discrete distribution (Tobin, 1958). Tobin suggested modeling the i^{th} individual in the sample as having a desired level of spending, Y_i^*, that is related to the regressors (for example, family size) according to a linear regression model. That is, when there is a single regressor, the desired level of spending is

$$Y_i^* = \beta_0 + \beta_1 X_i + u_i, i = 1, \ldots, n. \tag{11.21}$$

If Y_i^* (what the consumer wants to spend) exceeds some cutoff, such as the minimum price of a car, the consumer buys the car and spends $Y_i = Y_i^*$, which is observed. However, if Y_i^* is less than the cutoff, spending of $Y_i = 0$ is observed instead of Y_i^*.

When Equation (11.21) is estimated using observed expenditures Y_i in place of Y_i^*, the OLS estimator is inconsistent. Tobin solved this problem by deriving the likelihood function using the additional assumption that u_i has a normal distribution, and the resulting MLE has been used by applied econometricians to analyze many problems in economics. In Tobin's honor, Equation (11.21), combined with the assumption of normal errors, is called the *tobit* regression model. The tobit model is an example of a *censored regression model*, so-called because the dependent variable has been "censored" above or below a certain cutoff.

Sample Selection Models

In the censored regression model, there are data on buyers and nonbuyers, as there would be if the data were obtained by simple random sampling of the adult population. If, however, the data are collected from sales tax records, then the data would include only buyers: There would be no data at all for nonbuyers. Data in which observations are unavailable above or below a threshold (data for buyers only) are called truncated data. The *truncated regression model* is a regression model applied to data in which observations are simply unavailable when the dependent variable is above or below a certain cutoff.

The truncated regression model is an example of a sample selection model, in which the selection mechanism (an individual is in the sample by virtue of buying a car) is related to the value of the dependent variable (expenditure on a car). As discussed in the box in Section 11.4, one approach to estimation of sample selection models is to develop two equations, one for Y_i^* and one for whether Y_i^* is observed. The parameters of the model can then be estimated by maximum likelihood, or in a stepwise procedure, estimating the selection equation first and then estimating the equation for Y_i^*. For additional discussion, see Ruud (2000, Chapter 28), Greene (2000, Section 20.4), or Wooldridge (2002, Chapter 17).

Count Data

Count data arise when the dependent variable is a counting number, for example, the number of restaurant meals eaten by a consumer in a week. When these numbers are large, the variable can be treated as approximately continuous, but when they are small, the continuous approximation is a poor one. The linear regression model, estimated by OLS, can be used for count data, even if the number of counts is small. Predicted values from the regression are interpreted as the expected value of the dependent variable, conditional on the regressors. So, when the dependent variable is the number of restaurant meals eaten, a predicted value of 1.7 means, on average, 1.7 restaurant meals per week. As in the binary regression model, however, OLS does not take advantage of the special structure of count data and can yield nonsense predictions, for example, −0.2 restaurant meal per week. Just as probit and logit eliminate nonsense predictions when the dependent variable is binary, special models do so for count data. The two most widely used models are the Poisson and negative binomial regression models.

Ordered Responses

Ordered response data arise when mutually exclusive qualitative categories have a natural ordering, such as obtaining a high school degree, some college education (but not graduating), or graduating from college. Like count data, ordered response data have a natural ordering, but unlike count data, they do not have natural numerical values.

Because there are no natural numerical values for ordered response data, OLS is inappropriate. Instead, ordered data are often analyzed using a generalization of probit called the *ordered probit model*, in which the probabilities of each outcome (e.g., a college education), conditional on the independent variables (such as parents' income), are modeled using the cumulative normal distribution.

Discrete Choice Data

A *discrete choice* or *multiple choice* variable can take on multiple unordered qualitative values. One example in economics is the mode of transport chosen by a commuter: She might take the subway, ride the bus, drive, or make her way under her own power (walk, bicycle). If we were to analyze these choices, the dependent variable would have four possible outcomes (subway, bus, car, human-powered). These outcomes are not ordered in any natural way. Instead, the outcomes are a choice among distinct qualitative alternatives.

The econometric task is to model the probability of choosing the various options, given various regressors such as individual characteristics (how far the commuter's house is from the subway station) and the characteristics of each option (the price of the subway). As discussed in the box in Section 11.3, models for analysis of discrete choice data can be developed from principles of utility maximization. Individual choice probabilities can be expressed in probit or logit form, and those models are called *multinomial probit* and *multinomial logit* regression models.