23) Sample Selection

Vitor Kamada

August 2018

Exogenous Sample Selection

$$y_i = x_i\beta + u_i, \quad E(u|x) = 0$$

$$s_i y_i = s_i x_i \beta + s_i u_i$$

$s_i = 1$ if (y_i, x_i) is observed, and $s_i = 0$ otherwise

$$E(su) = 0 \text{ and } E[(sx)(su)] = E(sxu) = 0$$

If $s \perp (x, u)$, then
$$E(sxu) = E(s)E(xu) = 0$$

Example of Random Sample Selection

$$s = 1 \text{ if } IQ \geq r$$

$$s = 0 \text{ if } IQ < r$$

$$r \perp (IQ, x, u)$$

$$E(u|x_1, ..., x_k, s) = E(u|x_1, ..., x_k)$$

$$E(u^2|x, s) = E(u^2) = \sigma^2$$

Incidental Truncation: A Probit Selection Equation

$$y_1 = x_1\beta_1 + u_1$$

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

(x, y_2) are always observed

y_1 is observed when $y_2 = 1$

$$(u_1, v_2) \perp x$$

$$v_2 \sim N(0, 1)$$

$$E(u_1|v_2) = \gamma_1 v_2$$

Heckit Method

$$y_1 = x_1\beta_1 + u_1$$

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

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

$$= x_1\beta_1 + \gamma_1 v_2$$

$$E(v_2|x, y_2 = 1) = E(v_2|v_2 > -x\delta_2)$$

$$= \frac{\phi(x\delta_2)}{\Phi(x\delta_2)} = \lambda(x\delta_2)$$

Sample Selection Correction

1) Probit of y_2 on x and compute the inverse Mills Ratios, $\hat{\lambda} = \lambda(x\hat{\delta}_2)$

2) Regress y_1 on $x_1, \hat{\lambda}$

$$H_0 : \gamma_1 = 0$$

(no sample selection problem)

Mroz (1987)

Variable	Obs	Mean	Std. Dev.	Min	Max
lwage	428	1.190173	.7231978	-2.054164	3.218876
educ	753	12.28685	2.280246	5	17
exper	753	10.63081	8.06913	0	45
expersq	753	178.0385	249.6308	0	2025
nwifeinc	753	20.12896	11.6348	-.0290575	96
age	753	42.53785	8.072574	30	60
kidslt6	753	.2377158	.523959	0	3
kidsge6	753	1.353254	1.319874	0	8
inlf	753	.5683931	.4956295	0	1
motheduc	753	9.250996	3.367468	0	17
fatheduc	753	8.808765	3.57229	0	17

reg lwage $xlist

Source	SS	df	MS
Model	35.0222967	3	11.6740989
Residual	188.305144	424	.444115906
Total	223.327441	427	.523015084

Number of obs	=	428
F(3, 424)	=	26.29
Prob > F	=	0.0000
R-squared	=	0.1568
Adj R-squared	=	0.1509
Root MSE	=	.66642

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.1074896	.0141465	7.60	0.000	.0796837	.1352956
exper	.0415665	.0131752	3.15	0.002	.0156697	.0674633
expersq	-.0008112	.0003932	-2.06	0.040	-.0015841	-.0000382
_cons	-.5220406	.1986321	-2.63	0.009	-.9124667	-.1316144

heckman lwage $xlist, select(inlf = $xlist $slist) twostep

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lwage						
educ	.1090655	.015523	7.03	0.000	.0786411	.13949
exper	.0438873	.0162611	2.70	0.007	.0120163	.0757584
expersq	-.0008591	.0004389	-1.96	0.050	-.0017194	1.15e-06
_cons	-.5781032	.3050062	-1.90	0.058	-1.175904	.019698
inlf						
educ	.1309047	.0252542	5.18	0.000	.0814074	.180402
exper	.1233476	.0187164	6.59	0.000	.0866641	.1600311
expersq	-.0018871	.0006	-3.15	0.002	-.003063	-.0007111
nwifeinc	-.0120237	.0048398	-2.48	0.013	-.0215096	-.0025378
age	-.0528527	.0084772	-6.23	0.000	-.0694678	-.0362376
kidslt6	-.8683285	.1185223	-7.33	0.000	-1.100628	-.636029
kidsge6	.036005	.0434768	0.83	0.408	-.049208	.1212179
_cons	.2700768	.508593	0.53	0.595	-.7267473	1.266901
/mills						
lambda	.0322619	.1336246	0.24	0.809	-.2296376	.2941613
rho	0.04861					
sigma	.66362875					

Partial Maximum Likelihood Estimation

$$y_1 = x_1\beta_1 + u_1$$

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

$$\begin{bmatrix} u_1 \\ v_2 \end{bmatrix} \sim N\left[\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}\right]$$

$$L = \prod_{i=1}^{n} \{P[y_{2i} \leq 0]\}^{1-y_{2i}} \{f(y_{1i}|y_{2i} > 0) \cdot P[y_{2i} > 0]\}^{y_{2i}}$$

heckman lwage $xlist, select(inlf = $xlist $slist)

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lwage						
educ	.1083502	.0148607	7.29	0.000	.0792238	.1374767
exper	.0428369	.0148785	2.88	0.004	.0136755	.0719983
expersq	-.0008374	.0004175	-2.01	0.045	-.0016556	-.0000192
_cons	-.5526973	.2603784	-2.12	0.034	-1.06303	-.0423651
inlf						
educ	.1313415	.0253823	5.17	0.000	.0815931	.1810899
exper	.1232818	.0187242	6.58	0.000	.0865831	.1599806
expersq	-.0018863	.0006004	-3.14	0.002	-.003063	-.0007095
nwifeinc	-.0121321	.0048767	-2.49	0.013	-.0216903	-.002574
age	-.0528287	.0084792	-6.23	0.000	-.0694476	-.0362098
kidslt6	-.8673988	.1186509	-7.31	0.000	-1.09995	-.6348472
kidsge6	.0358723	.0434753	0.83	0.409	-.0493377	.1210824
_cons	.2664491	.5089578	0.52	0.601	-.7310898	1.263988
/athrho	.026614	.147182	0.18	0.857	-.2618573	.3150854
/lnsigma	-.4103809	.0342291	-11.99	0.000	-.4774687	-.3432931
rho	.0266078	.1470778			-.2560319	.3050564
sigma	.6633975	.0227075			.6203517	.7094303
lambda	.0176515	.0976057			-.1736521	.2089552

LR test of indep. eqns. (rho = 0): chi2(1) = 0.03 Prob > chi2 = 0.8577

Endogenous Explanatory Variables

$$y_1 = z_1\delta_1 + \alpha_1 y_2 + u_1$$

$$y_2 = z_2\delta_2 + v_2$$

$$y_3 = 1[z\delta_3 + v_3 > 0]$$

1) Run probit of y_3 on z using all observations, and get the inverse Mills ratios, $\hat{\lambda}_{i3} = \lambda(z_i\hat{\delta}_3)$

2) Estimate by 2SLS, using z_2 as IV:

$$y_1 = z_1\delta_1 + \alpha_1 y_2 + \gamma\hat{\lambda}_{i3} + error$$

probit inlf $xlist $slist motheduc fatheduc

predict xd3h, xb

gen phi3 = normalden(xd3h)

gen PHI3 = normal(xd3h)

gen lambda3 = phi3/PHI3

ivreg lwage exper expersq lambda3 (educ = nwifeinc age kidslt6

kidsge6 motheduc fatheduc)

lwage	Coef.	Std. Err.	t	P>|t|
educ	.1044079	.0175683	5.94	0.000
exper	.0435482	.0164173	2.65	0.008
expersq	-.0008552	.000442	-1.93	0.054
lambda3	.0241612	.136629	0.18	0.860
_cons	-.5113313	.3331186	-1.53	0.126