Limited Dependent Variable Models

Semester 2/2013

Part 5: Sample Selection Corrections

Chayanee Chawanote

Sample selection

- Population model: $y_i = \mathbf{x_i}\beta + u_i$
- Either y_i or some of the independent variables are not observed for certain i
- Define a selection indicator: $s_i = 1$ if we observe all of (y_i, x_i), $s_i = 0$ if we do not observe some of them, hence the observation will not be used used.
- Under which conditions OLS is consistent?

When OLS is consistent on the selected sample

- A: when it is exogenous sample selection.
- Consider $s_i y_i = s_i x_i \beta + s_i u_i$
- $\hat{\beta}_{OLS}$ is consistent if $E(su) = 0$ and $E[(sx_j)(su)] = E[sx_j u] = 0$
- If we have random sample and randomly drop observations, OLS is still consistent and unbiased.
- If s depends on the explanatory variables and additional random terms that are indepent of x and u, OLS is also consistent and unbiased.

When OLS is inconsistent on the selected sample

- A: when regression using a truncated sample
- For example, the truncation is from above: $s_i = 1$ if $y_i \leq c_i$, c_i is truncation threshold.
- This is the same as $s_i = 1$ if $u_i \leq c_i - \mathbf{x}_i\beta$. We have s_i depending directly on u_i
- $\hat{\beta}_{OLS}$ on the selected sample when s and u are correlated is inconsistent.
- For IV, 2SLS is consistency if $E[sz_h u] = 0$, which holds if $E(u|\mathbf{z}, s) = 0$

Incidental Truncation

- Assume we always observe the explanatory variables x_j
- But, we only observe y for a subset of the population. We observe y when it depends on other variable (not latent).
- Example, we observe wage if labor force participation = 1, while we have all other information on individual.
- Population model: $y = \mathbf{x}\beta + u$, $E(u|x) = 0$
  Selection equation: $s = 1[\mathbf{z}\gamma + v \geq 0]$
- The selection equation depends on observed variables z_h and unobserved error v
- We need assumption that $\mathbf{z}$ is exogenous: $E(u|\mathbf{x}, \mathbf{z}) = 0$
- We require that x be a strict subset of z, and that some variables in z are not also in x.

Incidental Truncation

- v is assumed to be independent of z, and therefore x.
- v is also assumed to have a standard normal distribution.
- (u,v) is independent of z.
- How can correlation between u and v causes a sample selection problem?
- If u and v are jointly normal with zero mean, $E(u|v) = \rho v$
- Then, $E[y|\mathbf{z}, s = 1] = \mathbf{x}\beta + \rho\lambda(z\gamma)$

Sample selection correction

1. Using all n observations, estimate a probit model os s_i on $\mathbf{z}_i$, and obtain the estimates $\hat{\gamma}_h$
   Compute the inverse Mills ratio, $\hat{\lambda}_i = \lambda(\mathbf{z}_i\hat{\gamma})$ for each i with $s_i = 1$
2. Using the selected sample (only observations that $s_i = 1$), run the regression of y_i on $\mathbf{x}_i, \hat{\lambda}_i$
   Then, $\hat{\beta}_j$ are consistent and approximately normally distributed.

Additional issues

- Simple test of selection bias:
  using t statistic on $\hat{\lambda}_i$ and test $H_0 : \rho = 0$ (no sample selection problem)
- When $\rho \neq 0$, the usual OLS standard errors from regressing y_i on $\mathbf{x}_i, \hat{\lambda}_i$ are not correct. Some econometrics packages compute corrected standard errors.
- We need at least one element of z that is not in x $\rightarrow$ we need a variable that affects selection but does not have a partial effect on y to avoid possible multicollinearity that lead to high standard errors in the second step.