Heckman Selection

- We generally assume that we are working with a random sample from an underlying population or that we have randomly sampled from a population that has exhibited some clustering.
  → what if this assumption is violated?

e.g. 1) We wish to estimate a savings function:

$$saving = \beta_0 + \beta_1 \, income + \beta_2 \, age + \beta_3 \, married + \beta_4 \, kids + u$$

But our data is only for HHs for whom the head is age ≥ 45. We are interested in a savings function for <u>all</u> families, but we only have a random sample about a subset of the population.

e.g. 2) Wage offer equation

$$w = X'\beta + u$$

But w is only observed if an individual works. Wage is missing as a result of the outcome of another variable: labor force participation.

Is this sample selection or self-selection?

When can we just ignore non-random sampling?

When random sample:

$$y = \alpha + \beta_1 S + \varepsilon \quad \text{where } S \text{ endogenous}$$

Instrument for S using z. Instrument is <u>valid</u> if $E(\varepsilon \mid z) = 0$

When non-random sample:

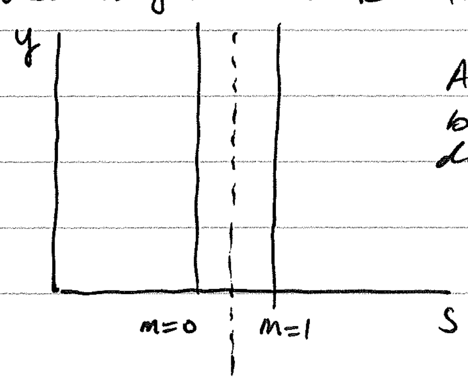
$$y = \alpha + \beta_1 S + \varepsilon \quad \text{where } S \text{ is endogenous.}$$

only observe sample for $S \geq n$

Instrument for S using z. Instrument is <u>valid</u> if $E(\varepsilon \mid z, m) = 0$ where $m = \begin{cases} 1 & \text{if } S \geq n \\ 0 & \text{if } S < n \end{cases}$

The errors cannot be correlated with the selection rule. i.e., whether or not $S \geq n$ or $S < n$, once we've instrumented, cannot be correlated with y.

When might this be true?



Around a narrow band around a discontinuity.

What can we do when $E(\varepsilon|z) = 0$ but $E(\varepsilon|z, m) \neq 0$? i.e., selection into sample is non-random.

We want to know $E(w_i|X_i)$. If w_i observed for everyone in the working age population, no problem. But potential sample selection because w_i only observed for people who work. Even though people who don't work technically have $w_i = 0$, this isn't really true $w^i = 0$ because their reservation wage is greater than the wage being offered to them $(w^r > w^o)$.

Suppose: $\overset{\text{wage}}{w} = x_1\beta + u$ structural eqn

$L = 1[x\delta + v > 0]$

labor force participation

Assumptions: 1) (x, L) are always observed

2) w is observed only when $L=1$

3) (u, v) are independent of x

4) $v \sim N(0,1)$

5) $E(u|v) = \gamma v$

What we can hope to estimate is

$E(w|x_1, L=1)$ <u>and</u> $P(L=1|x)$

First we have:

$$w = x_1\beta + u$$

$$E(w|x_1, v) = E(x_1|x, v)\beta + E(u|x, v)$$ Take expectation wrt x and v.

$$= x_1\beta + E(u|v)$$ since (u,v) are independent of x

$$E(w|x_1, v) = x_1\beta + \gamma v$$

If $\gamma = 0 \Rightarrow u, v$ are uncorrelated. If that's the case, $E(w|x_1, v) = x_1\beta + \gamma v$ becomes $E(w|x_1) = x_1\beta$. That's just OLS! And no sample selection problem.

What if $\gamma \neq 0$?

$$E(w|x_1, v) = x_1\beta + \gamma v$$

Take expectation wrt x, L

$$E(w|x_1, v, L) = E(x_1|x, L)\beta + \gamma E(v|x, L)$$

b/c (u, v) indep. of x and $L = 1[x\delta + v > 0]$ so L a function of v.

$$E(w|x_1, L) = x_1\beta + \gamma \underbrace{E(v|x, L)}_{= h(x, L)}$$

$$E(w|x_1, L) = x_1\beta + \gamma h(x, L).$$

If we knew $h(x, L)$, we could estimate both β and γ using only the selected sample.

Because the selected sample has $L = 1$, we need to find $h(x, 1)$.

$$L = 1 \quad \text{if} \quad x\delta + v > 0 \Rightarrow \text{if} \quad v > -x\delta$$

$$h(x, 1) = E(v | v > -x\delta) = \lambda(x\delta) \quad \text{where}$$

$\lambda(\cdot) = \dfrac{\phi(\cdot)}{\Phi(\cdot)}$ is the inverse Mills ratio

If x is a random variable distributed $N(\mu, \sigma^2)$

Then $$E(x \mid x > \alpha) = \mu + \sigma^2 \frac{\phi\left(\frac{\alpha-\mu}{\sigma}\right)}{1-\Phi\left(\frac{\alpha-\mu}{\sigma}\right)}$$

where ϕ is the standard normal density function and Φ is the cumulative distribution function.

Here v assumed $\sim N(0,1)$, so:

$$E(v \mid v > -x\delta) = 0 + 1 \cdot \frac{\phi(-x\delta)}{1-\Phi(-x\delta)}$$

$$E(v \mid v > -x\delta) = \frac{\phi(-x\delta)}{1-\Phi(-x\delta)}.$$

So substituting, we have:

$$E(w|x_1, L=1) = x_1\beta + \gamma\,\lambda(x\delta)$$

γ is the correlation between unobserved determinants of the propensity to work, V, and unobserved determinants of wage offers, u.

Note that if $\gamma \neq 0$, then running OLS on a selected sample results in omitted variable bias (OVB) because $\gamma\lambda(x\delta)$ is omitted.

How to run a Heckman selection procedure:

1) Estimate $P(L=1|x) = \Phi(x\delta)$ using all N observations. Use PROBIT!

   Obtain $\hat{\delta}$

2) Calculate Inverse Mills ratios $\lambda(x\hat{\delta})$ for $i=1,\ldots,N_1$, i.e., a subsample of N.

3) Run OLS regression to estimate $E(w|x_1, L=1)$:

$$w = x_1\beta + \gamma\hat{\lambda}$$

Estimates of $\hat{\beta}$ and $\hat{\gamma}$ are consistent (i.e., not biased).

Notes: x_1 does not need to be a strict subset of x. If $x_1 = x$, β is identified only due to the non-linearity of the inverse Mills ratio, λ.

The null is no selection bias.

$H_0: \gamma = 0$.

So a simple t-test on $\hat{\gamma}$ is a valid test of the null of no selection bias.