Applied Microeconometrics (L9): Selection models

Nicholas Giannakopoulos

University of Patras
Department of Economics

ngias@upatras.gr

December 17, 2019

Overview

Sample selection

Truncation and Censoring

Sample selection/incidental truncation

Heckman Two-Step Estimator

Heckman's sample selection model

- Addresses the problem of selection bias
- Contributions to program evaluation:
  - provided a theoretical framework that emphasized the importance of modeling the dummy endogenous variable
  - the first attempt that estimated the probability (i.e., the propensity score) of a participant being in one of the two conditions indicated by the endogenous dummy variable, and then used the estimated propensity score model to estimate coefficients of the regression model
  - treated the unobserved selection factors as a problem of specification error or a problem of omitted variables, and corrected for bias in the estimation of the outcome equation by explicitly using information gained from the model of sample selection
  - developed a creative two-step procedure by using the simple least squares algorithm

Limited dependent variables: characteristics

- Truncation
  - when sample data are drawn from a subset of a larger population of interest: is the effect of data gathering rather than data generation
  - a truncated distribution is the part of a larger, untruncated distribution
  - e.g., those whose incomes are above poverty threshold. Thus, using that limited information (truncated distribution) we can infer the income distribution for the entire population
- Censoring
  - censoring occurs when all values in a certain range of a dependent variable are transformed to a single value.
  - censoring differs from truncation in that the data collection may include the entire population, but below threshold the variable is coded as a scalar
  - e.g., those whose incomes are below poverty threshold are coded as zero.

Combination of truncation and censoring

- incidental truncation (sample selection)
- scope of analysing limited dependent variables: use the truncated distribution or censored data to infer the untruncated or uncensored distribution for the entire population.
- typical regression analysis: assume that the dependent variable follows a normal distribution
- then to develop moments (mean and variance) of the truncated or censored normal distribution
- key factor: inverse Mills ratio, or hazard function (λ)
- Heckman's sample selection model uses the inverse Mills ratio to estimate the outcome regression.
- sample selection or incidental truncation refers to a sample that is not randomly selected.
    - E1: Married women in the labor force
    - E2: Effects of unions on wages
    - E3: Returns of higher education

Why is important to model sample selction?

- researchers' aim: make causal inference
- randomized vs. nonrandomized experiments
- types of selection bias (Heckman, 1979):
  - self-selection bias
  - selection bias made by data analysts
- types of selection bias (Heckman and Smith, 1995):
  - individual selection
  - administrator selection
  - attrition selection
- when selectivity is inevitable, such as in observational studies, the parameter estimates from a naive ordinary least squares (OLS) regression model are inconsistent and biased

Three types of decisions that create selectivity

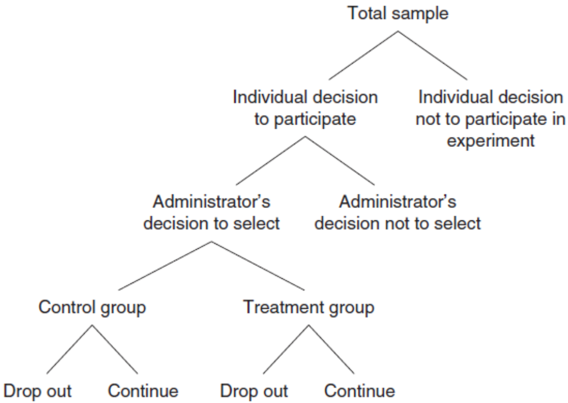
Total sample
Individual decision to participate
Individual decision not to participate in experiment
Administrator's decision to select
Administrator's decision not to select
Control group
Treatment group
Drop out
Continue
Drop out
Continue

Figure 4.1 Decision Tree for Evaluation of Social Experiments

SOURCE: Maddala (1983, p. 266). Reprinted with the permission of Cambridge University Press.

Moments of an incidentaly truncated bivariate normal distribution

- Suppose that y and z have a bivariate normal distribution with correlation ρ. We are interested in the distribution of y given that z exceeds a particular value α. The truncated joint density of y and z is:
$$f(y,z|z>\alpha)=\frac{f(y,z)}{Prob(z>\alpha)}$$
- Given the truncated joint density of y and z, given that y and z have a bivariate normal distribution with means μ_y and μ_z, standard deviations σ_y and σ_z, and correlation ρ, the moments (mean and variance) of the incidentally truncated variable y are as follows:
$$E(y|z>\alpha)=\mu_y+\rho\sigma_y\lambda(c_z)$$
$$Var(y|z>\alpha)=\sigma_y^2[1-\rho^2\delta(c_z)]$$
- where α is the cutoff threshold, $c_z=\frac{(\alpha-\mu_z)}{\sigma_z}$, $\lambda(c_z)=\frac{\phi(c_z)}{[1-\Phi(c_z)]}$, $\delta(c_z)=\lambda(c_z)[\lambda(c_z)-c_z]$, $\phi(c_z)$ is the standard normal density function, and $\Phi(c_z)$ is the standard cumulative distribution function

Inverse Mills ratio

- Inverse Mills ratio
$$\lambda(c_z) = \frac{\phi(c_z)}{[1-\Phi(c_z)]}$$
- λ is used in Heckman's derivation of his two-step estimator
- A sample selection model always involves two equations:
  1. the regression equation considering mechanisms determining the outcome variable
  2. the selection equation considering a portion of the sample whose outcome is observed and mechanisms determining the selection process

Modeling

- women's wage in the labor force
- Modeling:
  - Outcome equation: assume that the hourly wage (wage) of women is a function of education (educ) and age (age) $wage_i = \beta_0 + \beta_1 educ_i + \beta_2 age_i + u_{i1}$
  - Selection equation: whereas the probability of working (p^w, equivalent to the probability of wage being observed) is a function of marital status (married) and number of children at home (children). Thus, wage if observed if $p_i^w = \gamma_0 + \gamma_1 married_i + \gamma_2 children_i + \gamma_3 educ_i + \gamma_4 age_i + u_{i2} > 0$

Modeling

- The selection equation indicates that wage is observed only for those women whose wages were greater than 0 (i.e., women were considered as having participated in the labor force if and only if their wage was above a certain threshold value)
- Using a zero value in this equation is a normalization convenience and is an alternate way to say that the market wage of women who participated in the labor force was greater than their reservation wage (i.e., $y > y\star$)
- The fact that the market wage of homemakers (i.e., those not in the paid labor force) was less than their reservation wage (i.e., $y > y\star$) is expressed in the above model through the fact that these women's wage was not observed in the regression equation, that is it was incidentally truncated. The selection model further assumes that u_1 and u_2 are correlated to have a nonzero correlation ρ.

Modeling: general case

- Regression equation: $y_i = x_i\beta + \epsilon_i$ only if $w_i = 1$
- Selection equation: $w_i^\star = z_i\gamma + u_i$

  $$w_i = 1 \text{ if } w_i^\star > 0 \text{ and } w_i = 0 \text{ otherwise}$$

  where, where x_i is a vector of exogenous variables determining outcome y_i, and $w_i^\star$ is a latent endogenous variable.Thus, the selection rule is:

  $$Prob(w_i = 1|z_i) = \Phi(z_i\gamma)$$
  $$Prob(w_i = 0|z_i) = 1 - \Phi(z_i\gamma)$$

- Use ML estimation
- in Stata: heckman