Sample selection

- In many economic problems sample we use for estimation is a nonrandomly *selected sample*
- We will analyze three selection mechanisms
  - selection based on selection indicator
  - selection based on response variable
  - selection based on separate selection model (incidental truncation)
- It is always important to correctly define the population in question

Example 1. *(Wooldridge) Saving function*

$$\textit{saving} = \beta_0 + \beta_1 \textit{income} + \beta_2 \textit{age} + \beta_3 \textit{married} + \beta_4 \textit{kids} + u$$

but we only observe households for which age > 45

Example 2. *(Wooldridge) Estimate effect of eligibility in a pension plan on wealth.*

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

but we only have observations for people with wealth $< \$200,000$*.*

Example 3. *(Wooldridge) Wage offer function. By definition it should represent the (potential) wages of all people - not only the ones who are working. We only have data on wages for people who are actually working - the selection is based on another variable, labor force participation.*

Ignorability of selection

- Define selection indicator s ($s = 1$ if observation is in the sample)
- Selection problem can be ignored if in model

$$y = \beta_1 + \beta_2 x_2 + \ldots + \beta_K x_K + \varepsilon$$

$$\mathrm{E}\left(u \middle| \boldsymbol{z}, s\right) = 0$$

  where $\boldsymbol{z}$ is the vector of instruments. The consistent estimate of $\boldsymbol{\beta}$ can be obtained with standard $2SLS$.

- **Interpretation**: selection can be ignored if it is possible to find instruments $\boldsymbol{z}$ such that given $\boldsymbol{z}$ selection is not correlated with random term u.

- If selection is independent on $\boldsymbol{z}, u$ or based on deterministic function of $s = f(\boldsymbol{z})$ then if $\mathrm{E}(u|\,\boldsymbol{z}) = 0$ then $\mathrm{E}(u|\,\boldsymbol{z}, s) = \mathrm{E}(u|\,\boldsymbol{z}) = 0$

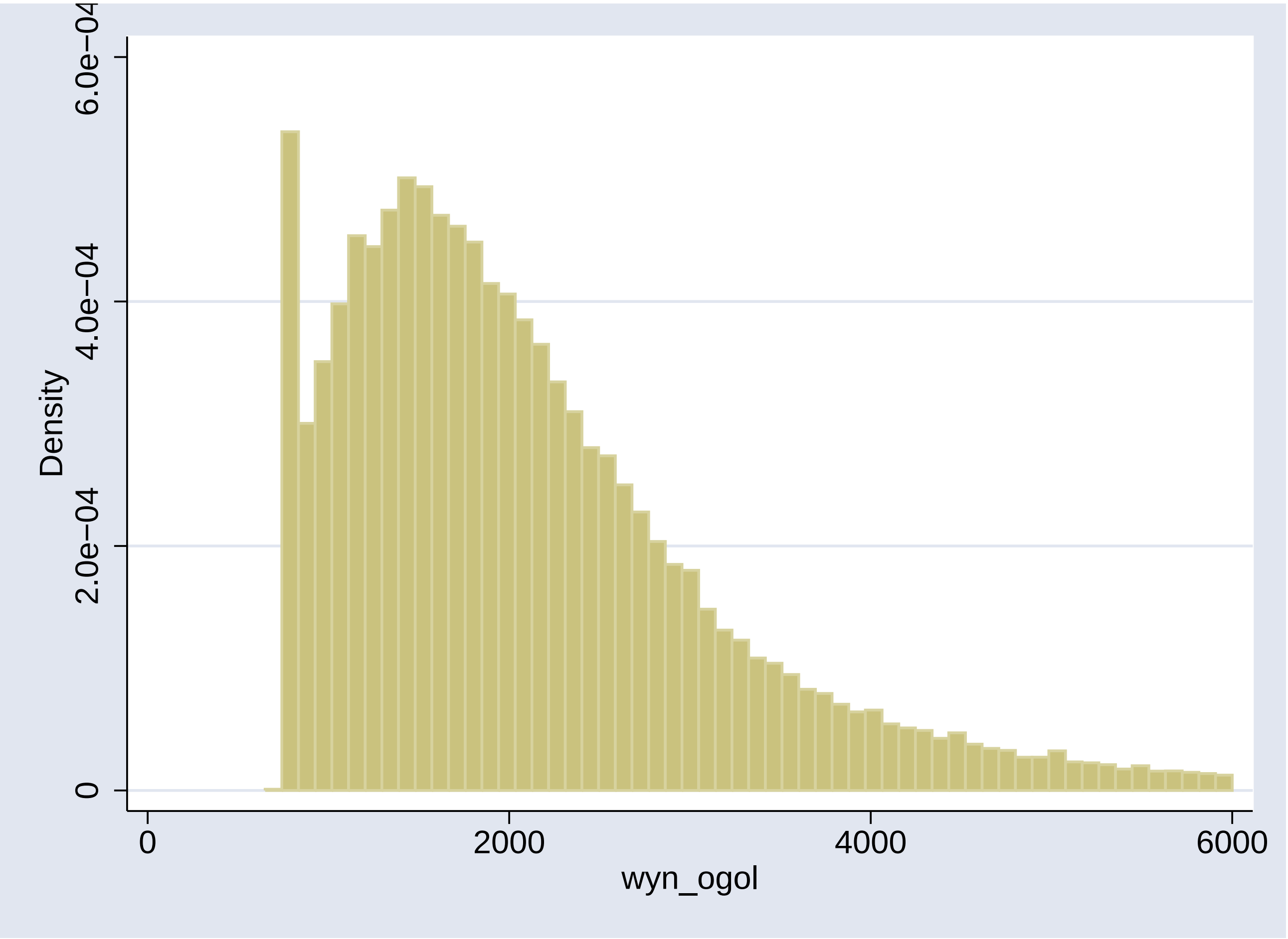
- The simplest case: $\boldsymbol{z} = \boldsymbol{x}$ (Classical Regression) and selection based on $\boldsymbol{x}$. In this case we can use OLS (Saving function example)

Selection based on response variable

- Hausman and Wise study (1977) - determinants of earnings, sample only for the people participating in negative income experiment, no data available for the people with income higher than same treshold

- Wealth example: we only observe the people whose wealth is smaller than $\$200,000$.

Example 4. *Gross wages in Poland, sample based on firm survey. .*



Only wages higher than minimal wage are declared!

- Difference with top coding - in the case of top coding we do not observe the response variable but we observe explanatory variable (see top coding for wealth example)

- In the case of sample selection (or truncated sample) - we observe nothing about the units not selected to the sample

- Assume that observation is selected to the sample if $s_i = 1$ and

$$s_i = 1\left[a_1 < y_i < a_2\right]$$

- cdf of observed y_i is given by

$$\Pr\left(y_i \leq Y \middle| \boldsymbol{x}_i, s_i = 1\right) = \frac{\Pr\left(y_i \leq Y, s_i = 1 \middle| \boldsymbol{x}_i\right)}{\Pr\left(s_i = 1 \middle| \boldsymbol{x}_i\right)}$$

where

- But $\Pr\left(s_i = 1 | \boldsymbol{x}_i\right) = \Pr\left(a_1 < y_i < a_2\right) = F\left(a_2 | \boldsymbol{x}_i; \boldsymbol{\beta}, \boldsymbol{\gamma}\right) - F\left(a_1 | \boldsymbol{x}_i; \boldsymbol{\beta}, \boldsymbol{\gamma}\right)$

- $\Pr\left(y_i \leq Y, s_i = 1 | \boldsymbol{x}_i\right) = \Pr\left(a_1 < y_i \leq Y | \boldsymbol{x}_i\right) = F\left(Y | \boldsymbol{x}_i; \boldsymbol{\beta}, \boldsymbol{\gamma}\right) - F\left(a_1 | \boldsymbol{x}_i; \boldsymbol{\beta}, \boldsymbol{\gamma}\right)$

- Density function is then

$$\frac{\partial \Pr\left(y_i \leq Y | \boldsymbol{x}_i, s_i = 1\right)}{\partial Y} = \frac{f\left(Y | \boldsymbol{x}_i; \boldsymbol{\beta}, \boldsymbol{\gamma}\right)}{F\left(a_2 | \boldsymbol{x}_i; \boldsymbol{\beta}, \boldsymbol{\gamma}\right) - F\left(a_1 | \boldsymbol{x}_i; \boldsymbol{\beta}, \boldsymbol{\gamma}\right)}$$

- If we assume that the distribution $F\left(\cdot\right)$ is normal distribution we will obtain the *truncated tobit* (called too truncated normal regression)

Incidental truncation

- In this case we have problem of self-selection of the units in the sample
- Selection is based on decision of the unit partially based on unobserved characteristics

Example 5. *(Wooldridge) Labor force participation and the wage offer. We are interested in* $\mathrm{E}\left(w_i^o \middle| \boldsymbol{x}_i\right)$ *and collect the sample for the people in working age. But only for the people actually working we obtain data on wage offer* w_i^o*. The utility maximization problem for individual is as follows:*

$$\max util\left(q, h\right) \ \textbf{\textit{s.t.}}\ 0 \le h \le 168$$

where h_i *are the hours worked and* $q_i = w_i^o h_i + a_i$ *is income and* a_i *is nonlabor income. We assume that marginal utility w.r.t income is positive and with* *respect to hours worked negative. Marginal utility of work is given by*

$$\begin{aligned}\frac{\partial util\left(q,h_i\right)}{\partial h_i} &= \frac{\partial util\left(q_i,h_i\right)}{\partial q_i}\frac{\partial q_i}{\partial h_i}+\frac{\partial util\left(q,h_i\right)}{\partial h_i}\\ &= \frac{\partial util\left(q_i,h_i\right)}{\partial q_i}w_i^o+\frac{\partial util\left(q,h_i\right)}{\partial h_i}\end{aligned}$$

First order condition is given by equality

$$\frac{\partial util\left(q,h_i\right)}{\partial h_i}=0$$

Individual will only work ($h_i>0$) if for $h_i=0$ his/her marginal utility of work is positive:

$$\frac{\partial util\left(a_i,0\right)}{\partial q_i}w_i^o+\frac{\partial util\left(a_i,0\right)}{\partial h_i}\geq 0$$

and so

$$w_i^o \geq -\frac{\partial util\,(a_i, 0)}{\partial h_i} \bigg/ \frac{\partial util\,(a_i, 0)}{\partial q_i} = w_i^r$$

where w_i^r is called reservation wage.

If wage offer is given by $w_i^o = \exp\left(\boldsymbol{x}_{i1}\boldsymbol{\beta} + u_{i1}\right)$ and reservation wage by $w_i^r = \exp\left(\boldsymbol{x}_{i2}\boldsymbol{\beta} + u_{i2}\right)$ then we have the following two equations in the model:

1. *wage equation*

$$\log w_i^o = \boldsymbol{x}_{i1}\boldsymbol{\beta} + u_{i1}$$

2. *participation equation*

$$\log w_i^o - \log w_i^r = \boldsymbol{x}_{i1}\boldsymbol{\beta} + u_{i1} - \boldsymbol{x}_{i2}\boldsymbol{\beta} - u_{i2} = \boldsymbol{x}_i\boldsymbol{\delta}_2 + v_{i2}$$

If individual is participating ($\log w_i^o - \log w_i^r > 0$) we observe w_i^o. We do not observe w_i^r but only $\boldsymbol{x}_i$ and the decision to participate or not.

Remarks:

- *as* $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2})$ *all the variables included in wage equation should also be included in participation equation*

- *as* $v_{i2} = u_{i1} - u_{i2}$ *we expect positive correlation between* u_{i1} *and* v_{i2}

- *Difference between the incidental truncation and top coding: we do not know the* $\log w_i^r$*, so we do not know exactly when the data is censored.*

- $\log w_i^o - \log w_i^r = \boldsymbol{x}_i \boldsymbol{\delta}_2 + v_{i2}$ *is a reduced form equation (wage is not included in this equation), the parameters of this equation can not be interpreted as parameters of the labor supply function.*

- The basic model

$$y_1 = x_1\beta_1 + u_1$$
$$y_2 = 1\left[\boldsymbol{x}\boldsymbol{\delta}_2 + v_2 > 0\right]$$

- This model is called Heckit or Tobit type II model

- Assumptions
  - $(\boldsymbol{x}, y_2)$ are always observed, y_1 is only observed when $y_2 = 1$
  - (u_1, v_2) have zero mean and are independent of $\boldsymbol{x}$
  - $v_2 \sim Normal\,(0, 1)$
  - $\mathrm{E}\left(u_1 | v_2\right) = \gamma_1 v_2$ (which is true e.g. if u_1, v_2 are bivariate normal - in this case $\gamma_1 > 0$ if u_1 and v_2 are positively correlated)

- The conditional expected value of y_1 given v_2 is equal to

$$\mathrm{E}\left(y_1 | \boldsymbol{x}, v_2\right) = \boldsymbol{x}_1\boldsymbol{\beta}_1 + \mathrm{E}\left(u_1 | v_2\right) = \boldsymbol{x}_1\boldsymbol{\beta}_1 + \gamma_1 v_2$$

- If u_1 and v_2 are not correlated ($\mathrm{E}\left(u_1 | v_2\right) = 0$ and $\gamma_1 = 0$) then $\mathrm{E}\left(y_1 | \boldsymbol{x}, v_2\right) = \boldsymbol{x}_1\boldsymbol{\beta}_1$ and as y_2 is a function of v_2, $\mathrm{E}\left(y_1 | \boldsymbol{x}, y_2\right) = \mathrm{E}\left(y_1 | \boldsymbol{x}, v_2\right) = \boldsymbol{x}_1\boldsymbol{\beta}_1$. In this case $\boldsymbol{\beta}$ can be consistently estimated with OLS (no sample selection bias - expected value of y_1 does not depend on unit being selected or not)

- However, if $\gamma_1 \neq 0$ then

$$\mathrm{E}\left(y_1 | \boldsymbol{x}, y_2\right) = \boldsymbol{x}_1\boldsymbol{\beta}_1 + \mathrm{E}\left(u_1 | \boldsymbol{x}, y_2\right) = \boldsymbol{x}_1\boldsymbol{\beta}_1 + \gamma_1 h\left(\boldsymbol{x}, y_2\right)$$

where $h\left(\boldsymbol{x}, y_2\right) = \mathrm{E}\left(u_1 | \boldsymbol{x}, y_2\right)$.

- $h(\boldsymbol{x},1) = \mathrm{E}(u_1 | v_2 > -\boldsymbol{x}\boldsymbol{\delta}_2) = \lambda(\boldsymbol{x}\boldsymbol{\delta}_2)$ where $\lambda(\cdot) = \frac{\phi(\cdot)}{\Phi(\cdot)}$ is inverse mils ratio

- As we only observe y_1 for $y_2 = 1$ we are interested in conditional expectation of y_1 given that individual was selected:

$$\begin{aligned} \mathrm{E}(y_1 | \boldsymbol{x}, y_2 = 1) &= \boldsymbol{x}_1\boldsymbol{\beta}_1 + \gamma_1 h(\boldsymbol{x},1) \\ &= \boldsymbol{x}_1\boldsymbol{\beta}_1 + \gamma_1 \lambda(\boldsymbol{x}\boldsymbol{\delta}_2) \end{aligned}$$

- Using that formula we can analyze the selection bias problem as omitted variable problem. Making OLS regression of y_1 on $\boldsymbol{x}_1$ for selected units we omit $\lambda(\boldsymbol{x}\boldsymbol{\delta}_2)$ which is correlated with $\boldsymbol{x}_1$, which cause the omitted variable bias.

- This equation also suggest how to estimate consistently $\boldsymbol{\beta}_1$ with OLS: we should add to the explanatory variables element $\lambda(\boldsymbol{x}\boldsymbol{\delta}_2)$.

- $\boldsymbol{\delta}_2$ is unknown but it can be consistently estimated with probit.

- Two stage Heckman procedure

  1. Estimate probit regression of y_{i2} on $\boldsymbol{x}_i$. Calculate $\widehat{\lambda}_{i2} = \lambda\left(\boldsymbol{x}\widehat{\boldsymbol{\delta}}_2\right)$
  2. Regress y_{i1} on $\boldsymbol{x}_{i1}$ and $\widehat{\lambda}_{i2}$

- Test of the existence of the selection bias ($H_0 : \gamma_1 = 0$) can be based on standard t-statistics

- If $\gamma_1 \neq 0$, the variance matrix should be adjusted to take into account that $\widehat{\boldsymbol{\delta}}_2$ is estimated

- Even if $\boldsymbol{x} = \boldsymbol{x}_1$ this model is identified. However in this case identification relies on nonlinearity of $\lambda(\cdot)$. As $\lambda(\cdot)$ may be often well approximated with linear function of $\boldsymbol{x}_1$, in such situation we have severe collinearity problem. Then in order to have precise estimate of $\boldsymbol{\beta}_1$ we should have some explanatory variables which belongs to $\boldsymbol{x}$ but does not belong to $\boldsymbol{x}_1$

Example 6. *(Wooldridge) Wage offer equation for married women. For* 753 *women in the sample,* 428 *are working. The labor force participation equation contains variables in wage equation plus income, age, number of young children, number of older children. Results*

	OLS	*Heckit*
educ	$.108$	$.109$
	$(.014)$	$(.016)$
exper	$.042$	$.044$
	$(.012)$	$(.016)$
exper^2	$-.00081$	$-.00086$
	$(.00039)$	$(.00044)$
constant	$-.522$	$-.578$
	$(.199)$	$(.307)$
$\widehat{\lambda}_2$	$-$	$.032$
		$(.134)$
Sample size	428	428
R-squared	$.157$	$.157$

Differences in estimates in this case are not great, inverse Mills ratio is not significant - no evidence for sample selection bias problem. If we estimate this heckit without exclusion restrictions the estimates become very imprecise

(e.g. se for educ becomes $.119$)

- Heckit model can also be estimated with ML. This procedure is efficient if u_1 and v_2 are bivariate normal.

- Sometimes we do not only observe the decision to participate but also y_2 (e.g. number of hours worked)

- In this case the selection equation has the form of the tobit

$$
\begin{aligned}
y_1 &= \boldsymbol{x}_1\boldsymbol{\beta}_1 + u_1 \\
y_2 &= \max\left[0, \boldsymbol{x}\boldsymbol{\delta}_2 + v_2\right]
\end{aligned}
$$

- This model is called type III tobit model

- Similarly as in heckit model

$$\mathrm{E}\left(y_1 \middle| \boldsymbol{x}, v_2, y_2 > 0\right) = \boldsymbol{x}_1\boldsymbol{\beta} + \gamma_1 v_2$$

- But in this case v_2 can be consistently estimated as residuals from tobit model.

- Two step procedure:

  1. estimate tobit regression of y_{i2} on $\boldsymbol{x}_i$. Calculate $\widehat{v}_{i2} = y_{i2} - \boldsymbol{x}\widehat{\boldsymbol{\delta}}_2$
  2. estimate OLS regression of y_1 on $\boldsymbol{x}_1$ and $\widehat{v}_{i2}$