Title

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

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 http://www.stata-press.com/data/r15/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

─────────────────────────────────────────────────────────────────────────────
             │      Coef.   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 8**, 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.

If we want to ask questions about the eventual level of college GPA, we must be specific about how we arrived at our values for `hsgpa`. Let's look at a single observation; we will pretend it is for Billy.

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

. list income if name=="Billy"

      ┌────────┐
      │ income │
      ├────────┤
537.  │      2 │
      └────────┘
```

What if we don't have records from Billy's high school and all we know about Billy is his parents' income? 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. These are values we are choosing, not the value that Billy arrived at through his own actions. We'll let `margins` give us the expected values for college GPA under these two counterfactuals.

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

Predictive margins                              Number of obs     =          1
Model VCE    : OIM

Expression   : mean of gpa, predict(fix(hsgpa))

1._at        : hsgpa           =           2

2._at        : hsgpa           =           3

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   1.369625   .1251674    10.94   0.000     1.124301    1.614948
          2  |   2.605493   .0190405   136.84   0.000     2.568174    2.642811
------------------------------------------------------------------------------
```

When we set Billy's high school GPA to 2.00 and consider his parents' income of $20,000, Billy's expected college GPA is 1.37. More correctly, this is the expected GPA for anyone whose parents' income is $20,000 and whose high school GPA is fixed at 2.00. Keeping his parents' income constant and fixing his high school GPA at 3.00, we see that Billy's expected college GPA rises to 2.61.

But in reality, we know more about Billy.

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

      ┌─────────────────────────────────┐
      │  gpa   hsgpa   income    hscomp │
      ├─────────────────────────────────┤
537.  │ 1.03       2        2      high │
      └─────────────────────────────────┘
```

And with this, we can ask a slightly different question. What is Billy's expected GPA given all that we know about him, including the competitiveness of his high school and the unobserved thing or things that drive the correlation between high school and college GPAs? What if we further ask how that expectation would change if we granted Billy one additional unit of high school GPA, taking him from 2.00 to 3.00. These are the same two counterfactuals for the value of high school GPA, but a different assumption about how Billy arrived at a 2.00. To obtain these counterfactuals, we run the same `margins` command, changing the `fix()` option to `base()`.

```
. generate hsgpaT = hsgpa                        // Observed ("True") H.S. GPA
. margins if name=="Billy", at(hsgpa=(2 3)) predict(base(hsgpa=hsgpaT))
Warning: prediction constant over observations.

Predictive margins                               Number of obs     =          1
Model VCE    : OIM

Expression   : mean of gpa, predict(base(hsgpa=hsgpaT))

1._at        : hsgpa           =           2

2._at        : hsgpa           =           3

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   1.044564   .1242365     8.41   0.000     .8010648    1.288063
          2  |   2.280432   .0207685   109.80   0.000     2.239726    2.321138
------------------------------------------------------------------------------
```

The numbers are not the same. The expected GPA of 1.04 is closer to Billy's true value of 1.03 than was the estimate using only `income`. That need not be the case for any individual, but given that we used more information, we would expect it to be true if we averaged over others with the same characteristics.

As discussed in [ERM] **intro 8**, we needed to save Billy's true value of `hsgpa` because `margins` manipulates the data to obtain its results. We did not need to do this with the `fix()` option because predictions using `fix()` do not care what Billy's true value of `hsgpa` is or how he arrived at that value. Predictions using `base()`, on the other hand, use Billy's true value of `hsgpa` and all information from the model about how Billy arrived at that GPA. The `base()` option instructs `margins` to use true `hsgpaT` when it formed both of its counterfactuals. Thus, both counterfactuals include information about his high school's competitiveness and information about the unobserved factor or factors creating the correlation between GPAs. The same values for this information are used when `margins` creates each counterfactual. We could say that, compared with the counterfactuals computed under `fix()`, these counterfactuals include more of what makes Billy, Billy. They are still the expected value for anyone with the same covariates, but they incorporate the fact that the GPA of 2.0 was arrived at through Billy's own actions and include the competitiveness of his high school.

In the parlance of treatment effects, our first set of estimates could be called the potential outcomes given the fixed treatment levels: 2.00 and 3.00. If that doesn't help your understanding, then skip this paragraph. The second set of values would be the counterfactuals required to estimate the treatment effect on the untreated (TEU). Why are we being so cagey with the language—"could be" instead of "are" and "counterfactual" instead of "potential outcome" in the second case? Experts in treatment effects don't like applying the term "potential outcome" when the treatment is continuous. That implies an infinite number of potential outcomes. They are even protective of the term when used to create the pieces needed for the TEU. Regardless, the computation is exactly what would be done to form these potential outcomes for a binary or ordinal treatment, and the interpretation conveys the same meaning.

Neither the `fix()` nor the `base()` counterfactuals can be said to be better. They simply answer different questions. When we consider exogenous changes to variables like high school GPA, the counterfactuals from `base()` will often be more relevant to answering many questions. Whether a guidance counselor or a policy maker is asking the question, both are likely to face the existing GPAs of individual students or those in the population.

Let's take the next step and estimate the resulting changes in expected college GPA for our two situations. We just need to add `contrast(at(r))` to each of our two `margins` commands.

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

Contrasts of predictive margins
Model VCE    : OIM

Expression   : mean of gpa, predict(fix(hsgpa))

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
------------------------------------------------------------------------------

. margins if name=="Billy", at(hsgpa=(2 3)) predict(base(hsgpa=hsgpaT))
> contrast(at(r) effects nowald)
Warning: prediction constant over observations.

Contrasts of predictive margins
Model VCE    : OIM

Expression   : mean of gpa, predict(base(hsgpa=hsgpaT))

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
------------------------------------------------------------------------------
```

As we have said repeatedly, the estimates of the effects are the same. It does not matter how Billy arrived at his 2.00. What's more, the standard errors are the same, and they are the same as the standard error of the regression coefficient from our `eregress` output. In this case, the additional information that was so important in getting the right GPA estimates is subtracted out when we compute the differences. That is a direct result of the model being linear and having additive errors. Stretching the parlance of treatment effects again, we could call our first contrast an estimate of the treatment effect and the second a treatment effect on the untreated. For linear models without interactions, these are always the same value.

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 commands.

```
. margins, at(hsgpa=(2 3)) predict(fix(hsgpa)) contrast(at(r) effects nowald)
Contrasts of predictive margins
Model VCE    : OIM
Expression   : mean of gpa, predict(fix(hsgpa))
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
------------------------------------------------------------------------------

. margins, at(hsgpa=(2 3)) predict(base(hsgpa=hsgpaT)) contrast(at(r) effects nowald)
  (output omitted)
```

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 don't gain or lose any information when we take an average over an estimate that is constant for all the observations.

We leave it to you to run the last command and see that `fix()` and `base()` produce the same results.

In linear models without interactions, we have just seen that the effects are the same for many questions, but the levels are often different. In nonlinear models, these differences in the levels will lead to differences in the effects.

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 do not affect either the relevant questions or how they are answered.

Also see

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

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

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

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