Heckman correction technique - a short introduction

Bogdan Oancea

INS Romania

Problem definition

- We have a population of ***N*** individuals and we want to fit a linear model to this population:

$$y_1 = x_1\beta_1 + \varepsilon$$

but we have only a sample of only ***n*** individuals **not randomly selected**

- Running a regression using this sample would give biased estimates.
- The link with our project – these ***n*** individuals are mobile phone owners connected to **one** MNO;
- One possible solution : ***Heckman correction***.

Heckman correction

- Suppose that the selection of individuals included in the sample is determined by the following equation:

$$y_2 = \begin{cases} 1, & if \quad x\delta + v \geq 0 \\ 0, & otherwise \end{cases}$$

- where $y_2=1$ if we observe y_1 and zero otherwise, $\boldsymbol{x}$ is supposed to contain all variables in $\boldsymbol{x_1}$ plus some more variables and $\boldsymbol{v}$ is an error term.
- Several MNOs use models like the above one.

Heckman correction

- Two things are important here:
  1. When OLS estimates based on the selected sample will suffer **selectivity bias?**
  2. How to obtain **non-biased estimates** based on the selected sample?

Heckman correction

- Preliminary assumptions:
  - (***x,y*$_2$**) are always observed;
  - ***y*$_1$** is observed only when ***y*$_2$**=*1* (*sample selection effect*);
  - (***ε, v***) is independent of ***x*** and it has zero mean *(E (**ε, v**) =0)*;
  - ***v*** *~ N(0,1);*
  - $E(\varepsilon \mid v) = \gamma \cdot v$, correlation between residuals (*it is a key assumption*) : it says that the errors are linearly related and specifies a parameter γ that control the degree to which the sample selection biases estimation of β_1 (i.e., $\gamma \neq 0$ will introduce the selectivity bias).

Heckman correction

- Derivation of the bias:

$$E(y_1 \mid x, v) = x_1\beta_1 + E(\varepsilon \mid x, v) = x_1\beta_1 + E(\varepsilon \mid v) = x_1\beta_1 + \gamma \cdot v$$

- We need the expected value of y_1 conditional on x and the selection outcome y_2:

$$E(y_1 \mid x, y_2) = E[(x_1\beta_1 + \gamma \cdot v) \mid x, v, y_2] = x_1\beta_1 + \gamma \cdot E(v \mid x, y_2) = x_1\beta_1 + \gamma \cdot g(x, y_2)$$

- For the selected sample we have $y_2=1$, so we need to find $g(x,1)$:

$$E(v \mid x, y_2 = 1) = E(v \mid v \geq -x\delta)$$

Heckman correction

- This means that v follows a truncated normal distribution and we can use a well know result:

$$E(z \mid z > a) = \frac{\varphi(a)}{1 - \phi(a)}$$

where ***z*** follows a standard normal distribution, ***a*** is a constant, φ is the standard normal ***pdf*** and ϕ the standard normal ***cumulative distribution function***

$$E(v \mid v \geq -x\delta) = \frac{\varphi(-x\delta)}{1 - \phi(-x\delta)} = \frac{\varphi(x\delta)}{\phi(x\delta)} = \lambda(x\delta)$$

- λ is called the *inverse Mills ratio*.

Heckman correction

- We obtained the parametric expression of the expected value of y_1 conditional on observable x and selection selectivity ($y_2=1$):

$$E(y_1 \mid x, y_2 = 1) = x_1\beta_1 + \gamma \cdot \lambda(x\delta)$$

- The last term in the above equation gives the selectivity correction.
- Plugging the term $\lambda(x\delta)$ into the initial equation we could get an ***unbiased estimation*** of β_1 (as well as γ).

A practical procedure (*heckit*)

1. Using all **N** observations (those for which $y_2=1$ and $y_2=0$) estimate a ***probit*** model with y_2 the dependent variable and ***x*** as explanatory variables to get an estimate $\hat{\delta}$

$$y_2 = x\delta + v$$

2. Use $\hat{\delta}$ to compute the *inverse Mills ratio* for each observation:

$$\lambda(x\hat{\delta}) = \frac{\varphi(x\hat{\delta})}{\phi(x\hat{\delta})}$$

3. Use the **selected sample** and run a regression:

$$y_1 = x_1\beta_1 + \gamma \cdot \lambda(x\hat{\delta}) + \xi$$

A practical procedure (*heckit*)

- This regression will give an ***unbiased estimate*** of $\boldsymbol{\beta_1}$ (as well as γ).
- In this last regression equation the dependent variables are $\boldsymbol{x_1}$ and $\lambda(x\delta)$.
- Support for practical implementation:
  - R package *sampleSelection*;
  - Stata *heckman* function;
  - Eviews implementation;

Steps forward

- This technique seems suitable to produce unbiased estimates in presence of selectivity which is the case of mobile phone subscribers (a mobile phone subscriber would have $y_2=1$ in our presentation);
- Investigate if this technique could be applied (and where) to our project;
- At a first glance eq. (1) from David's internal document could be a place where we can apply this correction technique since our estimations are only for the persons that have a phone registered to the MNO under consideration.