Tobit Models

1 Introduction

The next two sets of notes are going to look at two closely related topics: regression when the dependent variable is **incompletely observed** and regression when the dependent variable is completely observed but is observed in a **selected sample** that is not representative of the population. These models share the feature that OLS regression leads to inconsistent parameter estimates because the sample is not representative of the population. This week we are going to focus on incompletely observed data.

2 Truncation and Censoring

The leading causes of incompletely observed data are (i) truncation and (ii) censoring.

2.1 Truncation

Truncation occurs when some observations on both the dependent variable and regressors are lost. For example, income may be the dependent variable and only low-income people are included in the sample. In effect, truncation occurs when the sample data is drawn from a subset of a larger population.

2.2 Censoring

Censoring occurs when data on the dependent variable is lost (or limited) but not data on the regressors. For example, people of all income levels may be included in the sample, but for some reason the income of high-income people may be top-coded as, say, \$100,000. Censoring is a defect in the sample - if there were no censoring, then the data would be a representative sample from the population of interest.

Truncation entails a greater loss of information than censoring. Long (1997, 188) provides a nice picture of truncation and censoring.

3 Some Basics on Truncated and Censored Distributions

3.1 Normal and Standard Normal Distributions

y is normally distributed with mean μ and variance σ^2. Thus, the pdf of y is the familiar:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \tag{1}$$

This is often written as $y \sim N(\mu, \sigma^2 I)$. Any normal distribution like this, regardless of its mean and variance, can be rewritten as a function of the standard normal distribution i.e. N(0,1) (Greene 2003, 849-850). To see this, recall that normal distributions remain normal under linear transformations i.e. if $y \sim N(\mu, \sigma^2 I)$, then $(a + by) \sim N(a + b\mu, b^2\sigma^2 I)$. One particularly convenient transformation is $a = \frac{-\mu}{\sigma}$ and $b = \frac{1}{\sigma}$. The resulting variable $z = \frac{y-\mu}{\sigma}$ has a standard normal distribution, denoted N(0, 1) with a density

$$f(z) = \phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \tag{2}$$

The specific notation $\phi(z)$ is often used for this distribution and $\Phi(z)$ for its cdf. It follows from (2) that if $y \sim N(\mu, \sigma^2 I)$, then we can rewrite it as a function of the standard normal distribution in the following way:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma}\,\phi\left(\frac{y-\mu}{\sigma}\right) \tag{3}$$

The cdf can be written as

$$\begin{aligned} P(Y \leq y) &= \Phi\left(\frac{y-\mu}{\sigma}\right) \\ P(Y > y) &= 1 - \Phi\left(\frac{y-\mu}{\sigma}\right) \end{aligned} \tag{4}$$

3.2 Truncated Normal Distribution

Let y denote the observed value of the dependent variable. Unlike with normal regression, y is the incompletely observed value of a latent dependent variable y^*. Recall that with truncation, our sample data is drawn from a subset of a larger population. In effect, with truncation from below, we only observe $y = y^*$ if y^* is larger than the truncation point τ. In effect, we lose the observations on y^* that are smaller or equal to τ. When this is the case, we typically assume that the variable $y|y > \tau$ follows a truncated normal distribution. The issue that we need to address is that we have removed a part of the original distribution. As a result, the distribution no longer has an area equal to one. To make the area under what is left of the distribution equal to one, we have to re-scale it (Greene 2003, 757). Thus, if a continuous random variable y has a pdf $f(y)$ and τ is a constant, then we have

$$f(y|y > \tau) = \frac{f(y)}{P(y > \tau)} \tag{5}$$

Using the results from (4), we know that

$$P(y > \tau) = 1 - \Phi\left(\frac{\tau - \mu}{\sigma}\right) = 1 - \Phi(\alpha) \tag{6}$$

where $\alpha = \frac{\tau-\mu}{\sigma}$ and $\Phi(\cdot)$ is the standard normal cdf as before. Given this, the density of the truncated normal distribution is

$$f(y|y > \tau) = \frac{f(y)}{1 - \Phi(\alpha)} = \frac{\frac{1}{\sigma}\,\phi\left(\frac{y-\mu}{\sigma}\right)}{1 - \Phi\left(\frac{\tau-\mu}{\sigma}\right)} = \frac{\frac{1}{\sigma}\,\phi\left(\frac{y-\mu}{\sigma}\right)}{1 - \Phi(\alpha)} \tag{7}$$

where $\phi(\cdot)$ is the standard normal pdf as before.

The likelihood function for the truncated normal distribution is

$$L = \prod_{i=1}^{N} \frac{f(y)}{1 - \Phi(\alpha)} \tag{8}$$

or

$$\text{ln}L = \sum_{i=1}^{N} (\text{ln}[f(y)] - \text{ln}[1 - \Phi(\alpha)]) \tag{9}$$

Some results on truncated distributions are:

- If the truncation is from below, then the mean of the truncated variable is greater than the mean of the original one. If the truncation is from above, then the mean of the truncated variable is smaller than the mean of the original one.
- Truncation reduces the variance compared with the variance in the untruncated distribution.

3.2.1 Moments of the Truncated Normal Distribution

If $y \sim N(\mu, \sigma^2)$ and τ is the truncation point, then

$$E[y|y > \tau] = \mu + \sigma\lambda(\alpha) \tag{10}$$

and

$$\text{Var}[y|y > \tau] = \sigma^2[1 - \delta(\alpha)] \tag{11}$$

where $\alpha = \frac{\tau-\mu}{\sigma}$, $\phi(\alpha)$ is the standard normal density,

$$\delta(\alpha) = \lambda(\alpha)[\lambda(\alpha) - \alpha], \tag{12}$$

and

$$\lambda(\alpha) = \frac{\phi(\alpha)}{[1 - \Phi(\alpha)]} \tag{13}$$

$\lambda(\alpha)$ is called the inverse Mills ratio (IMR).[1] We can think of the inverse Mills ratio as measuring the amount of truncation – the higher λ, the more truncation.

3.2.2 Truncated Regression Model

We start with

$$y \sim N(X\beta, \sigma^2) \tag{14}$$

We are interested in the distribution of y given that y is greater than the truncation point τ. Thus, we have

$$E[y|y > \tau] = X\beta + \sigma \left[\frac{\phi\left[\frac{\tau - X\beta}{\sigma}\right]}{1 - \Phi\left[\frac{\tau - X\beta}{\sigma}\right]} \right] \tag{15}$$

As a result, the conditional mean is a nonlinear function of τ, σ, X, and β. Note that it is now easy to see why OLS is wrong. In effect, OLS omits everything after the $X\beta$ in (15) - there is omitted variable bias. It is also the case that the error term will be heteroskedastic.

[1] These equations are all assuming that truncation is from below i.e. $y > \tau$.

3.3 Censored Normal Distribution

When a distribution is censored on the left, observations with values at or below τ are set to τ_y.

$$y = \begin{cases} y^* & \text{if } y^* > \tau \\ \tau_y & \text{if } y^* \leq \tau \end{cases}$$

The use of τ and τ_y are just a generalization of having τ and τ_y set at 0. If a continuous variable y has a pdf f(y) and τ is a constant, then we have

$$f(y) = [f(y^*)]^{d_i}[F(\tau)]^{1-d_i} \tag{16}$$

In other words, the density of y is the same as that for y^* for $y > \tau$ and is equal to the probability of observing $y^* < \tau$ if $y = \tau$. d is an indicator variable that equals 1 if $y > \tau$ i.e. the observation is uncensored and is equal to 0 if $y = \tau$ i.e. the observation is censored. We know from earlier that

$$P(\text{censored}) = P(y^* \leq \tau) = \Phi\left(\frac{\tau - \mu}{\sigma}\right) = 1 - \Phi\left(\frac{\mu - \tau}{\sigma}\right) \tag{17}$$

and

$$P(\text{uncensored}) = 1 - \Phi\left(\frac{\tau - \mu}{\sigma}\right) = \Phi\left(\frac{\mu - \tau}{\sigma}\right) \tag{18}$$

Thus, the likelihood function can be written as

$$L = \prod_i^N \left[\frac{1}{\sigma}\phi\left(\frac{y - \mu}{\sigma}\right)\right]^{d_i} \left[1 - \Phi\left(\frac{\mu - \tau}{\sigma}\right)\right]^{1-d_i} \tag{19}$$

The expected value of a censored variable is just

$$\begin{aligned} E[y] &= (P(\text{uncensored}) \times E[y|y > \tau]) + (P(\text{censored}) \times E[y|y = \tau_y]) \\ &= \left\{\Phi\left(\frac{\mu - \tau}{\sigma}\right)[\mu + \sigma\lambda(\alpha)]\right\} + \Phi\left(\frac{\tau - \mu}{\sigma}\right)\tau_y \end{aligned} \tag{20}$$

For the special case of when $\tau = 0$, we have

$$E[y] = \Phi\left(\frac{\mu}{\sigma}\right)[\mu + \sigma\lambda] \tag{21}$$

where

$$\lambda = \frac{\phi\left(\frac{\mu}{\sigma}\right)}{\Phi\left(\frac{\mu}{\sigma}\right)} \tag{22}$$

4 Tobit Model

4.1 Introduction

With this in hand, we can now turn to the tobit model (or censored normal regression model). As Wooldridge (2002, 517-520) makes clear, censored regression applications fall into two categories.

1. **Censored Regression Applications:** In this application, we have true censoring as outlined above. There is a variable with quantitative meaning, y^* and we are interested in the population regression $E(y^*)$. If y^* were observed for everyone in the population, we could use OLS etc. However, a data problem arises in that y^* is censored from above and/or below i.e. it is not observed for some part of the population. This is the censoring issue we have been discussing.

2. **Corner Solution Models:** In the second application, though, it seems misleading to use the terminology of censoring. In this application, y is an observable choice or outcome describing some agent with the following characteristics: y takes on the value 0 with positive probability but is a continuous random variable over strictly positive values. In effect, we have an agent who is solving a maximization problem. For some of these individuals, the optimal choice will be the corner solution, $y = 0$. It seems better to refer to these types of models as corner solution models rather than censored regression models. Note that in the corner solution applications, the issue is NOT data observability: we are interested in features of the distribution of y such as E(y) and P(y=0). As Wooldridge points out, it is problematic to use OLS in this setting.

Both types of application - the censored regression application and the corner solution application - lead us to the standard censored Tobit model (type 1 Tobit model).[2]

4.2 Setup

The structural equation in the Tobit model is:

$$y_i^* = X_i\beta + \epsilon_i \tag{23}$$

where $\epsilon_i \sim N(0, \sigma^2)$. y^* is a latent variable that is observed for values greater than τ and censored otherwise.[3] The observed y is defined by the following measurement equation

$$y_i = \begin{cases} y^* & \text{if } y^* > \tau \\ \tau_y & \text{if } y^* \leq \tau \end{cases}$$

In the typical tobit model, we assume that $\tau = 0$ i.e. the data are censored at 0. Thus, we have

$$y_i = \begin{cases} y^* & \text{if } y^* > 0 \\ 0 & \text{if } y^* \leq 0 \end{cases}$$

4.3 Estimation

As we've seen from earlier, the likelihood function for the censored normal distribution is:

$$L = \prod_i^N \left[\frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)\right]^{d_i}\left[1 - \Phi\left(\frac{\mu-\tau}{\sigma}\right)\right]^{1-d_i} \tag{24}$$

[2] Note that Sigelman and Zeng (1999) imply that the Tobit model should not be used for things like the corner solution applications - this is where I think that they are wrong and Wooldridge is right.

[3] The tobit model can be generalized to take account of censoring both from below and/or from above. It can also take account of interval censored data.

where τ is the censoring point. In the traditional tobit model, we set $\tau = 0$ and parameterize μ as $X_i\beta$. This gives us the likelihood function for the tobit model:

$$L = \prod_i^N \left[\frac{1}{\sigma}\phi\left(\frac{y_i - X_i\beta}{\sigma}\right)\right]^{d_i}\left[1 - \Phi\left(\frac{X_i\beta}{\sigma}\right)\right]^{1-d_i} \tag{25}$$

The log-likelihood function for the tobit model is

$$\ln L = \sum_{i=1}^N \left\{d_i\left(-\ln\sigma + \ln\phi\left(\frac{y_i - X_i\beta}{\sigma}\right)\right) + (1-d_i)\ln\left(1 - \Phi\left(\frac{X_i\beta}{\sigma}\right)\right)\right\} \tag{26}$$

The overall log-likelihood is made up of two parts. The first part corresponds to the classical regression for the uncensored observations, while the second part corresponds to the relevant probabilities that an observation is censored.

To estimate a tobit model, type:

- tobit DV IVs, ll(0)

ll(0) indicates that the lower limit (censoring point) is 0. If censoring is from above, use ul(the censoring point). If the censoring point is not zero but must be estimated, use $\min(y_i|y_i > 0)$. This will exceed the true censoring point, but is better than using 0 - I think STATA does this automatically if you do not explicitly specify the lower limit.

4.4 Expected Values for Tobit Model

As Sigelman and Zeng (1999) point out, there are three expected values that we might be interested in. To simplify things, we'll keep looking at the case where censoring is at zero i.e. $\tau = 0$.

1. **Expected value of the latent variable y^*:**

$$E[y^*] = X_i\beta \tag{27}$$

2. **Expected value of $y|y > 0$:**[4]

$$E[y|y > 0] = X_i\beta + \sigma\lambda(\alpha) \tag{30}$$

where $\lambda(\alpha) = \frac{\phi(\frac{X_i\beta}{\sigma})}{\Phi(\frac{X_i\beta}{\sigma})}$ is the inverse Mills ratio.

[4] Where does this come from? Recall from (15) that the expected value of a truncated normal distribution is:

$$E[y|y > \tau] = X\beta + \sigma\left[\frac{\phi\left[\frac{\tau - X\beta}{\sigma}\right]}{1 - \Phi\left[\frac{\tau - X\beta}{\sigma}\right]}\right] \tag{28}$$

Replacing τ with 0, we have

$$E[y|y > \tau] = X\beta + \sigma\left[\frac{\phi\left[\frac{0 - X\beta}{\sigma}\right]}{1 - \Phi\left[\frac{0 - X\beta}{\sigma}\right]}\right] = X\beta + \sigma\left[\frac{\phi\left[\frac{X\beta}{\sigma}\right]}{\Phi\left[\frac{X\beta}{\sigma}\right]}\right] = X_i\beta + \sigma\lambda(\alpha) \tag{29}$$

3. **Expected value of y:**[5]

$$E[y] = \Phi\left(\frac{X_i\beta}{\sigma}\right)[X_i\beta + \sigma\lambda(\alpha)] \tag{32}$$

where $\lambda(\alpha) = \frac{\phi(\frac{X_i\beta}{\sigma})}{\Phi(\frac{X_i\beta}{\sigma})}$ is again the inverse Mills ratio. This is the probability of being uncensored multiplied by the expected value of y given y is uncensored.

Given that there are three expected values, which one should you report? As Greene (2003, 764) notes, there is no consensus on this and much will depend on the purpose of the analysis. He thinks that if the data is always censored, then focusing on the latent variable is not particularly useful. Wooldridge (2002, 520) also argues that you are probably not interested in the latent variable if you are employing a corner solution model. If we are interested in the effects of explanatory variables that may or may not be censored then we are probably interested in E[y]. If we are interested in just the uncensored observations, we will probably just want to look at $E[y|y > \tau]$. Greene seems to side with E[y] as the most useful but you should think about what is most useful for your particular purposes.

4.5 Marginal Effects for Tobit Model

Just as there are three expected values, there are three possible marginal effects.

1. **Marginal effect on the latent dependent variable, y^*:**

$$\frac{\partial E[y^*]}{\partial x_k} = \beta_k \tag{33}$$

Thus, the reported Tobit coefficients indicate how a one unit change in an independent variable x_k alters the latent dependent variable.

2. **Marginal effect on the expected value for y for uncensored observations:**

$$\frac{\partial E[y|y > 0]}{\partial x_k} = \beta_k\left\{1 - \lambda(\alpha)\left[\frac{X_i\beta}{\sigma} + \lambda(\alpha)\right]\right\} \tag{34}$$

where $\lambda(\alpha) = \frac{\phi(\frac{X_i\beta}{\sigma})}{\Phi(\frac{X_i\beta}{\sigma})}$. This indicates how a one unit change in an independent variable x_k affects uncensored observations.

3. **Marginal effect on the expected value for y (censored and uncensored):**[6]

$$\frac{\partial E[y]}{\partial x_k} = \Phi\left(\frac{X_i\beta}{\sigma}\right)\beta_k \tag{36}$$

[5] Where does this come from? Recall from (21) that the expected value of a censored normal distribution where the censoring occurs at 0 is:

$$E[y] = \Phi\left(\frac{\mu}{\sigma}\right)[\mu + \sigma\lambda] \text{ where } \lambda = \frac{\phi\left(\frac{\mu}{\sigma}\right)}{\Phi\left(\frac{\mu}{\sigma}\right)} \text{ and } \mu = X_i\beta \tag{31}$$

[6] It turns out that this equation can be written as

$$\frac{\partial E[y]}{\partial x_k} = P(y > 0)\frac{\partial E[y|y > 0]}{\partial x_k} + (E[y|y > 0]))\frac{\partial P(y > 0)}{\partial x_k} \tag{35}$$

This is called McDonald and Moffitt's decomposition. It allows us to see that a change in x_k affects the conditional mean of y^* in the positive part of the distribution and it affects the probability that the observation will fall in that part of the distribution.

Note that $\Phi\left(\frac{X_i\hat{\beta}}{\hat{\sigma}}\right)$ is simply the estimated probability of observing an uncensored observation at these values of X. As this scale factor moves closer to one - fewer censored observations - then the adjustment factor becomes unimportant and the coefficient β_k gives us the marginal effect at these particular values of X. Though not a formal result, this marginal effect suggests a reason why, in general, OLS estimates of the coefficients in a tobit model usually resemble the ML estimates multiplied by the proportion of uncensored observations in the sample.

Again, which of these marginal effects should be reported will depend on your purpose. Wooldridge recommends reporting both the marginal effects on E[y] and $E[y|y > 0]$.

4.6 Why not OLS?

Generally, OLS on the whole sample or just the uncensored sample will provide inconsistent estimates of β. This is relatively easy to see. Consider OLS on the uncensored sample. From (30), we have

$$\begin{aligned} y_i = X_i\beta + \sigma\lambda\left(\frac{X_i\beta}{\sigma}\right) + \epsilon_i \\ E[\epsilon_i|X_i, y_i > 0] = 0 \end{aligned} \tag{37}$$

This implies that $E[\epsilon_i|X_i, y_i > 0, \lambda_i] = 0$. Note that we mistakenly omit $\sigma\lambda\left(\frac{X_i\beta}{\sigma}\right)$ in our OLS regression. This implies that the effect of this omitted term will appear in the disturbance term, which means that the Xs will be correlated with the disturbance term, leading to inconsistent estimates.

Now consider OLS on the full sample. From the following equation,

$$E[y] = \Phi\left(\frac{X_i\beta}{\sigma}\right)[X_i\beta + \sigma\lambda(\alpha)] \tag{38}$$

we can see that OLS on the full sample will also produce inconsistent estimates because E[y] is a non-linear function of X, β, σ and OLS assumes linearity.

What's the relationship between OLS and the tobit model? If y and X are normally distributed and censoring is from below, it has been shown that the OLS slope parameters coverge to p times the true slope parameter, where p is the fraction of the sample that is uncensored. In practice, this proportionality result provides a good empirical approximation of the inconsistency of OLS if a tobit model is instead appropriate.

4.7 Tobit and Probit

The tobit and probit models are similar in many ways. They each have the same structural model, just different measurement models i.e. how the y^* is translated into the observed y is different. In the tobit model, we know the value of y^* when $y^* > 0$, while in the probit model we only know if $y^* > 0$. Since there is more information in the tobit model, the estimates of the βs should be more efficient. As Greene (2003, 776) points out, though, the probit estimates should be consistent for $\frac{1}{\sigma_{tobit}}(\beta_{tobit})$. In other words, if we multiply the probit coefficients by σ_{tobit}, we should get the tobit coefficients. Alternatively you can divide the tobit coefficients by σ_{tobit} to get the probit coefficients. This result will *only hold if* tobit is the correct model.

4.8 Some Assumptions

There are two basic assumptions underlying the tobit model. It turns out that if the disturbance ϵ_i is either heteroskedastic or non-normal, then the ML estimates are inconsistent. It is possible to get consistent estimates with heteroskedastic errors if the heteroskedasticity is modeled directly as in previous weeks i.e. $\sigma_i^2 = e^{Z_i \gamma}$. Note also that we are implicitly assuming that the same data generating process that determines the censoring is the same process that determines the outcome variable. The sample selection models that we are going to deal with next time allow you to specify a different model for the censoring and the outcome components.

References

Greene, William. 2003. *Econometric Analysis*. New Jersey: Prentice Hall.

Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. London: Sage Publications.

Sigelman, Lee & Langche Zeng. 1999. "Analyzing Censored and Sample-Selected Data with Tobit and Heckit Models." *Political Analysis* 8:167–182.

Wooldridge, Jeffrey. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press.