Title

Example 6b — Ordered probit regression with endogenous treatment and sample selection

Description   Remarks and examples   Also see

Description

Continuing from [ERM] **Example 6a**, we show you how to estimate and interpret the results of a model for an ordinal outcome when the model includes an endogenous treatment and the data are subject to endogenous sample selection.

Remarks and examples

stata.com

Suppose that we collected our data at doctors' offices and thus observe health score information only from women who visited their doctor in the study time frame (`drvisit` = 1). We suspect that unobserved factors that affect whether a woman visited the doctor are related to those that affect whether she has insurance and to those that affect her health status. Thus, we have an endogenously selected sample and an endogenously chosen treatment.

For our selection model, we use the endogenous treatment indicator for insurance status and regular checkups before the study (`regcheck`), which is excluded from the outcome model. Our command is otherwise exactly the same as specified in [ERM] **Example 6a**.

```
. eoprobit health i.exercise c.grade, entreat(insured = grade i.workschool)
> select(select = i.insured i.regcheck) vce(robust)
```

(iteration log omitted)

```
Extended ordered probit regression                  Number of obs    =   6,000
                                                          Selected   =   4,693
                                                       Nonselected   =   1,307

                                                    Wald chi2(4)     =  367.30
Log pseudolikelihood = -9806.1189                   Prob > chi2      =  0.0000
```

	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
health						
exercise#insured						
Yes#No	.4169984	.0851131	4.90	0.000	.2501798	.583817
Yes#Yes	.5399986	.037546	14.38	0.000	.4664098	.6135874
insured#c.grade						
No	.1317866	.0342405	3.85	0.000	.0646765	.1988967
Yes	.1343324	.0129342	10.39	0.000	.1089818	.159683
select						
insured						
Yes	1.01669	.092325	11.01	0.000	.8357364	1.197644
regcheck						
Yes	.5374105	.0397297	13.53	0.000	.4595417	.6152793
_cons	-.1690644	.0743716	-2.27	0.023	-.3148301	-.0232987
insured						
grade	.3057852	.0100116	30.54	0.000	.2861628	.3254076
workschool						
Yes	.5314797	.0452607	11.74	0.000	.4427703	.6201891
_cons	-3.584315	.1348183	-26.59	0.000	-3.848554	-3.320077
/health						
insured#c.cut1						
No	.7262958	.3313472			.0768673	1.375724
Yes	-.5450451	.3181876			-1.168681	.0785912
insured#c.cut2						
No	1.719809	.3129056			1.106526	2.333093
Yes	.5683456	.2464686			.085276	1.051415
insured#c.cut3						
No	2.620793	.3056038			2.021821	3.219766
Yes	1.442022	.2227768			1.005387	1.878656
insured#c.cut4						
No	3.48945	.3158536			2.870389	4.108512
Yes	2.391497	.2090187			1.981828	2.801166
corr(e.sel~t, e.health)	.496699	.0990366	5.02	0.000	.2795869	.665485
corr(e.ins~d, e.health)	.4032487	.121518	3.32	0.001	.1421331	.6118937
corr(e.ins~d, e.select)	.2661948	.0555596	4.79	0.000	.1543216	.3713287

At both levels of the treatment, exercise and education still have positive effects on health status.

The correlation between the errors from the selection equation and the errors from the main equation is 0.497. This is significantly different from zero, so we confirm our suspicion of endogeneity. Because it is positive, we conclude that unobservable factors that increase the chance of being in the study also tend to increase the chance of being in a higher health status category.

What are the expected average probabilities of being in each health status if every woman had insurance? If every woman did not have insurance? We can answer those questions using `estat teffects`.

```
. estat teffects, pomean

Predictive margins                                  Number of obs = 6,000

POmean_Pr1: Pr(health=1=Poor)
POmean_Pr2: Pr(health=2=Not good)
POmean_Pr3: Pr(health=3=Fair)
POmean_Pr4: Pr(health=4=Good)
POmean_Pr5: Pr(health=5=Excellent)
```

	Margin	Unconditional std. err.	z	P>|z|	[95% conf.	interval]
POmean_Pr1						
insured						
No	.1028382	.0327177	3.14	0.002	.0387126	.1669637
Yes	.0058955	.0033611	1.75	0.079	-.0006921	.0124831
POmean_Pr2						
insured						
No	.2621517	.0479497	5.47	0.000	.1681719	.3561314
Yes	.0618234	.0116191	5.32	0.000	.0390504	.0845965
POmean_Pr3						
insured						
No	.3216819	.0259933	12.38	0.000	.270736	.3726278
Yes	.1759926	.0100741	17.47	0.000	.1562478	.1957374
POmean_Pr4						
insured						
No	.2144017	.0402798	5.32	0.000	.1354547	.2933488
Yes	.3237595	.009282	34.88	0.000	.3055672	.3419519
POmean_Pr5						
insured						
No	.0989265	.0521147	1.90	0.058	-.0032163	.2010694
Yes	.4325289	.0165829	26.08	0.000	.400027	.4650309

These are the estimates of the average potential-outcome means for the population. We can consider the values in this table to be either the expected proportions of all women being in a status category or the average probabilities of being in a status category. If we multiply by 100, we can talk about the expected percentage of all women being in a status category. The first pair of rows shows the probabilities of being in the first health status, poor. If all women are uninsured, the probability of having a poor health status is 0.10. If all women are insured, that probability falls to 0.01. At the other end of the spectrum, only 9.9% of women are expected to have excellent health if no women are insured. That number rises to 43.3% if all women are insured.

If we sum all the proportions labeled `no`, that sum is 1.0. The same is true of the proportions labeled `yes`. The sum of the proportions must be 1.0 because each woman can be in only one health status.

In any health status, if we subtract the potential-outcome mean when assuming all women are uninsured from the mean when assuming all women to be insured, we estimate the average treatment effect (ATE). This is the ATE that being insured has on the probability of being in the health status category. Let's do that.

```
. estat teffects

Predictive margins                                    Number of obs = 6,000

ATE_Pr1: Pr(health=1=Poor)
ATE_Pr2: Pr(health=2=Not good)
ATE_Pr3: Pr(health=3=Fair)
ATE_Pr4: Pr(health=4=Good)
ATE_Pr5: Pr(health=5=Excellent)
```

	Margin	Unconditional std. err.	z	P>|z|	[95% conf.	interval]
ATE_Pr1 insured (Yes vs No)	-.0969427	.0333853	-2.90	0.004	-.1623767	-.0315086
ATE_Pr2 insured (Yes vs No)	-.2003283	.0552089	-3.63	0.000	-.3085358	-.0921207
ATE_Pr3 insured (Yes vs No)	-.1456893	.0322109	-4.52	0.000	-.2088216	-.082557
ATE_Pr4 insured (Yes vs No)	.1093578	.0437353	2.50	0.012	.0236382	.1950774
ATE_Pr5 insured (Yes vs No)	.3336024	.0637745	5.23	0.000	.2086066	.4585982

Looking at the last line, we see that the average probability of being in excellent health in the population of women aged 25 to 30 is 0.33 greater when all women have health insurance versus when no women have health insurance.

Because we specified `vce(robust)` at estimation, all of our estimates from `estat teffects` reported standard errors for the population ATE rather than standard errors that are conditional on the sample ATE.

Also see

[ERM] **eoprobit** — Extended ordered probit regression

[ERM] **eoprobit postestimation** — Postestimation tools for eoprobit and xteoprobit

[ERM] **estat teffects** — Average treatment effects for extended regression models

[ERM] **Intro 4** — Endogenous sample-selection features

[ERM] **Intro 5** — Treatment assignment features

[ERM] **Intro 9** — Conceptual introduction via worked example