Truncation and Selection

Quantitative Microeconomics

R. Mora

Department of Economics
Universidad Carlos III de Madrid

Outline

1. Introduction
2. Truncation
3. Incidental truncation: Heckman's model
4. The Roy Model

Example: Tobit vs. Truncation

Tobit: Investment in capital equipment

- $q_i^* = x_i\beta + \varepsilon_i$
- we observe $q_i = \begin{cases} q_i^* \text{ if } q_i^* > 0 \\ 0 \text{ if } q_i^* \leq 0 \end{cases}$
  (some firms have positive investments, some firms have zero investments)

Truncated Regression: wage data

- $w_i = x_i\beta + \varepsilon_i$
- for reasons of confidentiality, we only observe (w_i, x_i) if $w_i \leq \overline{W}$
  (the sampling design does not allow sufficiently large wages)

Heckman's Selection Model

- $w_i^* = x_i\beta + \varepsilon_i$
- $s_i = \begin{cases} 1 \text{ if } \gamma' z_i + \upsilon_i > 0 \\ 0 \text{ if } \gamma' z_i + \upsilon_i \leq 0 \end{cases}$
- we observe $w_i = w_i^*$ if $s_i = 1$

- the dependent variable of interest, w_i^*, is *incidental* in the sense that it depends on another condition (the participation equation)
- if (ε, υ) are jointly normally distributed, this is Heckman's Selection Model

The Truncated Normal Regression Model

- $y = \beta_0 + \beta x + \varepsilon, \quad \varepsilon | x \sim N(0, \sigma^2)$
- we observe only (y_i, x_i) if $y_i > 0$ (sample is not iid)

- In the Truncated model, we have only observations of a sample selected by the dependent variable

When is OLS inconsistent?

- consider a general truncation rule $s \in \{0,1\}$ such that $sy = \beta sx + s\varepsilon$
- since $s^2 = s$, $E\left[(sx)(s\varepsilon)\right] = E\left[sx\varepsilon\right]$

OLS is inconsistent when $E\left[sx\varepsilon\right] \neq 0$ $\left(E\left[s\varepsilon \left| x\right.\right] \neq 0 \Rightarrow E\left[sx\varepsilon\right] \neq 0\right)$

- when s is independent of ε, then OLS is consistent (even if s depends on x)
- in the truncation model, $s = 1\left(\beta x + \varepsilon > 0\right)$, so that $E\left[s\varepsilon \left| x\right.\right] \neq 0$ and OLS is inconsistent

ML Estimation

- The density of the sample is not a normal density because the population has been truncated
- We need the distribution of y_i given x_i AND given that $y_i > 0$
- Joint density for $(y_i, y_i > 0)$ given x_i : $\left(\frac{1}{\sigma}\right)\phi\left(\frac{\varepsilon_i}{\sigma}\right)$
- $Pr\left(y_i > 0|x_i\right) = \Phi\left(\frac{\beta x_i}{\sigma}\right)$

$$L_i(\beta, \sigma) = \frac{\left(\frac{1}{\sigma}\right)\phi\left(\frac{(y_i - \beta x_i)}{\sigma}\right)}{\Phi\left(\frac{\beta x_i}{\sigma}\right)}$$

Heckman's Selection Model

we observe w_i if $s_i = 1$

- output equation: $w = \beta_0 + \beta x + \varepsilon$
- participation equation: $s = 1(\gamma' z + v)$
- $\begin{bmatrix} u \\ v \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & \rho \\ \rho & 1 \end{bmatrix} \right)$

- we could generalize this model to include another output equation for those for whom $s = 0$

OLS is inconsistent

- note that $sw* = s\beta_0 + \beta sx + s\varepsilon$
- then $E\left[sx * s\varepsilon \,|x,z\right] = E\left[s\varepsilon \,|x,z\right]x$ because $s^2 = s$
- therefore, OLS will be biased if $E\left[s\varepsilon \,|x,z\right] \neq 0$

OLS is inconsistent if $\rho \neq 0$

Including Additional Regressors

- including z in the output equation does not solve the problem
- OLS fails because for individuals in the wage sample the conditional expectation of the error term is not zero
- intuitively, the workers are more likely to have large positive “errors” in the wages

ML Estimation

- it is possible to estimate the model by ML
- the actual expression for the likelihood is more complicated than that of the probit and tobit model as it requires obtaining the joint distribution of w and s
- `gretl` can implement Heckman's ML estimation
- in general, the likelihood function is not globally concave, and can have local maxima
- Heckman proposed a simple two-stage procedure based on the conditional expectation which gives consistent estimates

The Conditional Expectation

- from the Tobit model, we know that

$$E[w|x,z,s=1]=x\beta+\rho\lambda(z\gamma)$$

  - where $\lambda()$ is the inverse Mills ratio

- λ is like a missing variable which is correlated with ε
- if $\rho=0$, no problem with OLS

Two-step Sample Correction

Heckman's two-step sample selection correction

- First Step: Using all observations, estimate a probit model of *work* on z and compute the inverse of Mills ratio, $\hat{\lambda}_i = \frac{\hat{\phi}_i}{\hat{\Phi}_i}$
- Second Step: using the selected sample, ols *wage* on x and $\hat{\lambda}$

$\hat{\beta}$ is consistent and asymptotically normal

Why does this method work?

- ML estimates of the participation equation are consistent
- $\hat{\lambda}$ shifts the conditional expectations of those individuals more likely to work due to unobservable factors in the right direction. Assume that $\rho > 0$:
  - a wage observation with a low index $z\gamma$ (high λ_i) is likely to work due to unobservable factors and also more likely to have higher wages in the sample due to unobservable factors: λ_i should be large
  - a wage observation with a high index $z\gamma$ (low λ_i) is less likely to work due to unobservable factors and also less likely to have higher wages due to unobservable factors: λ_i should be small

Some Issues on Sample Selection

- OLS (Robust) Standard Errors in second step are invalid
- It is possible to test for sample selection: t test on $\hat{\rho}$ in second step
- If there are endogenous controls in wage equation, we replace OLS by 2SLS in second step
- The method works best if $x \subset z$ (i.e. some variables appear only in participation equation)

The Normality Assumption

- bad news: the procedure is asymptotically valid only if disturbances are normal
- good news: the procedure can be modified easily to account for
  - non-normality
  - heteroskedasticity in the errors

- the Roy model is a two-sector econometric model of self-selection
  - very influential in economics, especially in labor economics and the structural approach to policy evaluation.
  - first developed by Roy (1951) in his analysis of earnings in two occupational sectors, in which individuals self-select into the sector with the highest earnings.
- rational agents make optimizing decisions about what markets to participate in—job, education, marriage, crime, etc.
- makes direct comparisons of outcomes across individuals invalid to infer causal relations

Occupational gender segregation and wage gaps

- Many women work in “female” occupations (stable, flexible, no human capital depreciation after career interruptions) while many men work in “male” occupations (extra hours, firm-specific human capital)
  - On average, women earn less than men. Is the gender wage gap due to gender discrimination?

Migration and unobserved ability

- Migrants chose to go to host country (high skill premium) while non-migrants choose to remain in source country (high wage equality)
  - On average, migrants earn less than non-migrants in host country. Does this mean that they come from the low tail of the skill distribution from the source country?

The Roy Model

we observe w_i both if $s_i = 1$ and if $s_i = 0$

- output if $s_i = 1$: $w_{1i} = x_i'\beta_1 + \varepsilon_1$
- output if $s_i = 0$: $w_{0i} = x_i'\beta_0 + \varepsilon_0$
- Sector selection equation: $s_i = 1\left(z_i'\gamma + v_i\right)$
- $\begin{bmatrix} \varepsilon_1 \\ \varepsilon_0 \\ v \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{10} & \rho_1 \\ \sigma_{10} & \sigma_0^2 & \rho_0 \\ \rho_1 & \rho_0 & 1 \end{bmatrix} \right)$

Estimation

- ML estimation gives consistent estimates.
    - However, most standard statistical packages do not provide command: you need to program the likelihood
- Heckman's two step procedure still valid:
    - apply the procedure two workers in sector 1 and estimate β_1 and the parameter associated to λ_1
    - apply the procedure to workers in sector 0 and estimate β_0 and the parameter associated to λ_0
    - to estimate σ_1^2, σ_0^2, and σ_{10} we would also have to estimate models for $\text{Var}(w_1|s=1,x)$ and $\text{Var}(w_0|s=0,x)$ (beyond the scope of this course)

Summary

- there is a variety of ways to account for sample selection
- the Heckman model assumes normal errors and can be estimated by ML and using a two-step procedure
- in the two-step procedure the correct standard errors of the estimates must be obtained taking into account the two stages
- the Roy model is a basic model in structural policy evaluation