The Tobit Model

David A. Hughes, Ph.D.

Auburn University at Montgomery

david.hughes@aum.edu

March 27, 2020

Overview

Introduction

- Thus far, we have largely been discussing categorical variables.
- Generally, when we have continuous-level variables, OLS remains the best available estimator.
- Nevertheless, as we saw in the case of event-counts, we still might need to be wary of OLS given underlying conditions in our dependent variable.
- Today, we'll discuss another common type of constraint to using OLS: limited outcomes.

Truncation

- Suppose we are interested in the percentage of the vote candidates earn in elections.
- Naturally, percentages are bounded on their upper limit by 100 and on their lower limit by 0.
- Variables such as these, where observations are limited due to the very nature of a variable's measurement are said to be *truncated*.

Censoring

- Suppose we want to know how much a consumer will spend on a given commodity (a new television for example). She has a budget of $100. But suppose further that every television at the store costs more than $100. She leaves empty-handed.
- This presents a problem known as *censoring*. The consumer's demand doesn't appear in our data (or appears to be zero), not because she didn't have demand, but because she was censored out of expressing it.
- Truncation is a problem in that it limits observations in the dependent variable. Censoring is a problem in that it constrains observations to reflect values that poorly reflect the variable of interest.

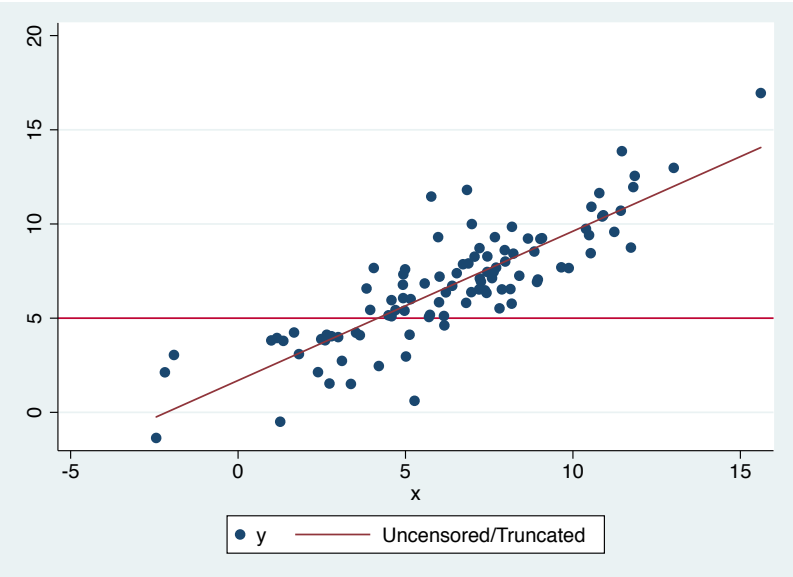
The basic setup

- Let y_i^* reflect an *uncensored* dependent variable that can take on any value over the real number line.
- Now suppose y_i is a *censored* dependent variable such that observations are censored if they are less than or equal to five.
- We can then characterize our uncertainty over the censored dependent variable as:

$$y_i = \begin{cases} y_i^*, & \text{if } y_i^* > 5 \\ 0, & \text{if } y_i^* \leq 5 \end{cases} \tag{1}$$

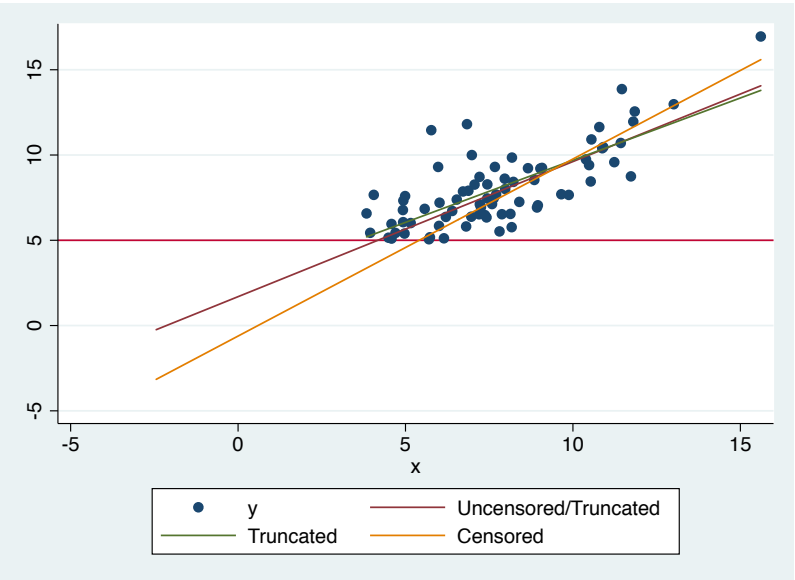
The problem

- Censoring and truncation can complicate inference with respect to the CLRM.
- Suppose we have an independent and dependent variable like those shown to the right.
- Without censoring or truncation, we get: $y_i = 1.69 + 0.79x_i + \epsilon_i$.



The problem (cont'd.)

- Suppose we truncate the data for all $y_i \leq 5$. OLS gives: $\hat{y}_i = 2.40 + 0.73x_i$.
- Now suppose we censor observations at five such that $y_i = 0, \forall y_i* \leq 5$. OLS gives: $\hat{y}_i = -0.62 + 1.04x_i$.



What is to be done?

- We could include the censored observations, but this has the effect of pulling down the intercept and increasing $\hat{\beta}_1$.
- We could omit the censored observations (i.e., truncate the data), but this has the effect of over-estimating the intercept and under-estimating $\hat{\beta}_1$.
- Or we could use maximum likelihood methods and model the problem directly.

The tobit model

- For the tobit model, we stick to the basic structure of the CLRM:

$$Y_i^* = \boldsymbol{X_i}\boldsymbol{\beta} + \epsilon_i \tag{2}$$

  where $Y_i^* \in \Re$ is a latent, variable.
- Let Y_i^* be observed for all values greater than τ such that:

$$Y_i = \begin{cases} Y_i^*, & \text{if } Y_i^* > \tau \\ \tau_Y, & \text{if } Y_i^* \leq \tau. \end{cases} \tag{3}$$

- This problem represents censoring from below, but we could just as easily rewrite it to reflect censoring from above (or both). We'll stick with below-censoring for simplicity.

The tobit model (cont'd.)

- Combining Equations 2 and 3, we get the following:

$$Y_i = \begin{cases} Y_i^* = \boldsymbol{X_i}\boldsymbol{\beta} + \epsilon, & \text{if } Y_i^* > \tau \\ \tau_Y, & \text{if } Y_i^* \leq \tau. \end{cases} \tag{4}$$

- Note that τ and τ_Y are conceptually distinct. The former is the threshold that establishes which observations are censored while the latter reflects the values the dependent variable takes when there is censoring.

The tobit model (cont'd.)

- The probability that an observation is censored depends upon the proportion of ϵ that falls below τ.
- Put differently, the probability of a case being censored for a given value of X is the area of the normal distribution less than or equal to τ:

$$\begin{aligned} Pr(\text{Censored} \mid \boldsymbol{X_i}) &= Pr(Y_i^* \leq \tau \mid \boldsymbol{X_i}) \\ &= Pr(\epsilon_i \leq \tau - \boldsymbol{X_i}\boldsymbol{\beta} \mid \boldsymbol{X_i}). \end{aligned} \quad (5)$$

The tobit model (cont'd.)

- Note that $\epsilon \sim N(0, \sigma^2)$. Therefore, $\frac{\epsilon}{\sigma}$ is distributed as: $\frac{\epsilon}{\sigma} \sim N(0, 1)$.
- We can rewrite Equation 5 as:

$$\begin{aligned} Pr(\text{Censored} \mid \boldsymbol{X_i}) &= Pr\left(\frac{\epsilon_i}{\sigma} \leq \frac{\tau - \boldsymbol{X_i\beta}}{\sigma} \mid \boldsymbol{X_i}\right) \\ &= \Phi\left(\frac{\tau - \boldsymbol{X_i\beta}}{\sigma}\right). \end{aligned} \tag{6}$$

The tobit model (cont'd.)

- To simplify Equation 6, let:

$$\delta_i = \frac{\boldsymbol{X_i}\boldsymbol{\beta} - \tau}{\sigma}.$$

- Then:

$$Pr(\text{Censored} \mid \boldsymbol{X_i}) = \Phi(-\delta_i) \quad (7)$$

$$Pr(\text{Uncensored} \mid \boldsymbol{X_i}) = \Phi(\delta_i). \quad (8)$$

The tobit model (cont'd.)

- The tobit model is therefore highly similar to the probit.
- In tobit, we know the value of Y_i^* for all values greater than τ while in probit, all observations are technically censored.
- Therefore, tobit is more efficient than probit is. Furthermore, we can estimate the variance in Y_i^* in tobit whereas we must assume it is equal to one in probit.

Estimating the tobit model

- To derive the maximum likelihood estimator, we divide the data into two sets: those than are uncensored, which ML treats in the same way as the CLRM, and those that are censored.
- For the latter group, we do not know the value of Y_i^*. Nevertheless, we can compute the probability of being in the censored group and use this quantity informatively in the likelihood function.

Estimating the tobit model (cont'd.)

- For uncensored observations:

$$Y_i = \boldsymbol{X_i}\boldsymbol{\beta} + \epsilon_i, \forall\ Y_i^* > \tau, \tag{9}$$

  where $\epsilon_i \sim N(0, \sigma^2)$.
- The log-likelihood function for uncensored observations can be expressed as:

$$lnL_U(\boldsymbol{\beta}, \sigma^2) = \sum_{\text{uncensored}} ln\frac{1}{\sigma}\phi\left(\frac{Y_i - \boldsymbol{X_i}\boldsymbol{\beta}}{\sigma}\right) \tag{10}$$

Estimating the tobit model (cont'd.)

- For censored observations:

$$Pr(Y_i^* \leq \tau \mid \boldsymbol{X_i}) = \Phi\left(\frac{\tau - \boldsymbol{X_i}\boldsymbol{\beta}}{\sigma}\right). \tag{11}$$

- We can express the likelihood function for censored observations as:

$$lnL_C(\boldsymbol{\beta}, \sigma^2) = \sum_{\text{censored}} ln\Phi\left(\frac{\tau - \boldsymbol{X_i}\boldsymbol{\beta}}{\sigma}\right). \tag{12}$$

Estimating the tobit model (cont'd.)

- Combining Equations 10 and 11, we get:

$$lnL(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{Y_i}, \boldsymbol{X_i}) = lnL_U(\boldsymbol{\beta}, \sigma^2) + lnL_C(\boldsymbol{\beta}, \sigma^2). \quad (13)$$

- So long as errors are homoskedastistic and and normally distributed, the standard ML assumptions hold.

Some example data

- Let's consider information about graduate school applicants' GRE scores.
- The range on these scores is 200 to 800. The data are censored because for all students who score an 800 or a 200, we can't distinguish among them.
- For predictor variables, we'll look at students' undergraduate GPAs and the reputation of their undergraduate institution (dichotomous).

Tobit in Stata

- We can estimate tobit regression models in Stata using the command `tobit`:
  `tobit y x1 x2 ... xk [if ], ul() ll() [options]`
- Using this template, we have `ul` to denote the upper-limit and `ll` to denote the lower-limit for the dependent variable.

Sample Stata output

```
. tobit gre top gpa, ll(200) ul(800)

Tobit regression                                  Number of obs     =        400
                                                  LR chi2(2)        =      70.93
                                                  Prob > chi2       =     0.0000
Log likelihood = -2331.4314                       Pseudo R2         =     0.0150

------------------------------------------------------------------------------
         gre |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    topnotch |   46.65774   15.75356     2.96   0.003     15.68716    77.62833
         gpa |   111.3085   15.19665     7.32   0.000     81.43273    141.1842
       _cons |   205.8515   51.24073     4.02   0.000     105.1152    306.5879
-------------+----------------------------------------------------------------
      /sigma |   111.4882   4.143727                      103.3419    119.6345
------------------------------------------------------------------------------
           0  left-censored observations
         375     uncensored observations
          25 right-censored observations at gre >= 800
```

Interpretation of tobit output

- Conveniently, interpreting changes in Y_i^* (the latent outcome) is the same as the CLRM:

$$E(Y_i^* \mid \boldsymbol{X_i}) \quad = \quad \boldsymbol{X_i}\boldsymbol{\beta}.$$

- Therefore, we can interpret the effect of a given variable, X_k on Y_i in the traditional way:

$$\frac{\partial \hat{Y}_i^*}{\partial X_k} \quad = \quad \hat{\beta}_k.$$

- Interpreting changes in the truncated or censored outcomes is a little trickier.

Changes in the truncated outcome

- The outcome, Y_i is undefined when it is truncated.
- The expected value of a truncated outcome is:

$$E(Y_i^T \mid Y_i > \tau, \boldsymbol{X_i}) = \boldsymbol{X_i\beta} + \sigma\lambda(\delta), \tag{14}$$

  where $\lambda(\cdot) = \frac{\phi(\cdot)}{\Phi(\cdot)}$ and $\delta = \frac{\boldsymbol{X_i\beta} - \tau}{\sigma}$.
- Then the effect of X_i on Y_i can be expressed as:

$$\frac{\partial \hat{Y}_i^T}{\partial X_k} = \beta_k\left[1 - \delta\lambda(\delta) - \lambda(\delta)^2\right]. \tag{15}$$

- The quantity in brackets in Equation 15 falls in the interval 0 to 1. It can be shown that as $\boldsymbol{X_i\beta}$ increases, $\frac{\partial Y^T}{\partial X_k} \approx \frac{\partial Y^*}{\partial X_k}$.

Changes in the censored outcome

- When the dependent variable is censored, observations of Y_i are equal to τ_Y.
- If, for example, $\tau_Y = 0$, then:

$$E(Y_i^C \mid \boldsymbol{X_i}) = \Phi(\delta)\boldsymbol{X_i}\boldsymbol{\beta}_\sigma\phi(\delta) + \Phi(-\delta)\tau_Y. \quad (16)$$

- Then the effect of X_i on Y_i can be expressed as:

$$\frac{\partial \hat{Y}_i^C}{\partial X_k} = \Phi(\delta)\beta_k + (\tau - \tau_Y)\phi(\delta)\frac{\beta_k}{\sigma}. \quad (17)$$

- If $\tau = \tau_Y$, then we get:

$$\frac{\partial \hat{Y}_i^C}{\partial X_k} = \Phi(\delta)\beta_k = Pr(\text{Uncensored} \mid \boldsymbol{X})\beta_k. \quad (18)$$

- As the probability a case is censored approaches 0, then $\frac{\partial Y^C}{\partial X_k} \approx \frac{\partial Y^*}{\partial X_k}$.

Discussion

- Censoring and truncation occur with many types of dependent variables we would ordinarily reach to OLS to examine.
- Nevertheless, failing to account for these limitations in the dependent variable can lead to inconsistent results using the CLRM.
- The tobit model addresses this problem and is desirable in that its coefficients are largely interpreted like OLS coefficients.