Censoring, Truncation, and Selection

Joan Llull

MOVE, Universitat Autònoma de Barcelona
and Barcelona School of Economics

I. Introduction

In this Chapter we review models that deal with censored, truncated and selected data. This introductory section, distinguishes the three concepts. In all cases, we consider a latent variable that is described by a linear model, and that is only partially observed. Hence, our latent variable y^* is given by:

$$y^* = \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon. \tag{1}$$

We define (left) *truncation* as the situation in which we only observe y^* if it is above certain threshold. In particular, we observe:

$$y = \begin{cases} y^* & \text{if } y^* > 0 \\ - & \text{if } y^* \leq 0, \end{cases} \tag{2}$$

where – indicates that the observation is missing. The threshold is normalized to zero without loss of generality, as in discrete choice.[1] What follows is for the case of left-truncation; results for right truncation are analogous.

Alternatively, we have (left) *censored data* in the similar situation in which, when y^* is below the threshold, we observe the individuals (and, eventually, the regressors), but not y^*:

$$y = \begin{cases} y^* & \text{if } y^* > 0 \\ 0 & \text{if } y^* \leq 0. \end{cases} \tag{3}$$

Finally, we have a *selected* sample if y^* is only observable for a particular (non-representative) subset of the population, which is observable. In this case:

$$y = \begin{cases} y^* & \text{if } \boldsymbol{z}'\boldsymbol{\gamma} + \nu > 0 \\ - & \text{otherwise,} \end{cases} \tag{4}$$

and $d \equiv \mathbb{1}\{\boldsymbol{z}'\boldsymbol{\gamma} + \nu > 0\}$ is observed, where $\boldsymbol{z}$ typically includes at least one variable that is not included in $\boldsymbol{x}$ (exclusion restriction). In this case, when the condition for observing y^* is not satisfied, we still observe the characteristics of the individual.

[1] The threshold could be individual-specific, $L = \boldsymbol{x}'\boldsymbol{\delta}$, in which case the normalization implies that we would identify $\boldsymbol{\beta}^* = \boldsymbol{\beta} - \boldsymbol{\delta}$ instead of $\boldsymbol{\beta}$.

The problem with these situations is that Least Squares estimation no longer provides consistent estimates of $\boldsymbol{\beta}$, even if the linear model is correctly specified. The bias comes from the fact that $\mathbb{E}[y|\boldsymbol{x}]$ is not be equal to $\boldsymbol{x}'\boldsymbol{\beta}$.[2] In the (left) truncation case:

$$\mathbb{E}[y|\boldsymbol{x}] = \mathbb{E}[y^*|\boldsymbol{x}, y^* \text{ is observed}] = \boldsymbol{x}'\boldsymbol{\beta} + \mathbb{E}[\varepsilon|\boldsymbol{x}, \varepsilon > -\boldsymbol{x}'\boldsymbol{\beta}]. \tag{5}$$

The bias comes from the fact that $\mathbb{E}[\varepsilon|\varepsilon > -\boldsymbol{x}'\boldsymbol{\beta}] > 0$. In the case of (left) censoring, defining $d = \mathbb{1}\{y^* > 0\}$:

$$\mathbb{E}[y|\boldsymbol{x}] = \mathbb{E}\left[\mathbb{E}(y|\boldsymbol{x}, d)\right] = \Pr[\varepsilon > -\boldsymbol{x}'\boldsymbol{\beta}|\boldsymbol{x}](\boldsymbol{x}'\boldsymbol{\beta} + \mathbb{E}[\varepsilon|\boldsymbol{x}, \varepsilon > -\boldsymbol{x}'\boldsymbol{\beta}]), \tag{6}$$

where we used the fact that $\mathbb{E}[y|\boldsymbol{x}, d = 0] = 0$. Finally, in the case of a selected sample, using $d = \mathbb{1}\{\boldsymbol{z}'\boldsymbol{\gamma} + \nu > 0\}$:

$$\mathbb{E}[y|\boldsymbol{x}] = \mathbb{E}[y^*|\boldsymbol{x}, d = 1] = \boldsymbol{x}'\boldsymbol{\beta} + \mathbb{E}[\varepsilon|\boldsymbol{z}'\boldsymbol{\gamma} + \nu > 0]. \tag{7}$$

In this latter case, there is only a bias if $\mathbb{E}[\varepsilon|\boldsymbol{z}'\boldsymbol{\gamma} + \nu > 0] \neq 0$, in which case we talk about *endogenous* selection; otherwise, selection would be exogenous and would cause no biases.

II. Censoring and Truncation. The Tobit Model

In this section we deal with the cases of censored and truncated data, and the next section covers the analysis of selected data. In general, unless otherwise noted, in this chapter we assume $\varepsilon|\boldsymbol{x} \sim \mathcal{N}(0, \sigma^2)$, but this assumption may be trivially relaxed.

A. Maximum Likelihood Estimation

Both for the cases of censored and truncated data, we can write the likelihood function if we assume a distribution (we assumed normality) for the error term. If we have truncated data, the density of y given $\boldsymbol{x}$ a truncated distribution:

$$g(y|\boldsymbol{x}, y > 0) = \frac{f(y|\boldsymbol{x})}{\Pr[y > 0|\boldsymbol{x}]} = \frac{f(y|\boldsymbol{x})}{1 - F(0|\boldsymbol{x})}. \tag{8}$$

Hence, the log-likelihood function is:

$$\mathcal{L}_{\mathrm{N}}(\theta) = \sum_{\mathrm{i}=1}^{\mathrm{N}} \left\{\ln f(y_i|\boldsymbol{x_i}) - \ln(1 - F(0|\boldsymbol{x}_i))\right\}, \tag{9}$$

[2] Consistency requires $\mathbb{E}[y|\boldsymbol{x}] = \boldsymbol{x}'\boldsymbol{\beta}$ so that $E[\hat{\boldsymbol{\beta}}_{OLS}] = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}']^{-1}\,\mathbb{E}[\boldsymbol{x}\,\mathbb{E}[y|\boldsymbol{x}]] = \boldsymbol{\beta}$.

where θ includes $\boldsymbol{\beta}$ and σ^2, and where 0 can instead be replaced by any threshold L_i (with the aforementioned identification considerations). In the case of the Tobit model, the assumption of $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ gives:

$$\mathcal{L}_{\mathrm{N}}(\theta) = \sum_{\mathrm{i}=1}^{\mathrm{N}} \left\{ -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 - \ln \Phi\left(\frac{\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right) \right\}, \tag{10}$$

where we exploited the fact that $\Phi(-\boldsymbol{x}_i'\boldsymbol{\beta}/\sigma) = 1 - \Phi(\boldsymbol{x}_i'\boldsymbol{\beta}/\sigma)$.

In the case of censored data, we have:

$$g(y|\boldsymbol{x}, y > 0) = f(y|\boldsymbol{x})^d F(0|\boldsymbol{x})^{1-d} = \begin{cases} f(y|\boldsymbol{x}) & \text{if } y^* > 0 \\ F(0|\boldsymbol{x}) & \text{if } y^* \leq 0, \end{cases} \tag{11}$$

where $d = \mathbb{1}\{y^* > 0\}$ as we defined before. Hence, the log-likelihood function is:

$$\mathcal{L}_{\mathrm{N}}(\theta) = \sum_{\mathrm{i}=1}^{\mathrm{N}} \left\{ d_i \ln f(y_i|\boldsymbol{x_i}) + (1 - d_i)\ln(F(0|\boldsymbol{x}_i)) \right\}, \tag{12}$$

and, in the Tobit case:

$$\mathcal{L}_{\mathrm{N}}(\theta) = \sum_{\mathrm{i}=1}^{\mathrm{N}} \left\{ d_i \left(-\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln \sigma^2 - \frac{(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2\sigma^2} \right) + (1 - d_i)\ln\left(1 - \Phi\left(\frac{\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right) \right\}. \tag{13}$$

B. Potential Inconsistency of the MLE

The maximum likelihood estimators described above are only consistent if the distribution of the residual is correctly specified. This is as usual in ML estimation, but in this case it is particularly severe. In particular, *heteroscedastic* or *non-normal* errors cause important biases.

In order to get a sense of how important are distributional assumptions in this context, consider the first order conditions for parameter $\boldsymbol{\beta}$ in the truncated sample case:

$$\frac{\partial \mathcal{L}_{\mathrm{N}}}{\partial \boldsymbol{\beta}} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{1}{\sigma^2}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta} - \sigma\lambda\left(\frac{\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right)\boldsymbol{x}_i = \mathbf{0}, \tag{14}$$

where $\lambda(z) = \phi(z)/\Phi(z) = \mathbb{E}[\epsilon|\epsilon > -z]$ if $\epsilon \sim \mathcal{N}(0,1)$.[3] Note that this coincides with the first order condition of a linear regression model in which we add $\sigma\lambda(\boldsymbol{x}_i'\boldsymbol{\beta}/\sigma)$ as a control. Hence, we are including two combinations of the same single index, and identification relies entirely on the choice of the functional form.

[3] This is known as the inverse Mills ratio, and it is centerpiece in the next section.

C. Alternative Methods for Censored Data

Heckman Two-Step Estimator The two-step procedure proposed by Heckman (1976, 1979) is commonly used in the context of self-selection. We cover it in detail when we talk about selection models. However, the analysis of censored samples can be seen as a special case of self-selection, and, hence, this method might be applied.

The mean of y conditional on observing $y > 0$ is:

$$\mathbb{E}[y|\boldsymbol{x}, y^* > 0] = \boldsymbol{x}'\boldsymbol{\beta} + \sigma\lambda\left(\frac{\boldsymbol{x}'\boldsymbol{\beta}}{\sigma}\right). \tag{15}$$

The two-step procedure is as follows: i) estimate $\boldsymbol{\alpha} = \boldsymbol{\beta}/\sigma$ from a Probit over $d \equiv \mathbb{1}\{y > 0\}$; and ii) 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.

Median Regression If we have censored data with less than half of the observations being censored, we can still make inference on the *median* of the distribution. Under a symmetric distribution for y, the median and the mean of the distribution is governed by the same single-index. This estimator, first proposed by Powell (1984), is known as the *censored least absolute deviations (CLAD)* estimator, and is given by:

$$\hat{\boldsymbol{\beta}}_{CLAD} = \arg\min_{\boldsymbol{\beta}} N^{-1} \sum_{\mathrm{i}=1}^{\mathrm{N}} |y_i - \max(\boldsymbol{x}_i'\boldsymbol{\beta}, 0)|. \tag{16}$$

Symmetrically Trimmed Mean Assume that $\varepsilon|\boldsymbol{x}$ is symmetrically distributed, and, again, that less than half of the observations are censored. Under this assumption we can use the information included in the observations with $\boldsymbol{x}'\boldsymbol{\beta} > 0$ (others do not include information on the central part of the distribution). The probability that, if $\boldsymbol{x}'\boldsymbol{\beta} > 0$, then $\boldsymbol{x}'\boldsymbol{\beta} + \varepsilon < 0 \Leftrightarrow \varepsilon < -\boldsymbol{x}'\boldsymbol{\beta}$ (i.e. the observation is censored) is equal to the probability of $\varepsilon > \boldsymbol{x}'\boldsymbol{\beta}$ or, equivalently, $y = \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon > 2\boldsymbol{x}'\boldsymbol{\beta}$.

The Symmetrically Trimmed Mean estimator keeps only observations with positive mean ($\boldsymbol{x}'\boldsymbol{\beta} > 0$), and artificially right-censors these data to compensate for the left censoring. Given the symmetry, the following moment conditions are hence satisfied:

$$\mathbb{E}[\mathbb{1}\{\boldsymbol{x}'\boldsymbol{\beta} > \mathbf{0}\}(\min(y, 2\boldsymbol{x}'\boldsymbol{\beta}) - \boldsymbol{x}'\boldsymbol{\beta})\boldsymbol{x}] = \mathbf{0}. \tag{17}$$

Note that $\min(y, 2\boldsymbol{x}'\boldsymbol{\beta})$ is equal to zero if $y^* \leq 0$, to y^* whenever $y^* \in [0, 2\boldsymbol{x}'\boldsymbol{\beta}]$, and to $2\boldsymbol{x}'\boldsymbol{\beta}$ if $y^* \geq 2\boldsymbol{x}'\boldsymbol{\beta}$, thus implying symmetric censoring from both tails.

A moment-based estimation relying on these moments, however, provides multiple solutions for $\hat{\boldsymbol{\beta}}$. Alternatively, it can be proved that the minimization of the following criterion delivers first order conditions that are sample analogs to the previous moments:

$$\hat{\boldsymbol{\beta}}_{STM} = \arg\min_{\boldsymbol{\beta}} \frac{1}{N} \sum_{\mathrm{i}=1}^{\mathrm{N}} \left\{ \left[y_i - \max\left(\frac{y_i}{2}, \boldsymbol{x}'\boldsymbol{\beta} \right) \right]^2 + \mathbb{1}\{y_i > 2\boldsymbol{x}_i'\boldsymbol{\beta}\} \left[\frac{y_i^2}{4} - \max(0, \boldsymbol{x}'\boldsymbol{\beta}) \right]^2 \right\}. \tag{18}$$

III. Selection

A. *The Sample Selection Model*

Consider the sample selection situation presented in the introduction. $d = \mathbb{1}\{\boldsymbol{z}'\boldsymbol{\gamma} + \nu > 0\}$ is an indicator variable that equals 1 if individual's outcome y^* is observed. Hence, we observe $y = y^* \times d$. Without loss of generality, $\boldsymbol{z}$ includes $\boldsymbol{x}$ and some other regressors. These additional regressors (and importantly the fact that $\nu \neq \varepsilon$) are the main difference between this model and the Tobit model. We discuss the importance of these additional regressors below.

We assume:

$$\begin{pmatrix} \varepsilon \\ \nu \end{pmatrix} \Big| \boldsymbol{z} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{bmatrix} \right), \tag{19}$$

where the normalization of the variance of ν is due to the identification problem of the Probit model for d. The likelihood of our sample is:

$$L_N(\theta) = \prod_{i=1}^{N} \left(1 - \Phi(\boldsymbol{z}_i'\boldsymbol{\gamma})\right)^{1-d_i} \left\{ f(y_i|\boldsymbol{z}_i) \Pr(d_i = 1|y_i, \boldsymbol{z}_i) \right\}^{d_i}, \tag{20}$$

where:

$$f(y|\boldsymbol{z}) = \frac{1}{\sigma}\phi\left(\frac{y - \boldsymbol{x}'\boldsymbol{\beta}}{\sigma} \right) \text{ and } \Pr(d = 1|y, \boldsymbol{z}) = \Phi\left(\frac{\boldsymbol{z}'\boldsymbol{\gamma} + \frac{\rho}{\sigma}(y - \boldsymbol{x}'\boldsymbol{\beta})}{\sqrt{1 - \rho^2}} \right). \tag{21}$$

Notice that we have made use of the conditional distribution of ν given $\boldsymbol{z}$ and ε. The parameters from the model can be estimated from this likelihood by ML. If $\rho = 0$, this likelihood boils down to the product of an OLS regression likelihood under normality and a Probit likelihood.

B. *Heckman Two-Step Estimator*

As we pointed out in the introduction, $\mathbb{E}[y|\boldsymbol{z}] = \mathbb{E}[y^*|d = 1, \boldsymbol{z}] \neq \boldsymbol{x}'\boldsymbol{\beta}$, and hence, OLS is inconsistent. Given joint normality, we can write $\varepsilon = \rho\sigma\nu + \xi$,

where $\xi \sim \mathcal{N}(0, \sigma^2(1-\rho^2))$ independent of ν; then, building on what we already developed, we have:

$$\begin{aligned}\mathbb{E}[y^*|d=1,\boldsymbol{z}] &= \boldsymbol{x}'\boldsymbol{\beta} + \mathbb{E}[\varepsilon|\boldsymbol{z}'\boldsymbol{\gamma}+\nu>0] \\ &= \boldsymbol{x}'\boldsymbol{\beta} + \mathbb{E}[\rho\sigma\nu+\xi|\boldsymbol{z}'\boldsymbol{\gamma}+\nu>0] \\ &= \boldsymbol{x}'\boldsymbol{\beta} + \rho\sigma\lambda\left(\boldsymbol{z}'\boldsymbol{\gamma}\right), \end{aligned} \tag{22}$$

where $\lambda(z) = \phi(z)/\Phi(z)$ as described above. The Heckman two-step procedure, also known as the *Heckit*, consists of a first stage in which $\lambda\left(\boldsymbol{z}'\hat{\boldsymbol{\gamma}}\right)$ is constructed through predictions of a Probit estimation of $\boldsymbol{\gamma}$, and a second stage which consists of an OLS estimation of the linear model augmented with the constructed regressor estimated in the first stage.

Both OLS and White's heteroscedasticity-robust standard errors are incorrect in this context. Consistent estimates of the standard errors need to take into account both the particular form of heteroskedasticity of the new error term (after controlling for $\lambda(\boldsymbol{z}'\hat{\boldsymbol{\gamma}})$), and, more importantly, for the fact that we used $\hat{\boldsymbol{\gamma}}$ instead of $\boldsymbol{\gamma}$. The asymptotic formulas for such standard errors are not straightforward, and it is common to proceed with bootstrap to estimate standard errors.

It is crucial for identification that $\boldsymbol{z}$ includes some regressors that are not in $\boldsymbol{x}$. Put differently, we need some variation in the selection equation that only affects the outcome through the selection, but not directly through the outcome equation. These are called *exclusion restrictions*. If we do not have them, we are in a situation that is analogous to the Tobit model. Although identification is theoretically achieved (given the functional form assumptions), the inverse Mills ratio $\lambda(.)$ is approximately linear over a wide range of its argument. Hence, especially when there is little variation in $\boldsymbol{z}'\boldsymbol{\gamma}$, we are approximately estimating:

$$\mathbb{E}[y|\boldsymbol{z}] \approx \boldsymbol{x}'\boldsymbol{\beta} + a + b\boldsymbol{z}'\hat{\boldsymbol{\gamma}}, \tag{23}$$

which, in the case of $\boldsymbol{z} = \boldsymbol{x}$ may induce multicollinearity. This need of an exclusion restriction may difficult the possibility of implementing the selection correction, because sometimes it is not easy to find such an excluded variable.

There are also other important remarks to make about this method. The first one is that this two-step method is nothing else than the LIML estimation of the problem in which we specify the likelihood for the observations in which the outcome is observed using the factorization $f(y, d=1|\boldsymbol{z}) = f(y|d=1,\boldsymbol{z})\Pr(d=1|\boldsymbol{z})$ instead of the one used in equation (20). As such, estimates are less precise than if it was estimated with FIML. However, estimation is much simpler. The

second one is that this estimator itself provides an interesting way of testing for endogenous selection. Endogenous selection is provided by $\rho \neq 0$. A simple t-test of the null hypothesis $\rho\sigma = 0$ delivers the result. Finally, the third remark is that we can either change the functional form assumptions or produce a semi-parametric estimate of $\lambda(\boldsymbol{z}'\boldsymbol{\gamma})$ to control for the endogeneity induced by selection.

References

Amemiya, Takeshi (1985), *Advanced Econometrics*, Blackwell, chapter 10.

Cameron, A. Colin and Pravin K. Triverdi (2005), *Microeconometrics: Methods and Applications*, Cambridge University Press, chapter 16.

Heckman James J. (1979), "Sample Selection Bias and Specification Error", *Econometrica*, 47, 153-161.

Powell, James L. (1986), "Symmetrically Trimmed Least Squares Estimation for Tobit Models", *Econometrica*, 54, 1435-1460.

Tobin, James (1958), "Estimation of Relationships for Limited Dependent Variables", *Econometrica*, 26, 24-36.

Wooldridge, Jeffrey M. (2002), *Econometric Analysis of Cross Section and Panel Data*, The MIT Press, chapters 16 and 17.