Title

stata.com

example 3b — Probit regression with endogenous covariate and treatment

Description Remarks and examples Also see

Description

We model a binary outcome that depends on a continuous endogenous covariate and has an endogenous treatment by using `eprobit` with the `endogenous()` and `entreat()` options.

Remarks and examples

stata.com

Continuing from [ERM] **example 3a**, State U administrators have implemented a voluntary program to increase retention freshman year. Whether a student chose to participate is stored in the indicator variable `program`. They are concerned that unobservable factors that influence a student's decision to participate in the college retention program also influence the probability of graduation. For example, students who have higher self-motivation may be more likely to join and also more likely to graduate without the program. Thus, they are concerned that participation in the program may be an endogenously chosen treatment. Further, they would like to control for the possibility that the unobserved factors affecting graduation have different relationships with the unobserved factors that affect participation and high school GPA for those who participated and those who did not.

The researchers believe the program was easier to access for students who lived on campus freshman year. They also think students who had scholarships may have been more motivated to attend the program. However, they do not believe either of these variables independently affects the probability of graduation after controlling for other covariates in the model. They use an indicator for on-campus residence during the freshman year (`campus`), having a scholarship of any kind (`scholar`), and parents' income in the treatment assignment model.

```
. eprobit graduate income i.roommate, endogenous(hsgpa = income i.hscomp)
> entreat(program = i.campus i.scholar income, pocorrelation) vce(robust)

Iteration 0:   log pseudolikelihood = -2793.4696
Iteration 1:   log pseudolikelihood = -2792.8365
Iteration 2:   log pseudolikelihood = -2792.7434
Iteration 3:   log pseudolikelihood = -2792.7433

Extended probit regression                      Number of obs     =      2,500
                                                Wald chi2(8)      =     335.99
Log pseudolikelihood = -2792.7433               Prob > chi2       =     0.0000
```

	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
graduate						
program# c.income						
0	.1824158	.0238431	7.65	0.000	.1356842	.2291475
1	.1865878	.0245008	7.62	0.000	.1385672	.2346084
roommate# program						
yes#0	.3099365	.0827593	3.75	0.000	.1477313	.4721418
yes#1	.2436647	.076438	3.19	0.001	.093849	.3934805
program# c.hsgpa						
0	1.083248	.6284794	1.72	0.085	-.1485491	2.315045
1	1.004868	.5841352	1.72	0.085	-.1400159	2.149752
program						
0	-4.201051	1.779367	-2.36	0.018	-7.688547	-.7135555
1	-3.590705	1.623489	-2.21	0.027	-6.772685	-.4087256
program						
campus						
yes	.7437785	.0734259	10.13	0.000	.5998663	.8876906
scholar						
yes	.8963839	.058676	15.28	0.000	.7813811	1.011387
income	-.0798981	.008895	-8.98	0.000	-.097332	-.0624643
_cons	-.3806292	.0859392	-4.43	0.000	-.5490669	-.2121916
hsgpa						
income	.0478622	.0016462	29.08	0.000	.0446358	.0510886
hscomp						
moderate	-.1351312	.0115348	-11.72	0.000	-.1577391	-.1125233
high	-.226768	.0194135	-11.68	0.000	-.2648178	-.1887181
_cons	2.794476	.0128195	217.99	0.000	2.769351	2.819602
var(e.hsgpa)	.0685876	.0019597			.0648522	.0725381
corr(e.pro~m, e.graduate)						
program						
0	.3223659	.1492073	2.16	0.031	.0079293	.5787898
1	.4280942	.1358716	3.15	0.002	.1307496	.6547793

corr(e.hsgpa, e.graduate) program						
0	.4241328	.1274031	3.33	0.001	.1471666	.6394236
1	.3792206	.1220983	3.11	0.002	.1190782	.5906426
corr(e.hsgpa, e.program)	-.0206714	.0264813	-0.78	0.435	-.0724717	.03124

The main equation output is slightly different from that in [ERM] **example 3a**. Because `program` was specified as a treatment, it was automatically interacted with each of the other covariates in the `graduate` equation.

We specified the `pocorrelation` suboption in `entreat()` so that we estimate separate correlation parameters for the two potential outcomes—for those who participated and those who did not. In the treated group, the correlation of the errors from the graduation equation and those from the program participation equation `corr(e.program,e.graduate)` is estimated to be 0.43 and is significantly different from zero. The researchers conclude that unobservable factors that increase the chance of participating in the program also increase the chance of graduating among the individuals that participate in the program.

Now, we use `estat teffects` to estimate the ATE of program participation on college graduation. We specified `vce(robust)` when we fit the model, so `estat teffects` reports standard errors and tests for the population ATE.

```
. estat teffects
Predictive margins                              Number of obs     =      2,500
```

	Unconditional Margin	Std. Err.	z	P>|z|	[95% Conf.	Interval]
ATE program (1 vs 0)	.1053155	.0492397	2.14	0.032	.0088075	.2018234

We estimate that the ATE is 0.11. In other words, the average probability of graduating increases by 0.11 when all students participate in the program versus when no students participate in the program.

We might be interested if those students who self-selected into the program increased their graduation probability by more than 0.11. We estimate the average treatment effect on the treated (ATET).

```
. estat teffects, atet
Predictive margins                              Number of obs     =      2,500
                                                Subpop. no. obs   =      1,352
```

	Unconditional Margin	Std. Err.	z	P>|z|	[95% Conf.	Interval]
ATET program (1 vs 0)	.1255127	.0497954	2.52	0.012	.0279154	.22311

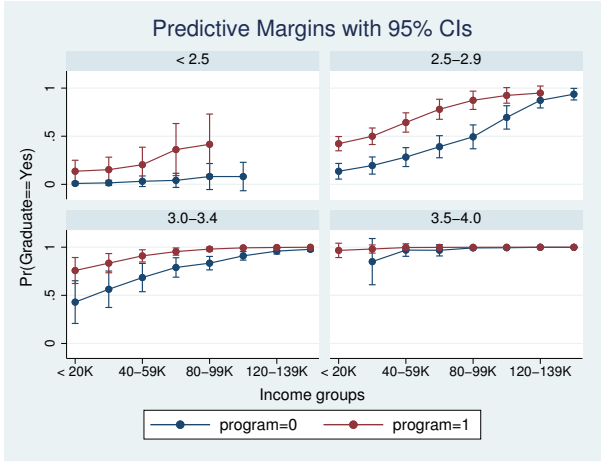
In this case, the program is only a little more effective on average for those who chose to participate than it would have been for everyone. The ATET is 0.13, only 0.02 higher than the ATE.

Those are the overall averages. Do graduation rates for participants and nonparticipants differ by high school GPA and parents' income? Our dataset has grouping variables, so we can let `margins` estimate graduation rates for subpopulations defined by all three covariates.

```
. margins, over(program incomegrp hsgpagrp) vce(unconditional)
```

The output is copious. You can type the command and see it if you like. The patterns are easier to see on a `marginsplot`.

```
. marginsplot, plot(program) xlabels(0 4 8 12)
```



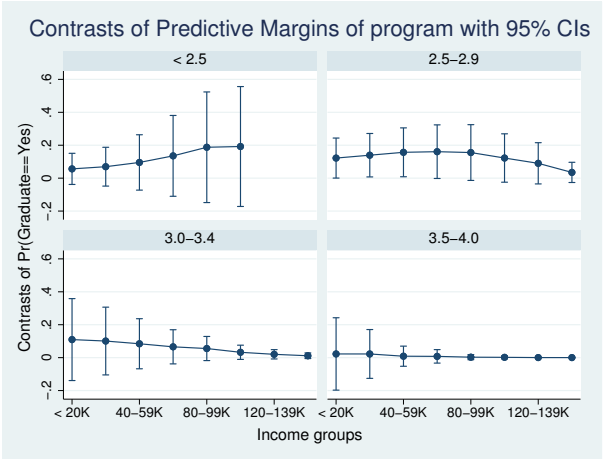
The red line shows expected graduation rates for those who participated in the program. The blue line shows rates for nonparticipants. Clearly, the differences between the groups in the program and those out of the program differ dramatically across GPA and family income. For GPAs at or above 3.5, the graduation rates are so high that there was no room for differences. For those with GPAs below 2.5, we see differences, with participation graduation rates being higher than nonparticipation, but lots of variation as income increases. For the other groups, the graduation rates are estimated to be substantially higher among those who participated.

We were careful not to call the comparisons above effects or attribute them directly to the program. They are indeed expected rates for the groups, but the students self-selected into program participation groups. If we want to compare graduation rates assuming all students do not participate and then assuming all students do participate, we need to instruct `margins` to `fix()` the values for program participation and also add the `r.` to `program`.

```
. margins r.program, over(incomegrp hsgpagrp) vce(unconditional)
> predict(fix(program)) contrast(nowald)
```
(*output omitted*)

The output is again long, so we leave you to see it for yourself. The graphs reveal the patterns across groups.

```
. marginsplot, by(hsgpagrp) xlabels(0 4 8 12)
```



These differences are close to what we would have seen had we differenced the red and blue lines of the first graph. In this graph, each point is an estimate of the average treatment effect for a subpopulation defined by a range of GPAs and a range of family income. We note that the confidence intervals, as represented by the capped lines, are fairly wide.

Also see

[ERM] **eprobit** — Extended probit regression

[ERM] **eprobit postestimation** — Postestimation tools for eprobit

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

[ERM] **intro 3** — Endogenous covariates features

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

[ERM] **intro 8** — Conceptual introduction via worked example