Sample Selection

(GB: Chapter 19.3-19.7) Isabel Casas

Previously...

- Selection on the basis of $\mathbf{X}$: exogenous sampling
  - Eg: Segmenting an existing sample by gender or status.
  - Sample selection problem can be ignored of the rule does not depend of the other exogenous variables nor of $\mathbf{y}$
  - If the selection depends on $\mathbf{y}$ then...
- Selection on the basis of $\mathbf{y}$ (the LHS variable)
  - More difficult to resolve

  1. Deterministic selection rule (known): truncated regression (Tobit model)
  2. Random selection rule (unknown): behavioural selection

2. If Selection is determined by behaviour
   - Probit models for selection (with and without endogeneity)
   - Tobit models for selection (with and without endogeneity)
   - Structural Tobits with selection

Selection determined by behaviour

The wage equation example (formally):

- Wage: $\log w_j = x_{j1}\beta_1 + \epsilon_{1j}$
- Reservation wage: $\log w_j^{res} = x_{j2}\beta_2 + \gamma_2\ a_j + \epsilon_{2j}$
- Selection rule: $\log w_j > \log w_j^{res}$
  - We don't know $w_j^{res} \Rightarrow$ not a truncated or censored regression model
  - Instead, selection depends on unobservables (that are correlated with the error term in the wage equation):

$$\log w_j > \log w_j^{res} \Rightarrow x_{j1}\beta_1 - x_{j2}\beta_2 - \gamma_2 a_j + \epsilon_{1j} - \epsilon_{2j} > 0$$

A) Probit selection

Model (type II Tobit):

$$\mathbf{y}_1 = \mathbf{X}_1\beta_1 + \boldsymbol{\epsilon}_1$$
$$\mathbf{y}_2 = \mathbf{1}[\mathbf{X}\boldsymbol{\delta}_2 + \boldsymbol{\nu}_2 > 0] \qquad \text{where } \mathbf{X}_1 \in \mathbf{X}$$

Assumptions:

- $(\mathbf{X}, \mathbf{y}_2)$ are always observed
- $\mathbf{y}_1$ only observed when $\mathbf{y}_2 = 1$
- $(\boldsymbol{\epsilon}_1, \boldsymbol{\nu}_2)$ independent of $\mathbf{X}$
- $\boldsymbol{\nu}_2 \sim N(0, 1)$
- $E(\boldsymbol{\epsilon}_1|\boldsymbol{\nu}_2) = \gamma_1\boldsymbol{\nu}_2$

$\mathbf{X}_1$ could in principle contain variables not in $\mathbf{X}$ (but undesirable)

A) Probit selection

Procedure 19.1 (Heckit):

1. Obtain probit estimate of $\boldsymbol{\delta}_2$ from: $P(y_{j2} = 1|\mathbf{x}_j) = \Phi(\mathbf{x}_j\boldsymbol{\delta}_2)$
   - The selection model using all observations
   - Construct λ from this estimate and $\mathbf{X}$: $\hat{\lambda}_j = \lambda(\mathbf{x}_j\hat{\boldsymbol{\delta}_2})$
2. With OLS:

$$y_{1j} = x_{1j}\boldsymbol{\beta}_1 + \gamma_1\hat{\lambda}_j + error_j$$

Estimators $\hat{\boldsymbol{\beta}}_1$ and $\hat{\gamma}_1$ are consistent and asymp. normal

A) Probit selection

NB!

- Test for sample selection bias: t-test for $H_0 : \gamma_1 = 0$ (no bias)
  - The asymptotic variance of β_1 and γ_1 are not affected by $\hat{\boldsymbol{\delta}}_2$ under the null
- If the t-test shows that $\gamma_1 \neq 0$
  - Asymptotic variance of the estimator of β_1 is complicated.
  - The robust standard errors are not enough
- Preferably, we should have $\mathbf{X} \neq \mathbf{X}_1$ in the selection model
  - But in principle works without this because of non-linearity in selection equation

Exercise 3 (15 minutes)

Example 19.6 (Wage offer equation). Estimate the log wage for married women.

- mroz.dat
- $n = 753$, but only 428 women work (look at variable $lwage$ and $inlf$)
- Model includes on RHS: $educ, exper, exper^2$
- Selection equation includes further: $age, kidslt6, kidsge6, nwifeinc$
- Write the selection equation $\mathbf{y}_2 =$?
- What is $\mathbf{X}$, $\mathbf{X}_1$?
- library $sampleSelection$
- function $heckit$

Exercise 3

In R,

$$heckit(selection, outcome, data, method = "2step")$$

or

$$selection(selection, outcome, data, method = "2step")$$

- $selection$ is the **formula** of $\mathbf{y}_2$ (variable defining the sample selection) and $\mathbf{X}$
- $outcome$ is the **formula** of $\mathbf{y}_1$ (what we are interested on)

Results of Exercise 3

Independent var	intercept	educ	exper	expersq	$\hat{\lambda}_2$	$\hat{\boldsymbol{\nu}}_2$	bias?
OLS	-0.5220 **	0.1075 ***	0.0416 **	-0.0008 *		NA	
Procedure 19.1	-0.5781 .	0.1091 ***	0.0439 **	-0.0008 .	0.0323		No

$$\mathbf{y}_1 = lwage$$

$$\mathbf{X}_1 = \{educ, exper, expersq\}$$

$$\mathbf{y}_2 = inlf$$

$$\mathbf{X} = \{\mathbf{X}_1, age, kidslt6, kidsge6, nwifeinc\}$$

B) Probit Selection with Endogenous RHS Variable

$$\mathbf{y}_1 = \mathbf{Z}_1\boldsymbol{\delta}_1 + \alpha_1\mathbf{y}_2 + \boldsymbol{\epsilon}_1$$
$$\mathbf{y}_2 = \mathbf{Z}_2\boldsymbol{\delta}_2 + \boldsymbol{\nu}_2$$
$$\mathbf{y}_3 = \mathbf{1}[\mathbf{Z}\boldsymbol{\delta}_3 + \boldsymbol{\nu}_3 > 0] \qquad \text{where } \mathbf{Z}_1 \in \mathbf{Z}$$

Assumptions:

- $(\mathbf{Z},\mathbf{y}_3)$ are always observed
- $(\mathbf{y}_1,\mathbf{y}_2)$ only observed when $\mathbf{y}_3 = 1$
- $(\boldsymbol{\epsilon}_1, \boldsymbol{\nu}_3)$ independent of $\mathbf{Z}$
- $\boldsymbol{\nu}_3 \sim N(0,1)$
- $E(\boldsymbol{\epsilon}_1|\boldsymbol{\nu}_3) = \gamma_1\boldsymbol{\nu}_3$ (always holds if $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\nu}_2$ are biv. normal)
- $E(\mathbf{Z}_2'\boldsymbol{\nu}_2) = 0 \leftarrow$ THE NEW ONE

$\mathbf{Z}$ contains should (preferably) contain at least 2 extra variables compared to $\mathbf{Z}_1$

B) Probit Selection with Endogenous RHS Variable

$$\mathbf{y}_1 = \mathbf{Z}_1\boldsymbol{\delta}_1 + \alpha_1\mathbf{y}_2 + \boldsymbol{\epsilon}_1$$
$$\mathbf{y}_2 = \mathbf{Z}_2\boldsymbol{\delta}_2 + \boldsymbol{\nu}_2$$
$$\mathbf{y}_3 = \mathbf{1}[\mathbf{Z}\boldsymbol{\delta}_3 + \boldsymbol{\nu}_3 > 0] \qquad \text{where } \mathbf{Z}_1 \in \mathbf{Z}$$

Procedure 19.2:

1. Obtain probit estimate of δ_3 from: $P(y_{j3} = 1|\mathbf{x}_j) = \Phi(Z_j\delta_3)$
   - Construct λ from this estimate and $\mathbf{Z}$: $\lambda(Z_j\hat{\delta}_3)$
2. Run 2SLS on:

$$y_{j1} = Z_{j1}\boldsymbol{\delta}_1 + \alpha_1 y_{j2} + \gamma_1\hat{\lambda}_j + error_j$$

using $\mathbf{Z}_2$ and $\hat{\lambda}$ as instruments.

Estimators $\hat{\boldsymbol{\delta}}_1$, $\hat{\alpha}_1$ and $\hat{\gamma}_1$ are consistent and asymp. normal

B) Probit Selection with Endogenous RHS Variable

$$\mathbf{y}_1 = \mathbf{Z}_1\boldsymbol{\delta}_1 + \alpha_1\mathbf{y}_2 + \boldsymbol{\epsilon}_1$$
$$\mathbf{y}_2 = \mathbf{Z}_2\boldsymbol{\delta}_2 + \boldsymbol{\nu}_2$$
$$\mathbf{y}_3 = \mathbf{1}[\mathbf{Z}\boldsymbol{\delta}_3 + \boldsymbol{\nu}_3 > 0] \qquad \text{where } \mathbf{Z}_1 \in \mathbf{Z}$$

NB! (as before)

- Test for sample selection bias: 2SLS t-test for $H_0 : \gamma_1 = 0$ (no bias)
- If the t-test shows that $\gamma_1 \neq 0$
  - Asymptotic variance of the estimator of $\boldsymbol{\beta}_1$ is complicated.
  - The robust standard errors are not enough

Exercise 5

Example 19.7 (Extension of 19.6))

- We allow $educ$ to be endogenous with IV $motheduc, fatheduc, huseduc$
- What are $\mathbf{y}_1, \mathbf{y}_2$ and $\mathbf{y}_3$?

- What are $\mathbf{Z}_1$, $\mathbf{Z}_2$, $\mathbf{Z}$?

- Run procedure 19.2: two steps
  - Run probit on the third equation, save λ with $invMillsRatio(probit.model)\$IMR1$.
  - Run *tsls* including lambdas in the first equation and the IV for education

Solution Exercise 5

Independent var	intercept	educ	exper	expersq	$\hat{\lambda}_2$	$\hat{\nu}_2$	bias?
OLS	-0.5220 **	0.1075 ***	0.0416 **	-0.0008 *			NA
Procedure 19.1	-0.5781 .	0.1091 ***	0.0439 **	-0.0008 .	0.0323		No
Procedure 19.2	-0.3249	0.0878	0.0457	-0.0009	0.0404		No

C) Tobit Selection

Model (type III Tobit):

$$\mathbf{y}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1$$
$$\mathbf{y}_2 = \max(0, \mathbf{X}\boldsymbol{\delta}_2 + \boldsymbol{\nu}_2) \qquad \text{where } \mathbf{X}_1 \in \mathbf{X}$$

Example:

- $\mathbf{y}_1$ is the log wage/hour
- $\mathbf{y}_2$ is the weekly hours of labour - extra info

Assumptions (almost as for probit selection):

- $(\mathbf{X},\mathbf{y}_2)$ are always observed
- $\mathbf{y}_1$ only observed when $\mathbf{y}_2 > 0$
- $(\boldsymbol{\epsilon}_1, \boldsymbol{\nu}_2)$ independent of $\mathbf{X}$
- $\boldsymbol{\nu}_2 \sim N(0, \tau_2^2) \leftarrow$ UNKNOWN VARIANCE
- $E(\boldsymbol{\epsilon}_1|\boldsymbol{\nu}_2) = \gamma_1\boldsymbol{\nu}_2$ (always holds if $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\nu}_2$ are biv. normal)

C) Tobit Selection

Model (type III Tobit):

$$\mathbf{y}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1$$
$$\mathbf{y}_2 = \max(0, \mathbf{X}\boldsymbol{\delta}_2 + \boldsymbol{\nu}_2) \qquad \text{where } \mathbf{X}_1 \in \mathbf{X}$$

Let s_2 be the selection indicator:

- $s_2 = 1$ if $\mathbf{y}_2 > 0$
- $s_2 = 0$ if $\mathbf{y}_2 = 0$

Just like in the probit (since s_2 is a function of $\boldsymbol{\nu}_2$ and $\mathbf{X}$):

$$E(\mathbf{y}_1|\mathbf{X}, \boldsymbol{\nu}_2, s_2) = \mathbf{X}_1\boldsymbol{\beta} + \gamma_1\boldsymbol{\nu}_2$$

Unfortunately $\boldsymbol{\nu}_2$ is unobservable

But when $\mathbf{y}_2 > 0 \Rightarrow \boldsymbol{\nu}_2 = \mathbf{y}_2 - X\boldsymbol{\delta}_2 \Rightarrow$ we can estimate $\boldsymbol{\delta}_2$ and derive $\hat{\boldsymbol{\nu}_2}$

C) Tobit Selection

Model (type III Tobit):

$$\begin{aligned} \mathbf{y}_1 &= \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1 \\ \mathbf{y}_2 &= \max(0, \mathbf{X}\boldsymbol{\delta}_2 + \boldsymbol{\nu}_2) \qquad \text{where } \mathbf{X}_1 \in \mathbf{X} \end{aligned}$$

Procedure 19.3:

1. Estimate selection equation using Tobit and get:

$$\hat{\nu}_{j2} = y_{j2} - \mathbf{x}_j\hat{\boldsymbol{\delta}}_2$$

2. Run OLS on selected sample: $y_{j1} = x_{j1}\beta_1 + \gamma_1\hat{\nu}_{j2}$

Estimators $\hat{\boldsymbol{\beta}}_1$ and $\hat{\gamma}_1$ are then consistent and asymp. normal

C) Tobit Selection

Model (type III Tobit):

$$\mathbf{y}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1$$
$$\mathbf{y}_2 = \max(0, \mathbf{X}\boldsymbol{\delta}_2 + \boldsymbol{\nu}_2) \qquad \text{where } \mathbf{X}_1 \in \mathbf{X}$$

NB!

- t-test on $H_0 : \gamma_1 = 0$ provides test of sample selection bias
- Likely to be more efficient than probit selection (exploits more information)
- $\mathbf{X} = \mathbf{X}_1$ causes no problem unlike in probit selection

Exercise 6

Example 19.8:

- Almost as example 19.6. Now $hours$ is the selection criteria.
- First run a tobit model on the selection equation

$$hours = educ+exper+expersq+age+kidslt6+kidsge6+nwifeinc$$

- Take the residuals of the above model res
- Run an OLS on

$$lwage = educ + exper + expersq + res$$

Solution Exercise 6

Independent var	intercept	educ	exper	expersq	$\hat{\lambda}_2$	$\hat{\nu}_2$	bias?
OLS	-0.5220 **	0.1075 ***	0.0416 **	-0.0008 *			NA
Procedure 19.1	-0.5781 .	0.1091 ***	0.0439 **	-0.0008 .	0.0323		No
Procedure 19.2	-0.3249	0.0878	0.0457	-0.0009	0.0404		No
Procedure 19.3	-0.3971 .	0.1033 ***	0.0368 **	-0.0007 *		-0.0001	No

D) Tobit Selection with Endogenous RHS Variable

$$\mathbf{y}_1 = \mathbf{Z}_1\boldsymbol{\delta}_1 + \alpha_1\mathbf{y}_2 + \boldsymbol{\epsilon}_1$$
$$\mathbf{y}_2 = \mathbf{Z}\boldsymbol{\delta}_2 + \boldsymbol{\nu}_2$$
$$\mathbf{y}_3 = \max(0, \mathbf{Z}\boldsymbol{\delta}_3 + \boldsymbol{\nu}_3) \qquad \text{where } \mathbf{Z}_1 \in \mathbf{Z}$$

Assumptions:

- $(\mathbf{Z},\mathbf{y}_3)$ are always observed
- $(\mathbf{y}_1,\mathbf{y}_2)$ only observed when $\mathbf{y}_3 > 0$
- $(\boldsymbol{\epsilon}_1, \boldsymbol{\nu}_3)$ independent of $\mathbf{Z}$
- $\boldsymbol{\nu}_3 \sim N(0, \tau_3^2)$ (unknown variance)
- $E(\boldsymbol{\epsilon}_1|\boldsymbol{\nu}_3) = \gamma_1\boldsymbol{\nu}_3$ (always holds if $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\nu}_3$ are biv. normal)
- $E(\mathbf{Z}'\boldsymbol{\nu}_2) = 0 \leftarrow$ THE NEW ONE

$\mathbf{Z}$ should contain at least 1 extra variable compared to $\mathbf{Z}_1$

D) Tobit Selection with Endogenous RHS Variable

$$\begin{aligned}
\mathbf{y}_1 &= \mathbf{Z}_1\boldsymbol{\delta}_1 + \alpha_1\mathbf{y}_2 + \boldsymbol{\epsilon}_1 \\
\mathbf{y}_2 &= \mathbf{Z}\boldsymbol{\delta}_2 + \boldsymbol{\nu}_2 \\
\mathbf{y}_3 &= \max(0, \mathbf{Z}\boldsymbol{\delta}_3 + \boldsymbol{\nu}_3) \qquad \text{where } \mathbf{Z}_1 \in \mathbf{Z}
\end{aligned}$$

Procedure 19.4:

1. Estimate selection equation using Tobit and get:

$$\hat{\nu}_{j3} = y_{j3} - Z_j\hat{\boldsymbol{\delta}}_3$$

2. Run 2SLS on

$$y_{1j} = Z_{1j}\boldsymbol{\delta}_1 + \alpha_1 y_{2j} + \gamma_1\hat{\nu}_{3j} + error_j$$

Estimators $\hat{\boldsymbol{\delta}}_1$, $\hat{\alpha}_1$ and $\hat{\gamma}_1$ are consistent and asymp. normal

D) Tobit Selection with Endogenous RHS Variable

$$\mathbf{y}_1 = \mathbf{Z}_1\boldsymbol{\delta}_1 + \alpha_1\mathbf{y}_2 + \boldsymbol{\epsilon}_1$$
$$\mathbf{y}_2 = \mathbf{Z}\boldsymbol{\delta}_2 + \boldsymbol{\nu}_2$$
$$\mathbf{y}_3 = \max(0, \mathbf{Z}\boldsymbol{\delta}_3 + \boldsymbol{\nu}_3) \qquad \text{where } \mathbf{Z}_1 \in \mathbf{Z}$$

NB!

- 2SLS t-test on $H_0 : \gamma_1 = 0$ provides test of sample selection bias
- If $\gamma \neq 0$ (statistically speaking) then standard errors should be corrected using two-step methods
- Likely to be more efficient than probit selection (exploits more information)
- $\mathbf{Z}$ should contain at least 1 extra variable compared to $\mathbf{Z}_1$

Exercise 7

- Using your own words, what is the difference between a probit and a tobit selection equation?
  - Compare (λ and ν) from the second stage.
  - Which one uses most information?
- Look at example 19.8.
  - How does it change results when using tobit instead of probit selection?

Solution Exercise 7

Independent var	intercept	educ	exper	expersq	$\hat{\lambda}_2$	$\hat{\nu}_2$	bias?
OLS	-0.5220 **	0.1075 ***	0.0416 **	-0.0008 *			NA
Procedure 19.1	-0.5781 .	0.1091 ***	0.0439 **	-0.0008 .	0.0323		No
Procedure 19.2	-0.3249	0.0878	0.0457	-0.0009	0.0404		No
Procedure 19.3	-0.3971 .	0.1033 ***	0.0368 **	-0.0007 *		-0.0001	No
Procedure 19.4	-0.1846	0.0874	0.0366	-0.0007		-0.0001	No

E) Structural Tobits with Selection

$$\mathbf{y}_1 = \mathbf{Z}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1$$
$$\mathbf{y}_2 = \max(0, \mathbf{Z}_2\boldsymbol{\beta}_2 + \alpha_2\mathbf{y}_1 + \boldsymbol{\epsilon}_2)$$

Assumptions:

- $(\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$,$\mathbf{y}_2)$ are always observed
- $\mathbf{y}_1$ only observed when $\mathbf{y}_2 > 0$
- $(\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2)$ independent of $\mathbf{Z}$ and bivariate normal
- $\mathbf{Z}_1$ contains at least one significant variable which is not in $\mathbf{Z}_2$

We have selection and SIMULTANEITY

E) Structural Tobits with Selection

$$\mathbf{y}_1 = \mathbf{Z}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1$$
$$\mathbf{y}_2 = \max(0, \mathbf{Z}_2\boldsymbol{\beta}_2 + \alpha_2\mathbf{y}_1 + \boldsymbol{\epsilon}_2)$$

Example:

- $\mathbf{y}_1$ is wage; and $\mathbf{y}_2$ is labour supply
- Wage only observed when supply$>$0
- But wage affects supply!

If $\mathbf{y}_1$ was always observed $\Rightarrow$ just run 2SLS on the first equation.

But $\mathbf{y}_1$ is not always observed, how do we estimate it?

E) Structural Tobits with Selection

$$\mathbf{y}_1 = \mathbf{Z}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1$$
$$\mathbf{y}_2 = \max(0, \mathbf{Z}_2\boldsymbol{\beta}_2 + \alpha_2\mathbf{y}_1 + \boldsymbol{\epsilon}_2)$$

Estimation:

- The reduced form of the model is:

$$\mathbf{y}_1 = \mathbf{Z}_1\beta_1 + \boldsymbol{\epsilon}_1$$
$$\mathbf{y}_2 = \max(0, \mathbf{Z}_2\boldsymbol{\beta}_2 + \alpha_2\mathbf{Z}_1\boldsymbol{\beta}_1 + \boldsymbol{\nu}_2)$$

E) Structural Tobits with Selection

Model:

$$\mathbf{y}_1 = \mathbf{Z}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1$$
$$\mathbf{y}_2 = \max(0, \mathbf{Z}_2\boldsymbol{\beta}_2 + \alpha_2\mathbf{y}_1 + \boldsymbol{\epsilon}_2)$$

Reduced form:

$$\mathbf{y}_1 = \mathbf{Z}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1$$
$$\mathbf{y}_2 = \max(0, \mathbf{Z}_2\boldsymbol{\beta}_2 + \alpha_2\mathbf{y}_1 + \mathbf{Z}_1\boldsymbol{\beta}_1 + \boldsymbol{\nu}_2)$$

Procedure 19.5 (three steps procedure):

- Use procedure 19.3 (two steps) to obtain $\hat{\beta}_1$
- Obtain $\hat{\boldsymbol{\beta}}_2$ and $\hat{\alpha}_2$ from tobit equation

$$y_{2j} = \max(0, Z_{2j}\boldsymbol{\beta}_2 + \alpha_2 y_{1j} + Z_{1j}\hat{\boldsymbol{\beta}}_1 + error_j)$$

Summary

- Non-random samples can result in biased estimates
  - Especially, if sample selection is correlated with unobservables
  - We often talk about "self-selection".
- We can tackle this, by modelling the selection process
  - As a probit or
  - As a tobit
  - We have seen several approaches: A)-D) and E)
- But as with IV, this requires some instruments/exclusion restrictions.
  - These can be hard to find.