Recent Advances in the Field of Trade Theory and Policy Analysis Using Micro-Level Data

July 2012
Bangkok, Thailand

Cosimo Beverelli
(World Trade Organization)
Content

a) Censoring and truncation

b) Tobit (censored regression) model

c) Alternative estimators for censored regression models

d) Sample selection models

e) An example with firm-level analysis
a) Censoring and truncation

- Censoring
- Truncation
Censoring

• We want to estimate the effect of $x$ on a continuous variable $y^*$ (latent dependent variable)

• We always observe $x$ but we observe the dependent variable only above a lower threshold $L$ (censoring from below) or below an upper threshold $U$ (censoring from above)

• Censoring from below (or left):

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

• Example: exports by firm $i$ are equal to the export value if the export value exceeds $L$, or equal to $L$ if the export value is lower than $L$

• Censoring from above (or right):

$$y = \begin{cases} y^* & \text{if } y^* < U \\ U & \text{if } y^* \geq U \end{cases}$$

• Example: recorded exports are top-coded at $U$. Exports by firm $i$ are equal to the export value if the export value is below $U$, or equal to $U$ if the export value is above $U$
Truncation

• We want to estimate the effect of $x$ on a continuous variable $y^*$ (latent dependent variable)

• Truncation from below (or left):

\[ y = y^* \text{ if } y^* > L \]

• All information below $L$ is lost
• Example: exports by firm $i$ are reported only if the export value is larger than $L$

• Truncation from above (or right):

\[ y = y^* \text{ if } y^* < U \]

• All information above $U$ is lost
• Example: in a consumer survey, only low-income individuals are sampled
b) Tobit (censored regression) model

- Assumptions and estimation
- Why OLS estimation is inconsistent
- Marginal effects (ME) in Tobit
- Problems with Tobit
- Tobit model with panel data
- Example: academic attitude
Assumptions and estimation

$$y^* = x' \beta + \varepsilon$$

where

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

- This implies that the latent variable is also normally distributed:
  $$y^* \sim \mathcal{N}(x' \beta, \sigma^2)$$
- We observe:
  $$y = \begin{cases} 
  y^* & \text{if } y^* > 0 \\
  0 & \text{if } y^* \leq 0 
  \end{cases}$$
- Tobit estimator is a MLE, where the log-likelihood function is detailed, for instance, here
Why OLS estimation is inconsistent

1. OLS estimation on the sample of positive observations:

\[ E[y|x] = E[y^*|x, y^* > 0] = x' \beta + E[\varepsilon|x, \varepsilon > -x' \beta] \]

- Under the normality assumption: \( \varepsilon|x \sim N(0, \sigma^2) \), the second term becomes \( \sigma \lambda \left( \frac{x' \beta}{\sigma} \right) \), where \( \lambda(\cdot) \equiv \frac{\phi(\cdot)}{\Phi(\cdot)} \) is the inverse Mills ratio
- If we run an OLS regression on the sample of positive observations, then we should also include in the regression the term \( \lambda \left( \frac{x' \beta}{\sigma} \right) \)
- A failure to do so will result in an inconsistent estimate of \( \beta \) due to omitted variable bias (\( \lambda(\cdot) \) and \( x \) are correlated in the selected sub-population)
2. OLS estimation on the censored sample (zero and positive observations)

\[ E[y|x] = \Pr[y^* > 0] \times E[y^*|x, y^* > 0] = \Pr[\varepsilon > -x'\beta]\{x'\beta + E[\varepsilon|\varepsilon > -x'\beta]\} \]

- Under the normality assumption: \( \varepsilon \sim N(0, \sigma^2) \), the first term is \( \Phi\left(\frac{x'\beta}{\sigma}\right) \) and the term in curly brackets is the same as in the previous slide.
- There is no way to consistently estimate \( \beta \) in a linear regression.
Marginal effects (ME) in Tobit

- For the latent variable:
  \[
  \frac{\partial E[y^*|x]}{\partial x_j} = \beta_j
  \]
  (1)

- This is the marginal effect of interest if censoring is just an artifact of data collection (for instance, top- or bottom-coded dependent variable)

- In a model of hours worked, (1) is the effect on the desired hours of work

- Two other marginal effects can be of interest:
  1. ME on actual hours of work for workers: \( \frac{\partial E[y,y>0|x]}{\partial x_j} \)
  2. ME on actual hours of work for workers and non-workers: \( \frac{\partial E[y|x]}{\partial x_j} \)

- The latter is equal to \( \Phi \left( \frac{x'\beta}{\sigma} \right) \beta_j \) and can be decomposed in two parts:
  - Effect on the conditional mean in the uncensored part of the distribution
  - Effect on the probability that an observation will be positive (not censored)
Problems with Tobit

- Consistency crucially depends on normality and homostkedasticity of errors (and of the latent variable)
- The structure is too restrictive: exactly the same variables affecting the probability of a non-zero observation determine the level of a positive observation and, moreover, with the same sign
- There are many examples in economics where this implication does not hold
  - For instance, the intensive and extensive margins of exporting may be affected by different variables
With panel data (each individual $i$ is observed $t$ times), the natural extension of the Tobit models is:

$$y_{it}^* = \alpha_i + x'_{it}\beta + \epsilon_{it}$$

where $\epsilon_{it} \sim N(0, \sigma^2)$ and we observe:

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

Due to the incidental parameters problem, fixed effects estimation of $\beta$ is inconsistent, and there is no simple differencing or conditioning method.

Honoré’s semiparametric (trimmed LAD) estimator (pantob in Stata)

Random effects estimation assumes that $\alpha_i \sim N(0, \sigma^2_{\alpha})$ (xttobit, re in Stata)
Example: academic attitude

- Hypothetical data file, with 200 observations
- The academic aptitude variable is apt, the reading and math test scores are read and math respectively
- The variable prog is the type of program the student is in, it is a categorical (nominal) variable that takes on three values, academic (prog = 1), general (prog = 2), and vocational (prog = 3)
- apt is right-censored:
  - Summarize apt, d
  - histogram apt, discrete freq
- Tobit model with right-censoring at 800:
  - tobit apt read math i.prog, ul(800) vce(robust)
c) Alternative estimators for censored regression models

- Two semi-parametric methods:
  
  1. Censored least absolute deviations (CLAD)
     - Based on conditional median (\textit{clad} in Stata)
  
  2. Symmetrically censored least squares (SCLS)
     - Based on symmetrically trimmed mean (\textit{scls} in Stata)
d) Sample selection models

- Sources of selection
- Bivariate sample selection model (Type 2 Tobit)
- Heckman two-step estimator
- Identification issues
- Selection models with panel data
Sources of selection

• Self-selection occurs when the outcome of interest is determined in part by individual choice of whether or not to participate to the activity of interest (e.g., exporting)

• Sample selection occurs when there is over-sampling of those who participate in the activity of interest
  • Extreme case: there is sampling of participants only

Consider exporting activity

• The Melitz model implies that there is self-selection of most productive firms in exporting
Bivariate sample selection model (Type 2 Tobit)

- Participation equation (e.g., decision to export):
  \[ y_1 = \begin{cases} 
  1 & \text{if } y_1^* > 0 \\
  0 & \text{if } y_1^* \leq 0 
  \end{cases} \]

- Outcome equation (e.g., export value):
  \[ y_2 = \begin{cases} 
  y_2^* & \text{if } y_1^* > 0 \\
  - & \text{if } y_1^* \leq 0 
  \end{cases} \]

- Linear model for the latent variables \( y_1^* \) and \( y_2^* \):
  \[ y_1^* = x_1' \beta_1 + \varepsilon_1 \]
  \[ y_2^* = x_2' \beta_2 + \varepsilon_2 \]

- Tobit is a special case where \( y_1^* = y_2^* \) (participation and outcome are determined by the same factors)

- This model is also called probit selection equation
Bivariate sample selection model (ct’d)

• We assume bivariate normality for $\varepsilon_1, \varepsilon_2$:

$$\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \sim \mathcal{N}_2 \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} \right)$$

where $\sigma_1^2 = 1$ by normalization

• The (truncated) mean of the sample selection model where only positive values of $y_2$ are used is:

$$E[y_2|x, y_1^* > 0] = E[x'_2 \beta_2 + \varepsilon_2 | x'_1 \beta_1 + \varepsilon_1 > 0]$$

$$= x'_2 \beta_2 + E[\varepsilon_2 | \varepsilon_1 > -x'_1 \beta_1]$$

• Notice that if $\varepsilon_1$ and $\varepsilon_2$ are independent, the second term is equal to zero and OLS regression of $y_2$ on $x_2$ gives consistent estimates of $\beta_2$
Heckman estimator

• Heckman (1979) showed that:

$$E[y_2|x, y_1^* > 0] = x_2' \beta_2 + \sigma_{12} \lambda(x_1' \beta_1)$$

where $\lambda(\cdot)$ is the inverse Mills ratio

• Heckman’s two step estimator is an augmented OLS regression on the model:

$$y_{2i} = x_{2i}' + \sigma_{12} \lambda(x_{1i}' \hat{\beta}_1) + u_i$$

where:

• Positive values of $y_2$ are used
• $\hat{\beta}_1$ is estimated by first-step probit regression

• A test on $\hat{\sigma}_{12}$ is a test of whether the errors are correlated and sample selection correction is needed
Identification issues

• The model is theoretically identified without any restriction on regressors.
• Exactly the same regressors can appear in $x_1$ and $x_2$...
• ...However, this leads to multicollinearity problems.
• The more variation in $x_1' \hat{\beta}_1$ (good performance of the Probit model) the less severe this issue.
• However, estimation may require that at least one regressor in the participation equation be excluded from the outcome equation (exclusion restriction).
• The exclusion restriction is strictly necessary in semi-parametric versions of the Heckman two-step method.
With panel data (each individual \(i\) is observed \(t\) times), the natural extension of the bivariate sample selection model models is:

\[
\begin{align*}
y_{it}^{*} &= \alpha_i + x'_{it} \beta + \varepsilon_{it} \\
d_{it}^{*} &= \delta_i + z'_{it} \gamma + \nu_{it}
\end{align*}
\]

where \(y_{it} = y_{it}^{*}\) is observed if \(d_{it}^{*} > 0\) and not observed otherwise.

Random effects estimation with ML (Hausman and Wise, 1979) assumes that the four unobservables are normally distributed.

Fixed effects is inconsistent in short panels...

However if we believe that \(d_{it}^{*} = \delta_i\) (selection only depends on observable and unobservable individual characteristics that do not vary over time), then the FE estimator in the model \(y_{it} = \alpha_i + x'_{it} \beta + \varepsilon_{it}\) is consistent.

Intuition: FE estimator “controls” for individual characteristics that determine selection.
e) An example with firm-level analysis

Sun (2009)

- Heckman two-stages model of export participation and export intensity of Chinese firms
- The focus is on the effects of innovation and foreign direct investment (FDI) on export decisions

- Two-step decision:
  - Whether to export (participation equation)
  - How much to export (outcome equation)

- In the sample, one-half of the firms report no exports
Sun (2009)

• For identification purposes, he assumes that the number of years participating in exporting between 2000 and 2003 (a variable ranging from 0 to 4) only explains export participation, not export intensity

• This variable signals the fixed export cost, and hence the more frequently the firm participates in exporting the more likely it will continue to export

• Nevertheless as the fixed export cost has been paid and become sunk, it should not affect how much the firm is willing to export

• Hence, it is reasonable (for the author) to exclude from the export intensity equation the number of firms participating in exporting in the four years
Sun (2009) (ct’d)

• He estimates three models:
  1. With the full set of explanatory variables (multicollinearity...)
  2. Because of the multicollinearity issue, he re-runs the model dropping some interactions terms
  3. As a robustness check, he also runs a Tobit model, which accounts for the non-participation of exporting but imposes the restriction that explanatory variables have equal effect on both export participation and export intensity decisions

• The magnitude of the estimated coefficients display some differences, but the signs do not change
1. The geographic location of firms determines whether their export intensity rises or falls with industry-level FDI (measured by foreign presence)

<table>
<thead>
<tr>
<th></th>
<th>Coastal China</th>
<th>Central China</th>
<th>West China</th>
</tr>
</thead>
<tbody>
<tr>
<td>Privately owned</td>
<td>-0.0408</td>
<td>0.7743</td>
<td>-1.7714</td>
</tr>
<tr>
<td>State and collectively owned</td>
<td>0.3892</td>
<td>1.2043</td>
<td>-1.3414</td>
</tr>
</tbody>
</table>

Note:
Foreign presence is measured in terms of output share.
2. FDI (measured by foreign presence) affects firms’ export intensity differently