

## intro 1 — An introduction to the ERM commands

[Description](#)[Remarks and examples](#)[Also see](#)

## Description

ERM stands for extended regression model. It is our term to designate commands for fitting linear regression, interval regression, probit, and ordered probit models that allow

- continuous, binary, and ordinal endogenous covariates,
- polynomials of endogenous covariates,
- interactions of endogenous covariates,
- interactions of endogenous with exogenous covariates,
- endogenous sample selection, and
- nonrandom exogenous or endogenous treatment assignment.

The features may be used separately or in any combination.

The estimation commands `eregress`, `eintreg`, `eprobit`, and `eoprobit` fit ERMs.

## Remarks and examples

stata.com

Remarks are presented under the following headings:

*The problems ERMs solve*  
*The simple syntax of ERMs*  
*Normality assumption underlying ERMs*  
*Learning more about ERMs*

## The problems ERMs solve

The ERM commands fit the following models:

Command	Purpose
<code>eregress</code>	linear regression
<code>eintreg</code>	interval regression
<code>eprobit</code>	binary-outcome probit regression
<code>eoprobit</code>	ordinal-outcome probit regression

These models are described in [\[ERM\] intro 2](#).

All the ERM commands provide the following features:

- **Endogenous covariates**

Explanatory variables in the model—covariates—can be exogenous or endogenous.

Endogenous covariates can themselves be continuous (linear), binary (probit), or ordinal (ordered probit).

Endogenous covariates can be interacted with other covariates, whether endogenous or exogenous. They can even be interacted with themselves to form polynomials.

Endogenous covariates can themselves be predicted by other endogenous covariates.

- **Endogenous selection**

Models can be adjusted for situations in which outcomes are unobserved for endogenous reasons.

In a medical trial, patients may skip the final visit, causing the final outcome to be unobserved. They may skip it for reasons correlated with the outcome.

In economic data, wages are observed only for those who have a job. Those who do not have a job may not for reasons correlated with the wage they would have received.

- **Exogenous or endogenous treatment assignment**

The purpose of models is often to measure the effect of a treatment, such as a drug that is administered or a training program that is attended. Ethics often prevent assignment from being random.

In a medical trial, doctors might assign patients most likely to benefit to a trial based on observed characteristics. That is called exogenous treatment assignment.

In another situation, subjects may volunteer, and subjects who perceive larger benefits will be more likely to benefit. If all the determinants of the perceptions are observed, then assignment is exogenous. It can be explained by the observed variables, just as in the previous case.

If the determinants are unobserved, then treatment is endogenous. Errors in the assignment equation will be correlated with errors in the outcome equation.

Stata has other commands that address each of these issues in the case of linear regression, and it has still other commands that can address some of these issues for interval regression, probit, and ordered probit. But Stata has no other commands that can adjust for all the above when they occur together. Even if your problem has only one of the issues, you may still prefer to use the ERM commands because they all have the same simple syntax.

## The simple syntax of ERMs

The basic syntax of the ERM commands is Stata's standard estimation syntax: the command followed by the dependent variable followed by the covariates. Typing

```
. eregress y1 x1 x2
```

fits a linear regression of  $y_1$  on  $x_1$  and  $x_2$ . If you need to use one or more ERM features, you add options to the command.

Option	Purpose
<code>endogenous()</code>	add endogenous covariates
<code>select()</code>	add endogenous sample selection
<code>tobitselect()</code>	add endogenous selection using tobit
<code>extreat()</code>	add exogenous treatment assignment
<code>entreat()</code>	add endogenous treatment assignment

You can type

```
. eregress y x1 x2, endogenous(w = x1 z1 z2)
```

to add endogenous covariate `w` to the right-hand side of the model. The option specifies that `w`'s instruments are variables `x1`, `z1`, and `z2`.

If you did not observe `y` but observed `y0` and `y1`, where  $y_0 \leq y \leq y_1$ , you could fit the equivalent interval regression by typing

```
. eintreg y0 y1 x1 x2, endogenous(w = x1 z1 z2)
```

If you observed `y` but it contained a 0/1 binary outcome, you could fit the equivalent probit model by typing

```
. eprobit y x1 x2, endogenous(w = x1 z1 z2)
```

If `y` contained 1, 2, or 3 for ordered categories, such as not ambulatory, partially ambulatory, and fully ambulatory, you could fit the equivalent ordered probit model by typing

```
. eoprobit y x1 x2, endogenous(w = x1 z1 z2)
```

Syntax is the same regardless of model fit.

Now, let's imagine that the outcome `y` is observed only when variable `selected` is true (that is, not equal to 0). Consider a case where the outcome is observed when

$$\gamma_0 + \gamma_1 x_2 + \gamma_2 w + e \cdot \text{selected} > 0$$

and, just to make the problem more complicated, assume that `w` is endogenous. To fit the model with this added complication, type

```
. eregress y x1 x2, endogenous(w = x1 z1 z2) select(selected = x2 w)
```

You would use the same syntax with the other ERM commands:

```
. eintreg y0 y1 x1 x2, endogenous(w = x1 z1 z2) select(selected = x2 w)
```

```
. eprobit y x1 x2, endogenous(w = x1 z1 z2) select(selected = x2 w)
```

```
. eoprobit y x1 x2, endogenous(w = x1 z1 z2) select(selected = x2 w)
```

Now, let's complicate the model even more. We also have the variable `treatment`, which records whether the observation was treated. `treatment` also affects `y`. In fact, measuring the effect of `treatment` is the primary reason we are fitting this model. Type

```
. eregress y x1 x2, endogenous(w = x1 z1 z2) select(selected = x2 w) ///
  extreat(treatment)
```

Option `extreat()` handles exogenous treatment. Exogenous treatment is more flexible than you might expect. It handles assignment based on all the covariates used in the model, which in this case are `x1`, `x2`, and `w`.

But let us assume in our data that subjects volunteered. Or perhaps health care professionals assigned subjects to being treated based on information not in the model. That would be reasonable: doctors meet their patients and so know more about them than what is recorded in our data. In any case, we will assume that treatment is a function of observed variables  $w$ ,  $z_2$ , and  $z_3$ , and we will assume that the error in the treatment equation is correlated with the error in the outcome equation. It is that last assumption that handles doctors knowing more about their patients than what is recorded in our data. To fit the model, we type

```
. eregress y x1 x2, endogenous(w = x1 z1 z2) select(selected = x2 w) ///
   entreat(treatment = w z2 z3)
```

We changed from exogenous to endogenous treatment by swapping option `extreat()` for `entreat()`.

Shall we continue? We are just trying to convince you how flexible ERMs are and how simple the syntax is to fit them. We will go one more step. Let's assume that  $y$  is not continuous but is ordinal.  $y$  contains 1, 2, and 3, meaning not ambulatory, partially ambulatory, and fully ambulatory. In that case, we type

```
. eoprobit y x1 x2, endogenous(w = x1 z1 z2) select(selected = x2 w) ///
   entreat(treatment = w z2 z3)
```

### Normality assumption underlying ERMs

If you are accustomed to fitting models with `regress` and `ivregress`, you expect that results do not require that the errors be normally distributed. They merely require that they be independent and identically distributed.

The results produced by ERMs share that feature when all the equations are linear. Linear excludes `eintreg`, `eprobit`, and `eoprobit`, as well as endogenous selection and endogenous treatment, both of which depend on a secondary probit model.

The nonlinear models that ERMs fit depend on normality.

### Learning more about ERMs

What follows is a useful footnote. Other Stata commands provide a subset of the features that ERMs provide. We list them below. We will discuss ERMs more in this manual, but ERMs provide so many statistical features that we do not tell you as much about them as you would like. If you would like to know more, read the documentation for the other commands and then use the ERM commands.

`eregress` provides the features of

Feature	Command
linear regression	<code>regress</code>
instrumental variables	<code>ivregress</code>
exogenous treatment assignment	<code>teffects ra</code>
endogenous treatment assignment	<code>eteffects</code> and <code>etregress</code>
endogenous sample selection	<code>heckman</code>

`eintreg` provides the features of

Feature	Command
interval regression	<code>intreg</code>
tobit regression	<code>tobit</code>
instrumental-variables interval regression	—
instrumental-variables tobit regression	<code>ivtobit</code>
exogenous treatment assignment	—
endogenous treatment assignment	—
endogenous sample selection	—

`eprobit` provides the features of

Feature	Command
probit regression	<code>probit</code>
instrumental variables	<code>ivprobit</code>
exogenous treatment assignment	<code>teffects ra</code>
endogenous treatment assignment	—
endogenous sample selection	<code>heckprobit</code>

`eoprobit` provides the features of

Feature	Command
ordered probit regression	<code>oprobit</code>
instrumental variables	—
exogenous treatment assignment	—
endogenous treatment assignment	—
endogenous sample selection	<code>heckkoprobit</code>

## Also see

[ERM] [intro 2](#) — The models that ERMs fit

[ERM] [intro 7](#) — A Rosetta stone for extended regression commands

[ERM] [intro 8](#) — Conceptual introduction via worked example