Title stata.com

intro 8 — Conceptual introduction via worked example

Description Remarks and examples References Also see

Description

This entry introduces the concepts of endogenous covariates, nonrandom treatment assignment, and endogenous sample selection through a series of examples. It also provides an overview of how to interpret the results of ERMs.

Remarks and examples stata.com

Remarks are presented under the following headings:

Introduction

In a perfect research world, several assumptions we conventionally make about our data and the data-collection process would be true. For example, we could gather data about all the variables that influence the outcome we want to study. These data would be collected on a random sample of the population of interest. Any inferences we made about a relationship between the dependent variable and an independent variable when studying one group would be just as valid if we studied this group again at a different time or even if we conducted the study for a different group.

Often, applied research is complicated when one or more of the classical assumptions are not true. For example, data on key variables of interest may be unavailable. Our interest may lie in a treatment that cannot be randomly assigned or may be endogenous. Or the subjects we have available to study are not representative of the population we want to study.

When any of these things is true, we cannot make accurate inferences using standard regression methods. Stata provides many commands that can be used when one of these complications occurs. The ERM commands allow you to address these problems in isolation and, more importantly, in combination—as they often occur.

Imagine that a large company is considering offering a workplace wellness program to its employees to help them lose weight. They have conducted a pilot study at one location, and all other locations are expected to be similar. In our dataset, the `wellpgm` variable records whether a given employee participated. After one year, the company wants to know whether the program was effective. Our outcome of interest is weight lost in kilograms. We have called this `weightloss0` to distinguish it from the observed `weightloss` later.

In our fictional data, the number of kilograms lost is also determined by the employee's age in years (`age`), the employee's sex (`sex`), and the employee's starting weight in kilograms (`weight`). Because this is an entirely fictitious example, we have a true measure of willingness to engage in healthy behaviors (`health`).

More formally, in our simulated data, the process that determines weight lost is

$$\texttt{weightloss0}_i = -4 - 0.1 \times \texttt{age}_i - 1.5 \times \texttt{sex}_i + 0.14 \times \texttt{weight}_i + 1.2 \times \texttt{wellpgm}_i \\ + 0.5 \times \texttt{health}_i + u_i$$

Suppose that we are in the situation described above. We observed complete information for all variables for all employees, and participation in the wellness program was unrelated to any employee attributes that we could not observe. In this case, we could fit our model by typing

```
. use http://www.stata-press.com/data/r15/wellness
(Fictional workplace wellness data)

. regress weightloss0 age i.sex weight i.wellpgm health

      Source |       SS           df       MS      Number of obs   =       545
-------------+----------------------------------   F(5, 539)       =    589.61
       Model |  2417.76071         5  483.552141   Prob > F        =    0.0000
    Residual |  442.044242       539  .820119187   R-squared       =    0.8454
-------------+----------------------------------   Adj R-squared   =    0.8440
       Total |  2859.80495       544  5.25699439   Root MSE        =     .9056

------------------------------------------------------------------------------
 weightloss0 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0991644   .0038045   -26.06   0.000    -.1066378   -.0916909
             |
         sex |
        male |  -1.481883   .0937504   -15.81   0.000    -1.666044   -1.297722
      weight |   .1359547   .0054405    24.99   0.000     .1252676    .1466419
             |
     wellpgm |
         yes |   1.254928   .1076792    11.65   0.000     1.043406     1.46645
      health |   .4814308   .0255931    18.81   0.000     .4311564    .5317053
       _cons |  -3.754726   .4432054    -8.47   0.000    -4.625348   -2.884105
------------------------------------------------------------------------------

. estimates store true
```

From this model, we can estimate the average treatment effect (ATE) of the wellness program by using the coefficient on `wellpgm`. We estimate that the ATE is 1.25 kg. In other words, the average weight lost over the course of the year would be 1.25 kg greater if all the company's employees participated in the program versus if no employees participated.

Because we simulated these data, we can confirm that all the confidence intervals contain the true values. If we continued to add more observations, our point estimates would become closer and closer to the real values. This is true because the coefficient estimates shown above are consistent. Because they are consistent, we can make inferences about the effects of each variable on the outcome. We `estimates store` these values as `true` for comparison with later models.

Complications

As discussed in [ERM] **intro 3**, a covariate is endogenous if it is correlated with the error term. Practically, this correlation arises for many reasons. For example, we may have omitted an important variable from our model that is correlated with a variable that we included, as we did here. Or we may not have accurately measured one of the covariates in our model. We could also have the case where a variable in the model and the outcome of interest are partially determined by the same unobserved factors. For concreteness, we focus on the role of a single omitted variable in this conceptual introduction.

Often in observational research, the treatment (participation in the wellness program) was not randomly assigned. As discussed in [ERM] **intro 5**, we might be able to ignore this issue if we do not suspect that unobserved factors that affect participation also affect the amount of weight loss. However, in this case, we believe participation in the wellness program is also likely to be determined by factors we cannot observe, such as the now-omitted `health` variable.

Further, suppose that the pilot study was structured such that baseline information about all employees was collected at a mandatory benefits meeting at the start of the year. At the end of the year, all employees were asked to go to the company gym during business hours to have their year-end weight recorded, regardless of program participation. Because employees were not required to have their final weight recorded, we observe only the weight of employees who voluntarily went to the gym. We have a selected sample in this case.

Whether an employee is observed in the study could be correlated with unobserved factors that also determine how much weight he or she lost. For example, employees with high values of the now-omitted `health` variable may have generally better diet and exercise habits (independent of the wellness program), leading to higher weight loss. Let's say that for bragging rights, they want to have their superior weight loss recorded, so they are more likely to show up at the end of the year. As discussed in [ERM] **intro 4**, if selection is related to unobserved factors that are correlated with the outcome, it cannot be ignored.

If we ignore all of these potential complications, we might erroneously fit the model below. In this model, we omit `health`, and `weightloss` records the observed weight loss only for employees who went to the gym at the end of the year.

$$\texttt{weightloss}_i = \beta_1 \times \texttt{age}_i + \beta_2 \times \texttt{sex}_i + \beta_3 \times \texttt{weight}_i + \beta_4 \times \texttt{wellpgm}_i + u_i$$

As before, we could fit the model using `regress`.

```
. regress weightloss age i.sex weight i.wellpgm

      Source |       SS           df       MS      Number of obs   =       337
-------------+----------------------------------   F(4, 332)       =    219.74
       Model |  1239.17345         4  309.793362   Prob > F        =    0.0000
    Residual |  468.060374       332   1.4098204   R-squared       =    0.7258
-------------+----------------------------------   Adj R-squared   =    0.7225
       Total |  1707.23382       336  5.08105304   Root MSE        =    1.1874

------------------------------------------------------------------------------
  weightloss |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0800928   .0063184   -12.68   0.000    -.0925219   -.0676637
             |
         sex |
       male  |  -1.023886    .146934    -6.97   0.000    -1.312925    -.734847
      weight |   .0803689   .0074025    10.86   0.000     .0658072    .0949305
             |
     wellpgm |
        yes  |   1.913531   .1596906    11.98   0.000     1.599398    2.227664
       _cons |  -.3699558   .6741312    -0.55   0.584    -1.696063    .9561513
------------------------------------------------------------------------------
```

None of the confidence intervals for our coefficient estimates contain the true values. We store the estimates so that we can compare them with estimates from other models later.

```
. estimates store base
```

Endogenous covariates

Continuing with our example, we suspect that `weight` is endogenous now that we cannot observe `health`. Employees who are predisposed to healthy behaviors will likely have a lower starting weight, and this could influence how much weight they lose over the course of the year-long study. If we have a suitable model for how `weight` relates to the unobserved `health`, we can still estimate the parameters consistently.

Let's suppose we believe that the employee's starting weight is a function of the employee's sex and the number of times the employee visits the company gym. We measure gym use as the employee's average number of visits per month to the company gym before the program (`gym`). This will be an instrumental variable for `weight`. Instrumental variables are exogenous covariates that are correlated with the endogenous covariate, not directly related to the outcome, and not correlated with the unobserved error. Because we are using preprogram gym use, we do not expect it to be related to weight loss during the year of the program.

We fit the model using `eregress`, storing the estimates for later comparison.

```
. eregress weightloss age i.sex i.wellpgm, endogenous(weight = i.sex gym)
  (output omitted)
. estimates store endog
```

Now, we view and compare the results from each of the commands. We focus on the coefficients here because our interest lies in illustrating how the point estimates change as we address different complications. At the end of the introduction, we show the full output of `eregress` and discuss its interpretation.

```
. estimates table true base endog, stats(N) equations(1) keep(#1:)

--------------------------------------------------------
    Variable |     true          base          endog
-------------+------------------------------------------
         age | -.09916437    -.08009282    -.07964086
             |
         sex |
       male  |  -1.481883     -1.023886    -1.6411717
             |
      weight |  .13595472     .08036889     .14701973
             |
     wellpgm |
        yes  |  1.2549281     1.9135311     1.9008534
             |
      health |  .48143082
       _cons | -3.7547263    -.36995584    -5.5172377
-------------+------------------------------------------
           N |        545           337           337
--------------------------------------------------------
```

Once we account for the endogeneity of `weight`, the coefficients for `sex` and `weight` are close to those of the `true` model and have the correct signs. The estimates for `age` and `wellpgm`, however, are close to each other in the `base` and `endog` models but not close to the `true` values. Our estimates remain inconsistent because we have not yet addressed the endogeneity of the `wellpgm` program indicator.

Endogenous covariates in ERMs need not be continuous. We could instead have an endogenous binary or ordinal covariate. To address the endogeneity of `wellpgm`, we could include an additional model by adding another `endogenous()` option; see [ERM] **intro 3** for more on specifying models with different types of endogenous covariates. Another way to approach the analysis of binary and ordinal endogenous covariates is in the potential-outcomes framework.

Nonrandom treatment assignment

Treatment-effect regressions model the effect of a discrete treatment or intervention on the outcome. In observational data, we cannot randomly assign a treatment of interest to individuals. Treatment status may be related to other covariates that we measure. It may even be related to the unobserved factors that affect the outcome and be endogenous. We cannot just take the sample means of the treated and untreated to estimate the ATE. Instead, we can use the potential-outcomes framework to estimate a treatment effect.

In the potential-outcomes framework, the treatment effect is the difference between the outcome that would occur when a given subject receives the treatment and the outcome that would occur when the subject receives the control instead. We only observe the potential outcome associated with that subject's observed treatment value (either treated or control). However, we can estimate both potential outcomes, conditional on covariates, by using information from the model. For more information about the potential-outcomes framework, see [TE] **teffects intro advanced**.

The ERM commands may be used with an exogenous or endogenous treatment where the treatment variable is binary or ordinal.

To address the endogenous selection of participation in the wellness program, we need a model for `wellpgm`. Whether the employee was a smoker at the beginning of the year (`smoke`) is an additional covariate in our treatment model. Because smoking signals a lower willingness to engage in healthy behaviors, it should be correlated with participation in the program, but smoking status measured before the program was offered should not be independently associated with weight loss during the program.

```
. eregress weightloss age i.sex, endogenous(weight = i.sex gym)
> entreat(wellpgm = age i.smoke, nointeract)
  (output omitted )
. estimates store entrt
```

By specifying `nointeract`, we keep the same coefficients for both treatment groups in the main equation. This is not the most common approach. However, we simulated the data this way to keep the `estimates table` results compact and easy to compare across models. We will show you a more interesting model later.

Now, we view and compare the results for the main equation for each of the models.

```
. estimates table true base endog entrt, stats(N) equations(1) keep(#1:)

------------------------------------------------------------------
    Variable |    true         base         endog        entrt
-------------+----------------------------------------------------
         age | -.09916437   -.08009282   -.07964086   -.10430319
             |
         sex |
       male  |  -1.481883    -1.023886   -1.6411717   -1.5995888
             |
      weight |  .13595472    .08036889    .14701973    .14151952
             |
     wellpgm |
        yes  |  1.2549281    1.9135311    1.9008534    .83556752
             |
      health |  .48143082
       _cons | -3.7547263   -.36995584   -5.5172377    -3.488841
-------------+----------------------------------------------------
           N |        545          337          337          337
------------------------------------------------------------------
```

In the `entrt` model, where we have accounted for the endogeneity of starting weight and the endogenous treatment assignment to the wellness program, we estimate that the effect of participating in the program is 0.84 kg lost. This is closer to the 1.25 kg we estimated in the `true` model than the 1.90 kg we estimated in the `endog` model that did not account for treatment assignment.

Endogenous sample selection

Sample selection is an ambiguous term because different authors have used it to mean different things. To add more ambiguity, sample selection has been equated with nonresponse bias and selection bias in some disciplines. Much of the ambiguity arises from authors not being precise about when sample selection is ignorable.

Sample selection is like treatment assignment: a process maps each individual into or out of the sample. This process depends on observable covariates and unobservable factors. When unobservable factors that affect who is in the sample are independent of unobservable factors that affect the outcome, then the sample selection is not endogenous. In this case, the sample selection is ignorable—our estimator that ignores sample selection is still consistent.

In contrast, when the unobservable factors that affect who is included in the sample are correlated with the unobservable factors that affect the outcome, the sample selection is endogenous and it is not ignorable, because estimators that ignore endogenous sample selection are not consistent.

The ERM commands may be used with endogenous sample selection with a probit or tobit selection model. A probit selection model is used when we have a binary indicator of selection. A tobit selection model is used when we have a continuous indicator for selection.

We suspect that unobserved factors that influence whether employees came to the gym for the year-end weigh-in also influence the amount of weight lost. In other words, we believe we may have endogenous sample selection. Our `true` model included all information on all 545 employees. In reality, only 337 completed the final weigh-in for our study. However, we still want to know what the potential effect of the program was for all employees. The 0.84 kg that we estimated in *Nonrandom treatment assignment* is not a consistent estimate of the program's ATE in the company if the 337 employees in our study are not representative of the population.

By modeling the sample-selection process, we can include all 545 employees in our estimation sample. The variable `completed` indicates whether the employee completed the final weigh-in. Employees with `completed` = 0 have missing values for `weightloss`. However, because all other data were gathered at a mandatory meeting at the start of the year (such as starting weight) or collected from administrative records (such as prior-year visits to the company gym), we have complete information for all other variables.

We include the employee's job classification (`salaried`) and years employed at the company (`experience`) as additional covariates in our selection model that are excluded from the main equation. `salaried` is 1 if the employee is salaried and is 0 if the employee is paid hourly. We anticipate that salaried employees will have more opportunity to visit the gym during the day and that employees who have been with the company longer will be more motivated to help complete the study. Aside from their effect on completing the weigh-in, we do not believe that `salaried` or `experience` have any direct effect on `weightloss`.

We fit our model, accounting for the potentially endogenous selection.

```
. eregress weightloss age i.sex, endogenous(weight = i.sex gym)
> entreat(wellpgm = age i.smoke, nointeract)
> select(completed = i.wellpgm experience i.salaried)
  (output omitted)
. estimates store endsel
```

Then, we compare these estimates with those from our previous models.

```
. estimates table true base endog entrt endsel, stats(N) equations(1) keep(#1:)

--------------------------------------------------------------------------
    Variable |    true         base         endog        entrt       endsel
-------------+------------------------------------------------------------
         age | -.09916437   -.08009282   -.07964086   -.10430319   -.11149981
             |
         sex |
       male  |  -1.481883    -1.023886   -1.6411717   -1.5995888   -1.5607651
             |
      weight |  .13595472    .08036889    .14701973    .14151952    .14353999
             |
     wellpgm |
        yes  |  1.2549281    1.9135311    1.9008534    .83556752    .92462755
             |
      health |  .48143082
       _cons | -3.7547263   -.36995584   -5.5172377    -3.488841   -3.6798876
-------------+------------------------------------------------------------
           N |        545          337          337          337          545
--------------------------------------------------------------------------
```

After accounting for the potentially endogenous selection that occurs because some employees chose not to complete the final weigh-in, we see that our estimated ATE is 0.925, which is closer to its true value than in the models that did not address selection.

Interpreting effects

In the previous sections, we showed only the coefficient estimates from the main outcome equation. The full output for `eregress` and the other ERM commands includes estimates of coefficients of covariates in the auxiliary models, error variances, and error correlation terms.

For many models, the coefficient estimates themselves are not directly useful. You will need to use `margins` or `estat teffects` to obtain interpretable effects. However, the correlation estimates always provide relevant information.

The full results for the last `eregress` command that we estimated are as follows:

```
. eregress weightloss age i.sex, endogenous(weight = i.sex gym)
> entreat(wellpgm = age i.smoke, nointeract)
> select(completed = i.wellpgm experience i.salaried)
  (iteration log omitted)
Extended linear regression                      Number of obs     =        545
                                                      Selected    =        337
                                                      Nonselected =        208

                                                Wald chi2(4)      =     749.04
Log likelihood = -2800.8318                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
weightloss   |
         age |  -.1114998   .0083531   -13.35   0.000    -.1278715   -.0951281
             |
         sex |
       male  |  -1.560765   .2062746    -7.57   0.000    -1.965056   -1.156474
      weight |     .14354   .0175073     8.20   0.000     .1092263    .1778537
             |
     wellpgm |
        yes  |   .9246275   .2750269     3.36   0.001     .3855848     1.46367
       _cons |  -3.679888   1.464123    -2.51   0.012    -6.549515     -.81026
-------------+----------------------------------------------------------------
completed    |
     wellpgm |
        yes  |   .6553902   .2263862     2.90   0.004     .2116814    1.099099
  experience |  -.8153984   .0617977   -13.19   0.000    -.9365196   -.6942772
             |
    salaried |
        yes  |   .4709859   .1419878     3.32   0.001      .192695    .7492768
       _cons |   4.902936   .3973849    12.34   0.000     4.124076    5.681796
-------------+----------------------------------------------------------------
wellpgm      |
         age |  -.0938617   .0072734   -12.90   0.000    -.1081173    -.079606
             |
       smoke |
        yes  |  -1.477078   .1772103    -8.34   0.000    -1.824404   -1.129752
       _cons |   4.228481    .337379    12.53   0.000      3.56723    4.889732
-------------+----------------------------------------------------------------
weight       |
         sex |
       male  |   9.506396   .6960864    13.66   0.000     8.142091     10.8707
         gym |  -.8184902   .0779351   -10.50   0.000    -.9712401   -.6657402
       _cons |   80.10245   .5407952   148.12   0.000     79.04251    81.16239
-------------+----------------------------------------------------------------
var(e.weig~s)|   2.015328    .263477                      1.559777    2.603927
var(e.weight)|   65.98395   3.997213                      58.59678    74.30241
-------------+----------------------------------------------------------------
corr(e.com~d,|
e.weightloss)|   .5434105   .0824836     6.59   0.000      .362338    .6849556
corr(e.wel~m,|
e.weightloss)|   .5878321   .1054098     5.58   0.000     .3440372    .7573749
corr(e.wei~t,|
e.weightloss)|  -.4801763    .089175    -5.38   0.000    -.6353685   -.2877017
corr(e.wel~m,|
 e.completed)|   .3753168   .1523364     2.46   0.014     .0470351    .6304273
corr(e.wei~t,|
 e.completed)|  -.0643813   .0718768    -0.90   0.370    -.2030702    .0768401
corr(e.wei~t,|
   e.wellpgm)|   -.096324   .0691411    -1.39   0.164    -.2292586    .0401382
------------------------------------------------------------------------------
```

The completed, wellpgm, and weight equations provide the coefficient estimates for the auxiliary endogenous selection, treatment assignment, and endogenous covariate models.

The correlation estimates tell us about the endogeneity in our model. For example, we speculated that we might have endogenous selection. The error correlation corr(e.completed,e.weightloss) is an estimate of the correlation between the error from the selection equation and the error from the outcome equation. The estimate is significant, so we reject the hypothesis that there is no endogenous selection. It is positive, so we conclude that unobserved factors that increase the likelihood of being in the sample tend to occur with unobserved factors that increase the amount of weight lost. Looking at the other correlations, we find that our suspicions of endogenous treatment choice and the endogeneity of initial weight are likewise confirmed.

We estimated an ATE in our running example. In our simple illustration, we were able to use the coefficient on wellpgm. If wellpgm had been interacted with other covariates in the model, we would have needed to use estat teffects. We also could have estimated the effect of the wellness program on just those employees who participated, the average treatment effect on the treated (ATET).

Using this regression, if we ask questions about how participating in wellpgm affects the expected change in weightloss, we will almost always get the same answer: 0.92 kg greater weight loss with the program than without. That is the coefficient on wellpgm in the main outcome equation. This model is linear and contains no interactions between the treatment and other covariates. So, whether we ask about the ATE or the ATET, the answer is 0.92. Whether we ask about the expected additional weightloss for a person who chose to participate or about all the women who chose to participate, the answer is the same. No matter what, the expected change is always 0.92.

To make this interesting, we will need a more complex model. We could take the nointeract suboption off the entreat() option. If we did that and refit the model, all the questions above would produce different answers. But, as we said, our data are simulated with no interaction. So let's use another artifice.

Let's assume that the clerk in charge of the final weigh-in overheard management discussing the new program. The managers seemed particularly interested in participants losing at least 4 kg (8.8 pounds). Thinking he was being helpful, our clerk decided to save everyone some effort and did not record actual weights. Instead, he recorded only whether employees were at least 4 kg lighter than they had been at the initial weigh-in.

We can no longer analyze weight loss, but we can analyze the probability of losing at least 4 kg. We fit the same full model but this time use eprobit, and our dependent variable becomes lost4, which is 0 if the employee lost less than 4 kg and is 1 if the employee lost 4 kg or more.

```
. eprobit lost4 age i.sex, endogenous(weight = i.sex gym)
> entreat(wellpgm = age i.smoke, nointeract)
> select(completed = i.wellpgm experience i.salaried) vce(robust)
  (iteration log omitted)
Extended probit regression                      Number of obs     =        545
                                                      Selected    =        337
                                                      Nonselected =        208

                                                Wald chi2(4)      =     184.27
Log pseudolikelihood = -2392.5364               Prob > chi2       =     0.0000

------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lost4        |
         age |  -.0461113   .0129744    -3.55   0.000    -.0715406    -.020682
             |
         sex |
       male  |  -1.192968   .1806428    -6.60   0.000    -1.547022   -.8389148
      weight |   .1131467   .0108868    10.39   0.000     .0918089    .1344844
             |
     wellpgm |
        yes  |   1.370215   .4048158     3.38   0.001     .5767905    2.163639
       _cons |  -8.034426   1.199574    -6.70   0.000    -10.38555   -5.683305
-------------+----------------------------------------------------------------
completed    |
     wellpgm |
        yes  |   .6534203   .2310957     2.83   0.005      .200481     1.10636
  experience |   -.801973   .0676059   -11.86   0.000    -.9344781   -.6694679
             |
    salaried |
        yes  |   .3955088   .1549943     2.55   0.011     .0917255    .6992921
       _cons |   4.862419   .4186367    11.61   0.000     4.041906    5.682932
-------------+----------------------------------------------------------------
wellpgm      |
         age |  -.0958611   .0071251   -13.45   0.000     -.109826   -.0818963
             |
       smoke |
        yes  |  -1.515911   .1754356    -8.64   0.000    -1.859758   -1.172063
       _cons |   4.310847    .338842    12.72   0.000     3.646728    4.974965
-------------+----------------------------------------------------------------
weight       |
         sex |
       male  |   9.501602   .6983151    13.61   0.000      8.13293    10.87028
         gym |  -.8162669   .0765488   -10.66   0.000    -.9662998    -.666234
       _cons |   80.09771   .5302486   151.06   0.000     79.05844    81.13697
-------------+----------------------------------------------------------------
var(e.weight)|   65.98399   3.805168                      58.93203     73.8798
-------------+----------------------------------------------------------------
corr(e.com~d,|
    e.lost4) |   .5236573   .1297834     4.03   0.000     .2268709    .7314522
corr(e.wel~m,|
    e.lost4) |    .249717   .2438804     1.02   0.306    -.2493086    .6439525
corr(e.wei~t,|
    e.lost4) |  -.6846067    .096236    -7.11   0.000    -.8314263    -.448426
corr(e.wel~m,|
 e.completed)|   .3678357   .1636913     2.25   0.025      .014886    .6392761
corr(e.wei~t,|
 e.completed)|  -.0821217    .074566    -1.10   0.271    -.2255026    .0647412
corr(e.wei~t,|
   e.wellpgm)|  -.0888819   .0671873    -1.32   0.186     -.218281    .0435887
------------------------------------------------------------------------------
```

These parameter estimates are pretty close to those from running eregress on weightloss. But unless you like thinking in terms of shifts along a standardized normal distribution, the coefficient of 1.37 on wellpgm is difficult to interpret. We still know that the effect of the program is statistically significant, but little more.

Note that we added vce(robust). This will allow us to treat our sample as a draw from a population when using estat teffects and margins, and thus make inferences about the population. Otherwise, we would be taking the sample as fixed and not as a draw from a population.

If management is thinking about expanding the program, they will want to evaluate its effectiveness. What proportion of employees across all facilities would lose 4 kg or more naturally, either through all employees not participating or through the program simply not being offered? What proportion would lose 4 kg or more if all employees participated? More to the point, what is the difference in those averages? We type

```
. estat teffects

Predictive margins                              Number of obs     =        545

---------------------------------------------------------------------------
             |            Unconditional
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+-------------------------------------------------------------
ATE          |
     wellpgm |
(yes vs no)  |   .3857447   .1195805     3.23   0.001     .1513712    .6201182
---------------------------------------------------------------------------
```

Only about 40% of employees would be expected to lose 4 kg; that is the ATE.

A related question is, What is the expected average increase in participants losing 4 kg? Let's estimate the expected effect of the wellness program on just those employees who choose to participate, the ATET.

```
. estat teffects, atet

Predictive margins                              Number of obs     =        545
                                                Subpop. no. obs   =        208

---------------------------------------------------------------------------
             |            Unconditional
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+-------------------------------------------------------------
ATET         |
     wellpgm |
(yes vs no)  |   .5335926   .1322879     4.03   0.000      .274313    .7928722
---------------------------------------------------------------------------
```

The ATET of 0.53 implies that just over half of those who choose to participate across all facilities would be expected to lose 4 kg. Recall that we believed success in the program would be positively correlated with employees' decision to participate. That is what made the decision endogenous. It is not surprising that we expect better results for participants than we do for all the employees as a whole.

We are going to need margins to answer some other questions, so let's introduce it by reestimating the ATET. First though, we generate a copy of the wellpgm variable in wellpgmT; margins will need it.

```
. generate wellpgmT = wellpgm

. margins r(0 1).wellpgm if  wellpgm, predict(base(wellpgm=wellpgmT))
> contrast(effects nowald)

Contrasts of predictive margins
Model VCE    : Robust

Expression   : Pr(lost4==1), predict(base(wellpgm=wellpgmT))

------------------------------------------------------------------------------
             |            Delta-method
             |   Contrast   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     wellpgm |
 (yes vs no) |   .5335926     .13197     4.04   0.000     .2749361    .7922491
------------------------------------------------------------------------------
```

We have reproduced the estimate.

There is a lot happening in that `margins` command.

`r(0 1).wellpgm` tells `margins` to form two counterfactuals—one at `wellpgm=0` and another at `wellpgm=1`—and to then take the reference (r) contrast (difference) of those two counterfactuals.

`if wellpgm` restricts the sample to those who participated in the wellness program.

`predict(base(wellpgm=wellpgmT))` specifies that each counterfactual prediction be conditioned on the employee's actual decision to participate in the program. These values are recorded in `wellpgmT`. Recall that `margins` changes the data to form the counterfactuals, and thus `predict` must be told where to find the employee's actual choice. The use and meaning of `predict(base())` are discussed more in [ERM] **intro 6**.

`contrast(effects nowald)` tells `margins` to report the z statistic and probability $> z$, which are not shown by default. It also tells `margins` to suppress the overall Wald statistic.

The standard errors are slightly smaller than those from `estat teffects`. If we wanted them to match exactly, we would use the `vce(unconditional)` option with `margins`. That option creates standard errors appropriate to make inferences about the population. The standard errors are so close that we will dispense with `vce(unconditional)` in this section.

Now, let's ask a series of different questions from a different perspective.

The physical trainer for our fictional company is having lunch with a new employee, Betty. The trainer mentions the wellness program, and Betty asks if it is likely to do her much good. Betty looks to be mid thirties and average weight. She says she goes to the gym a couple of times a month. The trainer recalls people with those characteristics doing well with the program. Betty's data are already in the company's database, so the trainer opens Stata on her laptop and types

```
. margins r(0 1).wellpgm if name=="Betty", predict(fix(wellpgm))
> contrast(effects nowald) noesample
Warning: prediction constant over observations.

Contrasts of predictive margins
Model VCE    : Robust

Expression   : Pr(lost4==1), predict(fix(wellpgm))

------------------------------------------------------------------------------
             |            Delta-method
             |   Contrast   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     wellpgm |
 (yes vs no) |   .6472455   .1350621     4.79   0.000     .3825287    .9119622
------------------------------------------------------------------------------
```

The trainer tells Betty that employees with her characteristics have about a 65% chance of losing 4 kg when they are in the program.

Later, another new employee, Fred, asks whether the program is likely to help him lose that last few kilograms. He is thin, in his upper fifties, and he already goes to the gym about twice a week. Our trainer types

```
. margins r(0 1).wellpgm if name=="Fred", predict(fix(wellpgm))
> contrast(effects nowald) noesample
Warning: prediction constant over observations.

Contrasts of predictive margins
Model VCE    : Robust

Expression   : Pr(lost4==1), predict(fix(wellpgm))

------------------------------------------------------------------------------
             |            Delta-method
             |   Contrast   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     wellpgm |
 (yes vs no) |   .0184247   .0225112     0.82   0.413    -.0256965    .0625459
------------------------------------------------------------------------------
```

She tells Fred that the program might be good for him but not to expect it to create much weight loss. Fred says he would like to sign up, just so he can meet some other employees.

When Fred leaves, our trainer calls her office mate and makes a wager that Fred will not lose 4 kg on the program. The trainer then realizes that she placed a bet on overall weight loss, not just the loss attributable to the wellness program. To be certain, she checks the potential outcomes of weight loss for Fred being in the program and for Fred being out of the program.

```
. margins i(0 1).wellpgm if name=="Fred", predict(fix(wellpgm)) noesample
Warning: prediction constant over observations.

Predictive margins                                Number of obs     =          1
Model VCE    : Robust

Expression   : Pr(lost4==1), predict(fix(wellpgm))

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     wellpgm |
          no |   .0000365   .0000718     0.51   0.612    -.0001043    .0001773
         yes |   .0184612   .0225357     0.82   0.413     -.025708    .0626304
------------------------------------------------------------------------------
```

With a negligible chance of losing 4 kg if Fred chooses not to participate and a slim 2% chance if Fred does participate, our trainer feels pretty good about her wager. Even the upper bound of the confidence intervals makes the trainer confident. Of course, these are the expected results for all employees with Fred's characteristics; Fred might be an overachiever.

Note that our trainer used `predict(fix(wellpgm))` to answer all of these questions. That is both the right and the only thing to do. Neither of these new employees has yet made a choice whether to participate. They have not revealed their unobserved characteristics that cause their weight loss and their decision to participate to be correlated. We called this unobserved characteristic the "willingness to engage in healthy behaviors" when we described the model for our data. Unlike when we computed ATET, we do not yet know Fred's and Betty's treatment choices, so we cannot use `base()` and thus condition our inferences on that additional information. We can make statements only about fixed levels of treatment.

The counterfactuals and contrasts that we computed for Betty and Fred are the expected values from our model conditioned only on the exogenous covariates in the main equation, `age` and `sex`, and on fixing the values of `wellpgm` first to 0 and then to 1. By “fixing”, we mean setting them to 0 and 1, not letting Betty or Fred choose 0 or 1. These estimates are no better or worse than the ATETs we estimated using `base()`. They are based on less information but use all the information we have about Betty and Fred. The estimates for Betty are what we would expect if we averaged over hundreds of employees who match Betty’s `age` and `sex`. The same applies to Fred.

Also note that we typed `r(0 1).`, rather than just `r.`. That is because we are operating on a single observation, and `margins` cannot determine the appropriate levels of `wellpgm` for which to form counterfactuals. We had to tell `margins` to use 0 and 1.

It is unlikely that our trainer has Stata on her laptop or has the inclination to type `margins` commands. As analysts, however, we might create a table for her that she can use to assess candidates and help employees form realistic expectations.

Our dataset already has grouping variables for `age`, `gym`, `weight`, and `sex`. We can estimate the expected additional probability of losing more than 4 kg for each combination of these groups by using an `over()` option.

```
. margins r.wellpgm, predict(fix(wellpgm)) contrast(effects nowald)
> over(agegrp gymgrp wtgrp sex)

Contrasts of predictive margins
Model VCE    : Robust

Expression   : Pr(lost4==1), predict(fix(wellpgm))
over         : agegrp gymgrp wtgrp sex

------------------------------------------------------------------------------
             |            Delta-method
             |   Contrast   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
wellpgm@ag~p#|
     gymgrp# |
   wtgrp#sex |
(yes vs no)  |
    20-29 0  |
      60-69  |
      female |   .6430664   .1601075     4.02   0.000     .3292614    .9568714
(yes vs no)  |
    20-29 0  |
  60-69 male |          0  (omitted)
  (output omitted)
(yes vs no)  |
60 up 11 up  |
< 60 female  |   .0032702   .0059912     0.55   0.585    -.0084723    .0150128
(yes vs no)  |
60 up 11 up  |
    < 60 male|          0  (omitted)
------------------------------------------------------------------------------
```

Those rows marked `(omitted)` represent combinations of characteristics for which we do not have any employees in our sample. We could use our model to extrapolate to those groups, but we are not going to do that. What we do have for each combination of groups is an estimate of the expected increase in the probability of losing 4 kg, a test that the probability is greater than 0, and a 95% confidence interval.

Those results will take a lot of transcription to create something compact for the trainer. And while our hearts are warmed by the tests and confidence intervals, the trainer might not feel the same way. If we wanted to be exceptionally helpful, we could build a table manually showing ATEs for each group.

```
. predict te, te

. table gymgrp wtgrp sex, by(agegrp) contents(mean te) format(%4.2f)

-------------------------------------------------------------------------------
Age    |
groups |
and    |
Gym    |            Employee sex; 0=female, 1=male and Weight groups
visit  | ------------- female ------------    -------------- male -------------
groups |  < 60  60-69  70-79  80-89  90 up     < 60  60-69  70-79  80-89  90 up
-------+-----------------------------------------------------------------------
20-29  |
     0 |         0.64   0.58   0.53   0.41                   0.63   0.62   0.52
   0-5 |         0.65   0.63   0.57                          0.64   0.65   0.59
  6-10 |         0.51   0.64   0.65                          0.55   0.58   0.64
 11 up |         0.40                                        0.35
-------+-----------------------------------------------------------------------
30-39  |
     0 |  0.48          0.65   0.62   0.61                   0.61   0.65   0.62
   0-5 |         0.48   0.64   0.64                                 0.56   0.62
  6-10 |  0.21   0.39   0.54   0.60                   0.27   0.31   0.51   0.48
 11 up |         0.34   0.43                                 0.24          0.53
-------+-----------------------------------------------------------------------
40-49  |
     0 |         0.45   0.57   0.62   0.65                   0.33   0.50   0.62
   0-5 |         0.39   0.48   0.59   0.61                   0.27   0.38   0.55
  6-10 |         0.22   0.33   0.38                   0.09   0.14   0.25   0.42
 11 up |  0.07   0.10          0.28                   0.09   0.08   0.19
-------+-----------------------------------------------------------------------
50-59  |
     0 |         0.26   0.35   0.46   0.62                   0.25   0.29   0.47
   0-5 |  0.06          0.22   0.40   0.62                   0.12   0.20   0.32
  6-10 |  0.05   0.04   0.22   0.16   0.39                   0.05   0.10   0.13
 11 up |  0.01   0.01   0.05                          0.01   0.01
-------+-----------------------------------------------------------------------
60 up  |
     0 |                0.17   0.25   0.37                   0.07   0.12   0.26
   0-5 |                0.07   0.33   0.28            0.02          0.03   0.11
  6-10 |         0.02   0.04   0.10                   0.01   0.02   0.03   0.08
 11 up |  0.00          0.02                                 0.00   0.02   0.03
-------------------------------------------------------------------------------
```

We first predicted the expected treatment effects for each observation in our sample. Then, we let `table` average those values for each combination of groups. For any combination of groups, these estimates match those from `margins`.

References

Cameron, A. C., and P. K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

——. 2010. *Microeconometrics Using Stata*. Rev. ed. College Station, TX: Stata Press.

Roodman, D. 2011. Fitting fully observed recursive mixed-process models with cmp. *Stata Journal* 11: 159–206.

Wooldridge, J. M. 2010. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. Cambridge, MA: MIT Press.

Also see

[ERM] **intro 1** — An introduction to the ERM commands

[ERM] **intro 6** — Model interpretation

[ERM] **Glossary**