Topics

Limited Dependent Variables

- Limited Distribution
    - Truncated Distribution
    - Censored Distribution
- Nonnormal Distribution
    - Generalized Linear Models
    - Models for Counted Data
        - Poisson Regression Models
        - Negative Binomial Regression Model
        - Zero Inflated Poisson Regression Model

By Tatre Jantarakolica

Limited Dependent Variables

1. Truncated Sample

   Some observations are not available for observe. However, those observations do exist but weren't observed. Only truncated data are available for the regression.

2. Censored Samples

   The observations can all be observed. However, some observations have no value for observe, hence, report as 0. Dependent variable may be censored, but censored observation can be included in regression.

Limited Dependent Variables

Why do we have truncation?

a. Truncation by survey design:

e.g., families whose incomes are greater than that threshold are dropped from the sample.

b. Incidental truncation:

e.g., wage offer married women, only those who are working has wage information. It is the people's decision, not the survey's design, that determines the sample selection.

Limited Dependent Variables

Types of Sample Selection → Truncation:

#1. Sample selection is done randomly.

s is sample selection variable and independent of ε and x. $E\left(\varepsilon \middle| x_i s\right) = E\left(\varepsilon \middle| \mathrm{x}\right) = 0$

OLS is unbiased.

#2. Sample is selected based only on value of x.

s is deterministic function of x, it drops out from the conditioning set.

OLS is unbiased.

Limited Dependent Variables

Types of Sample Selection → Truncation:

#3. Sample is selected based only on value of y.

$$E\left(\varepsilon \middle| x_i s\right) \neq E\left(\varepsilon \middle| \mathrm{x}\right) = 0$$

Sample Selection Biased. OLS is biased.

Truncated regression model.

#4. Sample selection is correlated with ε.

$$E\left(\varepsilon \middle| x_i s\right) \neq E\left(\varepsilon \middle| \mathrm{x}\right) = 0$$

Sample Selection Biased. OLS is biased.

Heckman Sample Correction method (Heckit model).

Topics include:

Sample Selection Biased caused by:

1. Truncated Data

#3. Sample is *selected* based only on value of y.

Variable x and y are both truncated (can't be observed) → Truncated regression model.

2. Censored Data

#4. Sample selection is correlated with ε.

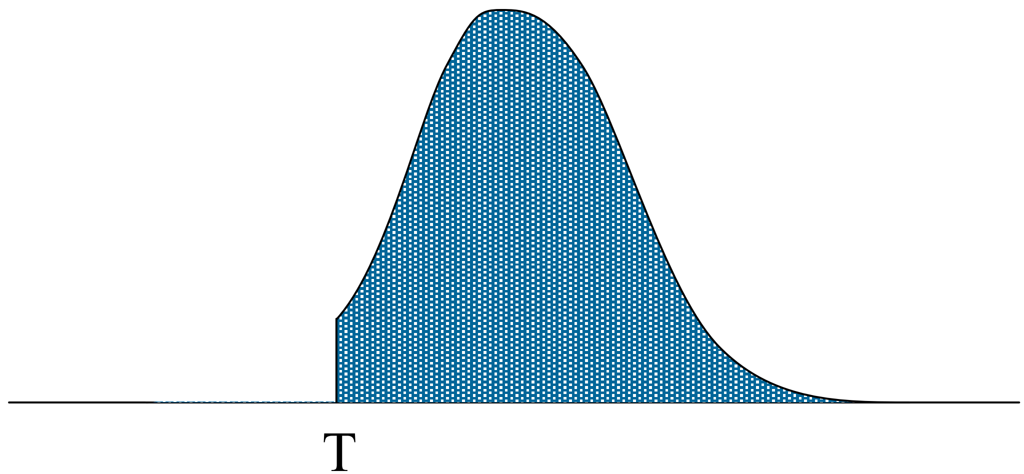
Only variable y is truncated (x can be observed) → Heckit model.

#3. Sample is *censored* based only on value of y.

Only variable y is censored (both variables can be observed) → Tobit model.

Truncated Sample

Truncated Normal Distribution

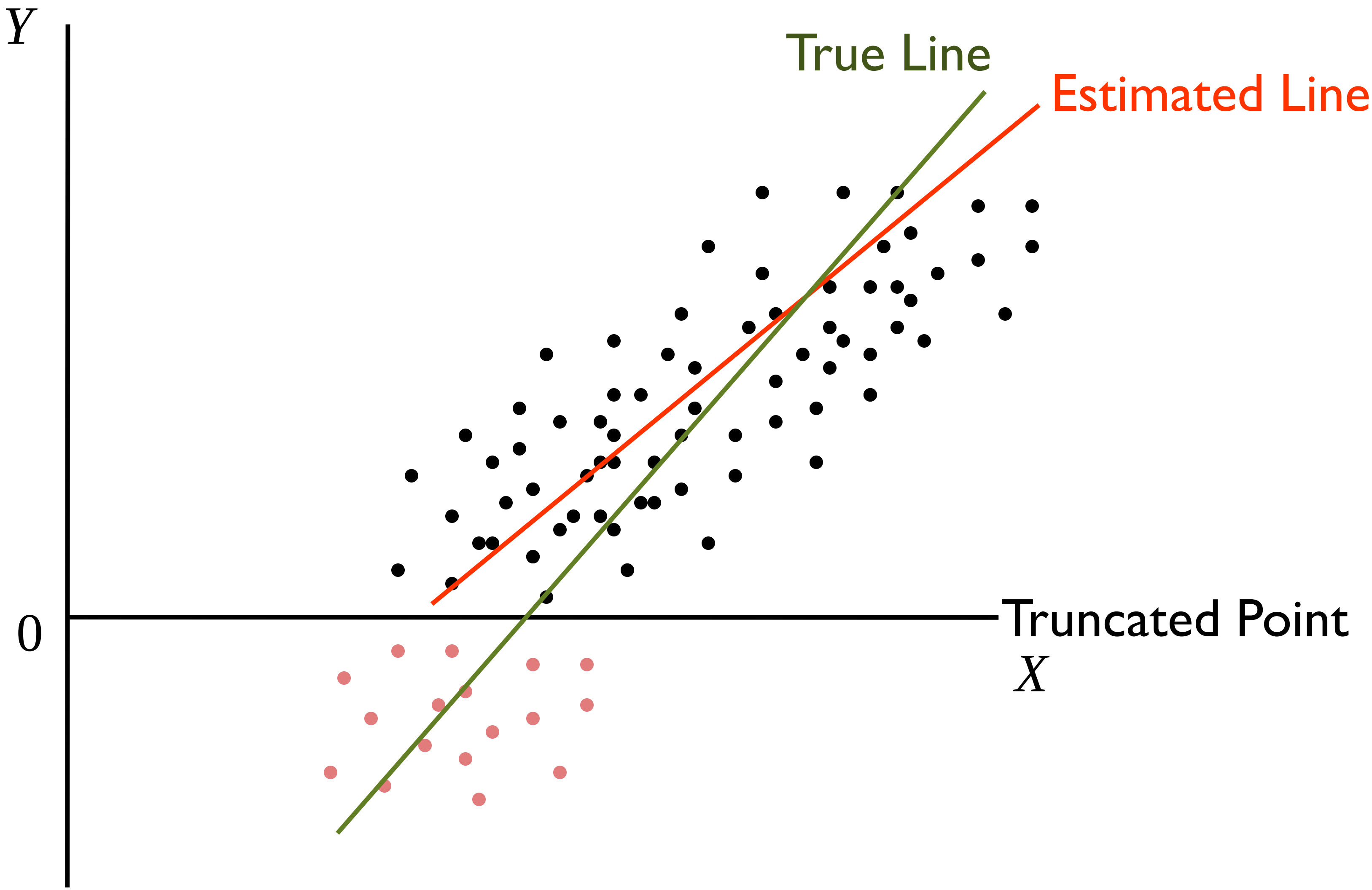


Probability Function:

$$\Pr(x > T) = 1 - \Phi\left(\frac{T-\mu}{\sigma}\right) = 1 - \Phi(\alpha)$$

$$f(x \mid x > T) = \frac{f(x)}{1-\Phi(\alpha)} = \frac{\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)}{1-\Phi(\alpha)}$$

Biased in Truncated Sample



Truncated Regression Model

Truncated model from below, we can only observe $y = y^{*} \;\; if \; y^{*} > T_L$

For example, consumption of durable goods where $T_L = 0$

Truncated model from above, we can only observe $y = y^{*} \;\; if \; y^{*} < T_U$

For example, only low-income individuals may be observed.

OLS → Sample Selection Biased

Truncated Regression Model

Model for truncated data from below can be stated as:

$$y_i = x_i'\beta + \varepsilon_i \quad if\ y^* > T_L \quad \text{and} \quad \varepsilon_i \big| x_i \sim N[0, \sigma^2]$$

Then, $$E\left[y_i \middle| y_i^* > T_L \right] = x_i'\beta + \sigma \frac{\phi\left[(T_L - x_i'\beta)/\sigma\right]}{1 - \Phi\left[(T_L - x_i'\beta)/\sigma\right]}$$

$$= x_i'\beta + \sigma\lambda(\alpha_i)$$

Marginal Effects (Partial Effects) of x_i

$$\frac{\partial E\left[y_i \middle| y_i^* > T_L \right]}{\partial x_i} = \beta\left(1 - \lambda_i^2 + \alpha_i\lambda_i\right) = \beta\left(1 - \delta_i\right)$$

where: $\delta_i = \lambda_i^2 + \alpha_i\lambda_i, \quad 0 < \delta_i < 1$

By Tatre Jantarakolica

Estimation Method for Truncated Regression Model

Truncated regression model can be estimated by maximizing the truncated log-likelihood function of the model.

e.g., truncated normal probability distribution can be stated as:

$$\ln L = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - x_i'\beta\right)^2$$
$$-\sum_{i=1}^{N}\log\left\{\Phi\left(\frac{T_U - x_i'\beta}{\sigma}\right) - \Phi\left(\frac{T_L - x_i'\beta}{\sigma}\right)\right\}$$

where: T_L is lower limit and T_U is upper limit.

Evaluation Criteria

Test of sample selection bias to confirm appropriateness of the Truncated regression Models.

1. Sign and meaning of the Coefficients.
   - Whether the estimated results are according to the theory.
   - Meaning – Marginal Effects for each case of the predicted y_j at …
2. Overall Test – Wald-Chi-squares-test.
   - Whether all explanatory variables can be used in explaining the dependent variable.

Evaluation Criteria

3. GOF.
 - Log-likelihood Value to make comparison.
4. Individual Test – z-test.
 - Whether each explanatory variables can explain the dependent variable.
 - Also significant of estimated sigma.
 - z-test – MLE assume Asymptotic Normal.

Heckit Model

Heckman Sample Correction method is based on truncated (unobserved) y but x can be observed. From truncated from below case:

$$y_i = x_i'\beta + \varepsilon_i \quad if\ y^* > T_L \quad \text{and} \quad \varepsilon_i \,|\, x_i \sim N[0, \sigma^2]$$

Then, $$E\left[y_i \,\middle|\, y_i^* > T_L\right] = x_i'\beta + \sigma \frac{\phi\left[(T_L - x_i'\beta)/\sigma\right]}{1-\Phi\left[(T_L - x_i'\beta)/\sigma\right]}$$

$$= x_i'\beta + \sigma\lambda(\alpha_i)$$

Heckman suggested two-step estimation:

1. Estimate Probit of the selection model, then determine value of λ.
2. Estimate $E\left[y_i \,\middle|\, y_i^* > T_L\right]$ with λ as additional independent variable.

Evaluation Criteria

Test of significance of estimated coefficient of λ to confirm appropriateness of Heckit Models.

1. Sign and meaning of the Coefficients.
 - Whether the estimated results are according to the theory.
 - Meaning – Marginal Effects for each case of the predicted y_j at …
2. Overall Test – Wald-Chi-squares-test.
 - Whether all explanatory variables can be used in explaining the dependent variable.

Evaluation Criteria

3. GOF.

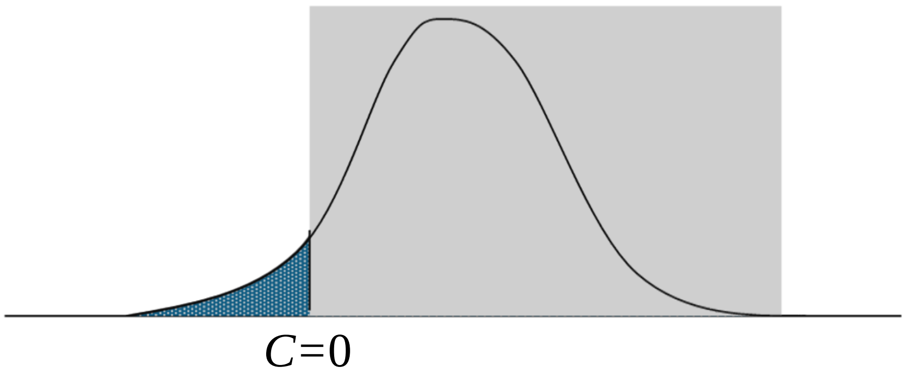
   -

4. Individual Test – z-test.

- Whether each explanatory variables can explain the dependent variable.
- Also significant of estimated sigma.

Censored Sample

Censored Normal Distribution



$$y = 0 \quad if \;\; y^* \le 0$$

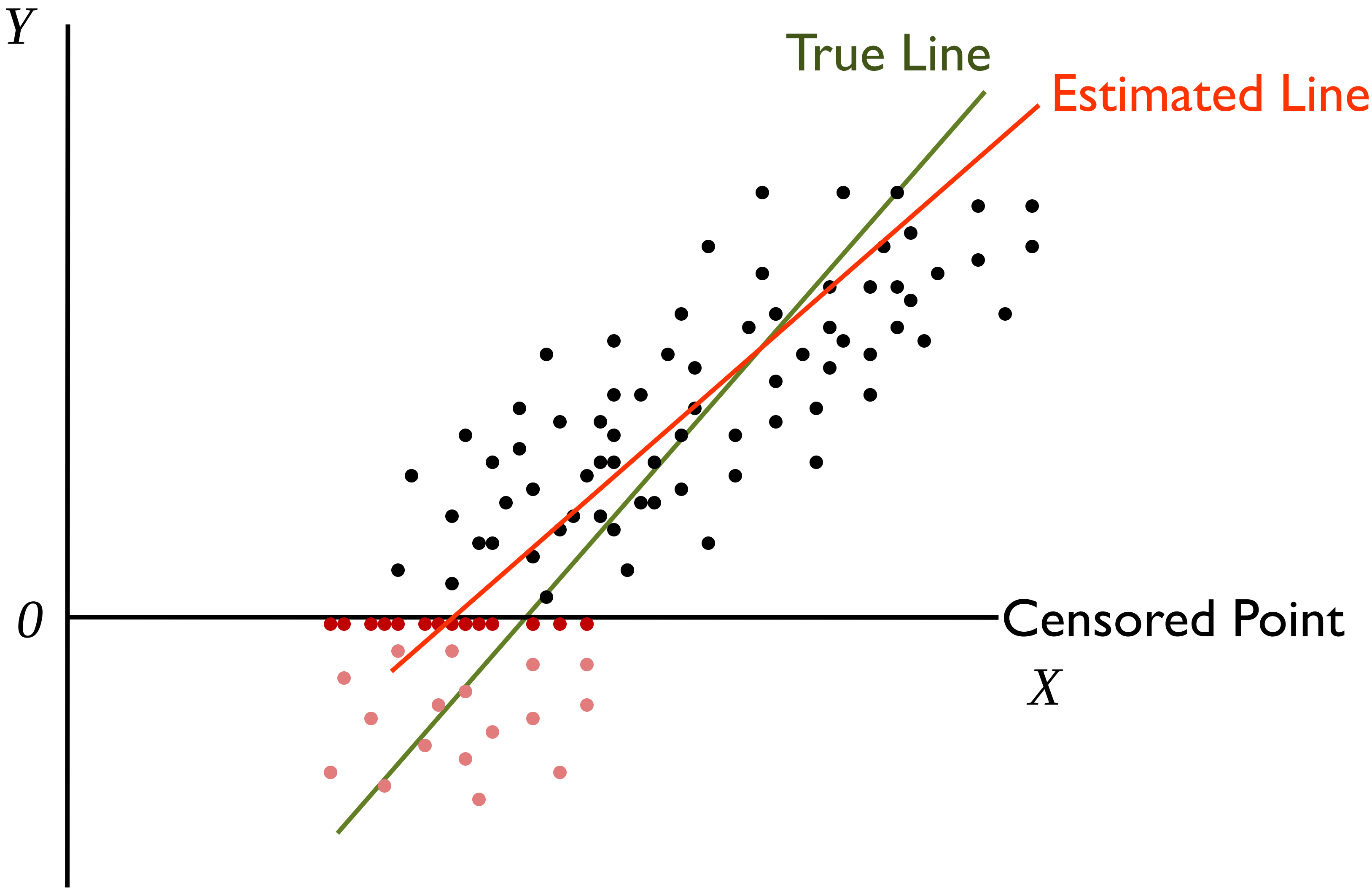
$$y = y^* \; if \;\; y^* > 0$$

Probability Function:

$$\Pr(y=0) = \Pr(y^* \le 0) = 1 - \Phi\left(\frac{\mu}{\sigma}\right)$$

$$E\left[y \middle| c=0\right] = \Phi\left(\mu/\sigma\right)\left(\mu + \sigma\lambda\right) \qquad where \;\; \lambda = \frac{\phi\left(\mu/\sigma\right)}{\Phi\left(\mu/\sigma\right)}$$

Biased in Censored Sample



By Tatre Jantarakolica

Model for Censored Data

Censoring from below, we can observe

$$y=\begin{cases} y^* & if \;\; y^* > C_L \\ C_L & if \;\; y^* \le C_L \end{cases}$$

For example, consumption of durable goods

Censoring from above, we can observe

$$y=\begin{cases} y^* & if \;\; y^* < C_U \\ C_U & if \;\; y^* \ge C_U \end{cases}$$

For example, maximum income of individuals may only be observed as >1 million.

Tobit Model for Censored Data

Model for censored data from below and $C_L = 0$ can be stated as: $y_i^* = x_i\beta + \varepsilon_i$

where $y = 0 \quad if \;\; y^* \le 0$ or $\quad y = y^* \; if \;\; y^* > 0$

Then, $E\left[y_i \left| x_i\right.\right] = \Phi\left(\dfrac{x_i\beta}{\sigma}\right)\left(x_i\beta + \sigma\lambda_i\right)$

where $\lambda_i = \dfrac{\phi\left[(0 - x_i\beta)/\sigma\right]}{1 - \Phi\left[(0 - x_i\beta)/\sigma\right]} = \dfrac{\phi\left(x_i\beta/\sigma\right)}{\Phi\left(x_i\beta/\sigma\right)}$

Marginal Effect of x_i

$$\frac{\partial E\left[y_i^* \left| x_i\right.\right]}{\partial x_i} = \beta \quad \text{and} \quad \frac{\partial E\left[y_i \left| x_i\right.\right]}{\partial x_i} = \beta\,\Phi\left(\frac{x_i\beta}{\sigma}\right)$$

Estimation Method for Tobit Model for Censored Data

Tobit Model can be estimated by maximizing the censored log-likelihood function of the model. For example, censored normal probability distribution when $C_L = 0$ can be stated as:

$$F^*(0) = \Pr\left[y^* \le 0\right] = \Pr\left[x_i\beta + \varepsilon_i\right] = \Phi\left(-x_i\beta/\sigma\right) = 1 - \Phi\left(x_i\beta/\sigma\right)$$

Censored density can be expressed as:

$$f(y) = \left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}\left(y - x'\beta\right)^2\right\}\right]^d \left[1 - \Phi\left(\frac{x'\beta}{\sigma}\right)\right]^{1-d}$$

where: $d = 1\ if\ y^* > 0\ and\ d = 0\ if\ y^* \le 0$

Estimation Method for Tobit Model for Censored Data

Then, log-likelihood function for Tobit Models assuming censored normal distribution can be stated as:

$$\ln L = \sum_{i=1}^{N} \left\{ \begin{array}{l} d_i \left(-\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}\left(y_i - x_i\beta\right)^2 \right) \\ +\left(1 - d_i\right)\ln\left(1 - \Phi y_i \left| x_i, \beta\right.\right) \end{array} \right\}$$

Evaluation Criteria

Test of sample selection bias to confirm appropriateness of the Tobit Models.

1. Sign and meaning of the Coefficients.
   - Whether the estimated results are according to the theory.
   - Meaning – Marginal Effects for each case of the predicted y_j at …
2. Overall Test – LR-Chi-squares-test.
   - Whether all explanatory variables can be used in explaining the dependent variable.

Evaluation Criteria

3. GOF.
  - Pseudo R^2
  - Log-likelihood Value to make comparison.
4. Individual Test – z-test.
  - Whether each explanatory variables can explain the dependent variable.
  - Also significant of estimated sigma.
  - z-test – MLE assume Asymptotic Normal.

Tobit Model using Panel Data

Tobit model with random effects can be stated as: $y_{it}^* = x_{it}\beta + \alpha_i + \varepsilon_{it}$

where $y_{it} = 0 \;\; if \;\; y_{it}^* \le 0$ or $y_{it} = y_{it}^* \;\; if \;\; y_{it}^* > 0$

$\alpha_i =$ Cross-sectional Random Effects

Joint density function: $\prod_{t=1}^{T}\left[\frac{1}{\sigma_\varepsilon}\phi_{it}\right]^{d_{it}}\left[1-\Phi_{it}\right]^{1-d_{it}}$

where

$$\phi_{it} = \phi\left(\left(y_{it} - \alpha_i - x_{it}\beta\right)/\sigma_\varepsilon\right), \quad \Phi_{it} = \Phi\left(\left(\alpha_i + x_{it}\beta\right)/\sigma_\varepsilon\right)$$

Log-likelihood $\sum_{i=1}^{N} \ln f\left(y_i \middle| x_{it}, \beta, \sigma_\varepsilon^2, \sigma_\alpha^2\right)$

where $f = \int f\left(y_i \middle| x_i, \alpha_i, \beta, \sigma_\varepsilon^2\right)\frac{1}{\sqrt{2\pi\sigma_\alpha^2}}\exp\left(\frac{-\alpha_i}{2\sigma_\alpha^2}\right)^2 d\alpha_i$

Test Pooled vs Panel Tobit Estimators

To test whether the pooled estimator or the panel estimator should be employed, the test whether $\rho = 0$ should be performed.

where: $$\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2}$$

If $\rho = 0$, it means that panel-level variance component σ_α^2 is unimportant.

Then, panel Tobit estimator is no different from the pooled Tobit estimator.

Evaluation Criteria

Test of random effects to confirm appropriateness of the RE Tobit Models.

1. Sign and meaning of the Coefficients.
   - Whether the estimated results are according to the theory.
   - Meaning – Marginal Effects for each case of the predicted y_j at …
2. Overall Test – Wald-Chi-squares-test.
   - Whether all explanatory variables can be used in explaining the dependent variable.

Evaluation Criteria

3. GOF.
  - Log-likelihood Value to make comparison.
4. Individual Test – z-test.
  - Whether each explanatory variables can explain the dependent variable.
  - Also significant of estimated sigma.
  - z-test – MLE assume Asymptotic Normal.