AAE 637 Lab 8: Limited Dependent Variables

TA: Charng-Jiun Yu

March 21, 2018

Limited dependent variables refer to the dependent variables that are restricted in certain ways. Here we introduce the regression models that deal with two types of the most common limited dependent variables.[1]

1 OVERVIEW

- **Truncated regression models** are used when we want to study the characteristics of the **full population** but our data do not include the observations below or above certain thresholds of the outcome variables.
    1. Truncation from below: we want to analyze the returns of education on income but our data only include people with income higher than $22,000.
    2. Truncation from above: we want to study the impact of government funds on high school classroom size but we only include the classes with no more than 30 students.
- **Censored regression models** are used when we are interested in the **subpopulation** where the outcome variables below or above certain thresholds are transformed into a single value. In this case, the conventional regression methods will generate biased estimates.
    1. Censoring from below: The observations with income below poverty line are all reported as if they are at the poverty line.
    2. Censoring from above: The capacity of Camp Randall Stadium is about 80,000. Therefore, we might have many observations with 80,000 tickets sold that generate a mass point in our distribution.
- Note that for truncated data, if we are interested the subpopulation rather than the full population, we do not need to apply the truncated regression models because the truncated data actually represent our subpopulation of interest.[2]
- The censoring problem is more popular in empirical works.

[1] This handout is largely based on Greene's Econometric Analysis, 7th Edition.

[2] For censoring data, if we are interested in the full population, say, the true demand for a Badgers home game without capacity constraint, we can apply the truncated regression model.

2 TRUNCATED REGRESSION MODEL

Let's look at how truncated data affect our estimation of marginal effect and how we deal with it using truncated regression model. Suppose that we have a classical linear regression model:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon$$

where

$$\epsilon_i|\mathbf{x}_i \sim N[0, \sigma^2]$$

In this model, the marginal effect of $\mathbf{x}$ on y is $\boldsymbol{\beta}$

Let $\tilde{y}$ be the truncated data where all data with $y \leq a$ are dropped. This gives

$$\tilde{y}_i = \mathbf{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{\epsilon}$$

As this model is estimated using the truncated data, intuitively the marginal effect of $\mathbf{x}$ for this subpopulation, $\tilde{\boldsymbol{\beta}}$, is in general not equal to $\boldsymbol{\beta}$, the marginal effect of $\mathbf{x}$ for the full population.

We can actually characterize the difference. The expected value of $\tilde{y}$ can be described as

$$E(\tilde{y}_i) = \mathbf{x}_i'\tilde{\boldsymbol{\beta}} = \mathbf{x}_i'\boldsymbol{\beta} + \sigma\frac{\phi[(a-\mathbf{x}_i'\boldsymbol{\beta})/\sigma]}{1-\Phi[(a-\mathbf{x}_i'\boldsymbol{\beta})/\sigma]} = E(y_i) + \sigma\frac{\phi[(a-\mathbf{x}_i'\boldsymbol{\beta})/\sigma]}{1-\Phi[(a-\mathbf{x}_i'\boldsymbol{\beta})/\sigma]}$$

and therefore

$$E(y_i) = \mathbf{x}_i'\tilde{\boldsymbol{\beta}} - \sigma\frac{\phi[(a-\mathbf{x}_i'\boldsymbol{\beta})/\sigma]}{1-\Phi[(a-\mathbf{x}_i'\boldsymbol{\beta})/\sigma]}$$

where ϕ and Φ are the standard normal pdf and cdf, respectively. From the above equation, we can see that the marginal effect of $\mathbf{x}$ on true y is generally not equal to $\tilde{\boldsymbol{\beta}}$. To get the marginal effect for the full population, our estimation needs to take the term $\sigma\frac{\phi[(a-\mathbf{x}_i'\boldsymbol{\beta})/\sigma]}{1-\Phi[(a-\mathbf{x}_i'\boldsymbol{\beta})/\sigma]}$ into account.

Let $\alpha_i = (a-\mathbf{x}_i'\boldsymbol{\beta})/\sigma$ and $\lambda_i = \lambda(\alpha_i) = \frac{\phi(\alpha_i)}{1-\Phi(\alpha_i)}$. The truncated regression model can be described as

$$\tilde{y}_i = \mathbf{x}_i'\boldsymbol{\beta} + \sigma\lambda_i + u_i$$

where u_i is the error term with zero mean and heteroskedastic variance equal to $\sigma^2(1-\lambda_i^2 + \lambda_i\alpha_i)$. We can get consistent estimates of $\boldsymbol{\beta}$ using the maximum likelihood estimator.

The log-likelihood function is

$$\log L = -\frac{T}{2}(\log 2\pi + \log\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{T}(\tilde{y}_i - \mathbf{x}_i'\boldsymbol{\beta})^2 - \sum_{i=1}^{T}\log[1-\Phi(\frac{a-\mathbf{x}_i'\boldsymbol{\beta}}{\sigma})]$$

where T is the number of observations in the truncated data.

The marginal effect of $\mathbf{x}$ for the truncated data (subpopulation) is

$$\frac{\partial E(\tilde{y})}{\partial \mathbf{x}_i} = \boldsymbol{\beta}(1-\delta_i)$$

where $\delta_i = \lambda_i^2 - \lambda_i\alpha_i$.

The marginal effect of $\mathbf{x}$ for full population is

$$\frac{\partial E(y)}{\partial \mathbf{x}_i} = \boldsymbol{\beta}$$

3 CENSORED REGRESSION MODEL

Now let's turn our attention to the censored regression model. You are probably more familiar with its other name called tobit model.

Again we consider the linear model in the previous section. This time, we have a censored dependent variable y^* where it transforms the values of the original dependent variable y to 0 for all $y \leq 0$. That is to say, we have

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i$$

$$y_i^* = 0 \quad \text{if } y_i \leq 0$$

$$y_i^* = y_i \quad \text{if } y_i > 0$$

Sometimes, however, the above regression with uncensored y is not our interest. When the censoring always exists, such as the capacity of Camp Randall Stadium, we might be more interested in knowing the marginal effect of $\mathbf{x}$ on the censored y^* where the capacity constraint is always there. How should we do our estimation?

The most intuitive but incorrect way is to estimate the below model using OLS:

$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta}^* + \epsilon_i^*$$

The problem with this is that, as there is a mass point at the probability where $y^* = 0$, our standard regression models will produce biased estimates. That's why we use the tobit model.

The basic idea of the tobit model is to decompose our regression into the unlimited part ($y^* > 0$) and the limited part ($y^* = 0$). The resulting log-likelihood function is:

$$\log L = -\frac{1}{2}[\sum_{y_i^*>0} \log 2\pi + \log\sigma^2 + \frac{(y_i^* - \mathbf{x}_i'\boldsymbol{\beta})}{\sigma^2}] + \sum_{y_i^*=0} \log[1 - \Phi(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma})]$$

The marginal effect of $\mathbf{x}$ on the censored y^* is

$$\frac{\partial E(y_i^*|\mathbf{x}_i)}{\partial \mathbf{x}_i} = \boldsymbol{\beta}\Phi(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma})$$