Title

stata.com

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

stata.com

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 grade point average (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 http://www.stata-press.com/data/r15/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
```

	Coef.	Robust 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 their current data, so we also create a variable holding the true values of `hsgpa` and specify `predict(base(hsgpa=hsgpaT))`.

```
. generate hsgpaT = hsgpa                    // True value of GPA for margins
. margins, at(hsgpa=generate(hsgpa)) at(hsgpa=generate(hsgpa+1))
> predict(base(hsgpa=hsgpaT)) subpop(if hsgpa <= 3) vce(unconditional)
Predictive margins                                Number of obs     =      2,500
                                                  Subpop. no. obs   =      1,430

Expression   : Pr(graduate==yes), predict(base(hsgpa=hsgpaT))
1._at        : hsgpa           = hsgpa
2._at        : hsgpa           = hsgpa+1
```

	Unconditional Margin	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_at						
1	.4315243	.0214675	20.10	0.000	.3894487	.4735998
2	.7737483	.0953191	8.12	0.000	.5869264	.9605702

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) predict(base(hsgpa=hsgpaT))
> contrast(at(r) nowald effects) vce(unconditional)
Contrasts of predictive margins

Expression   : Pr(graduate==yes), predict(base(hsgpa=hsgpaT))
1._at        : hsgpa           = hsgpa
2._at        : hsgpa           = hsgpa+1
```

	Unconditional Contrast	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_at						
(2 vs 1)	.342224	.113214	3.02	0.003	.1203287	.5641194

Giving students an additional point in their GPA increased graduation rates by just over 34%, with a 95% confidence interval from 12% to 56%.

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) predict(base(hsgpa=hsgpaT))
> contrast(at(r) nowald effects) noatlegend vce(unconditional) over(incomegrp)

Contrasts of predictive margins

Expression   : Pr(graduate==yes), predict(base(hsgpa=hsgpaT))
over         : incomegrp
```

	Unconditional					
	Contrast	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_at@ incomegrp						
(2 vs 1) < 20K	.3690987	.1359989	2.71	0.007	.1025457	.6356516
(2 vs 1) 20-39K	.3698609	.1273853	2.90	0.004	.1201903	.6195316
(2 vs 1) 40-59K	.3516159	.1103376	3.19	0.001	.1353581	.5678737
(2 vs 1) 60-79K	.3094611	.0927492	3.34	0.001	.1276761	.4912461
(2 vs 1) 80-99K	.255203	.0748521	3.41	0.001	.1084956	.4019105
(2 vs 1) 100-119K	.1829494	.0552683	3.31	0.001	.0746256	.2912732
(2 vs 1) 120-139K	.1238028	.0459416	2.69	0.007	.0337588	.2138467
(2 vs 1) 140K up	.0485429	.0207233	2.34	0.019	.0079259	.0891598

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
```

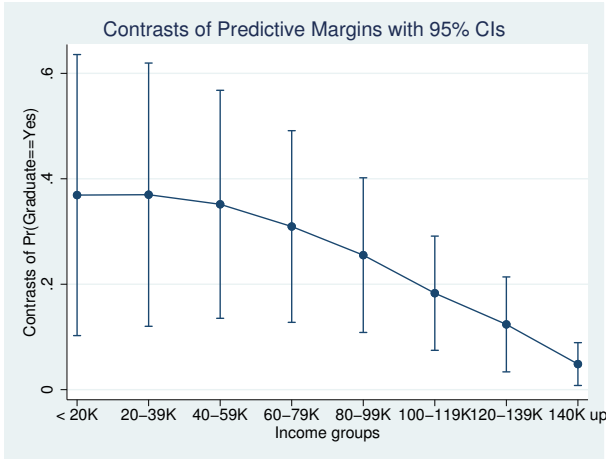

Figure 1.

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.

We can also examine averages and effects at specified values of the covariates in our model. Let's consider students who do not have roommates and evaluate them at 5 levels of high school GPA (2.0, 2.5, 3.0, 3.5, and 4.0) and at two levels of income ($30,000 and $110,000).

```
. margins, at(roommate=0 hsgpa=(2 2.5 3 3.5 4) income=(3 11)) noatlegend
Predictive margins                              Number of obs     =      2,500
Model VCE    : Robust
Expression   : Pr(graduate==yes), predict()
```

	Margin	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
_at						
1	.0068488	.0076828	0.89	0.373	-.0082092	.0219068
2	.1215437	.0353464	3.44	0.001	.052266	.1908213
3	.5517785	.0320675	17.21	0.000	.4889272	.6146297
4	.9232607	.043002	21.47	0.000	.8389784	1.007543
5	.9967789	.0051452	193.73	0.000	.9866944	1.006863
6	.0470211	.0496759	0.95	0.344	-.0503419	.144384
7	.3531365	.1042001	3.39	0.001	.1489081	.5573649
8	.8213242	.023535	34.90	0.000	.7751964	.867452
9	.9867056	.0071801	137.42	0.000	.9726328	1.000778
10	.9997797	.0003587	2787.22	0.000	.9990767	1.000483

Looking at all combinations of GPA and income, we see that graduation probabilities range from 0.0068 to 0.9998 for these values of the covariates.

We have suppressed the long legend that explains the `_at` levels in the table, so let's explain the lines. All results are for students without roommates. Lines 1–5 are for students with family incomes of $30,000 with the first line representing a GPA of 2, the second a GPA of 2.5, and so on. Lines 6–10

represent the same levels of GPA for students with a family income of $110,000. Because our model has only three covariates in the main equation and because we have specified values for each of the covariates, these can be considered fully conditional estimates. Even so, they are averages in the sense that they are expected values. Each probability represents what we would expect if hundreds of students were sampled who had the same values of the covariates as those on the corresponding line.

The patterns in these results are easier to see on a graph.

```
. marginsplot
```

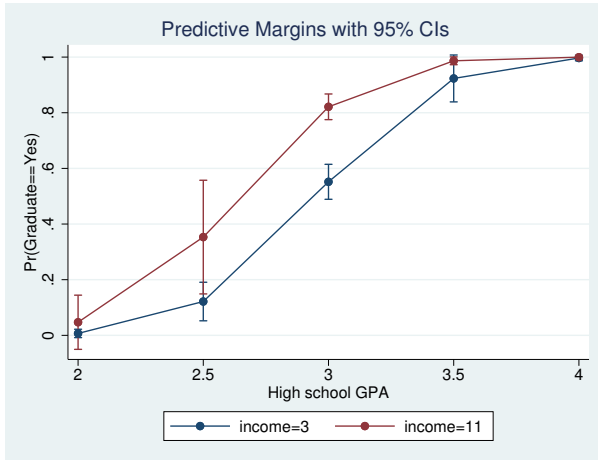

Figure 2.

Students with a GPA of 2.0 have nearly no chance of graduating, regardless of income. For those with a GPA between 2.5 and 3.0, the graduation rates differ sharply depending on income level. At GPAs of 3.5 and above, graduation rates are so high that there is again little difference due to income. These results aren't surprising; it's easier to struggle through school when you do not also have to worry over money issues.

What if we could grant the lower-income students a higher income? We would want to hold their unobservables at their initial level while moving them to the higher income. Perhaps they are adopted. Perhaps we are using this increase in income as a proxy for providing financial aid to lower-income students. Regardless, we use `predict(base(income=3))` to hold their unobservable characteristics to their initial level as we move income from $30,000 to $110,000.

```
. margins, at(roommate=0 hsgpa=(2 2.5 3 3.5 4) income=(3 11))
> noatlegend predict(base(income=3))
```

(*output omitted*)

We dispense with showing you the output and go straight to the graph. You can run the `margins` command if you wish.

```
. marginsplot
```

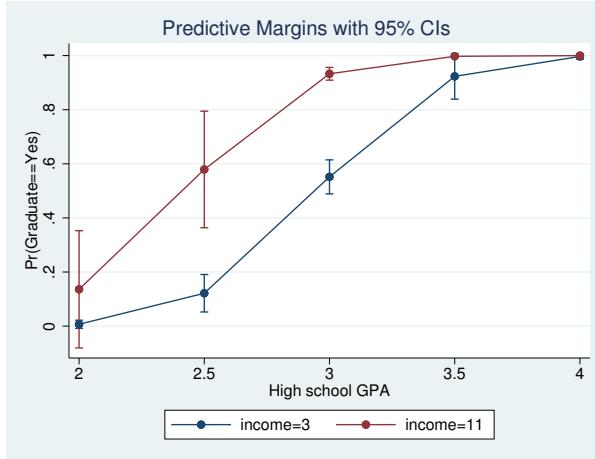

Figure 3.

The separation between graduation probabilities for incomes of $30,000 and $110,000 is even larger for those who obtain their high school GPA while in a family with $30,000 income and are then moved to $110,000.

Let's explore that a bit more, not because made-up data are interesting but because we have yet more tools to show you. `margins` will compute contrasts (differences) between our `at()` groupings but is an all-or-none proposition. It is either all levels or all differences. We want to see the differences in the lines we have been drawing while keeping our levels of GPA. We are going to estimate and graph the differences between the lines on the graph we just drew and also on the graph we drew before that. So we are going to compare the effects for those born with higher incomes and the effects with those granted higher incomes at entry to college. The latter is a proper effect due to an exogenous change. The former is just a comparison of two groups. We type

```
. margins, at(roommate=0 hsgpa=(2 2.5 3 3.5 4) income=3)
> predict(target(income=3)) predict(target(income=11) base(income=11))
> predict(target(income=11)) contrast(predict(r) nowald effects) noatlegend
  (output omitted)
```

Let's focus first on the syntax. The `predict(target())`s are new; see [ERM] **eprobit predict** for a detailed explanation. Briefly, `target()` specifies a counterfactual value directly. So `predict(target(income=3))` specifies an income of $30,000. Because that is the same value `margins` is specifying, that is more of a factual than a counterfactual. Well, it is a factual for low-income students and is shown as the blue line in figure 2 and figure 3.

`predict(target(income=11) base(income=11))` specifies that both the counterfactual income and the `base()` income from which the student's unobservable characteristics are obtained are $110,000. So it too is a factual. It is a factual for high-income students and is shown as the red line in figure 2. `predict(target(income=11))` specifies that our counterfactual income is 11, but because `margins` is setting the income to 3, the unobservable characteristics will be for a student whose parents earn $30,000. This is the red line in figure 3.

The results are

```
. margins, at(roommate=0 hsgpa=(2 2.5 3 3.5 4) income=3)
> predict(target(income=3)) predict(target(income=11) base(income=11))
> predict(target(income=11)) contrast(predict(r) nowald effects) noatlegend

Contrasts of predictive margins
Model VCE    : Robust

1._predict   : Pr(graduate==yes), predict(target(income=3))
2._predict   : Pr(graduate==yes), predict(target(income=11) base(income=11))
3._predict   : Pr(graduate==yes), predict(target(income=11))
```

	Contrast	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
_predict@_at						
(2 vs 1) 1	.0401723	.0421568	0.95	0.341	-.0424536	.1227981
(2 vs 1) 2	.2315929	.0725988	3.19	0.001	.0893018	.3738839
(2 vs 1) 3	.2695457	.0440405	6.12	0.000	.1832279	.3558636
(2 vs 1) 4	.0634449	.036527	1.74	0.082	-.0081467	.1350365
(2 vs 1) 5	.0030008	.0047942	0.63	0.531	-.0063957	.0123973
(3 vs 1) 1	.1292645	.1030033	1.25	0.209	-.0726182	.3311473
(3 vs 1) 2	.4575187	.078341	5.84	0.000	.3039732	.6110642
(3 vs 1) 3	.3809832	.0367368	10.37	0.000	.3089803	.4529861
(3 vs 1) 4	.0741338	.0414526	1.79	0.074	-.0071118	.1553795
(3 vs 1) 5	.0031996	.0051059	0.63	0.531	-.0068078	.0132071

And their graph is

```
. marginsplot
```

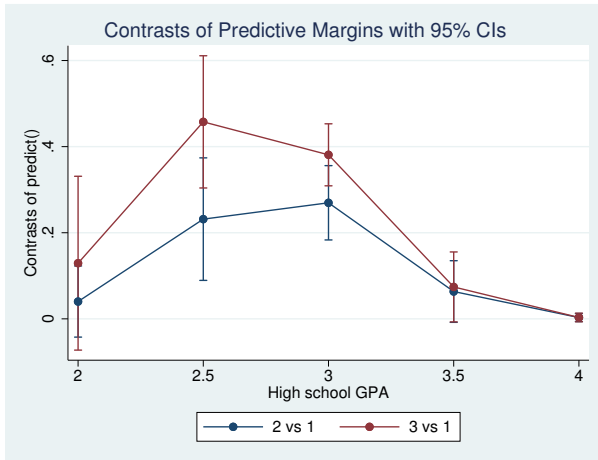

Figure 4.

The blue points and line represent the difference between a student from a family earning $30,000 and a student from a family earning $110,000. The red points and line represents the difference between the same student who started in a family earning $30,000 but was granted $110,000 family earnings on entry into college. The higher income means much more to those who achieved their GPA while in a lower-income family. This is particularly true for those with GPAs between 2.5 and 3.0.

Recall that our estimation results indicated a positive correlation between unobservable factors that increase a student's GPA and those that increase the probability that the student graduates. The

`margins` results above are driven by lower-income students having higher levels of these unobservable factors for any given level of high school GPA. In fact, the only thing that makes the two lines different is that the students who started with incomes of $30,000 have different unobservable characteristics from those who started with incomes of $110,000. All other covariates are the same. How important are those unobserved factors? We assess that directly by comparing our two counterfactuals that set income at $110,000.

We delete the line `predict(target(income=3))` so that we are comparing the two counterfactuals against each other, rather than each against the counterfactual of $30,000 family income.

```
. margins, at(roommate=0 hsgpa=(2 2.5 3 3.5 4) income=3)
> predict(target(income=11) base(income=11)) predict(target(income=11))
> contrast(predict(r) nowald effects) noatlegend
  (output omitted)
. marginsplot
```

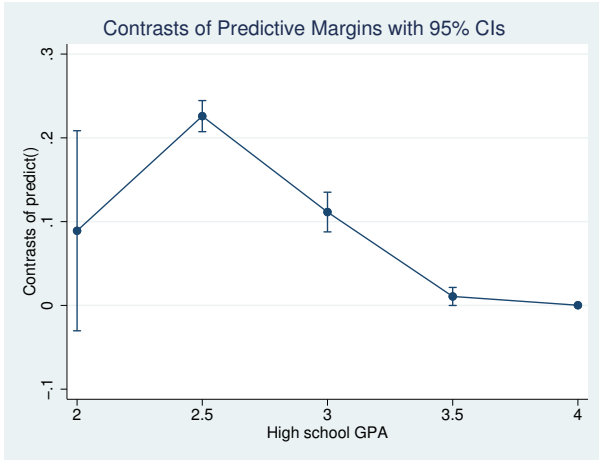

Figure 5.

These results directly measure the contribution of the student's unobservable characteristics to graduation rates. At a GPA of 2.0, a student from a family earning $30,000 and then being moved to a family income of $110,000 would be 10 percentage points more likely to graduate than a student from a family who always earned $110,000 .

That effect rises to over 20 percentage points if the student's GPA is 2.5.

So we can also analyze fully conditional counterfactuals and make complex inferences.

Also see

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

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

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

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