MICROECONOMETRICS
CLASS 7

Wiktor Budziński

SAMPLE SELECTION

Sample selection is kind of similar to the censoring of the dependent variable

For part of the sample we do not observe exact values of the dependent variable

- In this case the fact whether we observe it or not is endogenous – it often depends on some decision of the consumer

Example: We run a survey in which we ask about the household's energy use

- Some respondents will not want to disclose their energy bills
- If individuals with high energy use are more likely to refuse to answer we will have a sample selection issue

EXAMPLE – WAGE OFFER EQUATION

Let say we want to investigate how some variables (e.g. education, gender) affect the wage that individuals are offered in the labor market: $w_i^O = \mathbf{X_i}\boldsymbol{\beta} + \varepsilon_i$

- We only observe the wage offer for individuals who actually work
- People who do not work will usually have wage equal 0 - it does not mean that their wage offer is equal to 0
  - For example, educated person who decided to stay home to care for children
- Formally, we can write utility: $U_i = f(w_i^O h + hinc_i, h)$
  - Person will not work if:

$$w_i^O \leq -\frac{f_q'(hinc_i, 0)}{f_h'(hinc_i, 0)} = w_i^R = \mathbf{X_i}\boldsymbol{\alpha} + \eta_i$$

  - Decision to work can therefore be modelled as

$$P(w_i^O - w_i^R > 0) = P\left(\mathbf{X_i}\boldsymbol{\beta} + \varepsilon_i - \mathbf{X_i}\boldsymbol{\alpha} - \eta_i > 0\right) = P\left(\mathbf{X_i}\boldsymbol{\gamma} + v_i > 0\right)$$

- It is called a **selection equation** - endogeneity arises because error term is correlated with epsilon in main equation

SAMPLE SELECTION

More generally, sample selection model is a two-equation model

So called *outcome equation* is given by the usual formula: $y_1 = \mathbf{X}_1\boldsymbol{\beta} + \varepsilon$

Where y_1 is not observed for some consumers

- Whether y_1 is observed or not depends on the *selection equation*

Selection equation will be defined in terms of an index variable: $y_2^* = \mathbf{X}_2\boldsymbol{\alpha} + \omega$

- Analogously as in ordered or binary models

We assume that y_1 is observed if $y_2^* > 0$

- This is modelled with a probit model, where the dependent variable is $y_2 = \mathbf{1}_{\{y_2^* > 0\}}$

SAMPLE SELECTION

Sample selection becomes an issue if the error terms are correlated

Usually assumed to come from bivariate normal distribution: $(\varepsilon, \omega) \sim \mathrm{BN}(\mathbf{0}, \Sigma)$

With covariance matrix

$$\Sigma = \begin{bmatrix} \sigma^2 & \sigma\rho \\ \sigma\rho & 1 \end{bmatrix}$$

Model estimation

- 2 step (Heckman model, Heckit, Tobit type-2)
- Maximum Likelihood Estimation

2 STEP ESTIMATION

As a first step we estimate probit model for $y_2 = \mathbf{1}_{\{y_2^* > 0\}}$

- We need to assume normality of the error term in the selection equation

Using first step estimates we calculate so called inverse Mills ratio: $\lambda\left(\mathbf{X}_2\boldsymbol{\alpha}\right) = \dfrac{\phi\left(\mathbf{X}_2\boldsymbol{\alpha}\right)}{\Phi\left(\mathbf{X}_2\boldsymbol{\alpha}\right)}$

It is then used as a correction in a main regression (second step):

$$y_1 = \mathbf{X}_1\boldsymbol{\beta} + \sigma\rho\lambda\left(\mathbf{X}_2\boldsymbol{\alpha}\right) + \varepsilon$$

Where $\lambda\left(\mathbf{X}_2\boldsymbol{\alpha}\right) = \mathbf{E}\left(\eta \mid y_2 = 1, \mathbf{X_2}\right)$

Exclusion restriction is usually recommended for this model

- Additional covariates in selection equation, similar to instrumental variables
- Otherwise identified only through the nonlinearity of Mills ratio

EXERCISE 1: WAGE OFFER EQUATION

1. Use femlab.rds data to estimate wage offer equation
2. Estimate Heckman model on wages and logarithm of wages
   - See example 1 in *Sim_examples7*.r to see whether exclusion restriction is necessary
3. Compare marginal effects between Heckman and Tobit

WORKBOOK 7

Conduct a similar analysis in *Workbook7.R*

1. Analyze medical expenditures in the US with the sample selection model
   - You may think about this in the similar way as in the case of the wage equation
     - For some individuals, the 0 expenditures maybe caused by some unobserved socio-economic covariates, for example, access to healthcare, which can also affect the non-zero expenditures

SAMPLE SELECTION AND ENDOGENEITY

Unfortunately, it is often the case that sample selection and endogeneity are present in the data jointly

This can be accommodated easily with a 2-step estimation

- 1. Estimate selection equation
- 2. Estimate 2SLS regression in which inverse Mills ratio will be used as an explanatory variable with some additional covariate used as iv

EXERCISE 2: WAGE OFFER EQUATION

1. Consider education as an endogenous variable in the data
2. Estimate sample selection model with endogeneity, use parents' education as instrumental variables

SAMPLE SELECTION

Sample selection can also be present for not continuous variables

- Count data, ordered or binary

These models can be estimated with MLE, but are usually not as easily available in statistical software

- As these are bivariate nonlinear models it is likely that there will be some issues with convergence

For example, for binary data we would simply assume that outcome equation is given by:

$$y_1^* = \mathbf{X}_1\boldsymbol{\beta} + \varepsilon$$

But we only observe: $y_1 = \mathbf{1}_{\{y_1^* > 0\}}$

- Of course, outcome is only observed if $y_2^* > 0$

EXERCISE 3: SAMPLE SELECTION

1. Investigate the sample selection issue in the case of HIV testing
   - It is important to understand what factors affect the probability of infection
   - Some individual will not want to get tested
   - The issue arises when individuals who have HIV, or are more likely to have it, are less inclined to get tested