Censored and Truncated Outcome Panel Data Models

230347 Advanced Microeconometrics
Tilburg University

Christoph Walsh

Censored Data

- y_{it} is censored when it is partly continuous but has positive probability mass at one or more points.
  - For example, y_{it} is continuous when $y_{it} > 0$ but has a large mass at $y_{it} = 0$.
- We can sometimes think of the underlying model as:

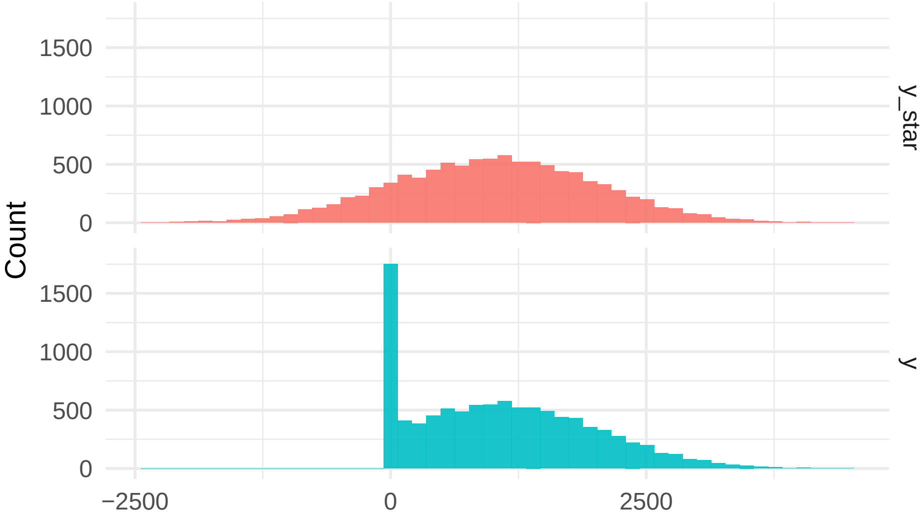
$$y_{it}^{\star} = \alpha_i + \boldsymbol{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}$$

  but we observe:

$$y_{it} = \begin{cases} y_{it}^{\star} & \text{if } y_{it}^{\star} > \underline{y} \\ \underline{y} & \text{if } y_{it}^{\star} \leq \underline{y} \end{cases} \qquad \text{or} \qquad y_{it} = \begin{cases} \overline{y} & \text{if } y_{it}^{\star} \geq \overline{y} \\ y_{it}^{\star} & \text{if } y_{it}^{\star} < \overline{y} \end{cases}$$

  - For example, top-coded income.
- Other times we can think of $\overline{y}$ or $\underline{y}$ as a corner solution of an optimization problem.
  - For example, hours worked, firm expenditure on R&D.

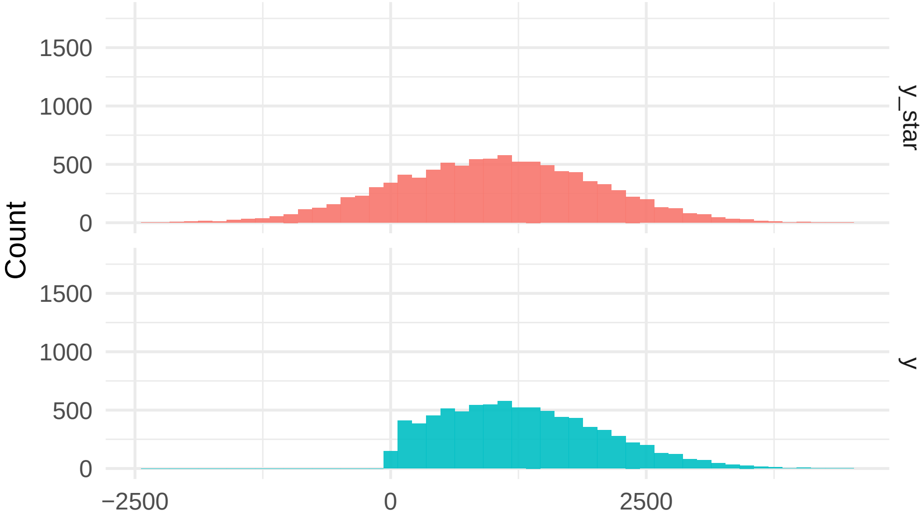
Histogram of censored y_{it} left of zero



Truncated Data

- Our sample may be **truncated**, where our sample only has observations where $y_{it} > \underline{y}$ or $y_{it} < \overline{y}$
- For example, we may only observe people who work.

Histogram of truncated y_{it}



Models in this Topic

- Static Censored Random Effects
- Static Truncated Fixed Effects
- It is possible to estimate Static & Dynamic Censored Fixed Effects models, but we won't cover them here.

Censored Data: Panel Random Effects Tobit Model

- We consider the left-censored data case where $\underline{y} = 0$.
- We observe:

$$y_{it} = \begin{cases} y_{it}^{\star} & \text{if } y_{it}^{\star} > 0 \\ 0 & \text{if } y_{it}^{\star} \leq 0 \end{cases}$$

- Let $d_{it} = \mathbb{1}\left\{y_{it} > 0\right\}$.
- If $\varepsilon_{it} \sim \mathcal{N}\left(0, \sigma_{\varepsilon}^2\right)$, then using $\varepsilon_{it} = y_{it}^{\star} - \alpha_i - \boldsymbol{x}_{it}'\boldsymbol{\beta}$, the joint conditional density of $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT})$ is

$$f\left(\boldsymbol{y}_i \,\middle|\, \boldsymbol{X}_i, \alpha_i, \boldsymbol{\beta}, \sigma_{\varepsilon}^2\right) = \prod_{t=1}^{T} \left[\frac{1}{\sigma_{\varepsilon}} \phi\left(\frac{y_{it} - \alpha_i - \boldsymbol{x}_{it}'\boldsymbol{\beta}}{\sigma_{\varepsilon}}\right)\right]^{d_{it}} \left[1 - \Phi\left(\frac{\alpha_i + \boldsymbol{x}_{it}'\boldsymbol{\beta}}{\sigma_{\varepsilon}}\right)\right]^{1-d_{it}}$$

where $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT})$ and ϕ and Φ are the pdf and cdf of the standard normal distribution respectively.

Censored Data: Panel Random Effects Tobit Model

- If we model $\alpha_i \sim \mathcal{N}\left(0, \sigma_\alpha^2\right)$, then we can integrate out the α_i:

$$f\left(\boldsymbol{y}_i \,\middle|\, \boldsymbol{X}_i, \boldsymbol{\beta}, \sigma_\varepsilon^2, \sigma_\alpha^2\right) = \int_{-\infty}^{\infty} f\left(\boldsymbol{y}_i \,\middle|\, \boldsymbol{X}_i, \alpha_i, \boldsymbol{\beta}, \sigma_\varepsilon^2\right) \frac{1}{\sqrt{2\pi\sigma_\alpha^2}} \exp\left(\frac{-\alpha_i^2}{2\sigma_\alpha^2}\right) d\alpha_i$$

- There is no closed-form solution for the likelihood and therefore needs to be computed using simulation methods.
- We can perform the same change of variables as with the probit random effects and approximate the integral with Gauss-Hermite quadrature.

Truncated Fixed Effects: Only observe y_{it} when $y_{it}^\star > 0$

- When data are truncated, we cannot eliminate the fixed effects by differencing or mean differencing.
- For observed y_{it}:

$$\begin{aligned} y_{it} &= \mathbb{E}\left[y_{it}^\star | \boldsymbol{x}_{it}, \alpha_i, y_{it}^\star > 0\right] + \nu_{it} \\ &= \alpha_i + \boldsymbol{x}_{it}'\boldsymbol{\beta} + \mathbb{E}\left[\varepsilon_{it} | \varepsilon_{it} > -\alpha_i - \boldsymbol{x}_{it}'\boldsymbol{\beta}\right] + \nu_{it} \end{aligned}$$

- Consider the $T = 2$ case. Taking differences:

$$\begin{gathered} y_{i2} - y_{i1} = \left(\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1}\right)' \boldsymbol{\beta} + \mathbb{E}\left[\varepsilon_{i2} | \varepsilon_{i2} > -\alpha_i - \boldsymbol{x}_{i2}'\boldsymbol{\beta}\right] - \\ \mathbb{E}\left[\varepsilon_{i1} | \varepsilon_{i1} > -\alpha_i - \boldsymbol{x}_{i1}'\boldsymbol{\beta}\right] + \nu_{i2} - \nu_{i1} \end{gathered}$$

- In general, this still depends on α_i (unless $\boldsymbol{x}_{i1} = \boldsymbol{x}_{i2}$)

Honoré (1992)

- Suppose we restricted our analysis to observations satisfying

$$y_{i1} \geq -\left(\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1}\right)' \boldsymbol{\beta} \qquad \text{and} \qquad y_{i2} \geq \left(\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1}\right)' \boldsymbol{\beta}$$

- Suppose that $\left(\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1}\right)' \boldsymbol{\beta} > 0$ ($\exists$ similar argument for the opposite case). Then:

$$\begin{aligned}
&\mathbb{E}\left[y_{i2} | \boldsymbol{x}_{i2}, \alpha_i, y_{i2} \geq \left(\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1}\right)' \boldsymbol{\beta}\right] \\
&= \alpha_i + \mathbf{x}_{i2}'\boldsymbol{\beta} + \mathbb{E}\left[\varepsilon_{i2} | \varepsilon_{i2} \geq -\alpha_i - \mathbf{x}_{i2}'\boldsymbol{\beta} + \left(\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1}\right)' \boldsymbol{\beta}\right] \\
&= \alpha_i + \mathbf{x}_{i2}'\boldsymbol{\beta} + \mathbb{E}\left[\varepsilon_{i2} | \varepsilon_{i2} \geq -\alpha_i - \mathbf{x}_{i1}'\boldsymbol{\beta}\right]
\end{aligned}$$

- Since $\left(\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1}\right)' \boldsymbol{\beta} > 0$, the restriction doesn't bind for y_{i1}:

$$\begin{aligned}
\mathbb{E}\left[y_{i1} | \boldsymbol{x}_{i1}, \alpha_i, y_{i1} \geq -\left(\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1}\right)' \boldsymbol{\beta}\right] &= \mathbb{E}\left[y_{i1} | \boldsymbol{x}_{i1}, \alpha_i, y_{i1} \geq 0\right] \\
&= \alpha_i + \mathbf{x}_{i1}'\boldsymbol{\beta} + \mathbb{E}\left[\varepsilon_{i1} | \varepsilon_{i1} \geq -\alpha_i - \mathbf{x}_{i1}'\boldsymbol{\beta}\right]
\end{aligned}$$

Honoré (1992)

- If we assume the $\varepsilon_{it}|\boldsymbol{x}_{it}, \alpha_i$ are iid, then:

$$\mathbb{E}\left[\varepsilon_{i1}|\varepsilon_{i1} \geq -\alpha_i - \boldsymbol{x}'_{i1}\boldsymbol{\beta}\right] = \mathbb{E}\left[\varepsilon_{i2}|\varepsilon_{i2} \geq -\alpha_i - \boldsymbol{x}'_{i1}\boldsymbol{\beta}\right]$$

- Therefore

$$\begin{aligned}
\mathbb{E}\left[y_{i1}|\boldsymbol{x}_{i1}, \alpha_i, y_{i1} \geq -\left(\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1}\right)' \boldsymbol{\beta}\right] &= \alpha_i + \boldsymbol{x}'_{i1}\boldsymbol{\beta} + \mathbb{E}\left[\varepsilon_{i1}|\varepsilon_{i1} \geq -\alpha_i - \boldsymbol{x}'_{i1}\boldsymbol{\beta}\right] \\
\mathbb{E}\left[y_{i2}|\boldsymbol{x}_{i2}, \alpha_i, y_{i2} \geq \left(\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1}\right)' \boldsymbol{\beta}\right] &= \alpha_i + \boldsymbol{x}'_{i2}\boldsymbol{\beta} + \mathbb{E}\left[\varepsilon_{i1}|\varepsilon_{i1} \geq -\alpha_i - \boldsymbol{x}'_{i1}\boldsymbol{\beta}\right]
\end{aligned}$$

- Together:

$$\begin{aligned}
&\mathbb{E}\left[y_{i2} - y_{i1}|\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \alpha_i, y_{i1} \geq -\left(\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1}\right)' \boldsymbol{\beta}, y_{i2} \geq \left(\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1}\right)' \boldsymbol{\beta}\right] \\
&= \left(\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1}\right)' \boldsymbol{\beta}
\end{aligned}$$

  which no longer depends on the fixed effect α_i.

- This only requires the iid assumption. We don't assume anything about the distribution of ε_{it}.

Honoré (1992): Estimation when $T = 2$

- If we knew the true $\boldsymbol{\beta}$, we could estimate it with OLS in the model:

$$y_{i2} - y_{i1} = (\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1})' \boldsymbol{\beta} + \nu_{i2} - \nu_{i1}$$

  using the sample where:
  - $y_{i1} \geq -(\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1})' \boldsymbol{\beta}$
  - $y_{i2} \geq (\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1})' \boldsymbol{\beta}$
- However, we do not know $\boldsymbol{\beta}$.

Honoré (1992): Estimation

- Honoré (1992) proposes the following objective:

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \Big\{ & \left[y_{i2} - y_{i1} - (\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1})' \boldsymbol{\beta} \right]^2 \\
& \times \mathbb{1} \left\{ y_{i1} \geq -(\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1})' \boldsymbol{\beta}, y_{i2} \geq (\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1})' \boldsymbol{\beta} \right\} \\
& + y_{i1}^2 \mathbb{1} \left\{ y_{i1} \geq -(\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1})' \boldsymbol{\beta}, y_{i2} < (\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1})' \boldsymbol{\beta} \right\} \\
& + y_{i2}^2 \mathbb{1} \left\{ y_{i1} < -(\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1})' \boldsymbol{\beta}, y_{i2} \geq (\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1})' \boldsymbol{\beta} \right\} \Big\}
\end{aligned}
$$

Honoré (1992): Estimation

- Why the 2nd and 3rd term?
- Consider the single regressor case.
- Suppose we estimated β by minimizing:

$$\sum_{i=1}^{N} \left[y_{i2} - y_{i1} - (x_{i2} - x_{i1})\,\beta\right]^2 \mathbb{1}\left\{y_{i1} \geq -(x_{i2} - x_{i1})\,\beta, y_{i2} \geq (x_{i2} - x_{i1})\,\beta\right\}$$

- By setting β sufficiently large or small, no y_{i1} and y_{i2} will satisfy $y_{i1} \geq -(x_{i2} - x_{i1})\,\beta$ and $y_{i2} \geq (x_{i2} - x_{i1})\,\beta$ simultaneously for any i.
- The objective function would then be zero, its lowest possible value.
- The inclusion of the 2nd and 3rd term excludes these trivial solutions.

Reading and References

- Cameron and Trivedi 23.5 for Random effects Tobit.
- Hsiao 8.4 and Honoré (1992) for Truncated Fixed Effects.

References:

HONORÉ, B. E. (1992): "Trimmed LAD and least squares estimation of truncated and censored regression models with fixed effects," *Econometrica*, 533–565.