Title

stata.com

Example 4b — Probit regression with endogenous treatment and sample selection

Description Remarks and examples Also see

Description

Continuing from [ERM] **Example 4a**, we show you how to estimate and interpret the results of a model for a binary outcome when the model includes an endogenous treatment and the data are subject to endogenous sample selection.

Remarks and examples

stata.com

In [ERM] **Example 4a**, we ignored the possibility that regular exercise was an endogenous treatment. However, we suspect that unobserved factors that influence the choice to exercise may be correlated with the unobserved factors that affect the chance of having another heart attack.

We would like to know the average expected change in probability of having a subsequent heart attack for those who exercise. That is, we are interested in estimating the average treatment effect on the treated (ATET). We continue to include BMI and age in our outcome model, and to account for endogenous sample selection, we specify the same auxiliary model for selection we did in [ERM] **Example 4a**. We add a third equation to account for endogenous treatment assignment. Whether a man ever joined a gym is an instrumental variable predicting exercise that we do not expect to otherwise affect `attack`, so we include it in our model for regular exercise.

```
. eprobit attack age bmi, select(full = age bmi i.checkup)
> entreat(exercise = bmi i.gym) vce(robust)
  (iteration log omitted)
Extended probit regression                      Number of obs   =       625
                                                      Selected  =       458
                                                  Nonselected   =       167
                                                Wald chi2(6)    =    111.78
Log pseudolikelihood = -711.90507               Prob > chi2     =    0.0000
```

	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
attack						
exercise# c.age						
No	.2156634	.0550909	3.91	0.000	.1076872	.3236397
Yes	.221641	.0423742	5.23	0.000	.1385891	.3046928
exercise# c.bmi						
No	.1925833	.04278	4.50	0.000	.108736	.2764306
Yes	.2134441	.038381	5.56	0.000	.1382186	.2886696
exercise						
No	-16.07086	3.282712	-4.90	0.000	-22.50486	-9.636863
Yes	-17.84655	2.61864	-6.82	0.000	-22.97899	-12.71411
full						
age	-.1650386	.0321825	-5.13	0.000	-.228115	-.1019621
bmi	-.1143184	.0206726	-5.53	0.000	-.154836	-.0738008
checkup						
Yes	2.315167	.1639928	14.12	0.000	1.993747	2.636587
_cons	11.92957	1.898426	6.28	0.000	8.208727	15.65042
exercise						
bmi	-.1815549	.0211349	-8.59	0.000	-.2229786	-.1401313
gym						
Yes	1.517225	.1248316	12.15	0.000	1.27256	1.761891
_cons	3.941703	.5728064	6.88	0.000	2.819023	5.064383
corr(e.full, e.attack)	-.5338178	.1584217	-3.37	0.001	-.7737932	-.1598432
corr(e.exe~e, e.attack)	-.435728	.1467897	-2.97	0.003	-.676196	-.1113554
corr(e.exe~e, e.full)	.3212358	.0928654	3.46	0.001	.1293396	.4899396

The correlation between the errors that affect having a subsequent heart attack and the errors that affect staying in the study is estimated to be -0.53 and is significant. So we do have endogenous selection and conclude that unobservable factors that increase the chance of staying in the study also tend to decrease the chance of having a subsequent heart attack.

Increases in age and BMI increase the chance of having another heart attack. This is true both for those who exercise, coefficients marked `yes`, and for those who do not, coefficients marked `no`.

We use `estat teffects` to estimate the ATET of regular exercise on having a subsequent heart attack. We specified `vce(robust)` when we fit the model so that `estat teffects` will report unconditional standard errors for the population ATET rather than the sample ATET.

```
. estat teffects, atet
Predictive margins                                      Number of obs   = 625
                                                        Subpop. no. obs = 291
```

	Unconditional Margin	std. err.	z	P>|z|	[95% conf.	interval]
ATET						
exercise (Yes vs No)	-.2993399	.0773753	-3.87	0.000	-.4509927	-.1476871

The estimated ATET is −0.30. Thus, for those who exercise regularly, the average probability of having a subsequent heart attack is 0.30 lower than it would be if they did not exercise regularly.

Also see

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

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

[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