Title

Example 3a — Probit regression with continuous endogenous covariate

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 continuous endogenous covariate.

Remarks and examples

In [ERM] **Example 1a** through [ERM] **Example 1c**, we showed how researchers at the fictional State University might approach an investigation of the relationship between the high school grade point average (GPA) of the students the university admits and their final college GPA. Suppose instead that they would like to know how the probability of college graduation is related to high school GPA. They again suspect that high school GPA is endogenous in a model of the probability of college graduation.

Their model for graduation includes parental income in $10,000s and whether the student had a roommate who also went to State U. The State U researchers expect that the effect of high school competitiveness on the probability of graduating from college is negligible once the other covariates are controlled for. So they use the ranking of the high school (`hscomp`) as the instrumental variable for high school GPA. They also include parental income in the auxiliary model for high school GPA.

We want to make inferences about how our covariates affect graduation rates in the population, not just in our sample. We add `vce(robust)` so that subsequent calls to `estat teffects` and `margins` will be able to consider our sample as a draw from the population.

```
. use https://www.stata-press.com/data/r18/class10
(Class of 2010 profile)

. eprobit graduate income i.roommate, endogenous(hsgpa = income i.hscomp)
> vce(robust)

Iteration 0:  Log pseudolikelihood = -1418.5008
Iteration 1:  Log pseudolikelihood = -1418.4414
Iteration 2:  Log pseudolikelihood = -1418.4414

Extended probit regression                             Number of obs =  2,500
                                                       Wald chi2(3)  = 326.79
Log pseudolikelihood = -1418.4414                      Prob > chi2   = 0.0000

------------------------------------------------------------------------------
              |               Robust
              | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
--------------+---------------------------------------------------------------
graduate      |
       income |   .1597677   .0158826    10.06   0.000     .1286384    .1908969
              |
     roommate |
          Yes |   .2636312   .0563563     4.68   0.000     .1531748    .3740876
        hsgpa |    1.01877   .4324788     2.36   0.018     .1711273    1.866413
        _cons |  -3.647166   1.204728    -3.03   0.002    -6.008389   -1.285943
--------------+---------------------------------------------------------------
hsgpa         |
       income |    .047859   .0016461    29.07   0.000     .0446327    .0510853
              |
       hscomp |
     Moderate |   -.135734   .0114717   -11.83   0.000     -.158218   -.1132499
        High  |   -.225314   .0195055   -11.55   0.000    -.2635441   -.1870838
              |
        _cons |   2.794711   .0127943   218.43   0.000     2.769634    2.819787
--------------+---------------------------------------------------------------
 var(e.hsgpa) |   .0685893   .0019597                       .064854    .0725398
--------------+---------------------------------------------------------------
corr(e.hsgpa,|
  e.graduate) |   .3687006   .0919048     4.01   0.000     .1765785    .5337596
------------------------------------------------------------------------------
```

The estimate of the correlation between the errors of our two equations is 0.37 and is significantly different from zero, so we have endogeneity. Because the correlation is positive, we conclude that the unobservable factors that increase high school GPA also increase the probability of graduation.

The results for the main equation are interpreted as you would those from probit. We can obtain directions but not effect sizes from the coefficients in the main equation. For example, we see that family income and high school GPA are positively associated with the probability that a student graduates.

Let's ask something more interesting. What if we could increase each student's high school GPA by one point, moving a 2.0 to a 3.0, a 2.5 to a 3.5, and so on? We obviously cannot increase anyone's GPA by one point if he or she is already above a 3.0; so we restrict our population of interest to students with a GPA at or below 3.0. margins will give us the population-average expected graduation rate given each student's current GPA if we specify at(hsgpa=generate(hsgpa)). It will also give us the population-average expected graduation rate with an additional point in each student's GPA if we specify at(hsgpa=generate(hsgpa+1)). We want to hold each student's unobservable characteristics to be those that are implied by his or her current data by using the default average structural function prediction.

```
. margins, at(hsgpa=generate(hsgpa)) at(hsgpa=generate(hsgpa+1))
> subpop(if hsgpa <= 3) vce(unconditional)

Predictive margins                                     Number of obs   = 2,500
                                                       Subpop. no. obs = 1,430

Expression: Average structural function probability, predict()
1._at: hsgpa =   hsgpa
2._at: hsgpa = hsgpa+1

------------------------------------------------------------------------------
             |            Unconditional
             |     Margin   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   .4315243   .0125571    34.37   0.000     .4069129    .4561357
          2  |   .7737483   .1057771     7.31   0.000     .5664289    .9810677
------------------------------------------------------------------------------
```

For students with a high school GPA at or below 3.0, the expected graduation rate is 43%. If those same students are given an additional point in their GPA, the graduation rate rises to 77%.

By adding `contrast(at(r))` to our `margins` command, we can difference those two counterfactuals and estimate the average effect of giving an additional point of GPA. We also added `effects` to add test statistics and `nowald` to clean up the output.

```
. margins, at(hsgpa=generate(hsgpa)) at(hsgpa=generate(hsgpa+1))
> subpop(if hsgpa <= 3) contrast(at(r) nowald effects) vce(unconditional)

Contrasts of predictive margins                        Number of obs   = 2,500
                                                       Subpop. no. obs = 1,430

Expression: Average structural function probability, predict()
1._at: hsgpa =   hsgpa
2._at: hsgpa = hsgpa+1

------------------------------------------------------------------------------
             |            Unconditional
             |   Contrast   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         _at |
   (2 vs 1)  |    .342224   .1070061     3.20   0.001     .1324959    .5519521
------------------------------------------------------------------------------
```

Giving students an additional point in their GPA increased graduation rates by just over 34 percentage points, with a 95% confidence interval from 13 to 55 percentage points.

Does this effect differ across any of our other covariates? Our dataset has a grouping variable for family income `incomegrp`, so let's estimate the effect within each income grouping. We just add `over(incomegrp)` to our prior `margins` command.

```
. margins, at(hsgpa=generate(hsgpa)) at(hsgpa=generate(hsgpa+1))
> subpop(if hsgpa <= 3) contrast(at(r) nowald effects) noatlegend
> vce(unconditional) over(incomegrp)

Contrasts of predictive margins                      Number of obs    = 2,500
                                                     Subpop. no. obs = 1,430

Expression: Average structural function probability, predict()
Over:       incomegrp

------------------------------------------------------------------------------
             |               Unconditional
             |   Contrast    std. err.       z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
        _at@ |
   incomegrp |
    (2 vs 1) |
       < 20K |   .3690987    .1367297     2.70   0.007     .1011134    .6370839
    (2 vs 1) |
     20-39K  |   .3698609    .1254644     2.95   0.003     .1239552    .6157667
    (2 vs 1) |
     40-59K  |   .3516159    .1026905     3.42   0.001     .1503463    .5528855
    (2 vs 1) |
     60-79K  |   .3094611    .0798654     3.87   0.000     .1529277    .4659944
    (2 vs 1) |
     80-99K  |    .255203    .0572386     4.46   0.000     .1430174    .3673887
    (2 vs 1) |
    100-119K |   .1829494     .038826     4.71   0.000     .1068519    .2590469
    (2 vs 1) |
    120-139K |   .1238027    .0344788     3.59   0.000     .0562255      .19138
    (2 vs 1) |
     140K up |   .0485429    .0139112     3.49   0.000     .0212775    .0758083
------------------------------------------------------------------------------
```

The effect is largest for the low-income groups and declines as income goes up. It becomes almost negligible for students from households whose income is above $140,000.

We can see this relationship more clearly if we graph the results.

```
. marginsplot
```

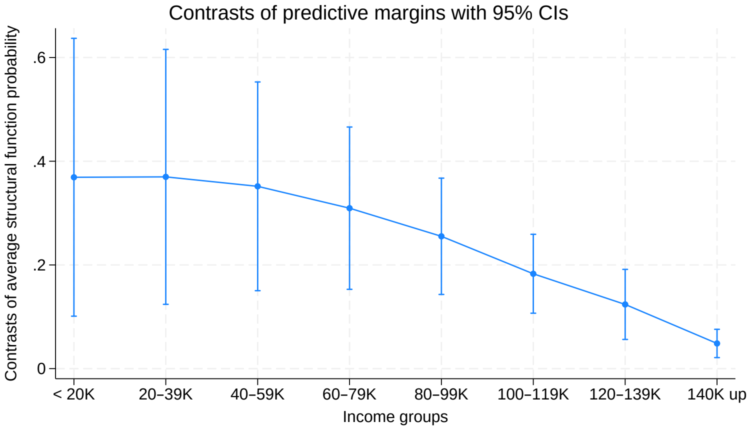

Figure 1.

Our point estimates of the effect on the probability of graduating are near 0.4 for the lowest-income groups and fall below 0.2 for incomes over $100,000.

So we can examine subpopulation averages and effects and make inferences about their values.

Let's see whether we can observe the effects of the unobservables that are affecting both graduation probability and high school GPA. We will restrict our attention to those with very low and very high incomes. We do not want any confounding from the individual's roommate status, so we will also include only those who have a roommate. Before we start, we modify our data to simplify our analysis.

```
. generate smpl = roommate==1 & (income < 3 | income > 10)
. generate byte hlincome = 1 if income < 3
(1,870 missing values generated)
. replace hlincome = 2 if income > 10
(216 real changes made)
. label define hiloinc 1 "Income < $30,000" 2 "Income > $100,000"
. label values hlincome hiloinc
```

We would like to know the expected graduation rates for those with high and low income over the groupings of high school GPA.

```
. margins, subpop(smpl) over(hsgpagrp hlincome) vce(unconditional)
  (output omitted)
. marginsplot
  (output omitted)
```

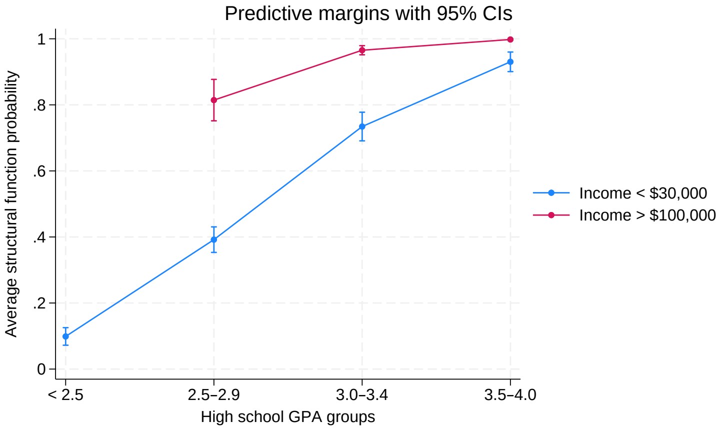

Figure 2.

Clearly, those with high incomes have much higher expected graduation rates.

What if we could level the playing field and give everyone the same family income level? We will give everyone $100,000. That is a substantial increase for those in the low-income group and somewhat of a reduction for most in the high-income group. We form these counterfactuals by adding `at(income=10)` to our `margins` command and then graph them.

```
. margins, subpop(smpl) over(hsgpagrp hlincome) at(income=10) vce(unconditional)
  (output omitted)
. marginsplot
  (output omitted)
```

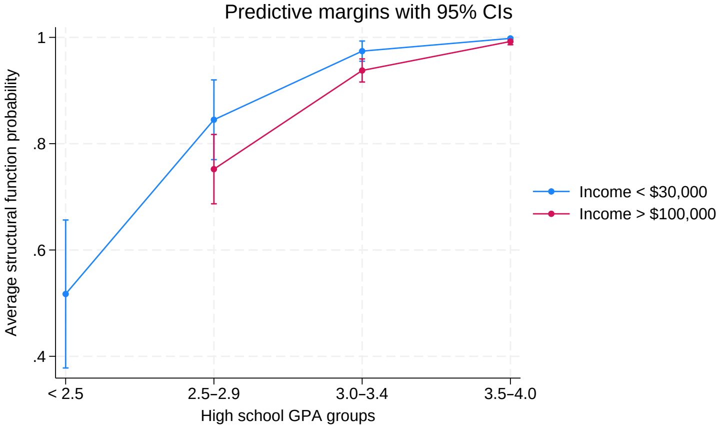

Figure 3.

Now, we have given everyone $100,000 of family income, but each person has retained his or her unobservable characteristics. The red line for those who originally had incomes over $100,000 is still pretty much where it was before. The line for those with incomes less than $30,000 is far higher than it was. It is now above the line for the high income students. For any level of GPA, the low-income group now has a higher graduation rate than the high-income group. We have locked up all the regressors in the main model by either putting them in groups or assigning them a counterfactual. So any differences that we see must be attributable to the individual's unobserved characteristics. If these were real data, the results would not be surprising. One way to view it is that any level of high school GPA is more difficult to obtain for individuals in a low-income family. That would mean that their unobservables will also tend to increase their graduation rate.

The results validate our previous conclusions. The important message is that we can analyze fully conditional counterfactuals and make complex inferences. These inferences account for not only observable characteristics but also unobservable traits and thus have a structural interpretation.

Also see

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

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

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

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

For suggested citations, see the FAQ on citing Stata documentation.