Example 2c — Linear regression with endogenous treatment

Description

Continuing from [ERM] **Example 2b**, we now consider the case where the treatment is endogenous and the variance and correlation parameters differ by treatment group.

Remarks and examples

In [ERM] **Example 2b**, we assumed that graduating from college was an exogenous treatment. However, unobserved factors such as ability may affect whether individuals graduate from college and also affect their wage. Thus, it may be more appropriate for us to treat having a college degree as an endogenous treatment. We found endogeneity in [ERM] **Example 2a**, which analyzes the treatment instead as a binary endogenous covariate. You may want to compare the result of this example with the results from [ERM] **Example 2b**.

Because college graduation is now assumed to be endogenous, we must specify a model for `college`. We model graduation as a function of the level of parental education (`peduc`), which we further assume does not have a direct effect on wage. The endogenous treatment equation is specified in option `entreat()`.

```
. eregress wage c.age##c.age tenure, entreat(college = i.peduc) vce(robust)
Iteration 0:   Log pseudolikelihood = -17382.446
Iteration 1:   Log pseudolikelihood = -17381.922
Iteration 2:   Log pseudolikelihood =  -17381.92
Extended linear regression                         Number of obs =      6,000
                                                   Wald chi2(8)  = 348743.60
Log pseudolikelihood = -17381.92                   Prob > chi2   =     0.0000

                           Robust
              Coefficient  std. err.      z    P>|z|     [95% conf. interval]

wage
     college#
        c.age
          No    .2338084   .0176633    13.24   0.000      .199189    .2684279
         Yes    .6777385   .0219827    30.83   0.000     .6346531    .7208239

     college#
  c.age#c.age
          No   -.0018611     .00019    -9.79   0.000    -.0022335   -.0014887
         Yes   -.0052533   .0002372   -22.14   0.000    -.0057183   -.0047883

     college#
     c.tenure
          No    .3948863   .0207452    19.04   0.000     .3542263    .4355462
         Yes    .5883544   .0257213    22.87   0.000     .5379415    .6387673

      college
          No    10.86301   .3675208    29.56   0.000     10.14268    11.58333
         Yes    3.184255   .4612019     6.90   0.000     2.280316    4.088194

college
        peduc
     College     .849575   .0356419    23.84   0.000     .7797181    .9194318
    Graduate    1.347272   .0491996    27.38   0.000     1.250843    1.443701
   Doctorate    1.541025   .1174797    13.12   0.000     1.310769    1.771281

        _cons   -.973061   .0292791   -33.23   0.000    -1.030447   -.9156749

  var(e.wage)   7.629807   .2245651                      7.202122    8.082889

corr(e.col~e,
      e.wage)    .623109   .0267317    23.31   0.000     .5679046    .6727326
```

As in [ERM] **Example 2b**, we can interpret the coefficients in the `wage` equation as coefficients in separate models for the two potential outcomes—the models for those with and without a college degree. The estimated correlation between the errors from the main and auxiliary equations is 0.62. We could use the z statistic for the correlation to test for endogeneity. We could also use the `estat teffects` and `margins` commands to answer questions related to the entire population or specific subpopulations. However, we will not interpret the results of this model any further because we will first extend it.

Above, we assumed that the relationship between the unobserved factors that affect wage and the unobserved factors that affect whether individuals graduate from college was the same for those individuals with a college degree and those without. We do not have a good reason to believe that these will be the same, so we specify the suboption `pocorrelation` within the option `entreat()` to model separate correlation parameters for the two potential outcomes. We also assumed that the unobserved factors affecting wage were equally variable for those who had a college degree and those who did not. We can relax this assumption and model different variances for the two potential outcomes by specifying the suboption `povariance` within the option `entreat()`.

```
. eregress wage c.age##c.age tenure,
> entreat(college = i.peduc, povariance pocorrelation) vce(robust)
Iteration 0:  Log pseudolikelihood = -17382.446
Iteration 1:  Log pseudolikelihood = -17381.327
Iteration 2:  Log pseudolikelihood = -17381.319
Iteration 3:  Log pseudolikelihood = -17381.319
Extended linear regression                          Number of obs =      6,000
                                                    Wald chi2(8)  = 104887.19
Log pseudolikelihood = -17381.319                   Prob > chi2   =     0.0000

------------------------------------------------------------------------------
             |               Robust
             | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
wage         |
    college# |
       c.age |
         No  |    .234277   .0176793    13.25   0.000     .1996261    .2689278
        Yes  |   .6759938   .0220455    30.66   0.000     .6327854    .7192021
             |
    college# |
 c.age#c.age |
         No  |  -.0018627     .00019    -9.80   0.000    -.0022351   -.0014902
        Yes  |  -.0052427   .0002376   -22.07   0.000    -.0057084   -.0047771
             |
    college# |
    c.tenure |
         No  |   .3917974   .0211184    18.55   0.000      .350406    .4331887
        Yes  |   .5951107   .0264841    22.47   0.000     .5432027    .6470187
             |
     college |
         No  |   10.82487   .3712505    29.16   0.000     10.09723    11.55251
        Yes  |   3.097338   .4678998     6.62   0.000     2.180271    4.014405
-------------+----------------------------------------------------------------
college      |
       peduc |
    College  |   .8482632   .0356294    23.81   0.000     .7784309    .9180955
   Graduate  |   1.343223   .0493492    27.22   0.000       1.2465    1.439945
  Doctorate  |   1.538188   .1162237    13.23   0.000     1.310393    1.765982
             |
       _cons |  -.9715507   .0292856   -33.18   0.000    -1.028949   -.9141521
-------------+----------------------------------------------------------------
 var(e.wage) |
     college |
         No  |    7.46846   .2657898                      6.965275    8.007997
        Yes  |    7.98125   .3990003                      7.236315    8.802871
-------------+----------------------------------------------------------------
corr(e.col~e,|
      e.wage)|
     college |
         No  |   .6057846   .0374579    16.17   0.000     .5271994    .6740954
        Yes  |   .6518029   .0359868    18.11   0.000     .5755573    .7168138
------------------------------------------------------------------------------
```

We see separate variance and correlation parameters for those with a college degree and those without. The estimated correlation between the errors from the main and auxiliary equation is 0.61 for individuals without a college degree and 0.65 for those with a college degree. The z statistics may be used for Wald tests of the null hypothesis that there is no endogenous treatment. For both treatment groups, we reject this hypothesis and conclude that having a college degree is an endogenous treatment. Because the estimates are positive, we conclude that unobserved factors that increase the chance of having a college degree also tend to increase wage.

We can use `estat teffects` to estimate the average effect of a college degree on wage. We use the `atet` option to estimate the ATET.

```
. estat teffects, atet
Predictive margins                              Number of obs    = 6,000
                                                Subpop. no. obs = 2,234

------------------------------------------------------------------------------
             |            Unconditional
             |     Margin   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
ATET         |
     college |
 (Yes vs No) |   5.238589   .2047014    25.59   0.000     4.837382    5.639797
------------------------------------------------------------------------------
```

We estimate that the average wage for those who graduated from college is $5.24 higher than it would have been had those same individuals not graduated from college. This is $2.39 less than the result from our model in [ERM] **Example 2b** that did not account for the endogeneity of college graduation. We said "same individuals" to emphasize that $5.24 is a treatment effect on those who chose to attend college and graduated. More formally, it is our estimate of what the average increase in wage is in the whole population for everyone who chose to attend college and graduated.

But we do not need to stop there. We can analyze how college education affects different subpopulations. Usually, when we zoom into the different subpopulations, we can see effects that otherwise would be averaged out over the population. Before I do that though, I am going to change the data slightly to get nicer output:

```
. label drop tengrp
. label define tengrp 0 "0-3 years tenure" 4 "3-7 years tenure"
>                     8 "Over 7 years tenure"
. label values tenuregrp tengrp
```

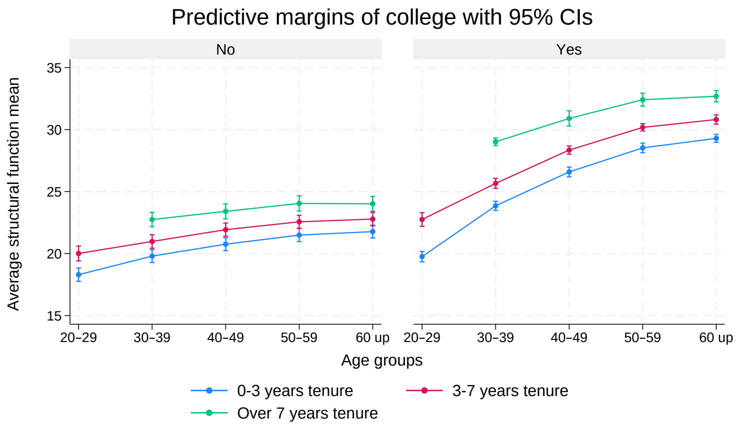
Now, I am ready to use `margins` to zoom into some interesting subgroups of my data.

```
. margins college, over(agegrp tenuregrp) subpop(if college==1 & peduc==1)
> vce(unconditional)
  (output omitted)
```

Here we have narrowed our focus on individuals with a college degree whose parents have a college education, using the `subpop()` option, for different subpopulations defined by our age and tenure groupings, using the `over()` option. Also, by using the default average structural function prediction `asf`, we condition on the unobservable factors that increase the probability of graduating from college.

If you run the margins command, you will see that it takes a few seconds and that it produces a lot of output. Let's graph the results,

```
. marginsplot, by(college) byopts(style(altleg))
```



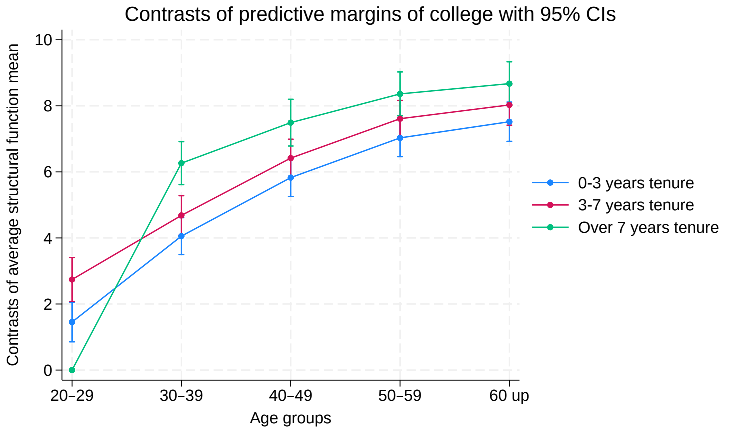
The age–earnings profiles on the left, where we have taken the degrees away from our college graduates, are distinctly different from those on the right, where they get to retain their degrees. We see that tenure does have an effect, and if we look closely, it has a larger effect on college graduates: the profiles are further apart on the right. What do the points on this graph represent? Each point in the panel on the right is the expected wage for someone who graduated from college, whose parents graduated from college, and who has the age and tenure shown on the graph. Each point on the left is a counterfactual where we assume those same people did not graduate from college.

Seeing that, we have to ask, what are the profiles of the effect of college? To find those, we just add an r. to college on our previous margins command.

```
. margins r.college, over(agegrp tenuregrp) subpop(if college==1 & peduc==1)
> vce(unconditional)
  (output omitted )
```

Again, the output is long, so we graph the results.

```
. marginsplot
```



College affects wages the least when people are young and have no tenure. The largest effects are seen for those older than 50 and even more so when they also have long tenure. Each point represents the expected increase in wages due to graduating from college among those who chose to attend college and graduated. So each is an average treatment effect on the treated (ATET). Unlike overall average ATETs, these are conditioned on being at a specific age and having a specific tenure. Each point is bracketed by a pointwise 95% confidence interval. The confidence intervals reveal that we have pretty tight estimates for each of the ATETs. Note that the previous graph also displayed 95% confidence intervals. They were just so narrow that they are difficult to see.

As we saw above, `margins` allows us to answer numerous questions depending on the counterfactuals and subpopulations we defined.

See *Treatment* under *Methods and formulas* in [ERM] **eregress** and *Estimating treatment effects with margins* in [R] **margins, contrast** for additional information about calculating the ATET.

Video example

Extended regression models: Nonrandom treatment assignment

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