Microeconometrics
Self-Selectivity

Pedro Portugal

NOVA School of Business and Economics

Spring 2024

Examples of Censored Variables

Selection may be due to:

- Self-selection, with the outcome of interest determined in part by individual choice of whether or not to participate in the activity of interest
- Sample selection, with those who participate in the activity of interest deliberately oversampled

We present two of the many selection models in the literature

- Tobit model (presented previously)
- Roy model
  - considers an outcome that takes one of two values depending on the value taken by a censoring variable

Truncated Sample

Consider the latent variable model

$$y^* = \mathbf{x}'\beta + \varepsilon$$

where $\varepsilon \sim \mathcal{N}[0, \sigma^2]$ has variance σ^2 constant across observations

If the selection rule is given by $s_i = \mathbf{1}[a_1 < y_i < a_2]$, then

$$P(s_i = 1) = \Phi(a_2|\mathbf{x}) - \Phi(a_1|\mathbf{x})$$

And the log-likelihood function is

$$\ln L_N(\beta, \sigma) = \sum_{i=1}^{N} \ln \phi\left(\frac{y_i - \mathbf{x}_i'\beta}{\sigma}\right) - \ln \sigma - \ln \left[\Phi\left(\frac{a_2 - \mathbf{x}_i'\beta}{\sigma}\right) - \Phi\left(\frac{a_1 - \mathbf{x}_i'\beta}{\sigma}\right)\right]$$

Roy Model

- 1 – Hunters

- 2 – Fishermen

Roy Model

The starting point is Roy's (1951) "Thoughts on the Distribution of Earnings", which discusses the optimizing choices of "workers" selecting between fishing and hunting

Roy considered the consequences for the occupational distribution of earnings when there is individual heterogeneity in skills and individuals self-select into occupations

Consider two outcome variables Y_1 (output in hunting) and Y_2 (output in fishing) where (Y_1, Y_2) follows a bivariate normal distribution with $E(Y_1) = \mu_1$, $E(Y_2) = \mu_2$, and

$$\text{Var-cov}(Y_1, Y_2) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

Roy Model

Consider the following two equations:

$$u_1 = Y_1 - \mu_1$$

$$u_2 = Y_2 - \mu_2$$

Then,

$$Y_1 - Y_2 = (\mu_1 - \mu_2) - (u_2 - u_1)$$

And

$$\sigma^2 = var(u_1 - u_2)$$

Define

$$Z = \frac{\mu_1 - \mu_2}{\sigma} \text{ and } u = \frac{u_2 - u_1}{\sigma}$$

Roy Model

The condition $Y_1 > Y_2$ implies $u < Z$. Then the mean income of hunters is given by

$$E[Y_1|u < Z] = \mu_1 - \sigma_{1u}\frac{\phi(Z)}{\Phi(Z)}$$

where

$$\sigma_{1u} = cov(u_1, u) = \frac{\sigma_{12} - \sigma_1^2}{\sigma}$$

and $\phi(.)$ and $\Phi(.)$ are the density function and the distribution function of the standard normal, respectively

- $u < Z$ implies that $u_2 - u_1 < \mu_1 - \mu_2$

Roy Model

And the mean income of fishermen is given by

$$E[Y_2|u > Z] = \mu_2 + \sigma_{2u}\frac{\phi(Z)}{1-\Phi(Z)}$$

where

$$\sigma_{2u} = cov(u_2, u) = \frac{\sigma_2^2 - \sigma_{12}}{\sigma} > \sigma_{1u}$$

- $u > Z$ implies that $u_2 - u_1 > \mu_1 - \mu_2$ or, equivalently, $y_2 > y_1$

Roy Model

We have four possible cases:

- $\sigma_{1u} < 0$ and $\sigma_{2u} > 0$: the mean income of hunters is greater than μ_1 and the mean income of fishermen is greater than μ_2
  - That is those who chose hunting are better than average hunters and those who chose fishing are better than fishermen
- $\sigma_{1u} < 0$ and $\sigma_{2u} < 0$: the mean income of hunters is greater than μ_1 and the mean income of fishermen is less than μ_2
  - That is those who chose hunting are better than average in both hunting and fishing, but they are better in hunting than in fishing; those who chose fishing are below average in both hunting and fishing, but they are better in fishing than in hunting
- $\sigma_{1u} > 0$ and $\sigma_{2u} > 0$: Reverse case of previous case
- $\sigma_{1u} > 0$ and $\sigma_{2u} < 0$: This is not possible, given the definitions of σ_{1u} and σ_{2u}

Self-selection

Consider the following equations:

$$y_1 = \mathbf{x}_1'\beta + u_1$$

$$y_2 = \mathbf{1}[\mathbf{x}_2'\delta + \nu_2 > 0]$$

where $E(u_1) = 0$, $E(\nu_2) = 0$, $\nu_2 \sim \mathcal{N}(0,1)$ and $E(u_1|\nu_2) = \gamma_1\nu_2$

Then

$$E[y_1|\mathbf{x}, y_2 = 1] = \mathbf{x}_1'\beta + \gamma_1\lambda(\mathbf{x}_2'\delta)$$

where $\lambda(\mathbf{x}_2'\delta) = E[\nu_2|\nu_2 > -\mathbf{x}_2'\delta]$

Heckman Two-Step Procedure

The **Heckit estimator** augments the OLS regression by an estimate of the omitted regressor $\lambda(\mathbf{x}_2'\delta)$

- $\widehat{\delta}$ is obtained by first-step probit regression of y_2 on $\mathbf{x}_2$
- Compute the estimated inverse Mills ratio

$$\lambda(\mathbf{x}_2'\widehat{\delta}) = \frac{\phi(\mathbf{x}_2'\widehat{\delta})}{\Phi(\mathbf{x}_2'\widehat{\delta})}$$

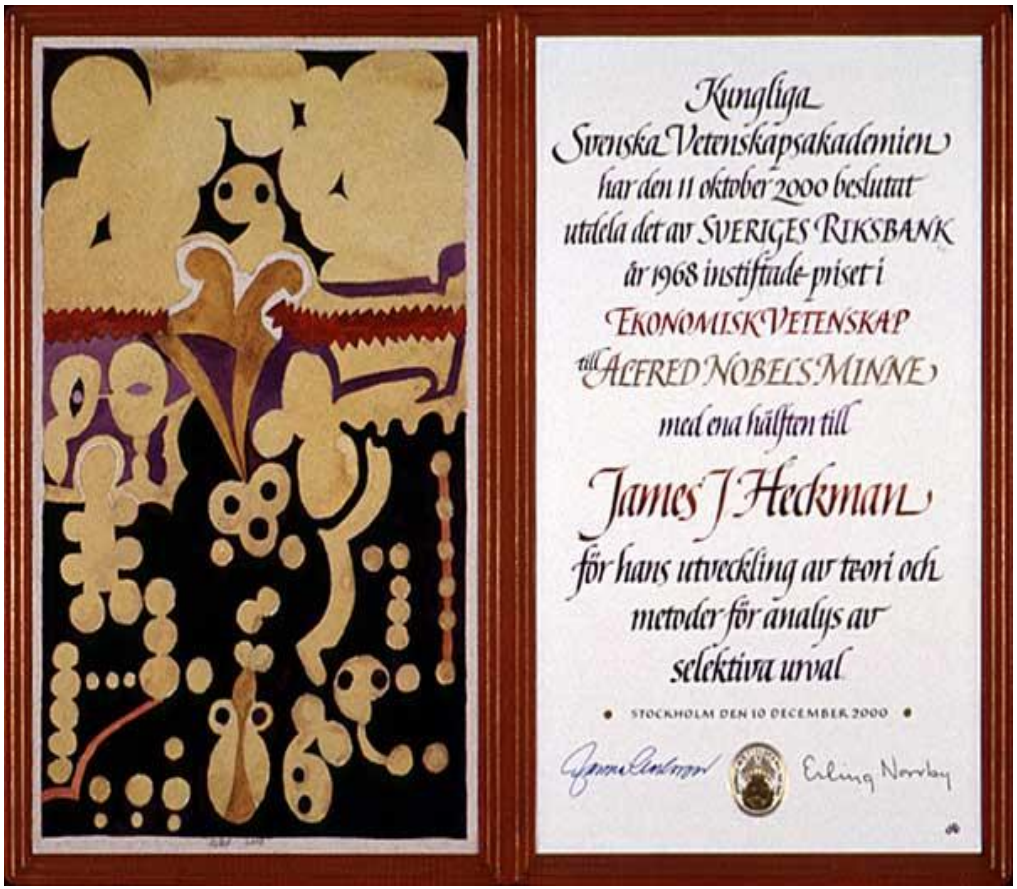
- Estimate by OLS the model

$$y_{1i} = \mathbf{x}_{1i}'\beta + \gamma_1\lambda(\mathbf{x}_2'\widehat{\delta}) + v_i$$

where v is an error term

James Heckman (Nobel 2000)

James Heckman (Nobel 2000)



Maximum Likelihood for Self-Selectivity

Consider the following:

- $(u_1, \nu_2) \sim BNormal$
- $E(u_1) = E(\nu_2) = 0$
- $Var(u_1) = \sigma_1^2$ and $Var(\nu_2) = 1$
- $Cov(u_1, \nu_2) = \sigma_{12}$
- $Corr(u_1, \nu_2) = \frac{\sigma_{12}}{\sigma_1} = \rho$

The unconditional density is:

$$f(y_2|\mathbf{x}) = \Phi(\mathbf{x}'\delta)^{y_2}[1 - \Phi(\mathbf{x}'\delta)]^{1-y_2}$$

Maximum Likelihood for Self-Selectivity

And the conditional density:

$$f(y_1|y_2,\mathbf{x}) = f(y_2|y_1,\mathbf{x})\frac{f(y_1|\mathbf{x})}{f(y_2|\mathbf{x})}$$

$$f(y_1|y_2=1,\mathbf{x}) = P(y_2=1|y_1,\mathbf{x})\frac{f(y_1|\mathbf{x})}{P(y_2=1|\mathbf{x})}$$

$$P(y_2=1|y_1,\mathbf{x}) = \Phi\left(\frac{\mathbf{x}'\delta + \frac{\sigma_{12}}{\sigma_1^2}\nu_1}{\sqrt{1-\rho^2}}\right)$$

Maximum Likelihood for Self-Selectivity

Then the log-likelihood function is:

$$\ln L_N(\beta, \delta, \sigma | \mathbf{x}) = \sum_{i=1}^{N} \mathbf{1}[y_{2i} = 0] \ln[1 - \Phi(\mathbf{x}_{2i}'\delta)] +$$
$$+ \mathbf{1}[y_{2i} = 1] \left\{ \ln \Phi \left[\frac{\mathbf{x}_{2i}'\delta + \frac{\rho}{\sigma_1}(y_{i1} - \mathbf{x}_i'\beta)}{\sqrt{1-\rho^2}} \right] + \ln \phi \left(\frac{y_{1i} - \mathbf{x}_i'\beta}{\sigma} \right) - \ln \sigma \right\}$$