STATA Features

ERMs
Endogeneity + Selection + Treatment + Panel Data

- Extended regression model commands fit
  - Linear models
  - Linear models with interval-censored outcomes, including tobit models
  - Probit models
  - Ordered probit models
- With any combination of
  - Endogenous covariates—continuous, binary, or ordinal
  - Sample selection
  - Nonrandom treatment assignment, both exogenous and endogenous
  - Panel data

Features in discipline-specific terminology:
  - Bias due to unmeasured confounding
  - Trials with informative dropout
  - Average causal effects (ACEs)
  - Average treatment effects (ATEs)
  - Simultaneous causality in linear models
  - Outcomes that are missing not at random (MNAR)
  - Nonignorable nonresponse
  - Selection on unobservables
  - Heckman selection
  - Longitudinal data
  - Two-level data
  - Random effects

Easy syntax

The syntax of ERMs is a command, such as **eregress**, followed by the main equation and then by one or more of the options **endogenous()**, **select()**, and **entreat()** or **extreat()**. The options may be specified in any combination. For instance,

```
                    Linear regression of y on x1 and x2
. eregress y x1 x2

                    Make covariate x2 endogenous
. eregress y x1   ,  endogenous(      x2 = x3 x4)

                    Add sample selection
. eregress y x1   ,  endogenous(      x2 = x3 x4)
                         select(selected = x2 x6)

                    Add exogenous treatment & drop sample selection
. eregress y x1   ,  endogenous(      x2 = x3 x4)
                         extreat(treated)

                    Replace exogenous with endogenous treatment
. eregress y x1   ,  endogenous(      x2 = x3 x4)
                         entreat( treated = x2 x3 x5)

                    Add sample selection
. eregress y x1   ,  endogenous(      x2 = x3 x4)
                         entreat( treated = x2 x3 x5)
                         select(selected = x2 x6)

                    Add panel data
. xteregress y x1 ,  endogenous(      x2 = x3 x4)
                         entreat( treated = x2 x3 x5)
                         select(selected = x2 x6)
```
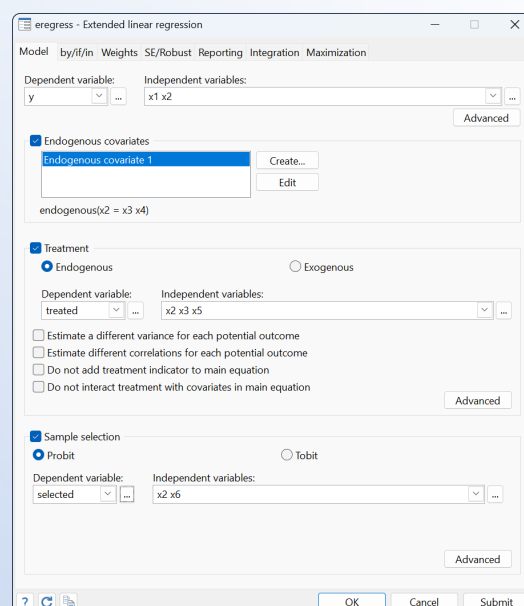

The same syntax that works with **eregress** and **xteregress** to fit linear regression models also works with **eintreg** and **xteintreg** to fit interval regression models, **eprobit** and **xteprobit** to fit probit models, and **eoprobit** and **xteoprobit** to fit ordered probit models. For instance,

```
. eprobit y x1 , endogenous(   x2 = x3 x4)
              entreat( treated = x2 x2 x5)
              select( selected = x2 x6)
```

Fit your model

We could fit a model of graduation rates by typing

```
. eprobit graduate income i.roommate,
        entreat(program = i.campus income)
        endogenous(hsgpa = income i.hscomp)
```

Our main equation models college graduation (0/1 variable **graduate**) as determined by parent's income (**income**), whether the student had a roommate who was also a student (**i.roommate**), and high school GPA (**hsgpa**). High school GPA is an endogenous covariate that itself is determined by parent's income and an ordinal variable measuring high school competitiveness (**i.hscomp**). The school offers a voluntary freshman program designed to increase graduation rates (**program**) that is modeled as an endogenous treatment determined by living on campus (**i.campus**) and parent's income.

Explore the results

After fitting this model, we could estimate the graduation rates for the subpopulations of students in four high school GPA groups by typing

```
Viewer - view erm1.smcl                                     —   □   ×
view erm1.smcl  ×
+                                   Dialog ▾  Also see ▾  Jump to ▾

. margins, over(hsgpagrp)

Predictive margins                          Number of obs = 2,500
Model VCE: OIM

Expression: Average structural function probability, predict()
Over:       hsgpagrp

                      Delta-method
            Margin    std. err.      z    P>|z|   [95% conf. interval]

hsgpagrp
   < 2.5   .1104557   .0122844     8.99   0.000   .0863787   .1345327
 2.5-2.9   .4648163   .0105969    43.86   0.000   .4440468   .4855859
 3.0-3.4   .8370201   .0086732    96.51   0.000   .8200209   .8540194
 3.5-4.0   .9882793   .0027071   365.06   0.000   .9829734   .9935852

                                                      CAP  NUM  INS
```

We see that the graduation rate is estimated to be 11% for those with high school GPAs below 2.5, with a 95% CI from 8.6% to 13.5%. That estimate rises all the way to 99% for those with GPAs from 3.5 to 4.0.

We could even explore the graduation rates as a function of program participation, parent's income, and high school GPA.

```
. margins, over(program incomegrp hsgpagrp)
  (output omitted)
. marginsplot
```

The resulting graph shows the expected graduation rates for those who took the program (in red) and those who did not take the program (in blue). The four panels are GPA groups. The *x* axis of each graph is parent's income.

Advanced inferences

- **Inference statistics**
  - Expected means
  - Expected probabilities
  - Contrasts (differences) of expected means and probabilities (also called effects)
  - Marginal effects
  - Average treatment effects (ATEs)
  - ATEs on the treated (ATETs)
  - Potential-outcome means (POMs)
  - Average structural function (ASF) means and effects
  - Average structural probability (ASP) means and effects
- **Estimates of statistics are available for**
  - Full population
  - Subpopulations
  - Expected values for specific covariate values
  - Censored and uncensored outcomes
- **Conditional analysis—specify values of all covariates**
- **Population-averaged analysis—specify values of some or no covariates, and average (margin) over the rest**
- **Inference types**
  - Tests against zero or any other value
  - Test of equality
  - Contrasts
  - Pairwise comparisons
  - Confidence intervals for every statistic
- **Profile plots**
  - Any inference statistic
  - Any statistic over subpopulations or subgroups (e.g., age groups or treatment levels)
  - Any statistic at multiple fixed levels of one or more covariates
  - Confidence intervals

Probability of graduating with 95% CI
By retention program, high school GPA, and parent's income