Sample Selection

(GB: Chapter 19.3-19.7)

Isabel Casas

SYDDANSK UNIVERSITET

Motivation

- So far we have assumed a random sample from the population.
  - Each individual of the population has the same chance to be selected

Motivation

- So far we have assumed a random sample from the population.
  - Each individual of the population has the same chance to be selected
- Often this is not the case (truncated samples):
  - Self selection: individual choice of whether or not to participate in the activity of interest
  - Sample selection: Those who participate in the activity of interest are oversampled (only participants)

Motivation

We define:

- Specify population from where the sample has to be obtained
- *Random sample* a sample where all individuals have the same probability to be chosen
- The *selected sample* is a nonrandom sample following a selection rule. The nonrandom nature is due to
  - sample design
  - behaviour of the units being sample
  - nonresponse in surveys
  - Selection mechanisms that result into nonrandom samples

Nonrandomly selected samples

Example 19.1 (Saving Function). We have only families whose household head is greater than 45 old

$$saving = \beta_0 + \beta_1\ income + \beta_2\ age + \beta_3\ married + \beta_4\ kid + \boldsymbol{\epsilon}$$

We are interested in results of every family but we only have data for $age \geq 45 \Rightarrow$ sample selection issue.

- Not such a big problem. We define the population as the group of families greater than 45 and we have a random sample of that subset

Nonrandomly selected samples

Example 19.2 (Truncation based on wealth). We are interested in estimating how a worker eligibility in a certain plan affects the wealth of the family.

$$wealth = \beta_0 + \beta_1\ plan + \beta_2\ educ + \beta_3\ age + \beta_4\ income + \boldsymbol{\epsilon}$$

We only have data for people with a net wealth less than \$200000. We are interested in results of every family but we have a selected sample depending of the $\mathbf{y}$ variable

- This is a more serious sampling problem

Nonrandomly selected samples

Example 19.3 (Wage offer function). We are interested in estimating the wage offer function for people of working age.

We are interested in results for every working age worker but we only have the ones who have took a job offer.

This is a self-selection example because people self-select into employment. It makes it harder to determine the causation.

More examples

- Self-selection example: sample only individuals who stay in a diet program (the drop outs are not taken into account)
- Sample selection example: we record wages from individuals in the labour force. We are only picking the sample of "industrious" people rather than the population as a whole
- Surviving selection: We are interested in data on assets for 15 years, we only take the ones who have been traded in the market for the last 15 years. We do not take into account the drop outs.

It doesn't matter what is the name we give to the sample selection problem, the result is that it can be difficult to determine the causation because of the bias induced.

Motivation

- How does sample selection affect estimation results?
- Should we do something about it?

Outline

- Selection on the basis of $\mathbf{X}$: exogenous sampling
  - Eg: Segmenting an existing sample by gender or status.
  - Sample selection problem can be ignored
- Selection on the basis of $\mathbf{y}$ (the LHS variable)
  - More difficult to resolve

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

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

Cases where the sample selection can be ignored

- Selection on the basis of $\mathbf{X}$ (no endogeneity)
  - The selection rule is independent of $\mathbf{X}$ and ϵ or
  - The selection rule is deterministic

$$\Downarrow$$

OLS yields consistent estimators

- Selection on the basis of $\mathbf{X}$ (with endogeneity)
  - The selection rule is independent of $\mathbf{z}$ (instruments) and ϵ or
  - The selection rule is deterministic

$$\Downarrow$$

2SLS yields consistent estimators

Selection on the basis of $\mathbf{X}$

Assume the following model:

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \ldots + \beta_k \mathbf{x}_k + \boldsymbol{\epsilon} \qquad E(\boldsymbol{\epsilon}|\mathbf{X}) = 0$$

- Let $s = 0, 1$ be the selection indicator.
  - Rule: $20 < age < 45$ ($s = 1$), otherwise $s = 0$
  - $age \in \mathbf{X}$
- The crucial assumption is that $\boldsymbol{\epsilon}$ (the unobserved part) is randomly distributed in the population.
- Sample selection is a problem if we get a non-random sample of the $\boldsymbol{\epsilon}$'s (e.g. all those with high $\boldsymbol{\epsilon}$'s)

Selection on the basis of $\mathbf{X}$

- If $\mathbf{X}$ are exogenous variables and
- If $E(\boldsymbol{\epsilon}|\mathbf{X}, s) = 0$, then sample selection can be ignored
- I.e. in particular if:
  - s is independent of $(\mathbf{X}, \boldsymbol{\epsilon})$: $E(\boldsymbol{\epsilon}|\mathbf{X}, s) = E(\boldsymbol{\epsilon}|\mathbf{X}) = 0$ or
  - s is a deterministic function of $\mathbf{X}$: $E(\boldsymbol{\epsilon}|\mathbf{X}, s) = E(\boldsymbol{\epsilon}|\mathbf{X}) = 0$
- Intuition:
  - If s is independent of $(\mathbf{X},\boldsymbol{\epsilon})$ and hence $(\mathbf{X},\mathbf{y}) \Rightarrow$ the selected sample is also a random sample
  - If s is a deterministic function of $\mathbf{X} \Rightarrow$ the selection does not give us those with particularly high (or low) $\boldsymbol{\epsilon}$'s, as $\mathbf{X}$ contain no info about the $\boldsymbol{\epsilon}$'s
- Estimation:
  - OLS yields consistent estimates

Selection on the basis of $\mathbf{X}$

- This result can be extended to the case with endogenous $\mathbf{X}$'s as long as:

$$E(\boldsymbol{\epsilon}|\mathbf{z}, s) = 0$$

- Where $\mathbf{z}$ is the vector of instruments (including exogenous $\mathbf{X}$'s), i.e. if
  - s is independent of ($\mathbf{z}$,$\boldsymbol{\epsilon}$), or
  - s is a deterministic function of $\mathbf{z}$
- This is stronger than what we need for random sample 2SLS ($E(\mathbf{z}, \boldsymbol{\epsilon}) = 0$)
- Estimation by 2SLS still yields consistent and asymptotically normal estimates.
- But not if s is a function of $\mathbf{y}$!! (we return to this case)

Selection on the basis of $\mathbf{X}$

Example:

$$wage = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma}\ age + \boldsymbol{\epsilon}$$
$$E(\boldsymbol{\epsilon}|\mathbf{X}, age) = 0$$

Selection rules:

- Deterministic: $k_1 < age < k_2 \Rightarrow$ determined by RHS $\Rightarrow$ OLS is consistent
- Random: $age > k_1 + r$ where r is a random variable, then OLS still ok if r does not contain info about $\boldsymbol{\epsilon}$:

$$E(\boldsymbol{\epsilon}|\mathbf{X}, age, r) = 0$$

Selection on the basis of $\mathbf{X}$

The result can also be extended to non-linear models:

- Probit
- Logit

If selection is ignorable: $E(\mathbf{y}|\mathbf{X}, s) = E(\mathbf{y}|\mathbf{X})$. This occurs if:

- s is a deterministic function of $\mathbf{X}$, or
- s is independent of ($\mathbf{X}$,$\mathbf{y}$)

Exercise 1 (5 minutes)

Discuss the following questions with your colleagues:

- In the linear model: When can selection be ignored? What is the intuition for this?
- In the probit/logit models: When can selection be ignored? What is the intuition for this?
- What if we select only those with:
  - $(\mathbf{x}_1 + \mathbf{x}_2)^2 > 5$?
  - $(\mathbf{x}_1 + \mathbf{x}_2)^2 > Y$?
  - $(\mathbf{x}_1 + \mathbf{x}_2)^2 >$ average of Y ?

Look at simulations!!!

Deterministic selection of $\mathbf{y}$: Truncated regression

- Selection on the basis of $\mathbf{y}$
- Selection rule is deterministic and known so it can be accounted for
- Find the log-lik function of the truncated distribution and obtain estimates by ML.
- If $\mathbf{y}|\mathbf{X} \sim N(0, \sigma^2)$ then we can use the truncated Tobit model.
- It is less efficient than censored regression
- If there is heteroskedasticity or the distribution is misspecified $\Rightarrow$ inconsistency

Truncated regression

A standard linear model:

$$E(y_j|\mathbf{x}_j) = \mathbf{x}_j\boldsymbol{\beta}$$

Selection rule:

$$s_j = \mathbf{1}[a_1 < y_j < a_2]$$

- Rule is known and observable and it depends on $\mathbf{y}$
  - Income equations: Individuals in certain income brackets
  - Labour supply: normal time workers
- The value of $\mathbf{X}\boldsymbol{\beta}$ will impose the observed $\boldsymbol{\epsilon}$'s in our sample and vice versa because the value of $\mathbf{y}$ is bounded.
- In our estimation, we need to condition on the rule

Truncated regression

- We know the probability of $y_j < y$ conditional on $\mathbf{x}_j$: $F(y|\mathbf{x}_j; \boldsymbol{\beta}, \boldsymbol{\gamma})$
- What is the distribution of y_j conditional on $\mathbf{x}_j$ and selection $a_1 < y < a_2$?

Truncated regression

- We know the probability of $y_j < y$ conditional on $\mathbf{x}_j$: $F(y|\mathbf{x}_j; \boldsymbol{\beta}, \boldsymbol{\gamma})$
- What is the distribution of y_j conditional on $\mathbf{x}_j$ and selection $a_1 < y < a_2$?

$$
\begin{aligned}
P(y_j < y|\mathbf{x}_j, s = 1) &= \frac{P(y_j < y, s = 1|\mathbf{x}_j)}{P(s = 1|\mathbf{x}_j)} \\
&= \frac{P(y_j < y, a_1 < y_j < a_2|\mathbf{x}_j)}{P(a_1 < y_j < a_2|\mathbf{x}_j)} \\
&= \frac{P(a_1 < y_j < y|\mathbf{x}_j)}{P(a_1 < y_j < a_2|\mathbf{x}_j)} \\
&= \frac{F(y|\mathbf{x}_j; \boldsymbol{\beta}, \boldsymbol{\gamma}) - F(a_1|\mathbf{x}_j; \boldsymbol{\beta}, \boldsymbol{\gamma})}{F(a_2|\mathbf{x}_j; \boldsymbol{\beta}, \boldsymbol{\gamma}) - F(a_1|\mathbf{x}_j; \boldsymbol{\beta}, \boldsymbol{\gamma})}
\end{aligned}
$$

Truncated regression

Now the density function of y_j given $\mathbf{x}_j$ and $s_j = 1$ can be found by differentiating wrt y:

$$f(y|\mathbf{x}_j, s = 1) = \frac{f(y|\mathbf{x}_j; \boldsymbol{\beta}, \boldsymbol{\gamma})}{F(a_2|\mathbf{x}_j; \boldsymbol{\beta}, \boldsymbol{\gamma}) - F(a_1|\mathbf{x}_j; \boldsymbol{\beta}, \boldsymbol{\gamma})}$$

Which is just the unconditional density (the numerator) divided by the probability of being sampled (the denominator) I.e. we are simply rescaling the densities to the new interval: (a_1, a_2)

How to get the log-lik?

- Take logs of the above and
- sum over the selected sample.

Truncated regression

Special case: if $\mathbf{y}|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2) \Rightarrow$ use Φ and ϕ instead of F and f:

- The Truncated Tobit Model (or Truncated Normal Regression Model)

Censored regression models (Ch. 16): Only $\mathbf{y}$ was not observed for some individuals but information about $\mathbf{X}$ from these could still be used.

Truncated regression model (now): Neither $\mathbf{X}$ nor $\mathbf{y}$ is observed for some individuals

Same code but censoring is preferable over truncation

Truncated regression

If misspecification:

- Heteroscedasticity, or
- Non-normality

This implies:

- Changes the log-lik
- Inconsistent estimates (same as in Chapter 16)

Exercise 2 (5 minutes)

Discuss with your colleagues:

- Think of an example where selection on $\mathbf{y}$ is likely to appear.
- How will that ? intuitively affect your estimations if not corrected for?
- Explain ? using your own words ? what the correction does (to the likelihood function)? Consider, e.g., an observation with an extreme $\mathbf{X}$ but a $\mathbf{y}$ value between a_1 and a_2. How will its density be altered?

Selection determined by behaviour

Cases to consider:

A) Selection can be modelled as probit

B) ... and with endogeneous RHS variables

D) Selection can be modelled as tobit

D) ... and with endogeneous RHS variables

E) Structural Tobit with selection

Selection determined by behaviour

Classic example: Wage offer equation (Example 19.5)

- How education affects how much a person could earn in the labour market.
- Your first intuition is to run OLS as if you had a random sample
- But only individuals with wages above their reservation wage have jobs and can thus be sampled
- Reservation wage varies across individuals (otherwise, we could use approach from previous slides)

Other example:

- Non-response in surveys

Selection determined by behaviour

The wage equation example (formally):

- Wage: $\log w_j = \mathbf{x}_{j1}\boldsymbol{\beta}_1 + \epsilon_{1j}$
- Reservation wage:
  $\log w_j^{res} = \mathbf{x}_{j2}\boldsymbol{\beta}_2 + \gamma_2 \underbrace{a_j}_{\text{no wage income}} + \epsilon_{2j}$
- 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):

$$\log w_j > \log w_j^{res} \Rightarrow \mathbf{x}_{j1}\boldsymbol{\beta}_1 - \mathbf{x}_{j2}\boldsymbol{\beta}_2 - \gamma_2 a_j + \epsilon_{1j} - \epsilon_{2j} > 0$$

Selection determined by behaviour

Structural equation (what we are interested on studying):

$$y_{1j} = \log w_j = \mathbf{x}_{j1}\boldsymbol{\beta}_1 + \epsilon_{j1}$$

Selection equation:

$$\begin{aligned} y_{2j} =& \mathbf{1}[\mathbf{x}_{j1}\boldsymbol{\beta}_1 - \mathbf{x}_{j2}\beta_2 - \gamma_2 a_j + \epsilon_{1j} - \epsilon_{2j} > 0] \\ =& \mathbf{1}[\mathbf{x}_j\boldsymbol{\delta}_2 + \nu_{2j} > 0] \end{aligned}$$

A) Probit selection

Model (type II Tobit):

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

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 $\mathbf{X}$ (exogenous)
- $\boldsymbol{\nu}_2 \sim N(0, 1)$
- $E(\boldsymbol{\epsilon}_1|\boldsymbol{\nu}_2) = \boldsymbol{\gamma}_1\boldsymbol{\nu}_2$ (rule and error correlated)

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

A) Probit selection

Model (type II Tobit):

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

We know how to calculate:

$$E(\mathbf{y}_1|\mathbf{X}) =$$
$$P(\mathbf{y}_2 = 1|\mathbf{X}) =$$

A) Probit selection

Model (type II Tobit):

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

We know how to calculate:

$$\begin{aligned}E(\mathbf{y}_1|\mathbf{X}) &= E(\mathbf{y}_1|\mathbf{X}_1) = \mathbf{X}_1\boldsymbol{\beta}_1 \\ P(\mathbf{y}_2 = 1|\mathbf{X}) &= \Phi(\mathbf{X}\boldsymbol{\delta}_2)\end{aligned}$$

A) Probit selection

Model (type II Tobit):

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

We know how to calculate:

$$\begin{aligned}E(\mathbf{y}_1|\mathbf{X}) &= E(\mathbf{y}_1|\mathbf{X}_1) = \mathbf{X}_1\boldsymbol{\beta}_1 \\ P(\mathbf{y}_2 = 1|\mathbf{X}) &= \Phi(\mathbf{X}\boldsymbol{\delta}_2)\end{aligned}$$

How do we estimate:

$$E(\mathbf{y}_1|\mathbf{X}, \mathbf{y}_2 = 1)$$

A) Probit selection

Model (type II Tobit):

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

- How does $E(\mathbf{y}_1|\mathbf{X}, \mathbf{y}_2 = 1)$ depend on β_1?
- Notice that: $E(\mathbf{y}_1|\mathbf{X}, \boldsymbol{\nu}_2) = \mathbf{X}_1\boldsymbol{\beta}_1 + \gamma_1\boldsymbol{\nu}_2$ (sample bias?)
- Two cases to consider:
  1. $\gamma_1 = 0$ and
  2. $\gamma_1 \neq 0$

A) Probit selection

Model (type II Tobit):

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

Case 1:

- $\gamma_1 = 0$ ($\boldsymbol{\epsilon}_1$ and $\boldsymbol{\nu}_2$ are uncorrelated)

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

- Hence

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

- In this case there is no sample selection problem
- Estimate $\boldsymbol{\beta}$ consistently with OLS on the selected sample

A) Probit selection

Model (type II Tobit):

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

Case 2:

- $\boldsymbol{\gamma}_1 \neq 0$ ($\boldsymbol{\epsilon}_1, \boldsymbol{\nu}_2$ are correlated):

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

- Conditioning on the selection variable $\mathbf{y}_2$ changes the expected value of $\mathbf{y}_1|\mathbf{X}$
- Because $\mathbf{y}_2$ and $\mathbf{X}$ gives us info about $\boldsymbol{\nu}_2$ and hence $\boldsymbol{\epsilon}_1$
- If we knew $h()$ then we could use OLS on the selected sample and obtain consistent estimates of $\boldsymbol{\beta}_1$ and $\boldsymbol{\gamma}_1$
- Omitting $h() \Rightarrow$ biased or even inconsistent estimate of $\boldsymbol{\beta}_1$

A) Probit selection

Case 2 (continued): The selected sample has $\mathbf{y}_2 = 1$, $h(X, 1)$

$$\begin{aligned} h(X,1) = E(\boldsymbol{\nu}_2|\mathbf{X}, \mathbf{y}_2 = 1) &= E(\boldsymbol{\nu}_2|\mathbf{X}\boldsymbol{\delta}_2 + \boldsymbol{\nu}_2 > 0) \\ &= E(\boldsymbol{\nu}_2|\boldsymbol{\nu}_2 > -\mathbf{X}\boldsymbol{\delta}_2) \\ &= \lambda(\mathbf{X}\boldsymbol{\delta}_2) = \frac{\phi(\mathbf{X}\boldsymbol{\delta}_2)}{\Phi(\mathbf{X}\boldsymbol{\delta}_2)} \end{aligned}$$

- This is the inverse Mill's ratio (see chapter 17)
- We then have:

$$E(\mathbf{y}_1|\mathbf{X}, \mathbf{y}_2 = 1) = \mathbf{X}_1\boldsymbol{\beta}_1 + \gamma_1\lambda(X\delta_2)$$

- We know everything in this expression but $\boldsymbol{\delta}_2 \Rightarrow$ we must estimate it
- Estimation with 2 step procedure.

A) Probit selection

Model (type II Tobit):

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

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:

$$y_{1j} = x_{1j}\beta_1 + \gamma_1\hat{\lambda}_j + error_j$$

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

A) Probit selection

NB!

- Test for sample selection bias: t-test for $H_0 : \boldsymbol{\gamma}_1 = 0$ (no bias)
  - The asymptotic variance of $\boldsymbol{\beta}_1$ and $\boldsymbol{\gamma}_1$ are not affected by $\hat{\boldsymbol{\delta}}_2$ under the null
- If the t-test shows that $\boldsymbol{\gamma}_1 \neq 0$
  - Asymptotic variance of the estimator of $\boldsymbol{\beta}_1$ is complicated.
  - The robust standard errors is 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)

A) Probit selection – special case

Case 2 (special case) with stronger assumption: $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\nu}_2$ are bivariate normal

$$\begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\nu}_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & 1 \end{pmatrix}\right)$$

- We can then run partial maximum likelihood instead of two-step procedure.
- We cannot run full ML as $\mathbf{y}_1$ is not always observed
- Advantage of ML: More efficient, and we get standard errors directly from ML estimation
- In R: $heckit(selection, outcome, data)$ or $selection(selection, outcome, data)$

A) Probit selection – an extension

Under the assumption that $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\nu}_2$ are bivariate normal,

- The set-up can also be extended to binary response models:

$$\begin{aligned} \mathbf{y}_1 &= \mathbf{1}[\mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1 > 0] \\ \mathbf{y}_2 &= \mathbf{1}[\mathbf{X}\boldsymbol{\delta}_2 + \boldsymbol{\nu}_2 > 0] \quad \mathbf{X}_1 \in X \end{aligned}$$

- Example:
  - $\mathbf{y}_1$: Participation in on-the-job training
  - $\mathbf{y}_2$: Labour market participation
- Estimation method: Partial max-lik as above

Exercise 4 (3 minutes)

Discuss with your colleagues:

- When can't we just run OLS on the equation of interest when we have sample selection?
- Why?
- How can we fix it?
- In R?