Lecture 1:

- The course is going to be about the estimation of censored and sample selection model under a variety of situations
- First I am going to very quickly review parametric estimation of the binary choice model for three reasons:
- i) remind everyone of MLE estimation
- ii) computation of the generalized residual
- iii) introduction of testing procedures of discrete models.

- Remember that the binary discrete choice model has the following latent representation
-

$$
\begin{aligned}
y_i^* &= \beta' x_i + u_i \; i = 1...n && (1)\\
y_i &= 1 \text{ if } y_i^* > 0 && (2)\\
y_i &= 0 \text{ otherwise} && (3)
\end{aligned}
$$

  where the asterisks denotes a latent variable; x_i is a k vector of exogenous variables;β is a k vector of unknown parameters and $E[u_i] = 0$.
- The objective is to estimate the unknown parameter vector β and we employ maximum likelihood assuming that the errors are normal.

- Using the structure of the latent model outlined above we can write

$$\begin{aligned} P(y_i &= 1) = P(u_i > -\beta' x_i) \\ &= 1 - F(-\beta' x_i) \end{aligned}$$

where F is the cumulative distribution function for u_i. The associated likelihood function has the form

$$L = \prod_{y_i=0} (F(-\beta' x_i) \prod_{y_i=1} [1 - F(-\beta' x_i)].$$

The estimates for β are then obtained by appropriately choosing a form for F and then maximizing the likelihood function. The most commonly chosen forms for F are the normal and logistic distributions and these produce the logit and probit estimators.

0.1 Probit

- Let us first consider the probit estimator which is based on the use of the normal distribution where u_i is distributed $IN(0, \sigma^2)$. That is

$$F(-\beta' x_i) = \int_{-\infty}^{-\beta' x_i/\sigma} \frac{1}{(2\pi)^{1/2}} \exp(-\frac{t^2}{2}) dt.$$

- The log likelihood function then becomes

$$\log\ L = \sum_{i=1}^{n} y_i \log \Phi(\beta' x_i) + \sum_{i=1}^{n} (1 - y_i) \log \Phi[1 - (\beta' x_i)] \qquad (4)$$

where $\Phi(.)$ is the standard normal cdf. To maximize the likelihood function we differentiate 4 with respect to β. The first order condition, or score function, for the probit model has the form

$$S(\beta) = \sum_{i=1}^{n} \frac{[y_i - \Phi(\beta' x_i)]}{\Phi(\beta' x_i)[1 - \Phi(\beta' x_i)]} \phi(\beta' x_i) x_i. \tag{5}$$

- The estimates or β are chosen such that $S(\beta) = 0$. Note that 5 has an important role in some of the other topics that we will cover. Note that if we evaluate the 5 for the constant contained in x_i this score function is equal to the generalized residual for the probit model.
- Note that is also equal to the Inverse Mills ratio. In order to do inference we need to compute the covariance matrix for β. This is equal to

$$I(\beta) = \sum_{i=1}^{n} \frac{[\phi(\beta' x_i)]^2}{\Phi(\beta' x_i)[1 - \Phi(\beta' x_i)]} x_i x_i'$$

- The score for the intercept, or the generalized residual is very important because it has the interpretation of the residual.
- The residual is important as it also suits the purpose of a control function in models with endogeneity.
- For example, consider 2sls. There are multiple ways to account for the endogeneity but one of the ways is to include the residual from the reduced form as an additional regressor in the primary equation.

- The second important use of the generalized residuals is the role in the testing procedure for binary choice models and other models estimated by MLE.
- This is due to an approach of Newey (1985) and surveyed in Pagan and Vella (1989)
- Explain conditional moment tests and how they are implemented

- We now turn our attention to models where the dependent variable is partially censored. This model arises frequently in micro data sets as variables are frequently not recorded for all observations or zero values correspond to something which is below some threshold.
- This type of model was originally considered by Tobin in a model of household consumption. The general approach which is adopted is to assume that there is a latent variable y_i^* which is mapped into the observed data in the following way

$$\begin{aligned} y_i^* &= x_i'\beta + u \\ y_i &= y_i^* \text{ if } y_i^* > 0 \\ &= 0 \text{ otherwise} \end{aligned}$$

where x is a vector of exogenous variables; β is a vector of unknown parameters; and $u \sim N(0, \sigma^2)$.

- In this original formulation we assume that the threshold is zero although this is arbitrarily chosen. Before proceeding to estimation one can see that this type of model arises frequently in economics. Aside from the consumption example there is also the popularly implemented example of hours of work, noting that many observations report zero hours.
- Also note that an important feature of this model, in contrast to a model that we will consider below, is that although the dependent variable is reported as zero when the latent variable is below some threshold the x vector is observed for all observations. This model is known as the censored regression model.

- The obvious way to estimate the above model is to ignore the unusual structure of the dependent variable and simply do OLS.
- In doing OLS there are two ways to proceed. First, we can ignore the existence of the censored observations and estimate over the uncensored observations.
- Alternatively, we can ignore the fact the observations are zero and estimate over the whole sample with the zeros included. First, consider the regression using only positive observations of y_i. From some simple rewriting of the model we get:

$$\begin{aligned} E(y_i|y_i^* &> 0) = x_i'\beta + E[u_i|y_i^* > 0] \\ &= x_i'\beta + E[u_i|u_i > -x_i'\beta]. \end{aligned}$$

The issue of consistent estimation now focuses on the value of $E[u_i|u_i > -x_i'\beta]$.

- For OLS to be consistent we need that $E[u_i|u_i > -x_i'\beta]$ is uncorrelated with the $x's$.
- Otherwise the exclusion of this variable will lead to biased estimates. If we assume normality it is straightforward to show that

$$E[u_i|u_i > -x_i'\beta] = \sigma \frac{\phi(x_i'\beta/\sigma)}{\Phi(x_i'\beta/\sigma)}$$

  where ϕ and Φ denote the normal pdf and cdf respectively.
- Note, from our discussion of the estimation of the probit model the term in brackets is the Inverse Mills ratio. As we discussed earlier the Inverse Mill ratio is an important concept in this literature so it useful to consider it the expected value of the error which is the generalized value when we evaluate it at the parameter estimates.
- However, it is very unlikely that this term will be uncorrelated with the regressors and thus the OLS estimates will be biased if the later term is not accounted for. Later on we will discuss two-step procedures which account for this latter term.

- It is also immediately clear from a diagrammatic representation that estimation over the whole sample will lead to biased estimates.
- Given the distributional assumptions one can also estimate this model by MLE.
- Before turning to a discussion of the tobit likelihood function it is clear that one could estimate this model by probit by just treating the uncensored observations as being equal to 1 and the censored observations as 0's.
- However, one suspects that the estimates would be more efficient if we incorporate the variation in the values of the uncensored observations in the likelihood function. Also, as we saw in probit estimation we are only able to estimate $\frac{\beta}{\sigma}$ and it is likely that using the actual values for the uncensored observations will help us identify these two parameters separately.

- The likelihood function will thus have the following components. First, like the probit likelihood function the first part will comprise the censored observations while the second part will explain the variation in the uncensored observations conditional on the observations being uncensored. The log likelihood function thus has the following form:

$$\begin{aligned} L &= -\sum_{i=1}^{n_1} \ln[\frac{1}{(2\pi\sigma^2)^{1/2}}] - \frac{1}{2\sigma^2}\sum_{i=1}^{n_1}(y_i - x_i'\beta)^2 - \\ &\quad \sum_{i=1}^{n_0} \ln[1 - \Phi(\frac{x_i'\beta}{\sigma})] \end{aligned}$$

  where $\sum_{i=1}^{n_1}$ and $\sum_{i=1}^{n_0}$ denote summation over the uncensored and uncensored observations respectively.

- To estimate this model we simply set the first order conditions equal to zero. Note that in this instance there are first order conditions for the slope parameters and a first order condition for the variance.
- It is useful to look at the these first order conditions as they are useful for the estimation of sample selection models, which we will consider next, and also for diagnostic testing which we also consider soon.

$$\frac{\partial L}{\partial \beta} = -\sum_{i=1}^{n_0} \frac{\phi(\frac{x'\beta}{\sigma})}{1-\Phi(\frac{x_i'\beta}{\sigma})} + \frac{1}{\sigma^2}\sum_{i=1}^{n_1}(y_i - x_i'\beta)x_i = 0$$

$$\begin{aligned} \frac{\partial L}{\partial \sigma^2} &= \frac{1}{2\sigma^2}\sum_{i=1}^{n_0} \frac{(x_i'\beta)\phi(\frac{x'\beta}{\sigma})}{1-\Phi(\frac{x_i'\beta}{\sigma})} - \frac{n_1}{2\sigma^2} + \\ \frac{1}{2\sigma^4}\sum_{i=1}^{n_1}(y_i - x_i'\beta)^2 &= 0. \end{aligned}$$

- Note that for computational sake these derivatives are simple to calculate. It is also valuable to note that the derivative for the constant continues to have the interpretation of the generalized residual. Its also interesting to note that the residual has the probit residual for the censored observations and the usual OLS residual for the uncensored observations.

- Testing of the tobit model.