Title

stata.com

Example 2a — Linear regression with binary 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 continuous outcome and endogenous binary covariate.

Remarks and examples

stata.com

Suppose that we want to study the effect of having a college degree on wages. One way to approach the problem is to look at the coefficient on an indicator for whether an individual has a college degree. This gives us an idea of how different the average wage is for individuals with a college degree compared with those without one. However, as in [ERM] **Example 1a**, we suspect that unobserved factors such as ability affect both the probability of graduating from college and wage level. Thus, we need to account for the potential endogeneity of the indicator for having a college degree.

In our fictional study, we collect data on the hourly wages (`wage`) and educational attainment (`college`) of 6,000 adults. We believe that differences in job tenure (`tenure`) and age (`age`) may also affect wages. We can control for these covariates by specifying them in the main equation. We specify `college` in the `endogenous()` option, but this time we also include the `probit` suboption to indicate that the variable is binary. We model graduation as a function of the level of parental education (`peduc`), which we assume does not have a direct effect on wage.

```
. use https://www.stata-press.com/data/r18/wageed
(Wages for 20 to 74 year olds, 2015)
. eregress wage c.age##c.age tenure, endogenous(college = i.peduc, probit)
> vce(robust)
Iteration 0:  Log pseudolikelihood = -18063.148
Iteration 1:  Log pseudolikelihood =   -18060.2
Iteration 2:  Log pseudolikelihood = -18060.164
Iteration 3:  Log pseudolikelihood = -18060.164
Extended linear regression                      Number of obs =   6,000
                                                Wald chi2(4)  = 7584.74
Log pseudolikelihood = -18060.164               Prob > chi2   =  0.0000
```

	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
wage						
age	.4200372	.0163312	25.72	0.000	.3880286	.4520457
c.age#c.age	-.0033523	.0001759	-19.06	0.000	-.003697	-.0030075
tenure	.4921838	.0182788	26.93	0.000	.4563581	.5280095
college						
Yes	5.238087	.1721006	30.44	0.000	4.900776	5.575398
_cons	5.524288	.3428735	16.11	0.000	4.852268	6.196307
college						
peduc						
College	.8605996	.0361723	23.79	0.000	.7897032	.9314959
Graduate	1.361257	.0490862	27.73	0.000	1.26505	1.457465
Doctorate	1.583818	.119513	13.25	0.000	1.349577	1.818059
_cons	-.9731264	.0294779	-33.01	0.000	-1.030902	-.9153508
var(e.wage)	8.99487	.2465919			8.524314	9.491402
corr(e.col~e, e.wage)	.5464027	.0286061	19.10	0.000	.4879055	.600014

The estimated correlation between the errors from the main and auxiliary equations is 0.55 and is significantly different from 0. We conclude that having a college degree is endogenous and that unobservable factors that increase the probability of graduating from college tend to also increase wages.

We find that graduating from college increases the expected wage by \$5.24 given a person's age and employment tenure. This estimate is different than comparing the average wages for college graduates and noncollege graduates.

```
. tabulate college, summarize(wage)
```

Indicator for college degree	Summary of Hourly wage Mean	Std. dev.	Freq.
No	17.768516	3.0674174	3,766
Yes	25.520703	5.045888	2,234
Total	20.654913	5.4248886	6,000

The difference in the average wages is $7.75, but unlike our regression coefficient, that value does not adjust for the different distribution of ages and tenures among college graduates and noncollege graduates.

Another approach to this problem is the potential-outcomes framework. With this approach, we consider the expected wage for each individual without a college degree versus the expected wage for each individual with a college degree. Specifically, we might like to know the average expected change in wages for those who complete college. This is called the average treatment effect on the treated. We consider this approach in [ERM] **Example 2b** and [ERM] **Example 2c**.

[ERM] **Example 2c** also includes an interpretation of how the expected level of income varies by age, tenure, and whether one graduates from college. That analysis could also be applied to this model.

Also see

[ERM] **eregress** — Extended linear regression

[ERM] **eregress postestimation** — Postestimation tools for eregress and xteregress

[ERM] **estat teffects** — Average treatment effects for extended regression models

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