

Selection Model

Jason Qiang Guo, New York

Tobit Model: MLE

► Basic setup:

$$\underbrace{y_i^*}_{1 \times 1} = \underbrace{x_i'}_{1 \times K} \underbrace{\beta}_{K \times 1} + \underbrace{\varepsilon_i}_{1 \times 1}$$

where

$$\varepsilon_i | x_i \sim N(0, \sigma^2).$$

The dependent variable y_i^* is determined by

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

Tobit Model: MLE

► Basic setup:

$$\underbrace{y_i^*}_{1 \times 1} = \underbrace{x_i'}_{1 \times K} \underbrace{\beta}_{K \times 1} + \underbrace{\varepsilon_i}_{1 \times 1}$$

where

$$\varepsilon_i | x_i \sim N(0, \sigma^2).$$

The dependent variable y_i^* is determined by

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

- In order to estimate this model, we need to derive the conditional distribution density function of y_i , i.e. $f(y_i | x_i)$. And we have to consider two cases for this model: (i) $y > 0$ and (ii) $y = 0$.

Tobit Model: MLE

(i) Case of $y_i > 0$

The above definition of y_i indicates that if $y_i > 0$ the conditional distribution of y_i is the same as that of y_i^* . Therefore, if $y > 0$, we have

$$\begin{aligned} f(y_i | x_i) &= f^*(y_i | x_i) \\ &= f^*(y_i^* | x_i) \quad (\text{since } y_i = y_i^* \text{ if } y_i > 0) \\ &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{(y_i - x_i'\beta)^2}{\sigma^2}\right) \quad (\text{since } y_i^* \text{ is distributed normally } y_i^* \sim N(x_i'\beta, \sigma^2)) \\ &= \frac{1}{\sigma} \underbrace{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{y_i - x_i'\beta}{\sigma}\right)^2\right)}_{=\phi\left(\frac{y_i - x_i'\beta}{\sigma}\right)} \\ &= \frac{1}{\sigma} \phi\left(\frac{y_i - x_i'\beta}{\sigma}\right). \quad (\text{since pdf of standard normal is } \phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right), \text{ now } z = \frac{y_i - x_i'\beta}{\sigma}) \end{aligned}$$

Tobit Model: MLE

(ii) **Case of** $y_i = 0$

On the other hand, for $y_i = 0$, we have the mass conditional probability $\Pr(y_i = 0 | x_i)$ which is equal to

$$\begin{aligned}\Pr(y_i = 0 | x_i) &= \Pr(y_i^* < 0 | x_i) \\&= \Pr(x_i' \beta + \varepsilon_i \leq 0 | x_i) \quad (\text{by definition of latent variable } y_i^*, y_i^* = x_i' \beta + \varepsilon_i) \\&= \Pr\left(\underbrace{\varepsilon_i}_{\text{distributed as } N(0, \sigma^2)} \leq -x_i' \beta \middle| x_i\right) \\&= \Pr\left(\underbrace{\frac{\varepsilon_i}{\sigma}}_{\text{distributed as standard normal}} \leq -\frac{x_i' \beta}{\sigma} \middle| x_i\right) \\&= \Phi\left(-\frac{x_i' \beta}{\sigma}\right) \quad (\text{where } \Phi \text{ is the c.d.f. of standard normal}) \\&= 1 - \Phi\left(\frac{x_i' \beta}{\sigma}\right) \quad (\text{since standard normal distribution is symmetric, } \Phi(-z) = 1 - \Phi(z)).\end{aligned}$$

Therefore, according to the result of (i) and (ii), the conditional density function is expressed as

$$f(y_i | x_i) = \begin{cases} \text{continuous part} & f^*(y_i | x_i) &= \frac{1}{\sigma} \phi\left(\frac{y_i - x_i' \beta}{\sigma}\right) &: \text{ if } y_i > 0 \\ \text{mass part} & \Pr(y_i = 0 | x_i) &= 1 - \Phi\left(\frac{x_i' \beta}{\sigma}\right) &: \text{ if } y_i \leq 0 \end{cases}$$

Tobit Model: MLE

- It is now clear that we can write down our MLE object

$$L(\beta, \sigma^2) = \prod_{y_i > 0} \frac{1}{\sigma} \phi\left(\frac{y_i - x_i' \beta}{\sigma}\right) \prod_{y_i = 0} (1 - \Phi\left(\frac{x_i' \beta}{\sigma}\right))$$

and the covariance-variance matrix is given by

$$I^{-1} = -E\left(\frac{\partial \log L(\beta, \sigma^2)}{\partial \beta' \beta}\right)$$

Tobit Model: MLE

- It is now clear that we can write down our MLE object

$$L(\beta, \sigma^2) = \prod_{y_i > 0} \frac{1}{\sigma} \phi\left(\frac{y_i - x_i' \beta}{\sigma}\right) \prod_{y_i = 0} (1 - \Phi\left(\frac{x_i' \beta}{\sigma}\right))$$

and the covariance-variance matrix is given by

$$I^{-1} = -E\left(\frac{\partial \log L(\beta, \sigma^2)}{\partial \beta' \beta}\right)$$

- What about conditional expectation $E_{y|x}[y|x]$ and $E_{y|x, y>0}[y|x, y > 0]$? We are interested in the marginal effects of x on y .

Tobit Model: Conditional Expectation

- ▶ The calculation involves lengthy mathematical calculation, but the basic idea is that

$$E_{y|x}[y|x] = Pr(y^* \leq 0|x_i)E_{y|x,y=0}[y|x, y=0] + Pr(y^* > 0|x_i)E_{y|x,y>0}[y|x, y>0]$$

Tobit Model: Conditional Expectation

- ▶ The calculation involves lengthy mathematical calculation, but the basic idea is that

$$E_{y|x}[y|x] = Pr(y^* \leq 0|x_i)E_{y|x,y=0}[y|x, y=0] + Pr(y^* > 0|x_i)E_{y|x,y>0}[y|x, y>0]$$

- ▶ The conditional expectations are:

$$E_{y_i|x_i, y_i>0}[y_i|x_i, y_i>0] = x_i'\beta + \sigma \cdot \frac{\phi\left(\frac{x_i'\beta}{\sigma}\right)}{\Phi\left(\frac{x_i'\beta}{\sigma}\right)}.$$

and

$$\begin{aligned} E_{y_i|x_i}[y_i|x_i] &= Pr(y^* > 0|x_i) \cdot E_{y_i|x_i, y_i>0}[y_i|x_i, y_i>0] \\ &= \Phi\left(\frac{x_i'\beta}{\sigma}\right) \cdot \left[x_i'\beta + \sigma \cdot \frac{\phi\left(\frac{x_i'\beta}{\sigma}\right)}{\Phi\left(\frac{x_i'\beta}{\sigma}\right)} \right] \\ &= \Phi\left(\frac{x_i'\beta}{\sigma}\right) x_i'\beta + \phi\left(\frac{x_i'\beta}{\sigma}\right) \end{aligned}$$

Tobit: Heckman Two-Step Estimation

Define inverse Mills ratio:

$$\lambda(z) = \frac{\phi(z)}{\Phi(z)}$$
$$\lambda\left(\frac{x_i'\beta}{\sigma}\right) = \frac{\phi\left(\frac{x_i'\beta}{\sigma}\right)}{\Phi\left(\frac{x_i'\beta}{\sigma}\right)}.$$

Tobit: Heckman Two-Step Estimation

Define inverse Mills ratio:

$$\begin{aligned}\lambda(z) &= \frac{\phi(z)}{\Phi(z)} \\ \lambda\left(\frac{x'_i\beta}{\sigma}\right) &= \frac{\phi\left(\frac{x'_i\beta}{\sigma}\right)}{\Phi\left(\frac{x'_i\beta}{\sigma}\right)}.\end{aligned}$$

Based on above equation, Heckman constructed the model equation for data that satisfy $y_i > 0$

$$y_i = x'_i\beta + \sigma \cdot \lambda\left(\frac{x'_i\beta}{\sigma}\right) + u_i$$

where the error term u_i has zero conditional expectation

$$E[u_i | x_i, y_i > 0] = 0.$$

We implement estimation with two steps.

Tobit: Heckman Two-Step Estimation

Step 1: Using all data and implement probit (or logit) estimation by using dummy variable (rewriting the definition of dummy variable d_i)

$$d_i = \begin{cases} 1 & : \text{ if } y_i > 0 \\ 0 & : \text{ if } y_i = 0 \end{cases},$$

and construct the probit (or logit) model

$$\underbrace{d_i}_{0 \text{ or } 1} = x_i' \beta + \underbrace{\varepsilon_i}_{\text{distributed as } N(0, \sigma^2)}.$$

Then, we can estimate (you know, in binary choice model, we can estimate β up to scale)

$$\widehat{\left(\frac{\beta}{\sigma}\right)},$$

by probit (or logit) estimation.

Step 2: Calculate the estimate of hazard function by using $\widehat{\left(\frac{\beta}{\sigma}\right)}$ in step 1

$$\lambda\left(\widehat{\frac{x_i' \beta}{\sigma}}\right) = \lambda\left(x_i' \widehat{\left(\frac{\beta}{\sigma}\right)}\right) = \frac{\phi\left(x_i' \widehat{\left(\frac{\beta}{\sigma}\right)}\right)}{\Phi\left(x_i' \widehat{\left(\frac{\beta}{\sigma}\right)}\right)}.$$

Then, implement OLS for the model by only using $y_i > 0$ data

$$y_i = x_i' \beta + \sigma \cdot \lambda\left(\widehat{\frac{x_i' \beta}{\sigma}}\right) + u_i,$$

and obtain estimator $\hat{\beta}$ and $\hat{\sigma}$.

Heckman Two-Step Estimation: A generalized tobit model

- ▶ The differences between Heckman selection model and tobit model are: (i) the covariates \mathbf{Z} that determine the selection do not enter the second stage estimation and (ii) errors of two stages are correlated.

Heckman Two-Step Estimation: A generalized tobit model

- ▶ The differences between Heckman selection model and tobit model are: (i) the covariates \mathbf{Z} that determine the selection do not enter the second stage estimation and (ii) errors of two stages are correlated.



$$y_2 = \mathbf{Z}\gamma + \mu_1,$$
$$d = \begin{cases} 1 & \text{if } y_2 > 0 \\ 0 & \text{if } y_2 \leq 0 \end{cases}$$

and

$$Y_1 = \mathbf{X}\beta + \nu_2 \text{ if } d = 1,$$

where

$$\begin{pmatrix} v_1 \\ \nu_2 \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1 \\ \rho\sigma_1 & 1 \end{pmatrix} \right]$$

Heckman Two-Step Estimation: A generalized tobit model

- ▶ The differences between Heckman selection model and tobit model are: (i) the covariates \mathbf{Z} that determine the selection do not enter the second stage estimation and (ii) errors of two stages are correlated.



$$y_2 = \mathbf{Z}\gamma + \mu_1,$$
$$d = \begin{cases} 1 & \text{if } y_2 > 0 \\ 0 & \text{if } y_2 \leq 0 \end{cases}$$

and

$$Y_1 = \mathbf{X}\beta + \nu_2 \text{ if } d = 1,$$

where

$$\begin{pmatrix} v_1 \\ \nu_2 \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1 \\ \rho\sigma_1 & 1 \end{pmatrix} \right]$$

- ▶ We can derive the conditional expectation

$$E_{y_1|x, y_2 > 0}(y_1 | x, y_2 > 0) = \mathbf{X}\beta + \sigma_1\rho\lambda(\mathbf{Z}'\gamma)$$

Heckman Two-Step Estimation: A generalized tobit model

- ▶ The differences between Heckman selection model and tobit model are: (i) the covariates \mathbf{Z} that determine the selection do not enter the second stage estimation and (ii) errors of two stages are correlated.



$$y_2 = \mathbf{Z}\gamma + \mu_1,$$
$$d = \begin{cases} 1 & \text{if } y_2 > 0 \\ 0 & \text{if } y_2 \leq 0 \end{cases}$$

and

$$Y_1 = \mathbf{X}\beta + \nu_2 \text{ if } d = 1,$$

where

$$\begin{pmatrix} v_1 \\ \nu_2 \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1 \\ \rho\sigma_1 & 1 \end{pmatrix} \right]$$

- ▶ We can derive the conditional expectation

$$E_{y_1|x, y_2 > 0}(y_1 | x, y_2 > 0) = \mathbf{X}\beta + \sigma_1\rho\lambda(\mathbf{Z}'\gamma)$$

Still two step implementation!

Heckman Two-Step Estimation: full likelihood estimation

The key is to obtain the probability distribution of y_1 given $y_2 > 0$

$$\begin{aligned}\Pr(y_{1i}, y_{2i} > 0 | X, Z) &= f(y_{1i}) \Pr(y_{2i} > 0 | y_{1i}, X, Z) = f(\nu_{1i}) \Pr(\nu_{2i} > -Z_i \delta | \nu_{1i}, X, Z) \\&= \frac{1}{\sigma_1} \phi \left(\frac{y_{1i} - X_i \beta}{\sigma_1} \right) \cdot \int_{-Z_i \delta}^{\infty} f(\nu_{2i} | \nu_{1i}) d\nu_{2i} \\&= \frac{1}{\sigma_1} \phi \left(\frac{y_{1i} - X_i \beta}{\sigma_1} \right) \cdot \int_{-Z_i \delta}^{\infty} \phi \left(\frac{\nu_{2i} - \frac{\rho}{\sigma_1} (y_{1i} - X_i \beta)}{\sqrt{1 - \rho^2}} \right) d\nu_{2i} \\&= \frac{1}{\sigma_1} \phi \left(\frac{y_{1i} - X_i \beta}{\sigma_1} \right) \cdot \left[1 - \Phi \left(\frac{-Z_i \delta - \frac{\rho}{\sigma_1} (y_{1i} - X_i \beta)}{\sqrt{1 - \rho^2}} \right) \right] \\&= \frac{1}{\sigma_1} \phi \left(\frac{y_{1i} - X_i \beta}{\sigma_1} \right) \cdot \Phi \left(\frac{Z_i \delta + \frac{\rho}{\sigma_1} (y_{1i} - X_i \beta)}{\sqrt{1 - \rho^2}} \right)\end{aligned}$$

Heckman Two-Step Estimation: full likelihood estimation

2. Those where y_1 is not observed and we know that $y_2 \leq 0$. For these observations, the likelihood function is just the marginal probability that $y_2 \leq 0$. We have no independent information on y_1 . This probability is written as

$$\Pr(y_{2i} \leq 0) = \Pr(\nu_{2i} \leq -Z_i\delta) = \Phi(-Z_i\delta) = 1 - \Phi(Z_i\delta)$$

Therefore the log likelihood for the complete sample of observations is the following:

$$\begin{aligned} \log L(\beta, \delta, \rho, \sigma; \text{thedata}) &= \sum_{i=1}^{N_0} \log [1 - \Phi(Z_i\delta)] \\ &+ \sum_{i=N_0+1}^N \left[-\log \sigma_1 + \log \phi\left(\frac{y_{1i} - X_i\beta}{\sigma_1}\right) + \log \Phi\left(\frac{Z_i\delta + \frac{\rho}{\sigma_1}(y_{1i} - X_i\beta)}{\sqrt{1 - \rho^2}}\right) \right] \end{aligned}$$