SAMPLE SELECTION MODELS

14 September 2011

Yogyakarta, Indonesia

Cosimo Beverelli

(World Trade Organization)

STANDARD TOBIT

- Let us consisder a simple example: an individual maximizing his utility relative to expenditure on tobacco (y) and other goods (z) given disposable income (x).

$$\max U(z, y)$$
$$s.t. \begin{cases} z + y \leq x & (1) \\ z \geq 0 & (2) \\ y \geq 0 & (3) \end{cases}$$

  - We can ignore the constraint on z, but, on the other hand, many individuals will pick up the corner solution for y.

STANDARD TOBIT

- Define a latent variable y^* and assume it is linear in income:

$$y^* = \beta_0 + \beta_1 x + \varepsilon$$

- The solution of the original constrained problem is therefore:

$$y = \begin{cases} y^* \; if \quad y^* > 0 \\ 0 \quad if \quad y^* \leq 0 \end{cases}$$

- In this model, we observe two types of individuals:

1. Individuals who spend a positive amount on tobacco
2. Individuals who spend a non-positive amount (observed expenditure = zero)

STANDARD TOBIT

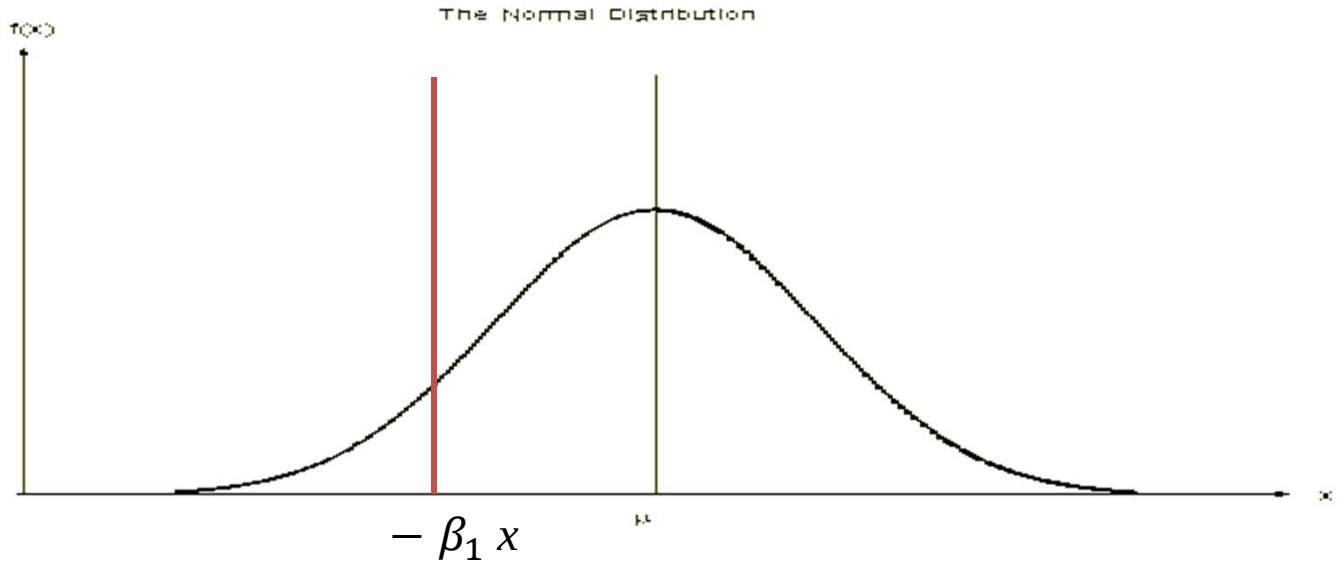
- Therefore:

$$\text{prob}(y = 0) = \text{prob}(y^* \leq 0) = \text{prob}(\varepsilon \leq -\beta_1 x)$$

- And:

$$\text{E}(y|y > 0) = \text{E}(\beta_1 x + \varepsilon|y^* > 0)$$

- The distribution of expenditure is truncated



STANDARD TOBIT

- In this case, OLS is not consistent
- Consistent estimation with censored or truncated sample is the Tobit model (`tobit` in Stata)
- Note that standard Tobit model is useful if the sample is randomly selected among the population: in case of tobacco expenditure, people simply choose whether to buy tobacco or not!
- In case of endogenous sample selection, we need another investigation method; the sample selection model (Tobit II).

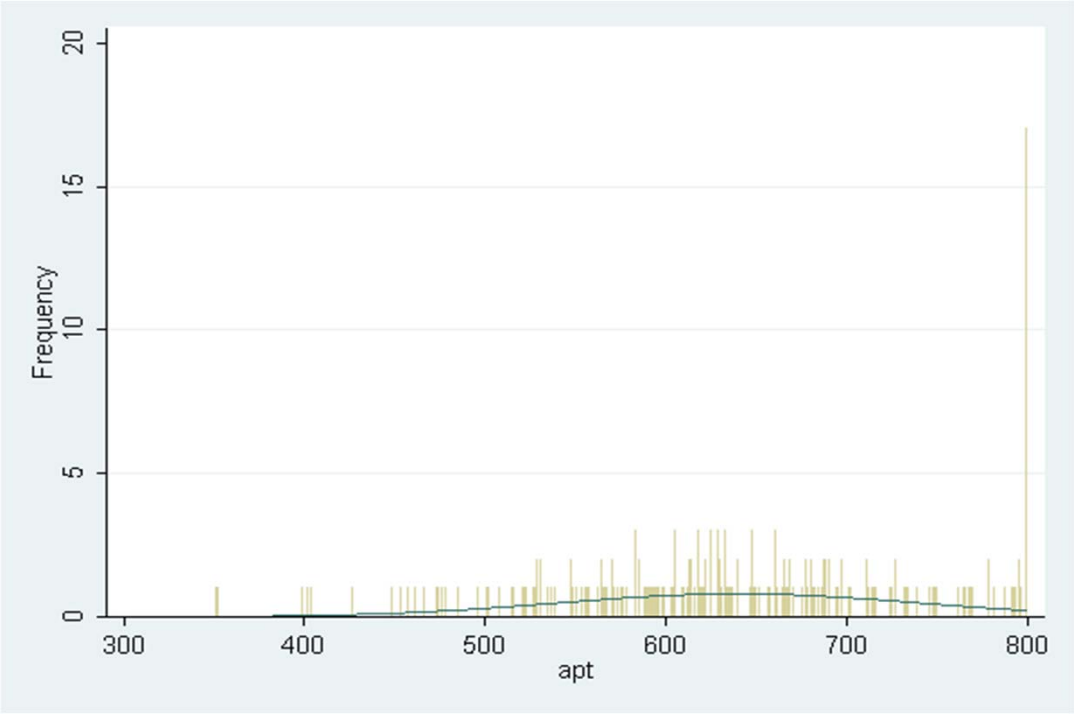
COMPUTATION

- In the following slides we illustrate an hypothetical data file from UCLA, with 200 observations. The academic aptitude variable is `apt`, the reading and math test scores are `read` and `math` respectively. The variable prog is the type of program the student is in, it is a categorical (nominal) variable that takes on three values, academic (`prog = 1`), general (`prog = 2`), and vocational (`prog = 3`).

```
summarize apt read math
```

Variable	Obs	Mean	Std. Dev.	Min	Max
apt	200	640.035	99.21903	352	800
read	200	52.23	10.25294	28	76
math	200	52.645	9.368448	33	75

TRUNCATION



- Looking at the histogram on the left showing the distribution of `apt`, we can see the censoring in the data, that is, there are far more cases with scores 800 than one would expect looking at the rest of the distribution.

TOBIT ANALYSIS

- The `ul( )` option in the `tobit` command indicates the value at which the right-censoring begins (i.e., the upper limit). There is also a `ll( )` option to indicate the value of the left-censoring (the lower limit) which was not needed in this example.

```
tobit apt read math i.prog, ul(800)

Tobit regression                                  Number of obs     =        200
                                                  LR chi2(4)        =     188.97
                                                  Prob > chi2       =     0.0000
Log likelihood = -1041.0629                       Pseudo R2         =     0.0832

------------------------------------------------------------------------------
         apt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        read |    2.697939    .618798     4.36   0.000     1.477582    3.918296
        math |    5.914485   .7098063     8.33   0.000     4.514647    7.314323
             |
        prog |
          2  |   -12.71476   12.40629    -1.02   0.307    -37.18173     11.7522
          3  |    -46.1439   13.72401    -3.36   0.001     -73.2096   -19.07821
             |
       _cons |     209.566   32.77154     6.39   0.000     144.9359    274.1961
-------------+----------------------------------------------------------------
      /sigma |    65.67672   3.481272                      58.81116    72.54228
------------------------------------------------------------------------------
  Obs. summary:          0  left-censored observations
                       183     uncensored observations
                        17 right-censored observations at apt>=800
```

SAMPLE SELECTION MODEL

- Let us consider the case of studying female's wages. Usually, in survey data wages are observed for a fraction of women in the sample, whereas the remaining part of women are observed as unemployed or inactive.
- If we run a linear regression using the observed wages, this would deliver consistent estimations only if working females are a random sample of the population.
- The point is that theory of labor supply suggests that this may not be the case, since (typically) female labor supply is sensitive to household decisions.
- That is, female workers self-select into employment, and the self-selection is not random.

SAMPLE SELECTION MODEL

- The consequence is that using observed wages to estimate the model equation will not deliver estimates that converge to the population parameter!

- The main problem with sample selection model is to establish if the rule determining whether or not observation are available is correlated with the process under investigation.

SAMPLE SELECTION MODEL

- Participation equation: $h = \begin{cases} 1 & if \quad h^* > 0 \\ 0 & if \quad h^* \leq 0 \end{cases}$
- where h* is a latent variable Therefore wages are:
- Outcome equation: $w = \begin{cases} w^* \; if h^* > 0 \\ unobserved \; otherwise \end{cases}$
- When h*>0, women are observed to work, and their wages w* are observed.

- Assume a linear model for latent variables:

$$h^* = x_1'\beta_1 + \varepsilon_1$$
$$w^* = x_2'\beta_2 + \varepsilon_2$$

Notice that for identification the estimation of the bivariate sample selection model may require at least one regressor in the participation equation be excluded from the outcome equation.

SAMPLE SELECTION MODEL

- We assume bivariate normality for $\varepsilon_1, \varepsilon_2$.

Normalization used since only the sign of h* is observed

$$(\varepsilon_1, \varepsilon_2) \sim N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \sigma_{12} \\ \sigma_{21} & {\sigma_2}^2 \end{bmatrix} \right)$$

- NOTE THAT, if $\beta_1 x_1 = \beta_2 x_2$ and $\varepsilon_1 = \varepsilon_2$, then h* = w* and we are back to the standard Tobit model.

SAMPLE ENDOGENEITY HIGHLIGHTED

- If we estimate the outcome (wage) equation conditional on labor market participation, we obtain:

$$E(w_i \mid h_i = 1) = \beta_2 x_2 + \sigma_{12} \cdot \frac{\phi(\beta_1 x_1)}{\Phi(\beta_1 x_1)}$$

If we use standard linear regression (OLS) on the observed wage earners → bias due to endogeneity → we "forget" this part!

- There would be no bias if:
  - $\sigma_{12} = 0$, i.e. the two error terms are not correlated, i.e. sample selection is not endogenous.

HECKMAN TWO-STEP ESTIMATOR

- Heckman argued that the presence of selection bias can be viewed as an omitted variable problem in the selected sample.
- We can consistently estimate the outcome equation through the following procedure:

1. Obtain a Probit estimate of $\widehat{\beta_1}$ from the model:

$$\text{prob}(h_i = 1|x_i) = \Phi(x_1'\beta_1)$$

using all observations N, then we obtain $\widehat{\lambda_{i1}} \equiv \lambda(x_1'\beta_1)$.

- λ is the inverse Mill's ratio: $\lambda = \frac{\phi(x_1'\beta_1)}{\Phi(x_1'\beta_1)}$

2. Obtain estimated coefficients from the OLS regression on the selected sample, w_i on $x_{i2}, \widehat{\lambda_{i2}}$.

These estimators are consistent.

FIRM LEVEL ANALYSIS

- In the previous session, we analysed the export probability of firms using the binary dependent variable “export Y/N”.
- A recent paper by Sun [The World Economy, 2009] on Chinese firms estimates a Heckman two-stages model of export participation and export intensity of Chinese firms.
- The focus is on the effects of innovations and foreign direct investments on export decisions.

SUN (1/4)

- The empirical approach of the paper by Sizhong on the effects of FDI on export behaviour of Chinese firms involves a two-step decision:

1. Whether to export
2. How much to export

- In the sample, one-half of the firms report no exports, therefore the impact of this unobserved export behaviour can by accounted for by the sample selection model.

SUN (2/4)

- For identification purposes, she assumes that the number of firms participating in exporting only explains export participation, not export intensity.
- The number of firms participating in exporting in the four years signals the fixed export cost, and hence the more frequently the firm participates in exporting the more likely it will continue to export. Nevertheless as the fixed export cost has been paid and become sunk, it should not affect how much the firm is willing to export. Hence, it is reasonable to exclude from the export intensity equation the number of firms participating in exporting in the four years.

SUN (3/4)

- Sun estimates three models:

1. With the full set of explanatory variables (multicollinearity...)
2. Because of the multicollinearity issue, he re-runs the model dropping some interactions terms.
3. As a robustness check, he also runs a Tobit model, which accounts for the non-participation of exporting but imposes a restriction that explanatory variables have equal effect on both export participation and export intensity decisions.

- The magnitude of the estimated coefficients display some differences, but the signs do not change.

SUN (4/4)

- The geographic location of firms determines whether their export intensity rises or falls with industry-level FDI

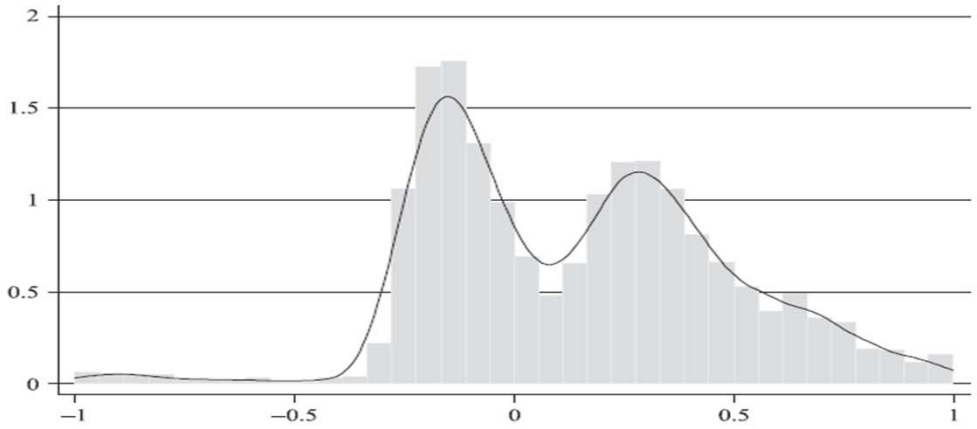
Marginal Effect of Foreign Presence on Export Intensity

	Coastal China	*Central China*	*West China*
Privately owned	−0.0408	0.7743	−1.7714
State and collectively owned	0.3892	1.2043	−1.3414

Note:
Foreign presence is measured in terms of output share.

- FDI affects firms' export intensity differently

Distribution of Marginal Effect of Foreign Presence on Export Intensity



Note:
Marginal effect computed from the estimation using the output share foreign presence as proxy for FDI.