Title

stata.com

Example 1a — Linear 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 continuous outcome and continuous endogenous covariate.

Remarks and examples

stata.com

The fictional State University is studying the relationship between the high school grade point average (GPA) of the students it admits and their final college GPA. They suspect that unobserved ability affects both high school GPA and college GPA. Thus, high school GPA is an endogenous covariate.

Using data on the 2,500 students in the cohort expected to graduate in 2010, the researchers at State U model college GPA (`gpa`) as a function of high school GPA (`hsgpa`). In both cases, GPA is measured in 0.01 increments, and we ignore complications due to the boundary points. We also ignore that, unfortunately, State U has a high dropout rate and college GPA is missing for these students, leaving the researchers with a sample of about 1,500 students.

The State U researchers expect that the effect of high school competitiveness on college GPA is negligible once high school GPA is controlled for. So they include a ranking of the high school (`hscomp`) as an instrumental covariate for high school GPA. They include parental income measured in $10,000s, which they believe may also influence student performance, in the main model and in the model for high school GPA.

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

. eregress gpa income, endogenous(hsgpa = income i.hscomp)

Iteration 0:  Log likelihood = -638.58598
Iteration 1:  Log likelihood = -638.58194
Iteration 2:  Log likelihood = -638.58194

Extended linear regression                          Number of obs =   1,528
                                                    Wald chi2(2)  = 1167.79
Log likelihood = -638.58194                         Prob > chi2   =  0.0000

-------------------------------------------------------------------------------
              | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
--------------+----------------------------------------------------------------
gpa           |
       income |   .0575145   .0055174    10.42   0.000     .0467007    .0683284
        hsgpa |   1.235868    .133686     9.24   0.000     .9738484    1.497888
        _cons |  -1.217141   .3828614    -3.18   0.001    -1.967535   -.4667464
--------------+----------------------------------------------------------------
hsgpa         |
       income |   .0356403   .0019553    18.23   0.000     .0318079    .0394726
              |
       hscomp |
    Moderate  |  -.1310549   .0136503    -9.60   0.000    -.1578091   -.1043008
        High  |  -.2331173   .0232712   -10.02   0.000     -.278728   -.1875067
              |
        _cons |   2.951233   .0164548   179.35   0.000     2.918982    2.983483
--------------+----------------------------------------------------------------
   var(e.gpa) |   .1436991   .0083339                      .1282592    .1609977
 var(e.hsgpa) |   .0591597   .0021403                        .05511     .063507
--------------+----------------------------------------------------------------
corr(e.hsgpa, |
       e.gpa) |   .2642138   .0832669     3.17   0.002     .0948986    .4186724
-------------------------------------------------------------------------------
```

The estimate of the correlation between the errors from the main and auxiliary equations is 0.26. The z statistic may be used for a Wald test of the null hypothesis that there is no endogeneity. The researchers reject this hypothesis. Because the estimate is positive, they conclude that unobservable factors that increase high school GPA tend to also increase college GPA.

Having satisfied themselves that it is appropriate to account for endogeneity of high school GPA, they examine the coefficient estimates. The estimates for the main equation are interpreted just like those from `regress`; see [R] **regress**. For example, the researchers expect the difference in college GPA is about 1.24 points for students with a difference of 1 point in high school GPA.

As we discussed in [ERM] **Intro 9**, the coefficients on `hsgpa` and `income` in this regression pretty much say everything there is to say about how college GPA changes when either high school GPA or parents' income changes. This is true because our model is linear and we have no interactions. We could make this the end of our story. But it is not the end if we want to ask questions about expected levels of college GPA.

Let's look at a single observation; we will pretend it is for Billy.

```
. generate str name = "Billy" in 537
(2,499 missing values generated)

. list gpa hsgpa income hscomp if name=="Billy"

      +--------------------------------+
      |  gpa   hsgpa   income   hscomp |
      |--------------------------------|
537.  | 1.03       2        2     High |
      +--------------------------------+
```

We have information on Billy's high school competitiveness, `hscomp`, and his parents' income. With this information, we could form counterfactuals about Billy. We could fix Billy's high school GPA at 2.00, and we could fix his high school GPA at 3.00. We will let `margins` give us the expected values for college GPA under these two counterfactuals.

```
. margins if name=="Billy", at(hsgpa=(2 3))
warning: prediction constant over observations.

Predictive margins                                       Number of obs = 1
Model VCE: OIM

Expression: Average structural function mean, predict()
1._at: hsgpa = 2
2._at: hsgpa = 3

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   1.044564    .039223    26.63   0.000     .9676881     1.12144
          2  |   2.280432   .1214604    18.78   0.000     2.042374     2.51849
------------------------------------------------------------------------------
```

When we set Billy's high school GPA to 2.00, Billy's expected college GPA is 1.04. Because we did not specify values for the other covariates, `margins` took them from the observation for Billy. So, more completely and correctly, this is the expected GPA for anyone whose high school GPA is 2.00 and whose parents' income is $20,000 and whose level of high school competitiveness is high.

Why do we care about high school competitiveness when it is not in the main equation? As discussed in [ERM] **Intro 7**, if you want to make inferences that have a structural interpretation with respect to the population (or data-generating process), you must include the level of endogeneity from the equation for high school GPA. In this case, that is an adjustment for Billy's unobserved ability.

When we fix Billy's high school GPA to 3.00, while keeping his parents' income constant at $20,000 and also keeping the adjustment for Billy's unobserved ability constant, we see that Billy's expected college GPA rises to 2.28.

As a sidebar, we note that our first counterfactual of high school GPA at 2.00 is really more of a factual than a counterfactual. Billy's observed GPA in the data is 2.00.

Let's take the next step and estimate the resulting difference in expected college GPA for our two counterfactuals. We just need to add `contrast(at(r))` to our `margins` command.

```
. margins if name=="Billy", at(hsgpa=(2 3)) contrast(at(r) effects nowald)
warning: prediction constant over observations.

Contrasts of predictive margins                            Number of obs = 1
Model VCE: OIM

Expression: Average structural function mean, predict()
1._at: hsgpa = 2
2._at: hsgpa = 3

------------------------------------------------------------------------------
             |            Delta-method
             |   Contrast   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         _at |
    (2 vs 1) |   1.235868    .133686     9.24   0.000     .9738484    1.497888
------------------------------------------------------------------------------
```

The estimated effect, its standard error, and all other associated statistics are identical to the coefficient on hsgpa in our eregress output.

Would we see anything different if we averaged the effects over the sample to get estimates of the effects in the population? Just remove Billy from the command.

```
. margins, at(hsgpa=(2 3)) contrast(at(r) effects nowald)

Contrasts of predictive margins                        Number of obs = 1,528
Model VCE: OIM

Expression: Average structural function mean, predict()
1._at: hsgpa = 2
2._at: hsgpa = 3

------------------------------------------------------------------------------
             |            Delta-method
             |   Contrast   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         _at |
    (2 vs 1) |   1.235868    .133686     9.24   0.000     .9738484    1.497888
------------------------------------------------------------------------------
```

Not surprisingly, the estimated effect is still 1.24—the same value we have gotten every time, the same value as the coefficient on hsgpa. Perhaps more surprisingly, the standard error of the population-average estimate is also the same as the standard error of the coefficient. We do not gain or lose any information when we take an average over an estimate that is constant for all the observations.

In linear models without interactions, we have just seen that the effects are the same for many questions. In nonlinear models, the effects usually differ.

The models in the remaining two examples in this series, [ERM] **Example 1b** and [ERM] **Example 1c**, have exactly the same interpretation we gave to the model in this entry. Adding interval-censoring and endogenous sample selection does not affect the relevant questions or how they are answered.

Video example

Extended regression models: Endogenous covariates

Also see

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

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

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

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

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