Estimating effects from extended regression models

David M. Drukker

Executive Director of Econometrics
Stata

2017 Nordic and Baltic Stata Users Group meeting
Karolinska Institutet
1 September 2017

Extended regression models

Extended regression model (ERM) is a Stata term for a class of regression models

- The outcome can be continuous (linear), probit, orderded probit, or censored (tobit)
- Some of the covariates may be endogenous
    - The endogenous covariates may be continuous, probit, or ordered probit
- Endogenous sample-selection may be modeled
- Exogenous or endogenous treatment assignment may be modeled
- The new-in-Stata-15 commands `eregress`, `eprobit`, `eoprobit`, and `eintreg` fit ERMs

Extended regression models

- Some of the covariates may be endogenous
  - The endogenous covariates may be continuous, binary, or ordinal
  - Polynomial terms and interaction terms constructed from the endogenous covariates are allowed
  - Interactions among the endogenous covariates and interactions between the endogenous covariates and the exogenous covariates are allowed

Outline

- I cannot do justice to ERMs in this short talk
- I discuss examples in which I
  - define some of the terms that I have already used
  - illustrate some command syntax
  - illustrate how to estimate some effects using postestimation commands

- Fictional data on wellness program from large company

```
. use wprogram
. describe
Contains data from wprogram.dta
  obs:         3,000
 vars:             6                          28 Jul 2017 07:13
 size:        72,000
------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------------
wchange         float   %9.0g      changel    Weight change level
age             float   %9.0g                 Years over 50
over            float   %9.0g                 Overweight (tens of pounds)
phealth         float   %9.0g                 Prior health score
prog            float   %9.0g      yesno      Participate in wellness program
wtprog          float   %9.0g      yesno      Offered work time to participate
                                                in program
------------------------------------------------------------------------------
Sorted by:
```

- Three levels of `wchange`

```
. tabulate wchange prog
     Weight |   Participate in
     change |  wellness program
      level |        No        Yes |     Total
------------+----------------------+----------
       Loss |       239        909 |     1,148
  No change |       468        605 |     1,073
       Gain |       593        186 |       779
------------+----------------------+----------
      Total |     1,300      1,700 |     3,000
```

- Data are observational
- Table does not account for how observed covariates and/or unobserved errors that affect program participation also affect the outcome variable

- I want a model that
  - allows observed covariates to affect both `wchange` and assignment to `prog`
  - allows the errors that affect assignment to `prog` to be correlated with the errors that affect `wchange`
  - I suspect that unobservables that increase program participation are negatively correlated with unobservables that affect weight gain
- In other words, I want allow `prog` to be endogenous

If prog is endogenous, I must model the dependence.
Consider

$$wchange = \begin{cases} \text{``}\textit{Loss}\text{''} & \text{if } \qquad \beta_1\texttt{prog} + \mathbf{x}\boldsymbol{\beta} + \epsilon \leq cut1 \\ \text{``}\textit{No change}\text{''} & \text{if } cut1 < \beta_1\texttt{prog} + \mathbf{x}\boldsymbol{\beta} + \epsilon \leq cut2 \\ \text{``}\textit{Gain}\text{''} & \text{if } cut2 < \beta_1\texttt{prog} + \mathbf{x}\boldsymbol{\beta} + \epsilon \end{cases}$$

$$prog = (\mathbf{x}\boldsymbol{\gamma} + \gamma_1\texttt{wtime} + \eta > 0)$$

$$\epsilon \text{ and } \eta \text{ are correlated and joint normal}$$

$$\mathbf{x}\boldsymbol{\beta} = \beta_2\texttt{age} + \beta_3\texttt{over} + \beta_4\texttt{phealth}$$

$$\mathbf{x}\boldsymbol{\gamma} = \gamma_2\texttt{age} + \gamma_3\texttt{over} + \gamma_4\texttt{phealth}$$

- wtime is an instrumental variable
    - It is included in the model for treatment
    - It is excluded from the model for the potential outcomes of wchange

$$
wchange = \begin{cases} \text{“}\textit{Loss}\text{”} & \text{if} \quad \beta_1\texttt{prog} + \mathbf{x}\boldsymbol{\beta} + \epsilon \leq cut1 \\ \text{“}\textit{No change}\text{”} & \text{if } cut1 < \beta_1\texttt{prog} + \mathbf{x}\boldsymbol{\beta} + \epsilon \leq cut2 \\ \text{“}\textit{Gain}\text{”} & \text{if } cut2 < \beta_1\texttt{prog} + \mathbf{x}\boldsymbol{\beta} + \epsilon \end{cases}
$$

$$
prog = (\mathbf{x}\boldsymbol{\gamma} + \gamma_1\texttt{wtime} + \eta > 0)
$$

ϵ and η are correlated and joint normal

$$
\mathbf{x}\boldsymbol{\beta} = \beta_2\texttt{age} + \beta_3\texttt{over} + \beta_4\texttt{phealth}
$$

$$
\mathbf{x}\boldsymbol{\gamma} = \gamma_2\texttt{age} + \gamma_3\texttt{over} + \gamma_4\texttt{phealth}
$$

Fit by:
```
eoprobit wchange age over phealth ,
      endog(prog = age over phealth wtime, probit)
```

```
. eoprobit wchange age over phealth ,                          ///
>         endog(prog = age over phealth wtprog, probit) ///
>         vsquish nolog
Extended ordered probit regression                Number of obs     =      3,000
                                                  Wald chi2(4)      =     409.97
Log likelihood = -4401.0952                       Prob > chi2       =     0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
wchange      |
         age |   .2155906   .0705048     3.06   0.002     .0774037    .3537776
        over |   .4349946   .0387185    11.23   0.000     .3591078    .5108814
     phealth |  -.4933361   .0411866   -11.98   0.000    -.5740603    -.412612
        prog |
         Yes |  -.3624996   .1031408    -3.51   0.000    -.5646519   -.1603473
-------------+----------------------------------------------------------------
prog         |
         age |  -.9341234   .0840002   -11.12   0.000    -1.098761   -.7694861
        over |  -1.058621   .0514252   -20.59   0.000    -1.159412   -.9578294
     phealth |   .9001108   .0504804    17.83   0.000      .801171    .9990507
      wtprog |   1.631615   .0780834    20.90   0.000     1.478574    1.784656
       _cons |   .0090842   .0535434     0.17   0.865     -.095859    .1140274
-------------+----------------------------------------------------------------
/wchange     |
        cut1 |  -.5897304   .0781626                     -.7429264   -.4365345
        cut2 |   .5029323    .068292                      .3690825    .6367821
-------------+----------------------------------------------------------------
  corr(e.prog,|
   e.wchange)|  -.3478179   .0604422    -5.75   0.000    -.4603282   -.2243109
------------------------------------------------------------------------------
```

```
Log likelihood = -4401.0952                         Prob > chi2       =     0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
wchange      |
         age |   .2155906   .0705048     3.06   0.002     .0774037    .3537776
        over |   .4349946   .0387185    11.23   0.000     .3591078    .5108814
     phealth |  -.4933361   .0411866   -11.98   0.000    -.5740603    -.412612
        prog |
         Yes |  -.3624996   .1031408    -3.51   0.000    -.5646519   -.1603473
-------------+----------------------------------------------------------------
prog         |
         age |  -.9341234   .0840002   -11.12   0.000    -1.098761   -.7694861
        over |  -1.058621   .0514252   -20.59   0.000    -1.159412   -.9578294
     phealth |   .9001108   .0504804    17.83   0.000      .801171    .9990507
      wtprog |   1.631615   .0780834    20.90   0.000     1.478574    1.784656
       _cons |   .0090842   .0535434     0.17   0.865     -.095859    .1140274
-------------+----------------------------------------------------------------
/wchange     |
        cut1 |  -.5897304   .0781626                     -.7429264   -.4365345
        cut2 |   .5029323    .068292                      .3690825    .6367821
-------------+----------------------------------------------------------------
  corr(e.prog,|
   e.wchange)|  -.3478179   .0604422    -5.75   0.000    -.4603282   -.2243109
------------------------------------------------------------------------------
```

- The coefficient on wtprog and its standard error give the impression that the instrument is relevant

```
           cut2 |   .5029323    .068292                          .3690825    .6367821
----------------+----------------------------------------------------------------
 corr(e.prog,   |
     e.wchange) |  -.3478179   .0604422     -5.75   0.000    -.4603282   -.2243109
---------------------------------------------------------------------------------
```

- The nonzero correlation between `e.prog` and `e.wchange` indicates that `prog` is endogenous
- Those who are more likely to participate are more likely to lose weight

```
. margins r.prog,                                    ///
>         predict(fix(prog) outlevel("Loss"))        ///
>         predict(fix(prog) outlevel("No change"))   ///
>         predict(fix(prog) outlevel("Gain"))        ///
>         contrast(nowald)
Contrasts of predictive margins
Model VCE    : OIM
1._predict   : Pr(wchange==Loss), predict(fix(prog) outlevel("Loss"))
2._predict   : Pr(wchange==No change), predict(fix(prog) outlevel("No
               change"))
3._predict   : Pr(wchange==Gain), predict(fix(prog) outlevel("Gain"))
--------------------------------------------------------------
               |             Delta-method
               |   Contrast   Std. Err.     [95% Conf. Interval]
---------------+----------------------------------------------
 prog@_predict |
(Yes vs No) 1  |   .1259899   .0356631      .0560914    .1958883
(Yes vs No) 2  |  -.0185024   .0055583     -.0293965   -.0076084
(Yes vs No) 3  |  -.1074874   .0306512     -.1675628   -.0474121
--------------------------------------------------------------
```

- When everyone joins the program instead of when no one participants in the program,
  - On average, the probablity of “Loss” goes up by .13
  - On average, the probablity of “No change” goes down by .02
  - On average, the probablity of “Gain” goes down by .11

- `fix(prog)` gets us the effect of the program that is not contaminated by the selection effect/correlation between ϵ and η that increases the participation among people more likely to lose weight
- `predict(fix(prog))` tells `margins` to specify `fix(prog)` to `predict` when computing each predicted probability

- `fix(prog)` causes the value of `prog` not to affect ϵ, even though they are correlated
    - `fix(prog)` specifies that the part of ϵ that is correlated with `y2` be integrated out

- This type of prediction is sometimes called the structural prediction or an average structural function; see Blundell and Powell (2003), Blundell and Powell (2004), Wooldridge (2010), and Wooldridge (2014),
- The difference between the mean of the average of the structural predictions when `prog=1` and the mean of the average of the structural predictions when `prog=0` is an average treatment effect (Blundell and Powell (2003) and Wooldridge (2014))

Standard errors for population versus sample

- The delta-method standard errors reported by `margins` hold the covariates fixed at their sample values
  - The delta-method standard errors are for a sample-average treatment effect instead of a population-averaged treatment effect
  - The sample-averaged treatment effect is for those individuals that showed up in that run of the treatment
  - The population-averaged treatment effect is for a random draw of individuals from the population
- To get standard errors for the population-average treatment effect, specify `vce(robust)` to the estimation command and specify `vce(unconditional)` to `margins`

```
. quietly eoprobit wchange age over phealth ,                   ///
>         endog(prog = age over phealth wtprog, probit) ///
>         vce(robust)
. margins r.prog,                                        ///
>         predict(fix(prog) outlevel("Loss"))            ///
>         predict(fix(prog) outlevel("No change")) ///
>         predict(fix(prog) outlevel("Gain"))            ///
>         contrast(nowald) vce(unconditional)
Contrasts of predictive margins
1._predict    : Pr(wchange==Loss), predict(fix(prog) outlevel("Loss"))
2._predict    : Pr(wchange==No change), predict(fix(prog) outlevel("No
                change"))
3._predict    : Pr(wchange==Gain), predict(fix(prog) outlevel("Gain"))
─────────────────────────────────────────────────────────────────
               │            Unconditional
               │   Contrast   Std. Err.     [95% Conf. Interval]
───────────────┼─────────────────────────────────────────────────
 prog@_predict │
(Yes vs No) 1  │   .1259899   .0349061      .0575753    .1944045
(Yes vs No) 2  │  -.0185024   .0054389     -.0291624   -.0078424
(Yes vs No) 3  │  -.1074874   .0300866     -.1664561   -.0485188
─────────────────────────────────────────────────────────────────
. matrix b = r(b)
```

More about ERM commands

- The commands `eregress`, `eprobit`, and `eintreg` fit ERMs handle continuous-and-unbounded, binary, and censored/corner outcomes
- Look at

  http://www.stata.com/manuals/erm.pdf

  for more examples and a wealth of details

Blundell, R. W., and J. L. Powell. 2003. Endogeity in nonparametric and semiparametric regression models. In *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, ed. M. Dewatripont, L. P. Hansen, and S. J. Turnovsky, vol. 2, 312–357. Cambridge: Cambridge University Press.

———. 2004. Endogeneity in semiparametric binary response models. *Review of Economic Studies* 71: 655–679.

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

———. 2014. Quasi-maximum likelihood estimation and testing for nonlinear models with endogenous explanatory variables. *Journal of Econometrics* 182: 226–234.