Censoring and Truncation

The Tobit and Heckman Models

November 20, 2014

Censoring, truncation, and sample selection

Occurs when a large portion of our sample's dependent variable is stacked on a particular value (often the value '0')[censoring], not measured at all[truncated], or not measured by a selection mechanism [sample selection]

- We don't observe anything if the dependent variable if the individual falls below (or above) a threshold level (truncation) Example: We only observe firm's profits if they are positive.
- We only observe a lower (or upper) threshold value for all dependent variables in sample if the "true" dependent variable is below (or above) a critical value (censoring). Example:The highest grade level I can assign is an "A". Different students may have different capabilities, but all the top students receive an "A".

For these kinds of problems, we will explore the truncated regression, the Tobit, and the Heckman models.

Types of Censoring/Truncations

The data has “meaningful” levels of observations being ‘stacked’ on some critical value of the dependent variable. This can take several forms:

- Lower Truncation: Wages are observed only if they are above minimum wage.
- Upper Censoring: Jury awards capped at $5 million dollars.
- Lower and Upper Truncation: Only children in the middle of the class ($\pm$ 1 Standard Deviation) are selected for an educational program.

OLS and Truncation Problems

Think of truncation as a sample selection issue: we only observe some part of the full sample. Three cases for consideration:

1. Sample is selected purely by random chance: OLS unbiased
2. Sample is selected based on value of x: OLS unbiased

$$\begin{aligned} E[\epsilon_i|X, s(X)] =& E[\epsilon_i|X, s(X)] \\ =& E[\epsilon_i|X] \\ =& 0 \end{aligned}$$

3. Sample is selected based on Y: OLS biased.

$$\begin{aligned} E[\epsilon_i|X, s(Y > a)] =& E[\epsilon_i|X, Y \geq a] \\ =& E[\epsilon_i|X, X\beta + \epsilon \geq a] \\ \neq& 0 \end{aligned}$$

Properties of Truncated Distributions

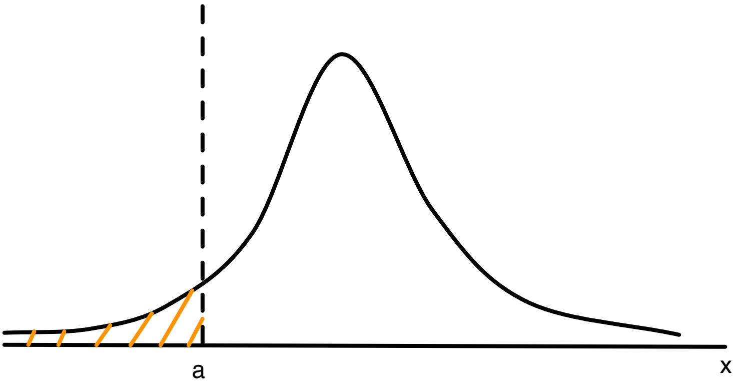
The PDF of a truncated distribution for the variable y can be written as (for Lower Truncation/Censoring)

$$f(y|y > a) = \frac{f(y)}{Prob(y > a)} \tag{1}$$

Most applications are based off the truncated normal distribution, making this expression:

$$f(y|y > a) = \frac{\frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)}{1 - \Phi(\frac{a-\mu}{\sigma})} \tag{2}$$

Note: Truncated Poissons have also been used for truncated count data problems.



$$Prob(y > a) = 1 - \int_{-\infty}^{a} f(y)dy \tag{3}$$

$$= 1 - \Phi\left(\frac{a-\mu}{\sigma}\right) \tag{4}$$

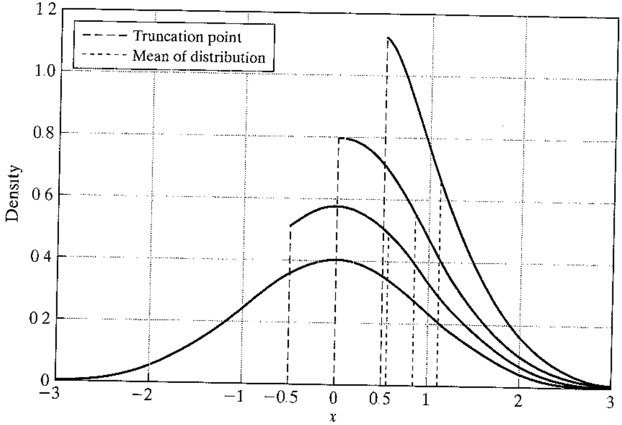


Figure: Truncated Normal

The expected value of a truncated distribution

Since we only observe y if y exceeds a, we can write

$$E(y|y>a) = \int_a^\infty yf(y|y>a)dy \tag{5}$$

$$= \int_a^\infty y \left[\frac{\frac{1}{\sigma}\phi(\frac{y-\mu}{\sigma})}{1-\Phi(\frac{a-\mu}{\sigma})} \right] dy \tag{6}$$

$$= \mu + \sigma \frac{\phi(\frac{a-\mu}{\sigma})}{1-\Phi(\frac{a-\mu}{\sigma})} \tag{7}$$

Note, the expected value of a truncated distribution is always greater (when truncated from below) than the non-truncated mean μ. It can also be shown that the variance of any truncated distribution is smaller than the variance of the non-truncated distribution σ^2.

Mills Ratio

The term $\frac{\phi(\frac{a-\mu}{\sigma})}{1-\Phi(\frac{a-\mu}{\sigma})}$ is known as the inverse of Mill's Ratio (or the Hazard Function) and provides information on the relevancy of the degree of truncation in the data. As the inverse of Mill's ratio gets larger, the mean of the truncated distribution diverges from the mean of the underlying full distribution. This occurs as the

1. Denominator gets smaller and smaller (as $Prob(y > a) \rightarrow 0$), or
2. μ occurs closer to the truncation point a.

Truncated Regression

If truncation is relevant, then an OLS model of the mean of a dependent variable is biased!! Instead, model the individual-specific mean of the truncated normal distribution as

$$\mu_i = \mathbf{x}_i \beta \tag{8}$$

where we are trying to uncover the relationship in the population regression model:

$$\mathbf{y} = \mathbf{x}\beta + \epsilon \tag{9}$$

But since we have a truncation problem, we need to write

$$E(y_i | y_i > a) = \mathbf{x}_i \beta + \sigma \frac{\phi(\frac{a - \mathbf{x}_i \beta}{\sigma})}{1 - \Phi(\frac{a - \mathbf{x}_i \beta}{\sigma})} \tag{10}$$

Likelihood Function for Truncated Regression Model

$$L = \prod_{i=1}^{N} \frac{\frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{x}_i\beta}{\sigma}\right)}{1 - \Phi(\alpha_i)} \tag{11}$$

Where

1. $\alpha_i = \frac{a - \mathbf{x}_i\beta}{\sigma}$
2. $\Phi(\cdot)$ and $\phi(\cdot)$ is the standard normal CDF and PDF respectively.

The log-likelihood is often expressed as:

$$LnL = \sum_{i=1}^{N} \left[log\left(\frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{x}_i\beta}{\sigma}\right)\right) - log(1 - \Phi(\alpha_i))\right] \tag{12}$$

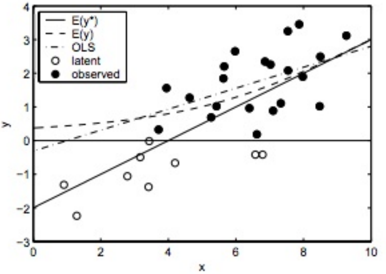
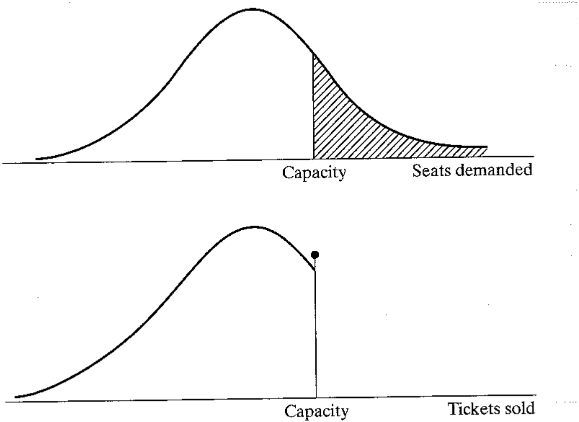


Figure: Truncated Data and Performance of OLS

Source: Schmidheiny.

Censoring



Note: Demand may have been higher for a game, but if so reported = Capacity

Model Setup

Here, we consider lower censoring at ‘a’. This can be generalized to any type of censoring. Let y_i^* reflect the unobserved dependent variable having the following relationship:

$$y_i^* = \mathbf{x_i}\beta + \epsilon_i \tag{13}$$

Except now we have a censoring problem, where the observed dependent variable, y_i is:

$$\begin{aligned} y_i =& a \text{ if } y_i^* \leq a \quad (14) \\ =& y_i^* \text{ if } y_i^* > a \quad (15) \end{aligned}$$

Expected Value of a Censored Variable

In general denote the lower censoring point as ‘a’

$$
\begin{aligned}
E(y_i) =& prob(y_i = a) \times E(y_i|y_i = a) && (16)\\
& + prob(y_i > a) \times E(y_i|y_i > a) && (17)\\
=& prob(y_i^* \leq a) \times a + prob(y_i^* > a) \times E(y_i^*|y_i^* > a) && (18)\\
=& \Phi\left(\frac{a-\mu}{\sigma}\right) a + \left(1 - \Phi\left(\frac{a-\mu}{\sigma}\right)\right) \left[\mu + \sigma \frac{\phi(\frac{a-\mu}{\sigma})}{1-\Phi(\frac{a-\mu}{\sigma})}\right] && (19)
\end{aligned}
$$

Note: the expression inside the large brackets is identical to that of a truncated distribution. Also note that in the censored regression model (The Tobit), we model $\mu_i = \mathbf{x}_i\beta$.

The Tobit Model

This framework leads us to the Tobit model, where (for the case of a=0) we can write the expected value of our dependent variable as

$$E(y_i|\mathbf{x}_i) = \left(1 - \Phi\left(\frac{0 - \mathbf{x}_i\beta}{\sigma}\right)\right)\left(\mathbf{x}_i\beta + \sigma\frac{\phi(\frac{0-\mathbf{x}_i\beta}{\sigma})}{1 - \Phi(\frac{0-\mathbf{x}_i\beta}{\sigma})}\right) \quad (20)$$

Likelihood Function for Tobit Censored Regression Model

$$lnL = \sum_{i=1}^{N} \left[d_i ln \left[\frac{1}{\sigma} \phi \left(\frac{y_i - \mathbf{x}_i \beta}{\sigma} \right) \right] + (1 - d_i) ln \left[1 - \Phi \left(\frac{\mathbf{x}_i \beta - a}{\sigma} \right) \right] \right] \tag{21}$$

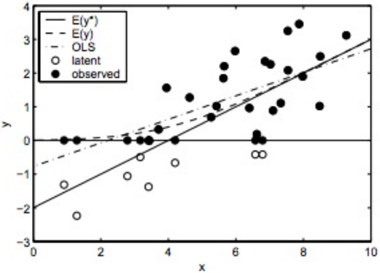


Figure: Censored Data and Performance of OLS

Source: Schmidheiny.

A Sample Selection Model

Consider a sample of women, some working and some not. We would like to estimate a labor supply equation for women, but want to be careful about non-working women. If we examine only those in the labor market, it may lead to results that *at best* are only applicable to the working population of women and *at worst* are biased!!

Model Setup

Consider a working/not working selection equation and, conditional on working, a wage equation.
Selection Equation:

$$z_i^* = \mathbf{w}_i\gamma + u_i, z_i = 1 \text{ if } z_i^* > 0 \text{ and } 0 \text{ otherwise} \tag{22}$$

$$Prob(z_i = 1|\mathbf{w}_i) = \Phi(\mathbf{w}_i\gamma) \tag{23}$$

$$Prob(z_i = 0|\mathbf{w}_i) = 1 - \Phi(\mathbf{w}_i\gamma) \tag{24}$$

Wage equation:

$$y_i = \mathbf{x}_i\beta + \epsilon_i \text{ observed if } , z_i = 1 \tag{25}$$

$$(u_i, \epsilon_i) \sim \text{ Bivariate Normal}(0, 0, 1, \sigma_\epsilon, \rho) \tag{26}$$

Conditional Expected Value of y_i in this setting

$$
\begin{aligned}
E(y_i|y_i \text{ observed}) =& E(y_i|z_i^* > 0) && (27)\\
=& E(y_i|u_i > -\mathbf{w}_i\gamma) && (28)\\
=& \mathbf{x}_i\beta + E(\epsilon_i|u_i > -\mathbf{w}_i\gamma) && (29)\\
=& \mathbf{x}_i\beta + \rho\sigma_\epsilon \frac{\phi(\mathbf{w}_i\gamma)}{\Phi(\mathbf{w}_i\gamma)} && (30)
\end{aligned}
$$

where the $Cov(\epsilon_i, u_i) = \rho$.

The *Unconditional* Expected Value of y_i in this setting

First, we need to define the selection probability:

$$P(y_i \text{ observed}) = \Phi(\mathbf{w}_i\gamma) \tag{31}$$

Then, the unconditional Expected value of y_i is

$$\begin{aligned} E(y_i) =& Prob(y_i \text{ observed}) E(y_i|z_i^* > 0) && (32) \\ =& \Phi(\mathbf{w}_i\gamma) \left[\mathbf{x}_i\beta + \rho\sigma_\epsilon \frac{\phi(\mathbf{w}_i\gamma)}{\Phi(\mathbf{w}_i\gamma)}\right] && (33) \end{aligned}$$

Note: The unconditional expectation of $E[y_i^*]$ (the full uncensored distribution is $\mathbf{x}_i\beta$). Green argues that for many contexts this won't be useful since we never observe y^* below the censoring point. There may be some applications where $E[y_i^*]$ is useful.

Marginal Effects and Stata

Note, there are several types of marginal effects one might consider:

1. The unconditional marginal effect: how y_i^* changes as x_k changes
2. The conditional marginal effect: how y_i changes as x_k changes
3. The probability of censoring marginal effect: how Φ changes as w_k changes

Make sure you know what stata is reporting when running mfx.

Mechanics of $\mathbf{z}$ and $\mathbf{x}$

Very important point about the Heckman Model:

- $\mathbf{w}_i$ and $\mathbf{x}_i$ may have some overlapping regressors. There must be some regressor that identifies the selection mechanism independent of the effect in the equation of interest to properly identify a model.
- Unlike the truncated regression model, we run this over the full sample of observations.

Likelihood Function for Heckman Sample Selection Regression Model

$$L = \prod_{i=1}^{N} \left[\frac{1}{\sigma} \phi \left(\frac{y_i - \mathbf{x_i b}}{\sigma} \right) \Phi \left(\frac{\mathbf{w}_i \gamma + \rho (\frac{y_i - \mathbf{x_i b}}{\sigma})}{\sqrt{1 - \rho^2}} \right) \right]^{d_i} \times \\ \left[(1 - \Phi \left(\mathbf{w}_i \gamma \right) \right]^{(1 - d_i)} \tag{34}$$

Note: Take the log of this expression for the log-likelihood:

$$LL = \sum_{i=1}^{N} d_i log \left[\frac{1}{\sigma} \phi \left(\frac{y_i - \mathbf{x_i b}}{\sigma} \right) \Phi \left(\frac{\mathbf{w}_i \gamma + \rho (\frac{y_i - \mathbf{x_i b}}{\sigma})}{\sqrt{1 - \rho^2}} \right) \right] + \\ (1 - d_i) log \left[(1 - \Phi \left(\mathbf{w}_i \gamma \right) \right] \tag{35}$$

Summary of Models

Rules of thumb for various models:

- If data exists for the entire sample (over the complete range of the DV) but the value of the dependent variable is transformed to some lower or upper limit because of an exogenous data collection rule, then the Tobit is the model to use.
- If data exists for only some subset of your sample based on values of the dependent variable and this was done based on an exogenous data collection rule, then the truncated regression model is appropriate.
- If data exists for the dependent variable on only some subset of your sample based on a sample selection rule, then the Heckman regression model is appropriate. This regression is run over the *full sample* of data.

Note:usually we are in a Tobit or Heckman world for most economic data.

Summary of Models, Cont.

- The Heckman Model allows the sample selection probability to have different regressors than the “continuous” equation for the dependent variable values that are not censored.
- The Tobit Model restricts the regressors in the censoring and “continuous” equation to be the same.
- The Truncated regression ignores sample selection and only leverages the “continuous” equation for recovering β
- If important economic phenomena occuring at the censoring point or there is endogenous selection, then the truncated regression model will lead to biased estimates of β.