Title

stata.com

example 4a — Probit regression with endogenous sample selection

Description Remarks and examples Also see

Description

In this example, we show how to estimate and interpret the results of an extended regression model with a binary outcome and endogenous sample selection.

Remarks and examples

stata.com

We are interested in whether regular exercise and body mass index (BMI) influence the chance of having a subsequent heart attack. In our fictional study, we collected data on 625 men who had a heart attack when they were between the ages of 50 and 55. Some men withdrew from the study before it completed, and we believe their reasons for leaving are related to unobserved factors that also affect their chances of having a second heart attack. We did, however, observe all cases where a second heart attack was fatal.

To account for the endogenous sample selection, we specify an auxiliary model for selection using a covariate that belongs in the auxiliary model and is excluded from the main equation. We expect that the direct effect of whether a man had regular checkups before the study is negligible after we condition on other covariates.

The outcome of interest is whether the man had another heart attack within five years of his first heart attack (`attack`). We believe that the man's current age is also an important exogenous covariate along with BMI. We model the indicator for whether the man was observed for the full five years of the study (`full`) as a function of an indicator for having regular checkups along with the covariates from the main equation.

```
. use http://www.stata-press.com/data/r15/heartsm
(Heart attacks)
. eprobit attack age bmi i.exercise, select(full = age bmi i.checkup) vce(robust)
Iteration 0:   log pseudolikelihood = -409.23137
Iteration 1:   log pseudolikelihood = -408.78569
Iteration 2:   log pseudolikelihood = -408.78452
Iteration 3:   log pseudolikelihood = -408.78452
Extended probit regression                    Number of obs     =        625
                                                    Selected    =        458
                                                    Nonselected =        167
                                              Wald chi2(3)      =     142.85
Log pseudolikelihood = -408.78452             Prob > chi2       =     0.0000
```

	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
attack						
age	.2237091	.0351334	6.37	0.000	.1548489	.2925693
bmi	.1760896	.0298853	5.89	0.000	.1175155	.2346636
exercise						
yes	-1.438937	.1515198	-9.50	0.000	-1.735911	-1.141964
_cons	-15.78445	2.105945	-7.50	0.000	-19.91202	-11.65687
full						
age	-.1599347	.032953	-4.85	0.000	-.2245214	-.095348
bmi	-.1146582	.0208896	-5.49	0.000	-.1556011	-.0737152
checkup						
yes	2.306638	.1660248	13.89	0.000	1.981236	2.632041
_cons	11.66488	1.942686	6.00	0.000	7.857284	15.47247
corr(e.full, e.attack)	-.4537026	.1636665	-2.77	0.006	-.71301	-.0852183

We estimate that the correlation between the errors from the outcome equation and the errors from the selection equation is -0.45. This is significantly different from zero, so selection into the study is endogenous. Because the correlation is negative, we conclude that unobserved factors that increase the chance of staying in the study tend to occur with unobserved factors that decrease the chance of having a subsequent heart attack.

The results for the main outcome equation (`attack`) and auxiliary selection equation (`full`) are interpreted just as you would those from `heckprobit`. Which is to also say that the results for the main equation can be interpreted as you would those from a probit regression using `probit` on uncensored data. The goal of including a selection model is to estimate the parameters of the main equation as though there were no selection.

Age and BMI have increased the chances of having another heart attack, while regular exercise decreases the chances. However, the magnitude of the effect on the probability of another heart attack cannot be determined from the coefficient estimates themselves. We can use `margins` to examine the effect of different covariates on the probability of having a second heart attack. But first we want to investigate a possible further complication in our data: regular exercise may be an endogenous treatment. We explore this in [ERM] **example 4b**.

Also see

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

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

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

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