Chapter 4: Censoring, Truncation, and Selection

Joan Llull

Microeconometrics
IDEA PhD Program

Introduction

Introduction

In this chapter we review models that deal with **censored**, **truncated** and **self-selected** data.

We consider a latent variable that is described by a **linear** model:

$$y^* = \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon.$$

Although this assumption can be trivially relaxed, we **assume** (unless when otherwise noted) that:

$$\varepsilon | \boldsymbol{x} \sim \mathcal{N}(0, \sigma^2).$$

Truncation

We say that our sample is **(left) truncated** whenever:

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

The **problem** with these data is that:

$$\mathbb{E}[y|\boldsymbol{x}] = \mathbb{E}[y^*|\boldsymbol{x}, y^* \text{ is observed}] = \boldsymbol{x}'\boldsymbol{\beta} + \mathbb{E}[\varepsilon|\boldsymbol{x}, \varepsilon > -\boldsymbol{x}'\boldsymbol{\beta}].$$

We analyze **left truncation**. Right truncation is analogous.

Censoring

We say that our sample is **(left) censored** whenever:

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

The **problem** with these data is that:

$$\mathbb{E}[y|\boldsymbol{x}] = \mathbb{E}\left[\mathbb{E}(y|\boldsymbol{x}, d)|\boldsymbol{x}\right] = \Pr[\varepsilon > -\boldsymbol{x}'\boldsymbol{\beta}|\boldsymbol{x}](\boldsymbol{x}'\boldsymbol{\beta} + \mathbb{E}[\varepsilon|\boldsymbol{x}, \varepsilon > -\boldsymbol{x}'\boldsymbol{\beta}]),$$

where $d = \mathbb{1}\{y^* > 0\}$.

Selection

We say that our sample is **self-selected** whenever:

$$y = \begin{cases} y^* & \text{if } \boldsymbol{z}'\boldsymbol{\gamma} + \nu > 0 \\ - & \text{otherwise,} \end{cases}$$

and $d \equiv \mathbb{1}\{\boldsymbol{z}'\boldsymbol{\gamma} + \nu > 0\}$ is observed.

The **problem** with these data is that:

$$\mathbb{E}[y|\boldsymbol{x}] = \mathbb{E}[y^*|\boldsymbol{x}, d = 1] = \boldsymbol{x}'\boldsymbol{\beta} + \mathbb{E}[\varepsilon|\boldsymbol{z}'\boldsymbol{\gamma} + \nu > 0].$$

Endogenous selection whenever $\mathbb{E}[\varepsilon|\boldsymbol{z}'\boldsymbol{\gamma} + \nu > 0] \neq 0$.

Censoring and truncation. The Tobit model

Maximum Likelihood Estimation (truncation)

The individual likelihood of an observation from a **(left) truncated** sample is:

$$g(y|\boldsymbol{x}, y > 0) = \frac{f(y|\boldsymbol{x})}{\Pr[y > 0|\boldsymbol{x}]} = \frac{f(y|\boldsymbol{x})}{1 - F(0|\boldsymbol{x})}.$$

Hence, the **log-likelihood** function is:

$$\mathcal{L}_{\mathrm{N}}(\theta) = \sum_{\mathrm{i=1}}^{\mathrm{N}} \left\{ \ln f(y_i|\boldsymbol{x_i}) - \ln(1 - F(0|\boldsymbol{x}_i)) \right\}.$$

In the **Tobit** model (given by **normality** assumption):

$$\mathcal{L}_{\mathrm{N}}(\theta) = \sum_{\mathrm{i=1}}^{\mathrm{N}} \left\{ -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 - \ln\Phi\left(\frac{\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right) \right\}.$$

Maximum Likelihood Estimation (censoring)

The individual likelihood of an observation from a **(left) censored** sample is:

$$g(y|\boldsymbol{x}, y > 0) = f(y|\boldsymbol{x})^d F(0|\boldsymbol{x})^{1-d} = \begin{cases} f(y|\boldsymbol{x}) & \text{if } y^* > 0 \\ F(0|\boldsymbol{x}) & \text{if } y^* \leq 0. \end{cases}$$

Hence, the **log-likelihood** function is:

$$\mathcal{L}_{\mathrm{N}}(\theta) = \sum_{\mathrm{i=1}}^{\mathrm{N}} \left\{ d_i \ln f(y_i|\boldsymbol{x_i}) + (1 - d_i) \ln(F(0|\boldsymbol{x}_i)) \right\}.$$

In the **Tobit** model (given by **normality** assumption):

$$\mathcal{L}_{\mathrm{N}}(\theta) = \sum_{\mathrm{i=1}}^{\mathrm{N}} \left\{ d_i \left(-\frac{1}{2} \ln 2\pi - \frac{1}{2} \ln \sigma^2 - \frac{(y_i - \boldsymbol{x}_i' \boldsymbol{\beta})^2}{2\sigma^2} \right) + (1 - d_i) \ln \left(1 - \Phi \left(\frac{\boldsymbol{x}_i' \boldsymbol{\beta}}{\sigma} \right) \right) \right\}.$$

Potential inconsistence of MLE

As usual, consistency requires that the likelihood is **correctly specified**.

Potentially more severe here. Intuition can be seen in the **FOC** for $\boldsymbol{\beta}$ in the **truncated data**:

$$\frac{\partial \mathcal{L}_{\mathrm{N}}}{\partial \boldsymbol{\beta}} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{1}{\sigma^2} \left(y_i - \boldsymbol{x}_i' \boldsymbol{\beta} - \sigma \lambda \left(\frac{\boldsymbol{x}_i' \boldsymbol{\beta}}{\sigma} \right) \right) \boldsymbol{x}_i = \mathbf{0},$$

where $\lambda(z) = \phi(z)/\Phi(z) = \mathbb{E}[\epsilon | \epsilon > -z]$ if $\epsilon \sim \mathcal{N}(0, 1)$.

Heteroscedastic or **non-normal** errors lead to biased estimates.

Alternatives (I): Heckman two-step procedure

This is usually used for **self-selected** samples —however, censoring is a **special case**.

Relies on:

$$\mathbb{E}[y|\boldsymbol{x}, y > 0] = \boldsymbol{x}'\boldsymbol{\beta} + \sigma\lambda\left(\frac{\boldsymbol{x}'\boldsymbol{\beta}}{\sigma}\right).$$

Two steps:

- Estimate $\boldsymbol{\alpha} = \boldsymbol{\beta}/\sigma$ from a **Probit** over $d = \mathbb{1}\{y^* > 0\}$.
- Use $\hat{\boldsymbol{\alpha}}$ to compute $\lambda(\boldsymbol{x}'\hat{\boldsymbol{\alpha}})$, and include it as a control in a **linear regression** estimated with **uncensored observations** to identify $\boldsymbol{\beta}$ and σ.

Alternatives (II): Median Regression

If censoring is **below the median** of y we can still make inference on the median.

If the distribution is **symmetric** (e.g. normal distribution), the mean and the median coincide.

The Censored Least Absolute Deviations estimator is:

$$\hat{\boldsymbol{\beta}}_{CLAD} = \arg\min_{\boldsymbol{\beta}} N^{-1} \sum_{\mathrm{i=1}}^{\mathrm{N}} |y_i - \max(\boldsymbol{x}_i'\boldsymbol{\beta}, 0)|.$$

Alternatives (III): Symmetrically Trimmed Mean

Key assumption: the distribution of $\varepsilon|\boldsymbol{x}$ is **symmetric** around 0.

We **do not consider** observations with $\boldsymbol{x}'\boldsymbol{\beta} < 0$.

Probability that $\boldsymbol{x}'\boldsymbol{\beta} + \varepsilon < 0$ (and hence the **observation is censored**) is equal to that of $\boldsymbol{x}'\boldsymbol{\beta} + \varepsilon > 2\boldsymbol{x}'\boldsymbol{\beta}$. Hence:

$$\mathbb{E}[\mathbb{1}\{\boldsymbol{x}'\boldsymbol{\beta} > \mathbf{0}\}(\min(y, 2\boldsymbol{x}'\boldsymbol{\beta}) - \boldsymbol{x}'\boldsymbol{\beta})\boldsymbol{x}] = \mathbf{0}.$$

This estimator delivers FOC that are **sample analogs** to the previous moment conditions:

$$\hat{\boldsymbol{\beta}}_{STM} = \arg\min_{\boldsymbol{\beta}} \frac{1}{N} \sum_{\mathrm{i=1}}^{\mathrm{N}} \left\{ \left[y_i - \max\left(\frac{y_i}{2}, \boldsymbol{x}'\boldsymbol{\beta} \right) \right]^2 + \mathbb{1}\{y_i > 2\boldsymbol{x}_i'\boldsymbol{\beta}\} \left[\frac{y_i^2}{4} - \max(0, \boldsymbol{x}'\boldsymbol{\beta}) \right]^2 \right\}.$$

Selection

The Sample Selection Model

The **model** is defined by:

$$y^* = \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon \qquad \begin{pmatrix} \varepsilon \\ \nu \end{pmatrix} \Big| \boldsymbol{z}, \boldsymbol{x} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{bmatrix} \right).$$

$$d = \mathbb{1}\{\boldsymbol{z}'\boldsymbol{\gamma} + \nu\}$$

We only **observe** $y = y^* \times d$.

Without loss of generality, $\boldsymbol{x}$ is **included** in $\boldsymbol{z}$.

The **likelihood** of our sample is:

$$L_N(\theta) = \prod_{i=1}^{N} \left(1 - \Phi(\boldsymbol{z}'\boldsymbol{\gamma})\right)^{1-d} \left\{ f(y^*|\boldsymbol{z}) \Pr(d = 1|y^*, \boldsymbol{z}) \right\}^d,$$

where:

$$f(y^*|\boldsymbol{z}) = \frac{1}{\sigma}\phi\left(\frac{y^* - \boldsymbol{x}'\boldsymbol{\beta}}{\sigma}\right) \text{ and } \Pr(d = 1|y^*, \boldsymbol{z}) = \Phi\left(\frac{\boldsymbol{z}'\boldsymbol{\gamma} + \frac{\rho}{\sigma}(y^* - \boldsymbol{x}'\boldsymbol{\beta})}{\sqrt{1-\rho^2}}\right).$$

(Note that when evaluating this, we observe y^* as we only have to evaluate it when $d = 1$).

Heckman two-step procedure (Heckit)

Same idea as we have seen for **Tobit**.

Relies on:

$$\mathbb{E}[y|\boldsymbol{x}, d = 1] = \boldsymbol{x}'\boldsymbol{\beta} + \rho\sigma\lambda\left(\boldsymbol{z}'\boldsymbol{\gamma}\right).$$

Two steps:

- Estimate $\boldsymbol{\gamma}$ from a **Probit** over $d = \mathbb{1}\{\boldsymbol{z}'\boldsymbol{\gamma} + \nu > 0\}$.
- Include $\lambda(\boldsymbol{z}'\hat{\boldsymbol{\gamma}})$ as a control in a **linear regression** with the observations for which the **outcome is observed** to estimate $\boldsymbol{\beta}$ and $\rho\sigma$.

Heckman two-step procedure: Remarks

Important remarks to make:

- **Standard errors** have to control for the particular form of heteroskedasticity and for using $\hat{\boldsymbol{\gamma}}$: **bootstrap**.

- **Credible identification** requires excluded variables in $\boldsymbol{z}$:

$$\mathbb{E}[y|\boldsymbol{z}] \approx \boldsymbol{x}'\boldsymbol{\beta} + a + b(\boldsymbol{z}'\hat{\boldsymbol{\gamma}}).$$

- It is a **LIML** estimation using, in the previous likelihood:

$$f(y^*, d = 1|\boldsymbol{z}) = f(y^*|d = 1, \boldsymbol{z}) \Pr(d = 1|\boldsymbol{z}).$$

- **Test** for selection: $\rho \neq 0$.

- Alternative: **semi-parametric** estimation of $g(\boldsymbol{z}'\boldsymbol{\gamma})$ vs $\lambda(\boldsymbol{z}'\boldsymbol{\gamma})$.