Title

Example 9 — Ordered probit regression with endogenous treatment and random effects

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 an ordinal outcome, an endogenous treatment, and random effects.

Remarks and examples

In [ERM] **Example 6a**, we examined fictional data on the health scores of women between the ages of 25 and 30. Each woman was observed at one time point. Our outcome was an ordinal health status ranging from 1 (poor) to 5 (excellent). We estimated the average treatment effect of having health insurance on the probabilities of having each health status.

Now suppose that we conduct a fictional study where we have collected data on 1,800 women between the ages of 25 and 30 annually from 2010 to 2013. We have measured the women's health status in each year. We want to estimate the average treatment effect (ATE) of having insurance on the probability of each of the five statuses. We suspect that our model needs to account for health insurance being an endogenous treatment. We also believe that unobserved characteristics of the individual might affect both health status and whether the woman has insurance, so we include random effects in both equations.

In addition to the insurance indicator, we include an indicator for whether the woman exercises regularly and the number of years of schooling she completed (`grade`) as exogenous covariates in the model for health status. For our treatment model, we use `grade` and an indicator for whether the woman is currently working or attending school (`workschool`), which is excluded from the outcome model.

Before we can fit our random-effects model, we need to specify the panel structure of the data using `xtset`. Our panel variable is `personid`, the identification code for the individual. The time variable is `year`, and it ranges from 2010 to 2013.

```
. use https://www.stata-press.com/data/r18/womenhlthre
(Women's health status panel)

. xtset personid year

Panel variable: personid (strongly balanced)
 Time variable: year, 2010 to 2013
         Delta: 1 unit
```

With the data xtset, we can estimate the parameters of the model.

```
. xteoprobit health exercise grade,
> entreat(insured = grade i.workschool) vce(robust)

(setting technique to bhhh)
Iteration 0:  Log pseudolikelihood = -12272.723
Iteration 1:  Log pseudolikelihood = -12256.949
Iteration 2:  Log pseudolikelihood = -12256.539
Iteration 3:  Log pseudolikelihood = -12256.478
Iteration 4:  Log pseudolikelihood = -12256.468
Iteration 5:  Log pseudolikelihood = -12256.466
Iteration 6:  Log pseudolikelihood = -12256.465
Iteration 7:  Log pseudolikelihood = -12256.465
Iteration 8:  Log pseudolikelihood = -12256.465
Iteration 9:  Log pseudolikelihood = -12256.465

Extended ordered probit regression                     Number of obs     =  7,200
Group variable: personid                               Number of groups =  1,800

                                                       Obs per group:
                                                                    min =      4
                                                                    avg =    4.0
                                                                    max =      4

Integration method: mvaghermite                        Integration pts. =      7

                                                       Wald chi2(4)     = 404.14
Log pseudolikelihood = -12256.465                      Prob > chi2      = 0.0000

                           (Std. err. adjusted for 1,800 clusters in personid)
------------------------------------------------------------------------------
             |               Robust
             | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
health       |
    insured# |
  c.exercise |
          No |    .356811   .0521592     6.84   0.000     .2545809     .459041
         Yes |   .4929456   .0360086    13.69   0.000     .4223701    .5635211
             |
    insured# |
     c.grade |
          No |   .0970783   .0198281     4.90   0.000     .0582159    .1359407
         Yes |    .130956   .0114576    11.43   0.000     .1084996    .1534124
-------------+----------------------------------------------------------------
insured      |
       grade |     .29484   .0100943    29.21   0.000     .2750555    .3146245
             |
  workschool |
         Yes |   .5841205   .0638709     9.15   0.000     .4589358    .7093052
       _cons |  -3.502613   .1377291   -25.43   0.000    -3.772557   -3.232669
```

/health						
insured#						
c.cut1						
No	.4910109	.1864684			.1255395	.8564823
Yes	-.2650117	.2049759			-.6667571	.1367337
insured#						
c.cut2						
No	1.388273	.1810191			1.033482	1.743064
Yes	.5527565	.1908832			.1786323	.9268806
insured#						
c.cut3						
No	2.192588	.1794012			1.840968	2.544207
Yes	1.381288	.1806265			1.027267	1.73531
insured#						
c.cut4						
No	2.994727	.1873594			2.627509	3.361945
Yes	2.297709	.1731544			1.958333	2.637086
corr(e.ins~d, e.health)	.3783935	.0770755	4.91	0.000	.2183033	.5186513
var(hea~h[per~d])	.379062	.0284741			.3271676	.4391877
var(ins~d[per~d])	.2436723	.0354709			.1831887	.3241259
corr(ins~d[per~d], hea~h[per~d])	.3251756	.0721159	4.51	0.000	.1774673	.458556

The estimated correlation between the observation-level errors is 0.38. The estimated correlation between the individual-level random effects affecting health status and the individual-level random effects affecting insurance status is 0.33. Both are significantly different from zero. We conclude that insurance status is endogenous and that the unobserved person-specific factors that increase the chance of having health insurance also tend to increase the chance of being in a high health status. Additionally, the unobserved observation-level (time-varying) factors that increase the chance of having health insurance also tend to increase the chance of being in a high health status.

We see estimates of both the coefficients and the cutpoints for two equations, one for insured women (`yes`) and one for uninsured women (`no`). For both insured and uninsured, exercise and education have positive effects on health status.

We can use `estat teffects` to estimate the ATE of insurance on the probabilities of each health category.

```
. estat teffects
Predictive margins                                   Number of obs = 7,200
                            (Std. err. adjusted for 1,800 clusters in personid)
```

	Margin	Unconditional std. err.	z	P>|z|	[95% conf.	interval]
ATE_Pr1 insured (Yes vs No)	-.1761541	.0279001	-6.31	0.000	-.2308372	-.1214709
ATE_Pr2 insured (Yes vs No)	-.1731894	.0227877	-7.60	0.000	-.2178525	-.1285264
ATE_Pr3 insured (Yes vs No)	-.0607013	.0127344	-4.77	0.000	-.0856602	-.0357424
ATE_Pr4 insured (Yes vs No)	.1145319	.0214062	5.35	0.000	.0725765	.1564874
ATE_Pr5 insured (Yes vs No)	.2955128	.0345022	8.57	0.000	.2278897	.3631359

We see that the treatment effect is negative on the probability of being in poor health. The treatment effect becomes more positive for each successive health status. 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 0.30 greater when all women have health insurance versus when no women have health insurance.

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 5** — Treatment assignment features

[ERM] **Intro 6** — Panel data and grouped data model features

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