样本选择问题与处理

王群勇 (经济学教授、博士生导师)

南开大学 数量经济研究所

2018 年 8 月 19-20 日, 广东 · 顺德

Contents

sample selection

- sample selection: sample is not representative of the population of interest.
- example: population equation

$$wage = \beta_0 + \beta_1 age + \beta_2 educ + u$$

define selection indicator $s = 1$ if in sample.
- exogenous sampling: sampling is based on conditioning variable (s is a deterministic function of x).
example: $s = 1(age < 65)$.
- endogenous sampling: sampling is based on response variable (s is a deterministic function of y).
example: $s = 1(wage < 10000)$.

sample selection

- incidental selection: s is a random function of x or y.

$$s = 1(\mathrm{z}\delta + v > 0).$$

- estimating equation

$$s_i y_i = s_i \mathrm{x}_i \beta + s_i u_i$$

Its OLS estimator

$$\hat{\beta} = \left(N^{-1} \sum s_i \mathrm{x}_i' \mathrm{x}_i\right)^{-1} \left(N^{-1} \sum s_i \mathrm{x}_i' y_i\right)$$
$$= \beta + \left(N^{-1} \sum s_i \mathrm{x}_i' \mathrm{x}_i\right)^{-1} \left(\sum s_i \mathrm{x}_i' u_i\right)$$

So,

$$\text{plim}(\hat{\beta}) = \beta + [E(s\mathrm{x}'\mathrm{x})]^{-1} E[s\mathrm{x}'u]$$

illustration

	lnwage	educ	age	informal	emp	ses	fatheduc
726.	.	13	58	.	0	Low SES	ISCED 1
727.	.	18	23	.	0	High SES	ISCED 1
728.	.	13	48	.	0	Middle SES	ISCED 2 & 3
729.	.	18	20	.	0	Middle SES	ISCED 2 & 3
730.	2.218704	6	52	Informal	1	Low SES	ISCED 1
731.	1.438545	13	42	Formal	1	High SES	ISCED 4 & higher
732.	2.154165	16	38	Formal	1	Middle SES	ISCED 2 & 3
733.	1.006763	6	53	Informal	1	Low SES	ISCED 1
734.	2.441847	16	33	Formal	1	Middle SES	ISCED 1
735.	3.001463	17	42	Formal	1	Low SES	ISCED 1

sample selection

- Assumption for consistency: $E(s\mathbf{z}'u) = 0$. A sufficient condition is

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

Proof:

$$E(s\mathbf{z}'u) = E[E(s\mathbf{z}'u)|\mathbf{z}, s] = E[s\mathbf{z}'E(u|\mathbf{z}, s)] = 0.$$

- If s is a deterministic function of $\mathbf{z}$ (exogenous selection) and $E(u|\mathbf{z}) = 0$, then

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

Contents

Contents

truncated regression

- sampling design

$$s_i = 1(a_1 < y_i < a_2)$$

- density of $f(y|\mathrm{x};\beta)$

$$f(y|\mathrm{x}, s=1) = \frac{f(y|\mathrm{x};\beta)}{P(a_1 < y < a_2|\mathrm{x})} = \frac{f(y|\mathrm{x};\beta)}{F(a_2|\mathrm{x},\beta) - F(a_1|\mathrm{x},\beta)}$$

probit selection

- population model (regression equation, selection equation):

$$y_1 = \mathrm{x}_1\beta_1 + u_1,$$
$$s = 1(\mathrm{x}\delta_2 + v_2 > 0)$$

- Assume $u_1 = \gamma_1 v_2 + e_1$,

$$E(y_1|\mathrm{x}, v_2) = \mathrm{x}_1\beta_1 + E(u_1|v_2) = \mathrm{x}_1\beta_1 + \gamma_1 v_2.$$

and

$$E(y_1|\mathrm{x}, s) = E[E(y_1|\mathrm{x}, v_2)|\mathrm{x}, s] = \mathrm{x}_1\beta_1 + \gamma_1 E(v_2|\mathrm{x}, s)$$

Heckman two-step

- With $s = 1$,

$$E(y_1|\mathrm{x}, s = 1) = \mathrm{x}_1\beta_1 + \gamma_1 E(v_2|\mathrm{x}, v_2 > -\mathrm{x}\delta_2)$$
$$\mathrm{x}_1\beta_1 + \gamma_1\lambda(\mathrm{x}\delta_2)$$

  where $\lambda(z)$ is the inverse Mills ratio $\lambda(z) = \phi(z)/\Phi(z)$.

- Heckman (1976) two-step method:
  (1) Probit of s on x using all data to get $\hat{\lambda}$.
  (2) Run OLS of y_1 on $\mathrm{x}_1, \hat{\lambda}$.
- Note:
  1. The se in the 2nd step should be adjusted.
  2. use t-test to test the sample selection problem.
  3. add at least one more variable in the selection equation.

Extensions

- some extension:

$$E(u_1|v_2) = \gamma_1 v_2 + \gamma_2(v_2^2 - 1)$$

can show that

$$E(v_2^2 - 1|\mathrm{x}, s = 1) = -\lambda(\mathrm{x}\delta_2)(\mathrm{x}\delta_2).$$

so, the mean equation is

$$E(y_1|\mathrm{x}, s) = \mathrm{x}_1\beta_1 + \gamma_1\lambda(\mathrm{x}\delta_2) - \gamma_2\lambda(\mathrm{x}\delta_2)(\mathrm{x}\delta_2)$$

Syntax

- command

. heckman *dep varlist*, select(*sel* = *varlit2*) twostep
. eregress *dep varlist*, select(*sel* = *varlit2*)

- option of predict

```
xb             linear prediction; the default
stdp           standard error of the prediction
stdf           standard error of the forecast
xbsel          linear prediction for selection equation
stdpsel        se of the linear prediction for selection equation
pr(a,b)        Pr(y | a < y < b)
e(a,b)         E(y | a < y < b)
ystar(a,b)     E(y*), y* = max{a,min(y,b)}
ycond          E(y | y observed)
yexpected      E(y*), y taken to be 0 where unobserved
mills          nonselection hazard (inverse of Mills's ratio)
psel           Pr(y observed)
```

Syntax

- example (“step China.dta”)

```
global xs "educ age age2 informal"
heckman lnwage $xs, select(emp=$xs cog)
margins, predict(pr(0,.))
margins, predict(e(0,.))
margins, predict(ystar(0,.))
eregress lnwage $xs, select(emp=$xs cog)
```

Contents

Tobit selection

- tobit selection model

$$\begin{aligned} y_1 &= \mathrm{x}_1\beta_1 + u_1, \\ y_2 &= \max(0, \mathrm{x}\delta_2 + v_2), \\ s &= 1(y_2 > 0) \end{aligned}$$

- Assume $u_1 = \gamma_1 v_2 + e_1$,

$$E(y_1|\mathrm{x}, v_2) = \mathrm{x}_1\beta_1 + E(u_1|v_2) = \mathrm{x}_1\beta_1 + \gamma_1 v_2.$$

Now v_2 can be effectively observed.

$$v_2 = y_2 - \mathrm{x}\delta_2$$

Tobit selection

- two-step method:
  (1) Tobit of y_2 on x using all data to get $\hat{v}_2$.
  (2) Run OLS of y_1 on $\mathrm{x}_1, \hat{v}_2$.
- Note:
  1. The se in the 2nd step should be adjusted.
  2. use t-test to test the sample selection problem.

Syntax

- command

```
. eregress dep varlist, options
```

some *options*:

1. tobitselect(*sel = varlist2*)
2. extreat(*tvar*) entreat(*tvar=varlist*)
3. endog(*endog=varlist, model*)

- option for `predict`

mean	mean; the default
pr	probability of binary or ordinal y
pomean	potential-outcome mean
te	treatment effect
tet	treatment effect on the treated
xb	linear prediction
pr(a,b)	Pr(a < y < b) for continuous y
e(a,b)	E(y | a < y < b) for continuous y
ystar(a,b)	E(y*), y* = max{a,min(y,b)} for continuous y

Syntax

- example ("step China.dta")

```
global xs "educ age age2 informal"
replace hours= 0 if mi(lnwage)
eregress lnwage $xs, tobitselect(hours=$xs cog)
```

Contents

Sample selection with EV

- sample selection with endogenous variable

$$\begin{aligned} y_1 &= \mathrm{z}_1\delta_1 + \alpha_1 y_2 + u_1, \\ y_2 &= \mathrm{z}_2\delta_2 + v_2, \\ s &= 1(\mathrm{z}\delta_3 + v_3 > 0) \end{aligned}$$

It is helpful to force oneself to include one at least more element in z_2 not in z_1, and then one more element in z not in z_2.

- Assume $E(u_1|v_3) = \gamma_1 v_3$,
  1. Probit of s on z using all data, obtain $\hat{\lambda}(\mathrm{z}\hat{\delta}_3)$.
  2. Apply 2SLS on

$$y = \mathrm{z}_1\delta_1 + \alpha_1 y_2 + \gamma_1\hat{\lambda}(\mathrm{z}\hat{\delta}_3) + \epsilon.$$

Syntax

- example ("step China.dta")

```
global xs "age age2"
eregress lnwage informal tenure $xs, ///
  endog(educ=heduc mothedu ses) ///
  select(emp=$xs cog)
```

Contents

Binary response model with sample selection

- For general model with sample selection

$$
\begin{aligned}
f(y_1|\mathrm{z}) &\sim \quad ...... \\
s &= 1(\mathrm{z}\delta + v_2 > 0)
\end{aligned}
$$

- binary response model:

$$
\begin{aligned}
y_1|\mathrm{z} &= 1(\mathrm{z}_1\delta_1 + u_1 > 0) \\
s &= 1(\mathrm{z}\delta + v_2 > 0)
\end{aligned}
$$

with $corr(u_1, v_2) = \rho$.

Syntax

- command for probit model with sample selection

```
. heckprobit dep varlist (sel = varlist2)
. eprobit dep varlist, select(sel = varlit2)
```

- options for `predict` after `heckprobit`

```
pmargin      Pr(depvar=1); the default
p11          Pr(depvar=1, depvar_s=1)
p10          Pr(depvar=1, depvar_s=0)
p01          Pr(depvar=0, depvar_s=1)
p00          Pr(depvar=0, depvar_s=0)
psel         Pr(depvar_s=1)
pcond        Pr(depvar=1 | depvar_s=1)
xb           linear prediction
stdp         standard error of the linear prediction
xbsel        linear prediction for selection equation
stdpsel      se of the linear prediction for selection equation
```

Syntax

- example ("step China.dta")

```
global xs "educ age age2"
heckprobit informal $xs, select(emp=$xs cog)
eprobit informal $xs, select(emp=$xs cog)
```

Ordinal response model with sample selection

- Ordinal response model with probit selection

$$
\begin{aligned}
Pr(y_i = j|z_1) &= Pr(c_{j-1} < z_1\delta_1 + u_1 \leq c_j), \;\; j = 1, 2, ..., J \\
s &= 1(z\delta + v_2 > 0)
\end{aligned}
$$

with (u_1, v_2) has bivariate normal distribution with correlation ρ.

Syntax

- command for ordinal probit model with sample selection

. heckoprobit *dep varlist* (*sel* = *varlist2*)

- options for `predict` after `heckoprobit`

```
pmargin    marginal probabilities; the default
p1         pr(y_i=j,s_i=1)
p0         pr(y_i=j,s_i=0)
pcond1     pr(y_i=j|s_i=1)
pcond0     pr(y_i=j|s_i=0)
psel       selection probability
xb         linear prediction
stdp       standard error of the linear prediction
xbsel      linear prediction for selection equation
stdpsel    se of the linear prediction for selection equation
outcome    which outcome
```

Syntax

- command for ordinal probit model with sample selection

. eoprobit *dep varlist*, select(*sel* = *varlit2*)
. eoprobit *dep varlist*, tobitselect(*sel* = *varlit2*)

- options for `predict`

```
pr              probability of each outcome; the default
outlevel(#)     calculate probability for m = # only
xb              linear prediction excluding all complications
```

Syntax

- example (“womensat.dta”)

```
global xs "age education"
heckoprobit satisfaction $xs, select(work=$xs married children)
eoprobit satisfaction $xs, select(work=$xs married children)
```

Count data model with sample selection

- count data model with probit selection

$$
\begin{aligned}
E(y_1|z,u_1) &= \exp(x_i\beta + u_1) \\
s &= 1(z\delta + v_2 > 0)
\end{aligned}
$$

with $corr(u_1, v_2) = \rho$.

Syntax

- command for probit model with probit selection

```
. heckpoisson depvar varlist (sel = varlist2)
```

- options for predict

```
n           E(y_i); the default
ir          incidence rate
ncond       E(y_i|s_i=1)
pr(n)       Pr(y = n)
pr(a,b)     Pr(a < y < b)
psel        Pr(y observed)
xb          linear prediction
xbsel       linear prediction for selection equation
```

Syntax

- example ("patent.dta")

```
heckpoisson npatents expenditure i.tech, select(applied = expenditure size i.tech)
margins i.tech, at(expenditure = generate(expenditure)) ///
  at(expenditure = generate(expenditure+1)) post
lincom (_b[2._at#1.tech] - _b[1._at#1.tech]) - (_b[2._at#0.tech] - _b[1._at#0.tech])
```

Contents

model

- model

$$s_{it}y_{it} = s_{it}x_{it}\beta + s_{it}(c_i + u_{it})$$

- consistency of POLS requires

$$E(s_{it}x_{it}c_i) = 0, E(s_{it}x_{it}u_{it}) = 0.$$

model

- FE estimation equation

$$s_{it}(y_{it} - \bar{y}_i) = s_{it}(x_{it} - \bar{x}_i)\beta + s_{it}(u_{it} - \bar{u}_i)$$

the consistency of FE estimator requires strict exogeneity

$$E(u_{it}|x_i, s_i, c_i) = 0$$

(1) This rules out selection in any time period depending on the shocks in any time period.
(2) s_{it} is allowed to depend on c_i in an unrestricted way.

model

- RE estimation equation

$$s_{it}(y_{it} - \lambda_i \bar{y}_i) = s_{it}(x_{it} - \lambda_i \bar{x}_i)\beta + s_{it}(1 - \lambda_i)c_i + s_{it}(u_{it} - \lambda_i \bar{u}_i)$$

the consistency of RE estimator requires strict exogeneity

$$\begin{aligned} E(u_{it}|x_i, s_i, c_i) &= 0 \\ E(c_i|x_i, s_i) &= E(c_i) \end{aligned}$$

test for selection

- model based on Mundlak-Chamberlan correlated random effect,

$$\begin{aligned} y_{it} &= x_{it}\beta + c_i + u_{it} \\ s_{it} &= 1(z_{it}\delta + \psi_2 + \bar{x}_i\xi_2 + v_{it}) \end{aligned}$$

with $v_{it}|x_i \sim Normal(0,1)$.

- (1) Estimate pooled probit model, get the IMR

$$\hat{\lambda}_{it} = \lambda(z_{it}\delta + \psi_2 + \bar{x}_i\xi_2)$$

(2) add $\hat{\lambda}_{it}$ into y_{it} equation,

$$y_{it} = x_{it}\beta + \gamma\hat{\lambda}_{it} + c_i + \epsilon_{it}$$

and use t-statistic in FE estimation to test selection. Or, interact $\hat{\lambda}$ with time dummies to get a joint test.

$$y_{it} = x_{it}\beta + d2_t\hat{\lambda}_{it} + \ldots + dT_t\hat{\lambda}_{it} + c_i + \epsilon_{it}$$

where $d2_t = 1$ if $t = 2$, ..., $dT_t = 1$ if $t = T$.

test for selection

- (1) For more flexibility, estimate the selection model separately for each t, and get $\hat{\lambda}_1, ..., \hat{\lambda}_T$.
  (2) add $\hat{\lambda}_1, ..., \hat{\lambda}_T$ into FE equation, and use F-statistic for selection test.

Heckman approach for selection

- model

$$
\begin{aligned}
y_{it} &= x_{it}\beta + \psi_1 + \bar{x}_i\xi_1 + u_{it} \\
s_{it} &= 1(z_{it}\delta + \psi_2 + \bar{z}_i\xi_2 + v_{it})
\end{aligned}
$$

  with $E(u_{it}|x_i, v_{it}) = \gamma_1 v_{it}$.

- then

$$
y_{it} = x_{it}\beta + \psi_1 + \bar{x}_i\xi_1 + \gamma_1 E(u_{it}|x_i, s_{it}) + e_{it}
$$

Heckman approach for selection

- two-step:
  (1) estimate pooled probit model

$$s_{it} = 1(z_{it}\delta + \psi_2 + \bar{z}_i\xi_2 + v_{it})$$

  and get $\hat{\lambda}_{it}$.
  (2) estimate pooled OLS model

$$y_{it} = x_{it}\beta + \psi_1 + \bar{x}_i\xi_1 + \gamma_1\hat{\lambda}_{it} + \epsilon_{it}$$

  or add interaction terms $d2_t\hat{\lambda}_{it}, d3_t\hat{\lambda}_{it}, ..., dT_t\hat{\lambda}_{it}$.
- a more general form

$$E(u_{it}|v_{it}) = \gamma_1 v_{it} + \eta_1(v_{it}^2 - 1).$$

Tobit selection

- tobit selection model

$$y_{it2} = \max(0, z_{it}\delta + \psi_2 + \bar{z}_i\xi_2 + v_{it})$$

- two-step:
  (1) estimate pooled tobit model, get $\hat{v}_{it}$.
  (2) estimate pooled ols

$$y_{it} = x_{it}\beta + \psi_1 + \bar{x}_i\xi_1 + \gamma_1\hat{v}_{it} + \epsilon_{it}$$

Attrition

- Assume a random sample from the population at time $t = 1$. In other words, $s_{i1} = 1$ for all i. With attrition, some units leave the sample in subsequent time periods.
- two-step procudure:
(1) starting with $t = 2$, estimate a estimate a sequence of probit models for the group of units in the sample at time $t - 1$: probit of s_{it} on z_{it} for the subsample with $s_{i,t-1} = 1$. The vector z_{it} grows as t increases. Obtain the inverse Mills ratios, $\hat{\lambda}_{it}$.
(2) Using the selected sample ($s_{it} = 1$), run the pooled OLS regression

$$\Delta y_{it} \text{ on } \Delta x_{it}, d2_t\hat{\lambda}_{it}, ..., dT_t\hat{\lambda}_{it}.$$

where allowing a different coefficient on $\hat{\lambda}$ in each time period is required because of the nature of the sequential procedure.

Contents

IPW

- IPW applies generally to any estimation problem that involves minimization or maximization.
- M-estimation

$$\min N^{-1} \sum s_i q(w_i, \theta)$$

Assumption:

$$P(s_i = 1 | w_i, z_i) = P(s_i = 1 | z_i)$$

- Let $\mu = E[g(w_i)]$. Using iterated expectation

$$\begin{aligned} E[s_i g(w_i)/p(z_i)] &= E[E(s_i g(w_i)/p(z_i) | w_i, z_i)] \\ &= E[E(s_i | w_i, z_i) g(w_i)/p(z_i)] \\ &= E[P(s_i = 1 | w_i, z_i) g(w_i)/p(z_i)] \\ &= E[P(z_i) g(w_i)/p(z_i)] = E[g(w_i)] \end{aligned}$$

IPW

- Weighting a function by $1/p(z_i)$ recover the population mean, and a consistent estimator of μ is

$$\hat{\mu} = N^{-1} \sum [s_i g(w_i)/p(z_i)].$$

- Based on $E(s_i/p(z_i) = 1$, a more common estimator is

$$\hat{\mu} = \left(\sum s_i/p(z_i) \right) \left(\sum [s_i g(w_i)/p(z_i)] \right).$$

- IPW M-estimator

$$\min N^{-1}\left[\sum s_i/p(z_i)q(w_i,\theta)\right].$$

Let $\hat{p}(z_i)=G(z_i,\hat{\gamma})$,

$$\min N^{-1}\left[\sum s_i/G(z_i,\hat{\gamma})q(w_i,\theta)\right].$$

the two-step estimator will get a consistent estimator of θ.

IPW

- IPW M-estimator can be used in linear and nonlinear models (such as probit or tobit models).
- Use bootstrap to make accurate inference.

Syntax

- example (“step china.dta”)

```
global x "educ age informal"
gen s = !mi(lnwage)
probit s educ age
predict p, pr
reg lnwage $x [pw=1/p]
```

Summarization of Stata commands

- Table.

dep.	Probit selection	Tobit selection
cont.	heckman eregress	eregress
binary	heckprobit eprobit	eprobit
ordinal	heckoprobit eoprobit	eoprobit
count	heckpoisson	–

References

- Vella, Francis (1998), Estimating Models with Sample Selection Bias: A Survey, Journal of Human Resources, 33, pp. 127-169.
  Kyriazidou, E. (1997). Estimation of a Panel Data Sample Selection Model. Econometrica. 65 (6), pp. 1335-1364.

  Wooldridge, J. (1995). Selection Corrections for Panel Data Models under Conditional Mean

  Independence Assumptions. Journal of Econometrics. 68, pp. 115-132.