Title

Example 5 — Probit regression with endogenous ordinal treatment

Description Remarks and examples Also see

Description

We model a binary outcome that depends on an endogenous ordinal treatment by using `eprobit` with the `entreat()` option.

Remarks and examples

stata.com

We are interested in estimating the average treatment effects (ATEs) of different levels of exercise intensity on 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. The outcome of interest is whether the man had another heart attack within five years of his first heart attack (`attack`). We believe that body mass index (BMI) and age are important covariates.

The `exintensity` variable records the intensity of exercise using the scale of `0` (no exercise), `1` (moderate), and `2` (heavy). We suspect that unobserved factors that influence the choice to exercise at a certain intensity level also affect the chance of having another heart attack, so we specify `exintensity` as an endogenous treatment. Whether an individual ever joined a gym is included as an instrumental covariate in the treatment model that we specify in `entreat()`.

```
. use https://www.stata-press.com/data/r17/heartsm
(Heart attacks)

. eprobit attack age bmi, entreat(exintensity = bmi i.gym) vce(robust)
  (iteration log omitted)
Extended probit regression                          Number of obs =    625
                                                    Wald chi2(9)  = 152.33
Log pseudolikelihood = -728.6686                    Prob > chi2   = 0.0000

─────────────┬────────────────────────────────────────────────────────────────
             │               Robust
             │ Coefficient  std. err.      z    P>|z|     [95% conf. interval]
─────────────┼────────────────────────────────────────────────────────────────
attack       │
 exintensity#│
       c.age │
        none │   .2118759   .0514612     4.12   0.000     .1110138     .312738
    moderate │   .2338466   .0425341     5.50   0.000     .1504813    .3172119
       heavy │   .2346887   .0805152     2.91   0.004     .0768818    .3924957
             │
 exintensity#│
       c.bmi │
        none │   .1948171   .0386314     5.04   0.000      .119101    .2705332
    moderate │   .2062276   .0405785     5.08   0.000     .1266952    .2857599
       heavy │   .2155222   .0765592     2.82   0.005     .0654689    .3655755
             │
 exintensity │
        none │  -15.90911   3.043587    -5.23   0.000    -21.87444   -9.943793
    moderate │   -18.2922   2.499325    -7.32   0.000    -23.19079   -13.39362
       heavy │  -18.61821   5.395246    -3.45   0.001     -29.1927   -8.043721
─────────────┼────────────────────────────────────────────────────────────────
exintensity  │
         bmi │  -.1720462   .0204172    -8.43   0.000    -.2120632   -.1320292
             │
         gym │
         yes │   1.518834   .1192361    12.74   0.000     1.285136    1.752532
─────────────┼────────────────────────────────────────────────────────────────
/exintensity │
        cut1 │  -3.677846   .5537938                     -4.763262    -2.59243
        cut2 │  -2.386538   .5372719                     -3.439572   -1.333505
─────────────┼────────────────────────────────────────────────────────────────
corr(e.exi~y,│
    e.attack)│  -.4722803   .1091789    -4.33   0.000    -.6575129   -.2332112
─────────────┴────────────────────────────────────────────────────────────────
```

The estimated correlation between the errors in the main outcome and auxiliary treatment equations is -0.47. This is significantly different from zero, so we confirm that the choice of exercise intensity level is endogenous. Because it is negative, we conclude that unobservable factors that increase the intensity of exercising tend to decrease the chance of having a subsequent heart attack. The cutpoints for the ordered probit model for the endogenous treatment are shown just beneath the treatment model.

The coefficients for `exintensity` in the main equation indicate that both moderate and heavy exercise have a negative effect because they are smaller, more negative, than the coefficient for no exercise. BMI has a positive effect on the chance of having another heart attack, regardless of exercise level. In fact, the values of the three coefficients for `bmi` are so close that we might not need separate parameters for the three levels of exercise. The same could be said of the three coefficients on `age`.

The coefficients for the intercepts of heavy and moderate exercise are close in magnitude. To test whether these two coefficients are equal, we can use test.

```
. test 1.exintensity == 2.exintensity
 ( 1)  [attack]1.exintensity - [attack]2.exintensity = 0
           chi2(  1) =    0.00
         Prob > chi2 =    0.9557
```

We cannot reject that the coefficients are equal.

We also have separate coefficients on age and bmi for heavy and moderate exercise. To jointly test the equality of each coefficient associated with heavy exercise with the corresponding coefficient associated with moderate exercise, we type

```
. test (1.exintensity == 2.exintensity)
>      (1.exintensity#c.bmi == 2.exintensity#c.bmi)
>      (1.exintensity#c.age == 2.exintensity#c.age)
 ( 1)  [attack]1.exintensity - [attack]2.exintensity = 0
 ( 2)  [attack]1.exintensity#c.bmi - [attack]2.exintensity#c.bmi = 0
 ( 3)  [attack]1.exintensity#c.age - [attack]2.exintensity#c.age = 0
           chi2(  3) =    0.04
         Prob > chi2 =    0.9983
```

We do not have any evidence that heavy and moderate exercise have a different effect on the probability of a second heart attack.

That was some pretty tricky coefficient referencing in our test command. We suggest you type

```
. eprobit, coeflegend
```

to see how to reference coefficients in test, nlcom, and other postestimation commands.

What if every man in the population did not exercise? What if they all exercised moderately? What if they all exercised heavily? estat teffects can estimate the average probability of a second heart attack over the five years for each of those counterfactuals.

```
. estat teffects, pomean
Predictive margins                                      Number of obs = 625
```

	Margin	Unconditional std. err.	z	P>|z|	[95% conf.	interval]
POmean						
exintensity						
none	.7918941	.0329342	24.04	0.000	.7273443	.856444
moderate	.5419335	.0326336	16.61	0.000	.4779728	.6058942
heavy	.5336232	.0767752	6.95	0.000	.3831466	.6840998

When no one in the population exercises, we estimate that 79% will have subsequent heart attacks. We are pretty confident in that number: the 95% confidence interval begins at 73% and ends at 86%. It does not matter much whether every man exercises moderately or heavily. Either intensity drops the expected rate of subsequent heart attacks to about 54%. These are the average potential-outcome means (POMs) under the three exercise-intensity regimes.

The difference between these POMs gives us estimates of the average treatment effects (ATEs) in the population. `estat teffects` will estimate those too.

```
. estat teffects
Predictive margins                                    Number of obs = 625
```

	Margin	Unconditional std. err.	z	P>|z|	[95% conf. interval]	
ATE						
exintensity						
(moderate vs none)	-.2499606	.0507776	-4.92	0.000	-.349483	-.1504383
(heavy vs none)	-.2582709	.0965797	-2.67	0.007	-.4475637	-.0689781

We estimate that the ATE for heavy intensity compared with no exercise is −0.26. So the average probability of a subsequent heart attack is 26 percentage points lower when all men in the population exercise with heavy intensity versus when none of them exercise at all. The estimated ATE for moderate intensity versus none is −0.25. We again see no substantive difference between moderate and heavy exercise.

We used `vce(robust)` at estimation so that `estat teffects` would report standard errors that account for sampling variability in our covariates and are therefore valid for inference about the POMs, ATEs, and ATETs in the population from which our sample was drawn.

We have established that men who choose to exercise have unobserved attributes that tend to decrease their chance of another heart attack beyond the direct effect of exercising and beyond the effect of the other covariates. We can include the effect of these attributes for men who exercise by estimating the average treatment effect on the treated (ATET).

```
. estat teffects, atet
(subpopulation of first non-control treatment level assumed)
Predictive margins                                  Number of obs    = 625
                                                    Subpop. no. obs = 201
```

	Margin	Unconditional std. err.	z	P>|z|	[95% conf. interval]	
ATET						
exintensity						
(moderate vs none)	-.2992132	.0531572	-5.63	0.000	-.4033994	-.195027
(heavy vs none)	-.309572	.100091	-3.09	0.002	-.5057467	-.1133973

The ATETs are both about 0.30, making them about 5 percentage points higher than the ATEs. We cannot, however, directly attribute that difference to the unobserved attributes. The ATETs are also averaged over subsamples and are therefore affected by any differences in the distribution of `age` or `bmi` in treated subsamples. The effect of those distributions could be either positive or negative.

With some care, we can extract just the effect of the unobserved attributes. It is a little tricky, both conceptually and syntactically. So continue reading only if you are truly interested.

Let's consider only the moderate exercisers. When we type

```
. margins r(0 1).exintensity, subpop(if exintensity == 1) vce(unconditional)
```

`margins` will produce the average difference for `exintensity` levels 0 and 1 (none and moderate). `subpop(if exintensity == 1)` restricts the average to men who exercised moderately.

`margins` would use the unobserved attributes associated with moderate exercise for both of the counterfactuals it requires to compute the contrast. Which is to say, it would use the true value of exercise intensity in the subpopulation we are averaging over. If you were to guess that this difference will be the ATET, you would be correct. For each man who chose moderate exercise, the ATET computation compares the man's expected probability of another attack using all the information on the man with that same man's expected probability if he instead did not choose to exercise. When we say "same man", we mean that he retains his original unobserved attributes when evaluating the counterfactual that he does not exercise. The ATET is then the average of that comparison over all those who exercise moderately.

We may also want to use `margins` to test whether the ATE for heavy exercise and the ATE for moderate exercise are equal. We specify two `predict()` options. On the first, we request treatment effects (`te`) for heavy exercisers (`tlevel(heavy)`). On the second, we request the treatment effects for moderate exercisers (`tlevel(moderate)`). We add `contrast(predict(r))` to request the difference between the predictions (their contrast). Finally, we use `vce(unconditional)` to request standard errors that account for sampling variability in the covariates and thus allow us to make inferences about the population.

```
. margins, predict(te tlevel(heavy)) predict(te tlevel(moderate))
> contrast(predict(r)) vce(unconditional)

Contrasts of predictive margins                          Number of obs = 625

1._predict: treatment effect Pr(attack==yes), exintensity: heavy vs. none,
            predict(te tlevel(heavy))
2._predict: treatment effect Pr(attack==yes), exintensity: moderate vs. none,
            predict(te tlevel(moderate))
```

	df	chi2	P>chi2
_predict	1	0.01	0.9085

	Contrast	Unconditional std. err.	[95% conf.	interval]
_predict (2 vs 1)	.0083103	.0722814	-.1333587	.1499793

We cannot reject that the ATE for heavy exercise is equal to the ATE for moderate exercise. This result agrees with what we saw when we tested the coefficients for heavy and moderate exercise.

As we have seen repeatedly in the examples in the manual, most of the interesting questions are answered by `estat teffects` and `margins` and not by the parameter estimates themselves. This is particularly true of models estimated using `eprobit` and `eoprobit`.

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 5** — Treatment assignment features

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