Heckman's Selection Model

Econometrics II

Ricardo Mora

Department of Economics
Universidad Carlos III de Madrid
Máster Universitario en Desarrollo y Crecimiento Económico

Notes

Outline

Notes

Example 1

Investment in capital equipment

- $q_i^* = x_i\beta + \varepsilon_i$
- we observe $q_i = \begin{cases} q_i^* \text{ if } q_i^* > 0 \\ 0 \text{ if } q_i^* \leq 0 \end{cases}$

- firms only carry out investment decisions if their net discounted value is positive
- the censored dependent variable is a latent variable which is the result of our economic model
- this is the Tobit model

Notes

Example 2

- $w_i = x_i\beta + \varepsilon_i$
- we only observe (w_i, x_i) if $w_i \leq \overline{W}$

- for reasons of confidentiality, the dataset does not report any information for individuals with a large wage
- the dependent variable is truncated to the right because of the data collection mechanism
- this is a truncated regression model

Notes

Example 3

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

- wages are only observed for individuals who work
- the dependent variable is only observed among those who work
- if (ε, υ) are jointly normally distributed, this is Heckman's Selection Model

Notes

The Truncated Normal Regression Model

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

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

Notes

OLS is inconsistent

- define $s = 1(\beta_0 + \beta x + \varepsilon > 0)$
- note that $sy = s\beta_0 + \beta sx + s\varepsilon$
- then $E[(sx)(s\varepsilon)] = E[sx\varepsilon]$ (note that $s^2 = s$)

OLS is inconsistent because $E[sx\varepsilon] \neq 0$

Notes

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 : $\left(\frac{1}{\sigma}\right)\phi\left(\frac{\varepsilon_i}{\sigma}\right)$
- $Pr(y_i > 0|x_i) = \Phi\left(\frac{\beta x_i}{\sigma}\right)$

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

Notes

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 can generalize this model to include another output equation for those for whom $s = 0$

Notes

OLS is inconsistent

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

OLS is inconsistent if $\rho \neq 0$

Notes

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

Notes

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
- Stata 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

Notes

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

Notes

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

Notes

Why Is this Method Good?

- 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

Notes

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)

Notes

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

Notes

A Simple Example

Participation

- $U_m - U_h = \beta_m + \beta_m educ + \beta_m kids + v$
- $Pr(work = 1) = \Phi(\beta_0 + \beta_e educ + \beta_k kids)$

Wage equation

- $wage = \beta_0 + \beta_1 educ + u$

- $cov(educ, u) = 0$
- $\begin{bmatrix} u \\ v \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \sigma_{uv} \\ \sigma_{uv} & 1 \end{bmatrix}\right)$

Notes

Summary

- there is a variety of ways to account for sample selection
- Stata allows for estimation of Heckman's Selection Model
- both two-stage and ML estimation
- testing and prediction is computed as usual

Notes

Notes