18) IV Probit

Vitor Kamada

August 2018

"In probit analysis, neglected heterogeneity is a much more serious problem than in linear models because, even if the omitted heterogeneity is independent of x, the probit coeficients are inconsistent."

$$P(y = 1|x, c) = \Phi(x\beta + \gamma c)$$

$$y^* = x\beta + \gamma c + e$$

$$c \perp x \text{ and } c \sim N(0, \tau^2)$$

$$(\gamma c + e) \sim N(0, \gamma^2\tau^2 + 1)$$

Attenuation Bias

$$P(y=1|x) = P(\gamma c + e > -x\beta|x) = \Phi(x\beta/\sigma)$$

$$plim\hat{\beta}_j = \frac{\beta_j}{\sigma}$$

$$\sigma = \sqrt{\gamma^2\tau^2+1} > 1$$

$$\frac{\partial P(y=1|x,c)}{\partial x_j} = \beta_j\phi(x\beta+\gamma c)$$

$$E[\beta_j\phi(x\beta+\gamma c)] = \frac{\beta_j}{\sigma}\phi(\frac{x\beta}{\sigma})$$

Probit of y on x consistently estimates the APE

Continuous Endogenous Explanatory Variables

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

$$y_2 = z_1\delta_{21} + z_2\delta_{22} + v_2 = z\delta_2 + v_2$$

$$y_1 = 1[y_1^* > 0]$$

$$z \perp (u_1, v_2) \sim N(0, \Sigma)$$

$$Var(u_1, v_2) = \Sigma = \begin{bmatrix} 1 & \Sigma_{21}' \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

As $v_2 \sim N(0, \Sigma_{22})$**,** y_2 **should not be Dummy**

Rivers and Vuong (1988): Control Function Approach

$$u_1 = \theta_1 v_2 + e_1$$

$$\theta_1 = \frac{\eta_1}{\tau_2^2}$$

$$\eta_1 = Cov(v_2, u_1) \text{ and } \tau_2^2 = Var(v_2)$$

$$Var(e_1) = Var(u_1) - \frac{\eta_1^2}{\tau_2^2} = 1 - \rho_1^2$$

$$\rho_1 = Corr(v_2, u_1)$$

$$e_1 | z, y_2, v_2 \sim N(0, 1 - \rho_1^2)$$

Two-Step Approach

$$y_1^* = z_1\delta_1 + \alpha_1 y_2 + \theta_1 v_2 + e_1$$

$$P(y_1 = 1|z, y_2, v_2) = \Phi[(\frac{z_1\delta_1+\alpha_1 y_2+\theta_1 v_2}{\sqrt{1-\rho_1^2}})]$$

1) Run OLS regression y_2 on z and get the $\hat{v}_2$

2) Run the probit y_1 on $z_1, y_2, \hat{v}_2$ to get:

$$\delta_{\rho 1} = \frac{\delta_1}{\sqrt{1-\rho_1^2}},\ \alpha_{\rho 1} = \frac{\alpha_1}{\sqrt{1-\rho_1^2}},\ \theta_{\rho 1} = \frac{\theta_1}{\sqrt{1-\rho_1^2}}$$

Conditional Maximum Likelihood Estimation (CMLE)

$$f(y_1, y_2|z) = f(y_1|y_2, z)f(y_2|z)$$

$$P(y_1 = 1|y_2, z) = \Phi[(\frac{z_1\delta_1+\alpha_1 y_2+(\rho_1/\tau_2)(y_2-z\delta_2)}{\sqrt{1-\rho_1^2}})]$$

$$[\{\Phi(w)\}^{y_1}\{1-\Phi(w)\}^{1-y_1}]\frac{1}{\tau_2}\phi[\frac{y_2-z\delta_2}{\tau_2}]$$

$$\ell_i(\delta_1, \alpha_1, \rho_1, \delta_2, \tau_2) =$$

$$y_{i1} log\Phi(w_i) + (1-y_{i1}) log[1-\Phi(w_i)]$$

$$-\frac{1}{2}log(\tau_2^2) - \frac{1}{2}(\frac{y_{i2}-z_i\delta_2}{\tau_2})^2$$

Sample is restricted to Medicare

ins: supplementary insurance

linc: log household income

hstatusg: health status is good

adl: # of limitations on activities of daily living

sretire: spouse retirement

probit ins linc $xlist2, vce(robust) nolog

```
Probit regression                               Number of obs     =      3,197
                                                Wald chi2(11)     =     366.94
                                                Prob > chi2       =     0.0000
Log pseudolikelihood = -1933.4275               Pseudo R2         =     0.0946
```

ins	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
linc	.3466893	.0402173	8.62	0.000	.2678648	.4255137
female	-.0815374	.0508549	-1.60	0.109	-.1812112	.0181364
age	.1162879	.1151924	1.01	0.313	-.109485	.3420608
age2	-.0009395	.0008568	-1.10	0.273	-.0026187	.0007397
educyear	.0464387	.0089917	5.16	0.000	.0288153	.0640622
married	.1044152	.0636879	1.64	0.101	-.0204108	.2292412
hisp	-.3977334	.1080935	-3.68	0.000	-.6095927	-.1858741
white	-.0418296	.0644391	-0.65	0.516	-.168128	.0844687
chronic	.0472903	.0186231	2.54	0.011	.0107897	.0837909
adl	-.0945039	.0353534	-2.67	0.008	-.1637953	-.0252125
hstatusg	.1138708	.0629071	1.81	0.070	-.0094248	.2371664
_cons	-5.744548	3.871615	-1.48	0.138	-13.33277	1.843677

ivprobit ins $xlist2 (linc = $ivlist2), twostep first

linc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
retire	-.0909581	.0288119	-3.16	0.002	-.1474499	-.0344663
sretire	-.0443106	.0317252	-1.40	0.163	-.1065145	.0178932
female	-.0936494	.0297304	-3.15	0.002	-.151942	-.0353569
age	.2669284	.0627794	4.25	0.000	.1438361	.3900206
age2	-.0019065	.0004648	-4.10	0.000	-.0028178	-.0009952
educyear	.094801	.0043535	21.78	0.000	.0862651	.1033369
married	.7918411	.0367275	21.56	0.000	.7198291	.8638531
hisp	-.2372014	.0523874	-4.53	0.000	-.3399179	-.134485
white	.2324672	.0347744	6.69	0.000	.1642847	.3006496
chronic	-.0388345	.0100852	-3.85	0.000	-.0586086	-.0190604
adl	-.0739895	.0173458	-4.27	0.000	-.1079995	-.0399795
hstatusg	.1748137	.0338519	5.16	0.000	.10844	.2411875
_cons	-7.702456	2.118657	-3.64	0.000	-11.85653	-3.548385

ivprobit ins $xlist2 (linc = $ivlist2), twostep

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
linc	-.6109088	.5723054	-1.07	0.286	-1.732607	.5107893
female	-.167917	.0773839	-2.17	0.030	-.3195867	-.0162473
age	.3422526	.1915485	1.79	0.074	-.0331756	.7176808
age2	-.0025708	.0014021	-1.83	0.067	-.0053188	.0001773
educyear	.13596	.0543047	2.50	0.012	.0295249	.2423952
married	.8351517	.441743	1.89	0.059	-.0306487	1.700952
hisp	-.6184546	.181427	-3.41	0.001	-.9740451	-.2628642
white	.1818279	.1528281	1.19	0.234	-.1177098	.4813655
chronic	.0095837	.0309618	0.31	0.757	-.0511004	.0702678
adl	-.1630884	.0568288	-2.87	0.004	-.2744709	-.0517059
hstatusg	.2809463	.1228386	2.29	0.022	.0401871	.5217055
_cons	-12.04848	5.928158	-2.03	0.042	-23.66746	-.4295071

```
Instrumented:  linc
Instruments:   female age age2 educyear married hisp white chronic adl
               hstatusg retire sretire
```

```
Wald test of exogeneity: chi2(1) = 3.57                  Prob > chi2 = 0.0588
```

ivprobit ins $xlist2 (linc = $ivlist2), vce(robust) mle

	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
linc	-.5338252	.3852132	-1.39	0.166	-1.288829	.2211788
female	-.1394072	.0494471	-2.82	0.005	-.2363218	-.0424926
age	.2862293	.1280821	2.23	0.025	.0351929	.5372656
age2	-.0021472	.0009318	-2.30	0.021	-.0039735	-.0003209
educyear	.1136881	.0237914	4.78	0.000	.0670579	.1603183
married	.7058309	.2377594	2.97	0.003	.239831	1.171831
hisp	-.5094514	.1049487	-4.85	0.000	-.715147	-.3037558
white	.1563454	.1035674	1.51	0.131	-.0466429	.3593338
chronic	.0061939	.027525	0.23	0.822	-.0477542	.060142
adl	-.1347664	.0349799	-3.85	0.000	-.2033258	-.0662071
hstatusg	.2341789	.0709755	3.30	0.001	.0950694	.3732883
_cons	-10.00787	4.065771	-2.46	0.014	-17.97664	-2.039107
corr(e.linc, e.ins)	.5879559	.2355329			-.0309872	.8809669
sd(e.linc)	.7177787	.0167816			.6856296	.7514352

```
Instrumented:  linc
Instruments:   female age age2 educyear married hisp white chronic adl
               hstatusg retire sretire
```

```
Wald test of exogeneity (corr = 0): chi2(1) = 3.51         Prob > chi2 = 0.0610
```

Married women in the United States who have at least two children

$y_1 = worked$: 59 % of the women report being in the labor force

$y_2 = morekids$: 1 if a woman has three or more children (49% of the sample)

samesex: 1 if first two children are of the same sex

Controls: "non-momi" income, educ, age, black, and hispanic

Estimated Effect of Having Three or More Children on Women's Labor Force Participation

Dependent Variable: *worked*

	(1)	(2)	(3)	(4)	(5)
Model	LPM	Probit	LPM	Bivariate probit	Bivariate probit
Estimation method	OLS	MLE	2SLS: *samesex* as IV	MLE: *samesex* as IV	MLE: no IV
Coefficient on *morekids*	−.109 (.006)	−.299 (.015)	−.201 (.096)	−.703 (.204)	−.966 (.243)
APE for *morekids*	−.109 (.006)	−.109 (.006)	−.201 (.096)	−.256 (.072)	−.349 (∗)
$\hat{\rho}$	—	—	—	.254 (.131)	.426 (.162)
Number of observations	31,857	31,857	31,857	31,857	31,857