Lecture 11: Sample Selection

March 27, 2025

Introduction

- Now turn to the problem of using only a subset of a random sample obtained from a well-defined population (presumably, the one of interest).
- Obvious but important point: There is not an issue of nonrandom sample selection if a random sample has been obtained from a given population. The population is not immutable. We can choose a population of interest from a bigger population.

Introduction

- For example, if we are interested in the effect of a job training program on a population of men with poor labor market histories, we can define the population based on observed past labor market outcomes, such as unemployment status or labor earnings. If we can collect a random sample from the defined population, we just apply standard methods.
- Sample selection becomes an issue when the sample we can obtain are not representative of the population of interest.

Example to discuss three cases of "selection"

- As an example, suppose we are interested in a wealth equation,

  $$\text{wealth} = \beta_0 + \beta_1 \text{plan} + \beta_2 \text{educ} + \beta_3 \text{age} + \beta_4 \text{income} + u$$

  which describes the population of all families in the United States (where *educ* and *age* are for the self-described “household head”). If we assume that u has zero mean and is uncorrelated with each explanatory variable, we would use OLS if we have a random sample from the population.
- Suppose, though, that only people less than 65 years old were sampled. What if we use OLS on the selected sample?

Example to discuss three cases of "selection"

- As we will see, OLS on the nonrandom sample nevertheless consistently estimates the β_j provided

$$E(u|\text{plan}, \text{educ}, \text{age}, \text{income}) = 0.$$

- Zero correlation is not enough! Must have the conditional mean correctly specified. This falls under “exogenous sampling.”
- Next suppose that only families with wealth greater than zero are included in the sample. Now, the data are selected on the basis of the response variable, wealth. As we will see later, using standard methods (including OLS) on such sample leads to biased and inconsistent estimators of the β_j, even under the zero conditional mean assumption.

Example to discuss three cases of "selection"

- Another example is when y is observed only when a certain event is true. A leading example is when y is $log(wage^o)$, the log of the "wage offer" - the hourly wage someone could get paid if in the work force. We observe $wage^o$ only if the person decides to enter the work force.
- Generally called the problem of **incidental truncation**.
- The hallmark of the incidental truncation problem is the notion of "self-selection". For example, we only observe the wage offer if the person "self-selects" into the workforce.
- Whether someone chooses to report, say, their annual income has a self-selection component.

When can sample selection be ignored?

Linear Model

- Assume there is a population represented by the random vector $(\mathbf{x}, y, \mathbf{z})$, where $\mathbf{x}$ is a $1 \times K$ vector of explanatory variables, y is the scalar response variable, and $\mathbf{z}$ is a $1 \times L$ vector of instrumental variables.
- Standard linear model with (possibly) endogenous explanatory variables:

$$\begin{aligned} y &= \beta_1 + \beta_2 x_2 + \ldots + \beta_K x_K + u = \mathbf{x}\beta + u \\ E(\mathbf{z}'u) &= 0, \end{aligned}$$

with $x_1 = 1$ (so z_1 is almost certainly equal to unity, too).

When can sample selection be ignored?

- Given a random sample from the population, we can, under a rank condition, use 2SLS to consistently estimate β.
- Unfortunately, the rank condition (essentially rank $E(\mathbf{z}'\mathbf{x}) = K$) and $E(\mathbf{z}'u) = 0$ are not usually enough to consistently estimate β with a selected sample.
- Analysis is simplified by thinking of drawing units randomly from the population, but now the random draw for unit i, $(\mathbf{x}_i, y_i, \mathbf{z}_i)$, is supplemented by drawing a selection indicator, s_i. By definition, $s_i = 1$ if unit i is used in the estimation, and $s_i = 0$ if we do not use random draw i.
- Therefore, our "data" consists of $\{(\mathbf{x}_i, y_i, \mathbf{z}_i, s_i) : i = 1, \ldots, N\}$, where the value of s_i determines whether we observe all of $(\mathbf{x}_i, y_i, \mathbf{z}_i)$.
- Because identification is properly studied in the population, let s denote a random variable with the distribution of s_i for all i. In other words, $(\mathbf{x}, y, \mathbf{z}, s)$ now represents the population.

When can sample selection be ignored?

- Consider the algebraically simpler case of just identification (in the population!), that is, $L = K$. Let $\{(\mathbf{x}_i, y_i, \mathbf{z}_i, s_i) : i = 1, \dots, N\}$ be a random sample from the population.
- The IV estimator using the selected sample can be written as

$$\hat{\beta}_{IV} = \left(\sum_{i=1}^{N} s_i \mathbf{z}_i' \mathbf{x}_i\right)^{-1} \left(\sum_{i=1}^{N} s_i \mathbf{z}_i' y_i\right)$$

$$= \beta + \left(\sum_{i=1}^{N} s_i \mathbf{z}_i' \mathbf{x}_i\right)^{-1} \left(\sum_{i=1}^{N} s_i \mathbf{z}_i' u_i\right).$$

When can sample selection be ignored?

- By the law of larger numbers for random samples,

$$\text{plim}_{N\to\infty}(\hat{\beta}_{IV}) = \beta + \left[E(s\mathbf{z}'\mathbf{x})\right]^{-1} E(s\mathbf{z}'u).$$

- Weak assumptions sufficient for consistency are

$$\text{rank } E(\mathbf{z}'\mathbf{x}|s=1) = K$$

and

$$E(s\mathbf{z}'u) = 0,$$

When can sample selection be ignored?

- Practically, for the rank condition to hold on the subpopulation, we should have it holding in the population and then the subpopulation not being "too small."
- More interesting is $E(s\mathbf{z}'u) = 0$. Holds is when s is independent of (z, u) along with zero correlation in the population:

  $$E(\mathbf{z}'u) = 0.$$

  Why? If s is independent of (z, u) then

  $$E(s\mathbf{z}'u) = E(s)E(\mathbf{z}'u) = \rho \cdot 0 = 0$$

  where $\rho = E(s)$ is the (unconditional) probability that a randomly draw observation is kept.
- In statistics, if s is independent of $(\mathbf{x}, y, \mathbf{z})$, the data are said to be missing completely **at random**.

When can sample selection be ignored?

- Another sufficient condition is

$$E(u|\mathbf{z}, s) = E(u|\mathbf{z}) = 0,$$

  where the second equality would be a strengthening of the exogeneity requirement on the instruments. The first equality rules out correlation between s and u.
- Sufficient for this latter condition is $E(u|\mathbf{z}) = 0$ and s is a deterministic function of $\mathbf{z}$, say $s = h(\mathbf{z})$. Then

$$E(u|\mathbf{z}, s) = E(u|\mathbf{z}, h(\mathbf{z})) = E(u|\mathbf{z}).$$

  This is the case of **exogenous sampling**.

When can sample selection be ignored?

- With $\mathbf{z} = \mathbf{x}$, a sufficient condition is

$$E(u|\mathbf{x}, s) = E(y|\mathbf{x}) = \mathbf{x}\beta,$$

  which means s can be an arbitrary function of the exogenous variables. The rank condition is that $E(\mathbf{x}'\mathbf{x}|s = 1)$ has rank K.
- If $y = \mathbf{x}\beta + u$, $E(\mathbf{x}'u) = 0$, and s is independent of $(\mathbf{x}, y)$, then OLS using $s_i = 1$ is consistent for β.
- The cases with $\mathbf{x}$ exogenous and with instruments are very important for sample selection corrections. If we can obtain an equation where the selection indicator is a function of the explanatory variables (or instruments), we can apply OLS or 2SLS to that equation for consistent estimation.

Incidental Truncation: A Probit Selection Equation

Exogenous Explanatory Variables

- Motivation: Interested in estimating $E(\text{wage}^o|x)$, where wage^o is the wage offer. But need to recognize that if we randomly sample adults, some will not be working, so wage^o is unobserved.
- Simple utility maximization approach (with w^o the wage offer) to choosing weekly hours:

  $$\max_h \text{util}_i(w_i^o h + a_i, h) \text{ subject to } 0 \leq h \leq 168$$

  Assume can rule out a solution at $h_i = 168$. Can show that if

  $$\frac{ds_i(0)}{dh} < 0, \text{ where } s_i(h) = \text{util}_i(w_i^o h + a_i, h)$$

  then $h_i = 0$ is the optimum.

Heckman Selection

- Equivalent to

  $$w_i^o \leq -\frac{mu_i^h(a_i, 0)}{mu_i^q(a_i, 0)}$$

  where $mu_i^h(\cdot, \cdot)$ is the marginal disutility of working and $mu_i^q(\cdot, \cdot)$ is the marginal utility of income.
- Can think of the right-hand side as the reservation wage, w_i^r.
- Assume the person works only if

  $$w_i^o > w_i^r$$

  (where we can ignore ties under continuity).

Heckman Selection

- This looks like censoring the wage offer from below, but there is a key difference: we do not observe w^r. Called incidental truncation. (Perhaps "incidental censoring" would be a better name, as we can generally draw a random sample from the population of working-age adults, and then observe other attributes.)
- Model the wage offer and reservation wages as

$$w_i^o = \exp(x_{i1}\beta_1 + u_{i1})$$

$$w_i^r = \exp(x_{i2}\beta_2 + \gamma_2 a_i + u_{i2})$$

Heckman Selection

- We observe w_i^o if $\log(w_i^o) - \log(w_i^r) > 0$ or

$$x_{i1}\beta_1 + u_{i1} - \mathbf{x}_{i2}\beta_2 - \gamma_2 a_i - u_{i2} > 0$$

  or

$$\mathbf{x}_i\delta_2 + v_{i2} > 0,$$

  where $\mathbf{x}_i$ includes all nonredundant elements of $\mathbf{x}_{i1}$ and $\mathbf{x}_{i2}$ and also a_i, nonwage income.

- Having a_i (at least) affect the reservation wage, and therefore the labor force participation decision, but having no affect on the wage offer, is very important for identification.

Heckman Selection

- In the population, we can write the Gronau-Heckman model as

  $$\log(\text{wage}^o) = x_1\beta_1 + u_1$$

  $$\text{inlf} = 1[x\delta_2 + v_2 > 0]$$

  where inlf is equal to unity if a person is in the labor force. We observe wage^o, and therefore $\log(\text{wage}^o)$, only if $\text{inlf} = 1$.
- We have some interest in estimating the factors that affect inlf, but we are primarily interested in the wage offer equation.

Heckman Selection

- Notation for the general population model

$$y_1 = x_1\beta_1 + u_1$$

$$y_2 = 1[x\delta_2 + v_2 > 0]$$

  where y_1 is the response that is only partially observed, and now y_2 is the selection indicator.

- Assumptions:
  (a) $(\mathbf{x}, y_2)$ are always observed, y_1 is observed only when $y_2 = 1$;
  (b) (u_1, v_2) is independent of x with zero mean;
  (c) $v_2 \sim \text{Normal}(0, 1)$;
  (d) $E(u_1|v_2) = \gamma_1 v_2$.

- So, we can think of a random draw $(\mathbf{x}_i, y_{i1}, y_{i2})$ from the population, but we only observe y_{i1} if $y_{i2} = 1$.

Heckman Selection

- Joint normality of (u_1, v_2) is not necessary for a two-step estimation method, but it is often imposed for a (partial) MLE analysis.
- Because v_2 is independent of $\mathbf{x}$ and standard normal, y_2 follows a probit: $P(y_2 = 1|\mathbf{x}) = \Phi(\mathbf{x}\delta_2)$.
- Because $(\mathbf{x}, y_2)$ is assumed to always be observed, δ_2 is identified, and so we can treat it as known for the purposes of deriving an estimating equation for β_1.

Heckman Selection

- How can we obtain an estimating equation for β_1? Under the previous assumptions,

$$E(y_1|\mathbf{x}, v_2) = x_1\beta_1 + E(u_1|\mathbf{x}, v_2)$$

$$= x_1\beta_1 + E(u_1|v_2) = x_1\beta_1 + \gamma_1 v_2.$$

- If we could observe (or, in effect, estimate) v_2, we could solve the selection problem by adding v_2 as a regressor and using OLS on the selected sample.

Heckman Selection

- But we only observe $y_2 = 1[\mathbf{x}\delta_2 + v_2 > 0]$. So we need to obtain $E(y_1|\mathbf{x}, y_2)$. But $(\mathbf{x}, y_2)$ is a function of $(\mathbf{x}, v_2)$, so we can apply iterated expectations:

$$E(y_1|\mathbf{x}, y_2) = E[E(y_1|\mathbf{x}, v_2)|\mathbf{x}, y_2] = x_1\beta_1 + \gamma_1 E(v_2|\mathbf{x}, y_2).$$

- When $y_2 = 1[\mathbf{x}\delta_2 + v_2 > 0]$ and $v_2|\mathbf{x} \sim \text{Normal}(0, 1)$, $E(v_2|\mathbf{x}, y_2)$ has a well-known form: it is the inverse Mills ratio. (Actually, its form depends on whether $y_2 = 1$ or $y_2 = 0$, and we only need the $y_2 = 1$ expression here.)

Heckman Selection

- For completeness (and because it is useful later for treatment effect estimation),

$$E(v_2|\mathbf{x}, y_2) = y_2\lambda(\mathbf{x}\delta_2) - (1 - y_2)\lambda(-\mathbf{x}\delta_2) = r(y_2, \mathbf{x}\delta_2)$$

  where

$$\lambda(\cdot) = \frac{\phi(\cdot)}{\Phi(\cdot)}$$

  is the IMR. The function $r(y_2, \mathbf{x}\delta_2)$ is sometimes called a generalized residual. Note that $E[r(y_2, \mathbf{x}\delta_2)|\mathbf{x}] = 0$ necessarily follows by iterated expectations because $E(v_2|\mathbf{x}) = 0$.

Heckman Selection

- Therefore, on the selected sample we have

$$E(y_1|\mathbf{x}, y_2 = 1) = x_1\beta_1 + \gamma_1\lambda(\mathbf{x}\delta_2)$$

- If we just regress y_1 on x_1 using the $y_2 = 1$ sample, then, in effect, we omit the variable $\lambda(x\delta_2)$ from the regression. (It is *possible* that, in the subpopulation with $y_2 = 1$, $\lambda(x\delta_2)$ is uncorrelated with x_1, in which case OLS would be consistent for the slopes in β_1. But this would be a fluke and cannot be relied on.)
- The equation for $E(y_1|\mathbf{x}, y_2 = 1)$ is properly viewed as an estimating equation, not a model that we begin with!

Heckman Selection

- The expression for $E(y_1|\mathbf{x}, y_2 = 1)$ suggests a simple two-step estimation method.
  - **(i)** Estimate probit of y_2 on x_i using all of the data, $i = 1, \ldots, N$, to obtain $\hat{\delta}_2$ and
  $$\hat{\lambda}_{i2} = \lambda(x_i \hat{\delta}_2).$$
  - **(ii)** Run OLS of y_{i1} on x_{i1}, $\hat{\lambda}_{i2}$, $i = 1, \ldots, N_1$ where the data have been ordered so that $y_{i2} = 1$ for $i = 1, \ldots, N_1$.
- This has been called the Heckit method after Heckman (1976).
  - Should adjust our standard errors and inference for two-step estimation. Many packages, including Stata, make the adjustment routinely. Bootstrapping is also valid.

Heckman Selection

- Technically, the procedure goes through with $x_1 = \mathbf{x}$, that is, without an exclusion restriction. But then identification of β_1 is possible only because $\lambda(\cdot)$ is a nonlinear function.
- Generally, one should be hesitant to achieve identification "off of a nonlinearity." Cannot really tell if $\lambda(x_i\hat{\delta}_2)$ is statistically significant because selection is an issue or the functional form $E(y|\mathbf{x}) = \mathbf{x}\beta_1$ is misspecified (in the population).
- If we write $\mathbf{x} = (x_1, x_2)$, we are assuming $E(y_1|\mathbf{x})$ (the population regression) does not depend on x_2. The only reason $E(y_1|\mathbf{x}, y_2 = 1)$ depends on x_2 is because x_2 predicts selection and selection is correlated with u_1.
- Often, over the range of $x_i\hat{\delta}_2$ in the data, $\lambda(\cdot)$ is pretty close to linear. Very high collinearity is usually present unless x_i contains something not in x_{i1} that is useful for predicting selection.

Heckman Selection

Bottom line: The Heckit approach is not believable unless one has at least one exclusion restriction in the regression equation. And, if we write $\mathbf{x} = (x_1, x_2)$, so that:

$$\mathbb{P}(y_2 = 1|\mathbf{x}) = \Phi(x_1\delta_{21} + x_2\delta 22)$$

them we must be able to reject $H_0 : \delta_{22} = 0$ at some low significance level. (Just like with instrumental variables.) What we cannot generally test is whether excluding x_2 from the regression equation is appropriate (just like with IV).