# Title

stata.com

**eregress —** Extended linear regression



## Description

`eregress` fits a linear regression model that accommodates any combination of endogenous covariates, nonrandom treatment assignment, and endogenous sample selection. Continuous, binary, and ordinal endogenous covariates are allowed. Treatment assignment may be endogenous or exogenous. A probit or tobit model may be used to account for endogenous sample selection.

## Quick start

Regression of `y` on `x` with continuous endogenous covariate `y2` modeled by `x` and `z`

```
eregress y x, endogenous(y2 = x z)
```

As above, but adding continuous endogenous covariate `y3` modeled by `x` and `z2`

```
eregress y x, endogenous(y2 = x z) endogenous(y3 = x z2)
```

Regression of `y` on `x` with binary endogenous covariate `d` modeled by `x` and `z`

```
eregress y x, endogenous(d = x z, probit)
```

Regression of `y` on `x` with endogenous treatment recorded in `trtvar` and modeled by `x` and `z`

```
eregress y x, entreat(trtvar = x z)
```

Regression of `y` on `x` with exogenous treatment recorded in `trtvar`

```
eregress y x, extreat(trtvar)
```

Regression of `y` on `x` with endogenous sample-selection indicator `selvar` modeled by `x` and `z`

```
eregress y x, select(selvar = x z)
```

As above, but adding endogenous covariate `y2` modeled by `x` and `z2`

```
eregress y x, select(selvar = x z) endogenous(y2 = x z2)
```

As above, but adding endogenous treatment recorded in `trtvar` and modeled by `x` and `z3`

```
eregress y x, select(selvar = x z) endogenous(y2 = x z2)  ///
     entreat(trtvar = x z3)
```

## Menu

Statistics > Endogenous covariates > Models adding selection and treatment > Linear regression

# Syntax

*Basic linear regression with endogenous covariates*

```
eregress depvar [indepvars] ,
  endogenous(depvars_en = varlist_en) [options]
```

*Basic linear regression with endogenous treatment assignment*

```
eregress depvar [indepvars] ,
  entreat(depvar_tr [= varlist_tr]) [options]
```

*Basic linear regression with exogenous treatment assignment*

```
eregress depvar [indepvars] ,
  extreat(tvar) [options]
```

*Basic linear regression with sample selection*

```
eregress depvar [indepvars] ,
  select(depvar_s = varlist_s) [options]
```

*Basic linear regression with tobit sample selection*

```
eregress depvar [indepvars] ,
  tobitselect(depvar_s = varlist_s) [options]
```

*Linear regression combining endogenous covariates, treatment, and selection*

```
eregress depvar [indepvars] [if] [in] [weight] [, extensions options]
```

| *extensions* | Description |
|---|---|
| **Model** | |
| `endogenous(`*enspec*`)` | model for endogenous covariates; may be repeated |
| `entreat(`*entrspec*`)` | model for endogenous treatment assignment |
| `extreat(`*extrspec*`)` | exogenous treatment |
| `select(`*selspec*`)` | probit model for selection |
| `tobitselect(`*tselspec*`)` | tobit model for selection |

| *options* | Description |
|---|---|
| **Model** | |
| `noconstant` | suppress constant term |
| `offset(`*varname*$_o$`)` | include *varname*$_o$ in model with coefficient constrained to 1 |
| `constraints(`*numlist*`)` | apply specified linear constraints |
| `collinear` | keep collinear variables |
| **SE/Robust** | |
| `vce(`*vcetype*`)` | *vcetype* may be `oim`, `robust`, `cluster` *clustvar*, `opg`, `bootstrap`, or `jackknife` |
| **Reporting** | |
| `level(#)` | set confidence level; default is `level(95)` |
| `nocnsreport` | do not display constraints |
| *display_options* | control columns and column formats, row spacing, line width, display of omitted variables and base and empty cells, and factor-variable labeling |
| **Integration** | |
| `intpoints(#)` | set the number of integration (quadrature) points for integration over four or more dimensions; default is `intpoints(128)` |
| `triintpoints(#)` | set the number of integration (quadrature) points for integration over three dimensions; default is `triintpoints(10)` |
| **Maximization** | |
| *maximize_options* | control the maximization process; seldom used |
| `coeflegend` | display legend instead of statistics |

*enspec* is *depvars*$_{\text{en}}$ `=` *varlist*$_{\text{en}}$ [ `,` *enopts* ]

where *depvars*$_{\text{en}}$ is a list of endogenous covariates. Each variable in *depvars*$_{\text{en}}$ specifies an endogenous covariate model using the common *varlist*$_{\text{en}}$ and options.

*entrspec* is *depvar*$_{\text{tr}}$ [ `=` *varlist*$_{\text{tr}}$ ] [ `,` *entropts* ]

where *depvar*$_{\text{tr}}$ is a variable indicating treatment assignment. *varlist*$_{\text{tr}}$ is a list of covariates predicting treatment assignment.

*extrspec* is *tvar* [ `,` *extropts* ]

where *tvar* is a variable indicating treatment assignment.

*selspec* is *depvar*$_s$ = *varlist*$_s$ [ , noconstant offset(*varname*$_o$) ]

where *depvar*$_s$ is a variable indicating selection status. *depvar*$_s$ must be coded as 0, indicating that the observation was not selected, or 1, indicating that the observation was selected. *varlist*$_s$ is a list of covariates predicting selection.

*tselspec* is *depvar*$_s$ = *varlist*$_s$ [ , *tselopts* ]

where *depvar*$_s$ is a continuous variable. *varlist*$_s$ is a list of covariates predicting *depvar*$_s$. The censoring status of *depvar*$_s$ indicates selection, where a censored *depvar*$_s$ indicates that the observation was not selected and a noncensored *depvar*$_s$ indicates that the observation was selected.

| *enopts* | Description |
|---|---|
| Model | |
| probit | treat endogenous covariate as binary |
| oprobit | treat endogenous covariate as ordinal |
| povariance | estimate a different variance for each level of a binary or an ordinal endogenous covariate |
| pocorrelation | estimate different correlations for each level of a binary or an ordinal endogenous covariate |
| nomain | do not add endogenous covariate to main equation |
| noconstant | suppress constant term |

| *entropts* | Description |
|---|---|
| Model | |
| povariance | estimate a different variance for each potential outcome |
| pocorrelation | estimate different correlations for each potential outcome |
| nomain | do not add treatment indicator to main equation |
| nointeract | do not interact treatment with covariates in main equation |
| noconstant | suppress constant term |
| offset(*varname*$_o$) | include *varname*$_o$ in model with coefficient constrained to 1 |

| *extropts* | Description |
|---|---|
| Model | |
| povariance | estimate a different variance for each potential outcome |
| pocorrelation | estimate different correlations for each potential outcome |
| nomain | do not add treatment indicator to main equation |
| nointeract | do not interact treatment with covariates in main equation |

| *tselopts* | Description |
|---|---|
| Model | |
| ll(*varname* \| #) | left-censoring variable or limit |
| ul(*varname* \| #) | right-censoring variable or limit |
| main | add censored selection variable to main equation |
| noconstant | suppress constant term |
| offset(*varname*$_o$) | include *varname*$_o$ in model with coefficient constrained to 1 |

*indepvars*, *varlist*$_{\text{en}}$, *varlist*$_{\text{tr}}$, and *varlist*$_s$ may contain factor variables; see [U] **11.4.3 Factor variables**.

*depvar*, *indepvars*, *depvars*$_{\text{en}}$, *varlist*$_{\text{en}}$, *depvar*$_{\text{tr}}$, *varlist*$_{\text{tr}}$, *tvar*, *depvar*$_s$, and *varlist*$_s$ may contain time-series operators; see [U] **11.4.4 Time-series varlists**.

`bootstrap`, `by`, `jackknife`, `rolling`, `statsby`, and `svy` are allowed; see [U] **11.1.10 Prefix commands**.

Weights are not allowed with the `bootstrap` prefix; see [R] **bootstrap**.

`vce()` and weights are not allowed with the `svy` prefix; see [SVY] **svy**.

`fweights`, `iweights`, and `pweights` are allowed; see [U] **11.1.6 weight**.

`coeflegend` does not appear in the dialog box.

See [U] **20 Estimation and postestimation commands** for more capabilities of estimation commands.

# Options

Model

`endogenous(`*enspec*`)`, `entreat(`*entrspec*`)`, `extreat(`*extrspec*`)`, `select(`*selspec*`)`, `tobitselect(`*tselspec*`)`; see [ERM] **erm options**.

`noconstant`, `offset(`*varname*$_o$`)`, `constraints(`*numlist*`)`, `collinear`; see [R] **estimation options**.

SE/Robust

`vce(`*vcetype*`)`; see [ERM] **erm options**.

Reporting

`level(#)`, `nocnsreport`; see [R] **estimation options**.

*display_options*: `noci`, `nopvalues`, `noomitted`, `vsquish`, `noemptycells`, `baselevels`, `allbaselevels`, `nofvlabel`, `fvwrap(#)`, `fvwrapon(`*style*`)`, `cformat(%`*fmt*`)`, `pformat(%`*fmt*`)`, `sformat(%`*fmt*`)`, and `nolstretch`; see [R] **estimation options**.

Integration

`intpoints(#)`, `triintpoints(#)`; see [ERM] **erm options**.

Maximization

*maximize_options*: `difficult`, `technique(`*algorithm_spec*`)`, `iterate(#)`, `[no]log`, `trace`, `gradient`, `showstep`, `hessian`, `showtolerance`, `tolerance(#)`, `ltolerance(#)`, `nrtolerance(#)`, `nonrtolerance`, and `from(`*init_specs*`)`; see [R] **maximize**.

Setting the optimization type to `technique(bhhh)` resets the default *vcetype* to `vce(opg)`.

The following option is available with `eregress` but is not shown in the dialog box:

`coeflegend`; see [R] **estimation options**.

# Remarks and examples

stata.com

`eregress` fits models that we refer to as "extended linear regression models". We use this term to mean linear regression models that accommodate endogenous covariates, nonrandom treatment assignment, and endogenous sample selection. `eregress` can account for these complications whether they arise individually or in combination.

In this entry, you will find information on the eregress command syntax. You can see *Methods and formulas* for a full description of the models that can be fit with eregress and details about how those models are fit.

More information on extended linear regression models is found in the separate introductions and example entries. We recommend reading those entries to learn how to use eregress. Below, we provide a guide to help you locate the ones that will be helpful to you.

For an introduction to eregress and the other extended regression commands (eintreg, eprobit, and eoprobit), see [ERM] **intro 1**–[ERM] **intro 8**.

[ERM] **intro 1** introduces the ERM commands, the problems they address, and their syntax.

[ERM] **intro 2** provides background on the four types of models—linear regression, interval regression, probit regression, and ordered probit regression—that can be fit using ERM commands.

[ERM] **intro 3** considers the problem of endogenous covariates and how to solve it using ERM commands.

[ERM] **intro 4** gives an overview of endogenous sample selection and using ERM commands to account for it.

[ERM] **intro 5** covers nonrandom treatment assignment and how to account for it using eregress or any of the other ERM commands.

[ERM] **intro 6** discusses interpretation of results. You can interpret coefficients from eregress in the usual way, but this introduction goes beyond the interpretation of coefficients. We demonstrate how to find answers to interesting questions by using margins. If your model includes an endogenous covariate or an endogenous treatment, the use of margins differs from its use after other estimation commands, so we strongly recommend reading this intro if you are fitting these types of models.

[ERM] **intro 7** will be helpful if you are familiar with heckman, ivregress, etregress, and other commands that address endogenous covariates, sample selection, or nonrandom treatment assignment. This introduction is a Rosetta stone that maps the syntax of those commands to the syntax of eregress.

[ERM] **intro 8** walks you through an example that gives insight into the concepts of endogenous covariates, treatment assignment, and sample selection while fitting models with eregress that address these complications. This intro also demonstrates how to interpret results by using margins and estat teffects.

Additional examples are presented in [ERM] **example 1a**–[ERM] **example 6b**. For examples using eregress, see

| | |
|---|---|
| [ERM] **example 1a** | Linear regression with continuous endogenous covariate |
| [ERM] **example 2a** | Linear regression with binary endogenous covariate |
| [ERM] **example 2b** | Linear regression with exogenous treatment |
| [ERM] **example 2c** | Linear regression with endogenous treatment |

See *Examples* in [ERM] **intro** for an overview of all the examples. These examples demonstrate all four extended regression commands, and all may be interesting because they handle complications in the same way.

You can also find in literature discussion and examples of many models that eregress can fit. For example, eregress can fit the linear regression model with endogenous sample selection (Heckman 1976), the linear regression model with an endogenous treatment (Heckman 1978; Maddala 1983), and the linear regression model with a tobit selection equation (Amemiya 1985; Wooldridge 2010, sec. 19.7). The linear regression model with endogenous regressors and endogenous sample selection

discussed in Wooldridge (2010, sec 19.6) is also supported, along with the tobit selection regression with endogenous regressors discussed in Wooldridge (2010, sec 19.7). Roodman (2011) investigated linear regression models with endogenous covariates and endogenous sample selection, and demonstrated how multiple observational data complications could be addressed with a triangular model structure. His work has been used to model processes like the effect of aphid infestations and virus outbreaks on crop yields (Elbakidze, Lu, and Eigenbrode 2011) and the effect of calorie intake per day on food security in poor neighborhoods (Maitra and Rao 2014).

## Stored results

`eregress` stores the following in `e()`:

Scalars

| | |
|---|---|
| `e(N)` | number of observations |
| `e(N_selected)` | number of uncensored observations |
| `e(N_nonselected)` | number of censored observations |
| `e(k)` | number of parameters |
| `e(k_cat#)` | number of categories for the #th *depvar*, ordinal |
| `e(k_eq)` | number of equations in `e(b)` |
| `e(k_eq_model)` | number of equations in overall model test |
| `e(k_dv)` | number of dependent variables |
| `e(k_aux)` | number of auxiliary parameters |
| `e(df_m)` | model degrees of freedom |
| `e(ll)` | log likelihood |
| `e(N_clust)` | number of clusters |
| `e(chi2)` | $\chi^2$ |
| `e(p)` | $p$-value for model test |
| `e(n_quad)` | number of integration points for multivariate normal |
| `e(n_quad3)` | number of integration points for trivariate normal |
| `e(rank)` | rank of `e(V)` |
| `e(ic)` | number of iterations |
| `e(rc)` | return code |
| `e(converged)` | `1` if converged, `0` otherwise |

Macros

| | |
|---|---|
| `e(cmd)` | `eregress` |
| `e(cmdline)` | command as typed |
| `e(depvar)` | names of dependent variables |
| `e(wtype)` | weight type |
| `e(wexp)` | weight expression |
| `e(title)` | title in estimation output |
| `e(clustvar)` | name of cluster variable |
| `e(offset#)` | offset for the #th *depvar*, where # is determined by equation order in output |
| `e(chi2type)` | `Wald`; type of model $\chi^2$ test |
| `e(vce)` | *vcetype* specified in `vce()` |
| `e(vcetype)` | title used to label Std. Err. |
| `e(opt)` | type of optimization |
| `e(which)` | `max` or `min`; whether optimizer is to perform maximization or minimization |
| `e(ml_method)` | type of `ml` method |
| `e(user)` | name of likelihood-evaluator program |
| `e(technique)` | maximization technique |
| `e(properties)` | `b V` |
| `e(estat_cmd)` | program used to implement `estat` |
| `e(predict)` | program used to implement `predict` |
| `e(marginsok)` | predictions allowed by `margins` |
| `e(marginsnotok)` | predictions disallowed by `margins` |
| `e(asbalanced)` | factor variables `fvset` as `asbalanced` |
| `e(asobserved)` | factor variables `fvset` as `asobserved` |

Matrices

| | |
|---|---|
| `e(b)` | coefficient vector |

| | |
|---|---|
| e(cat#) | categories for the #th *depvar*, ordinal |
| e(Cns) | constraints matrix |
| e(ilog) | iteration log (up to 20 iterations) |
| e(gradient) | gradient vector |
| e(V) | variance–covariance matrix of the estimators |
| e(V_modelbased) | model-based variance |
| Functions | |
| e(sample) | marks estimation sample |

# Methods and formulas

The methods and formulas presented here are for the linear model. The estimator implemented in `eregress` is a maximum likelihood estimator covered by the results in chapter 13 of Wooldridge (2010) and White (1996).

The log-likelihood function maximized by `eregress` is implied by the triangular structure of the model. Specifically, the joint distribution of the endogenous variables is a product of conditional and marginal distributions, because the model is triangular. For a few of the many relevant applications of this result in literature, see chapter 10 of Amemiya (1985); Heckman (1976, 1979); chapter 5 of Maddala (1983); Maddala and Lee (1976); sections 15.7.2, 15.7.3, 16.3.3, 17.5.2, and 19.7.1 in Wooldridge (2010); and Wooldridge (2014). Roodman (2011) used this result to derive the formulas discussed below.

Methods and formulas are presented under the following headings:



## Introduction

A linear regression of outcome $y_i$ on covariates $\mathbf{x}_i$ may be written as

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$$

where the error $\epsilon_i$ is normal with mean 0 and variance $\sigma^2$. The log likelihood is

$$\ln L = \sum_{i=1}^{N} w_i \ln \phi\left(y_i - \mathbf{x}_i\boldsymbol{\beta}, \sigma^2\right)$$

The conditional mean of $y_i$ is

$$E(y_i|\mathbf{x}_i) = \mathbf{x}_i\boldsymbol{\beta}$$

If you are willing to take our word for some derivations and notation, the following is complete. Longer explanations and derivations for some terms and functions are provided in *Methods and formulas* of [ERM] **eprobit**. For example, we need the two-sided probability function $\Phi_d^*$ that is discussed in *Introduction* in [ERM] **eprobit**.

If you are interested in all the details, we suggest you read *Methods and formulas* of [ERM] **eprobit** in its entirety, before reading this section. Here, we mainly show how the complications that arise in ERMs are handled in a linear regression framework.

## Endogenous covariates

### Continuous endogenous covariates

A linear regression of $y_i$ on exogenous covariates $\mathbf{x}_i$ and $C$ continuous endogenous covariates $\mathbf{w}_{ci}$ has the form

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \mathbf{w}_{ci}\boldsymbol{\beta}_c + \epsilon_i$$
$$\mathbf{w}_{ci} = \mathbf{z}_{ci}\mathbf{A}_c + \boldsymbol{\epsilon}_{ci}$$

The vector $\mathbf{z}_{ci}$ contains variables from $\mathbf{x}_i$ and other covariates that affect $\mathbf{w}_{ci}$. For the model to be identified, $\mathbf{z}_{ci}$ must contain one extra exogenous covariate not in $\mathbf{x}_i$ for each of the endogenous regressors in $\mathbf{w}_{ci}$. The unobserved errors $\epsilon_i$ and $\boldsymbol{\epsilon}_{ci}$ are multivariate normal with mean 0 and covariance

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & \boldsymbol{\sigma}_{1c}' \\ \boldsymbol{\sigma}_{1c} & \boldsymbol{\Sigma}_c \end{bmatrix}$$

The log likelihood is

$$\ln L = \sum_{i=1}^{N} w_i \ln\phi_{C+1}(\mathbf{r}_i, \boldsymbol{\Sigma})$$

where

$$\mathbf{r}_i = \begin{bmatrix} y_i - \mathbf{x}_i & \mathbf{w}_{ci} - \mathbf{z}_{ci}\mathbf{A}_c \end{bmatrix}$$

The conditional mean of $y_i$ is

$$E(y_i|\mathbf{x}_i, \mathbf{w}_{ci}, \mathbf{z}_{ci}) = \mathbf{x}_i\boldsymbol{\beta} + \mathbf{w}_{ci}\boldsymbol{\beta}_c + \boldsymbol{\sigma}_{1c}'\boldsymbol{\Sigma}_c^{-1}(\mathbf{w}_{ci} - \mathbf{z}_{ci}\mathbf{A}_c)'$$

### Binary and ordinal endogenous covariates

Here, we begin by formulating the linear regression of $y_i$ on exogenous covariates $\mathbf{x}_i$ and $B$ binary and ordinal endogenous covariates $\mathbf{w}_{bi} = [w_{b1i}, \ldots, w_{bBi}]$. Indicator (dummy) variables for the levels of each binary and ordinal covariate are used in the model. You can also interact other covariates with the binary and ordinal endogenous covariates, as in treatment-effect models.

The binary and ordinal endogenous covariates $\mathbf{w}_{bi}$ are formulated as in *Binary and ordinal endogenous covariates* in [ERM] **eprobit**.

The model for the outcome can be formulated with or without different variance and correlation parameters for each level of $\mathbf{w}_{bi}$. Level-specific parameters are obtained by specifying `povariance` or `pocorrelation` in the `endogenous()` option.

If the variance and correlation parameters are not level specific, we have

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \mathbf{wind}_{b1i}\boldsymbol{\beta}_{b1} + \cdots + \mathbf{wind}_{bBi}\boldsymbol{\beta}_{bB} + \epsilon_i$$

The $\mathbf{wind}_{bji}$ vectors are defined in *Binary and ordinal endogenous covariates* in [ERM] **eprobit**. The binary and ordinal endogenous errors $\epsilon_{b1i}, \ldots, \epsilon_{bBi}$ and outcome error $\epsilon_i$ are multivariate normal with mean 0 and covariance

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_b & \boldsymbol{\sigma}_{1b} \\ \boldsymbol{\sigma}_{1b}' & \sigma^2 \end{bmatrix}$$

From here, we discuss the model with ordinal endogenous covariates. The results for binary endogenous covariates are similar.

Using results from *Likelihood for multiequation models* in [ERM] **eprobit**, the joint density of $y_i$ and $\mathbf{w}_{bi}$ can be written using the conditional density of $\epsilon_{b1i}, \ldots, \epsilon_{bBi}$ on $\epsilon_i$.

Define

$$r_i = y_i - (\mathbf{x}_i\boldsymbol{\beta} + \mathbf{wind}_{b1i}\boldsymbol{\beta}_{b1} + \cdots + \mathbf{wind}_{bBi}\boldsymbol{\beta}_{bB})$$

Let

$$\boldsymbol{\mu}_{b|1,i} = \frac{\boldsymbol{\sigma}_{1b}'}{\sigma^2} r_i = [\, e_{b1i} \ldots e_{bBi} \,]$$

$$\boldsymbol{\Sigma}_{b|1} = \boldsymbol{\Sigma}_b - \frac{\boldsymbol{\sigma}_{1b}\boldsymbol{\sigma}_{1b}'}{\sigma^2}$$

For $j = 1, \ldots, B$ and $h = 0, \ldots, B_j$, let

$$c_{bjih} = \begin{cases} -\infty & h = 0 \\ \kappa_{bjh} - \mathbf{z}_{bji}\boldsymbol{\alpha}_{bj} - e_{bji} & h = 1, \ldots, B_j - 1 \\ \infty & h = B_j \end{cases}$$

So, for $j = 1, \ldots, B$, the probability for $w_{bji}$ has lower limit

$$l_{bji} = c_{bji(h-1)} \quad \text{if} \quad w_{bji} = v_{bjh}$$

and upper limit

$$u_{bji} = c_{bjih} \quad \text{if} \quad w_{bji} = v_{bjh}$$

Let

$$\mathbf{l}_i = [\, l_{b1i} \quad \ldots \quad l_{bBi} \,]$$

$$\mathbf{u}_i = [\, u_{b1i} \quad \ldots \quad u_{bBi} \,]$$

So, the log likelihood for this model is

$$\ln L = \sum_{i=1}^{N} w_i \ln \left\{ \Phi_B^*(\mathbf{l}_i, \mathbf{u}_i, \boldsymbol{\Sigma}_{b|1}) \phi\big(r_i, \sigma^2\big) \right\}$$

The expected value of $y_i$ conditional on $\mathbf{w}_{bi}$ can be calculated using the techniques discussed in *Predictions using the full model* in [ERM] **eprobit postestimation**.

When the endogenous ordinal variables are different treatments, holding the variance and correlation parameters constant over the treatment levels is a constrained form of the potential-outcome model. In an unconstrained potential-outcome model, the variance of the outcome and the correlations between the outcome and the treatments—the endogenous ordinal regressors $\mathbf{w}_{bi}$—vary over the levels of each treatment.

In this unconstrained model, there is a different potential-outcome error for each level of each treatment. For example, when the endogenous treatment variable $w_1$ has three levels (0, 1, and 2) and the endogenous treatment variable $w_2$ has four levels (0, 1, 2, and 3), the unconstrained model has $12 = 3 \times 4$ outcome errors. So there are 12 outcome error variance parameters. Because there is a different correlation between each potential outcome and each endogenous treatment, there are $2 \times 12$ correlation parameters between the potential outcomes and the treatments in this example model.

We denote the number of different combinations of values for the endogenous treatments $\mathbf{w}_{bi}$ by $M$, and we denote the vector of values in each combination by $\mathbf{v}_j$ ($j \in \{1, 2, \ldots, M\}$). Letting $k_{wp}$ be the number of levels of endogenous ordinal treatment variable $p \in \{1, 2, \ldots, B\}$ implies that $M = k_{w1} \times k_{w2} \times \cdots \times k_{wB}$.

Denoting the outcome errors $\epsilon_{1i}$, ..., $\epsilon_{Mi}$, we have

$$
\begin{aligned}
y_{1i} &= \mathbf{x}_i\boldsymbol{\beta} + \mathbf{wind}_{b1i}\boldsymbol{\beta}_{b1} + \cdots + \mathbf{wind}_{bBi}\boldsymbol{\beta}_{bB} + \epsilon_{1i} \\
&\vdots \\
y_{Mi} &= \mathbf{x}_i\boldsymbol{\beta} + \mathbf{wind}_{b1i}\boldsymbol{\beta}_{b1} + \cdots + \mathbf{wind}_{bBi}\boldsymbol{\beta}_{bB} + \epsilon_{Mi} \\
y_i &= \sum_{j=1}^{M} 1(\mathbf{w}_{bi} = \mathbf{v}_j) y_{ji}
\end{aligned}
$$

For $j = 1, \ldots, M$, the endogenous errors $\epsilon_{b1i}, \ldots, \epsilon_{bBi}$ and outcome error $\epsilon_{ji}$ are multivariate normal with 0 mean and covariance

$$
\boldsymbol{\Sigma}_j = \begin{bmatrix} \boldsymbol{\Sigma}_b & \boldsymbol{\sigma}_{j1b} \\ \boldsymbol{\sigma}'_{j1b} & \sigma_j^2 \end{bmatrix}
$$

Now, let

$$
\begin{aligned}
\sigma_{i,b} &= \sum_{j=1}^{M} 1(\mathbf{w}_{bi} = \mathbf{v}_j)\sigma_j \\
\boldsymbol{\Sigma}_{i,b|1} &= \sum_{j=1}^{M} 1(\mathbf{w}_{bi} = \mathbf{v}_j) \left( \boldsymbol{\Sigma}_b - \frac{\boldsymbol{\sigma}_{j1b}\boldsymbol{\sigma}'_{j1b}}{\sigma_j^2} \right)
\end{aligned}
$$

Now, the log likelihood for this model is

$$
\ln L = \sum_{i=1}^{N} w_i \ln \left\{ \Phi_B^*(\mathbf{l}_i, \mathbf{u}_i, \boldsymbol{\Sigma}_{i,b|1}) \phi\big(r_i, \sigma_{i,b}^2\big) \right\}
$$

As in the other case, the expected value of $y_i$ conditional on $\mathbf{w}_{bi}$ can be calculated using the techniques discussed in *Predictions using the full model* in [ERM] **eprobit postestimation**.

## Treatment

In the potential-outcomes framework, the treatment $t_i$ is a discrete variable taking $T$ values, indexing the $T$ potential outcomes of the outcome $y_i$: $y_{1i}, \ldots, y_{Ti}$.

When we observe treatment $t_i$ with levels $v_1, \ldots, v_T$, we have

$$
y_i = \sum_{j=1}^{T} 1(t_i = v_j) y_{ji}
$$

So for each observation, we only observe the potential outcome associated with that observation's treatment value.

For exogenous treatments, our approach is equivalent to the regression adjustment treatment-effect estimation method. See [TE] **teffects intro advanced**. We do not model the treatment assignment process. The formulas for the treatment effects and potential-outcome means (POMs) are equivalent to what we provide here for endogenous treatments. The treatment effect on the treated for $\mathbf{x}_i$ for an exogenous treatment is equivalent to what we provide here for the endogenous treatment when the correlation parameter between the outcome and treatment errors is set to 0. The average treatment effects (ATEs) and POMs for exogenous treatments are estimated as predictive margins in an analogous manner to what we describe here for endogenous treatments. We can also obtain different variance parameters for the different exogenous treatment groups by specifying `povariance` in `extreat()`.

From here, we assume an endogenous treatment $t_i$. As in *Treatment* in [ERM] **eprobit**, we model the treatment assignment process with a probit or ordered probit model, and we call the treatment assignment error $\epsilon_{ti}$. A linear regression of $y_i$ on exogenous covariates $\mathbf{x}_i$ and endogenous treatment $t_i$ taking values $v_1, \ldots, v_T$ has the form

$$
\begin{aligned}
y_{1i} &= \mathbf{x}_i\boldsymbol{\beta}_1 + \epsilon_{1i} \\
&\vdots \\
y_{Ti} &= \mathbf{x}_i\boldsymbol{\beta}_T + \epsilon_{Ti} \\
y_i &= \sum_{j=1}^{T} 1(t_i = v_j) y_{ji}
\end{aligned}
$$

This model can be formulated with or without different variance and correlation parameters for each potential outcome. Potential-outcome specific parameters are obtained by specifying `povariance` or `pocorrelation` in the `entreat()` option.

If the variance and correlation parameters are not potential-outcome specific, for $j = 1, \ldots, T$, $\epsilon_{ji}$ and $\epsilon_{ti}$ are bivariate normal with mean 0 and covariance

$$
\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & \sigma\rho_{1t} \\ \sigma\rho_{1t} & 1 \end{bmatrix}
$$

The treatment is exogenous if $\rho_{1t} = 0$. Note that we did not specify the structure of the correlations between the potential-outcome errors. We do not need information about these correlations to estimate POMs and treatment effects because all covariates and the outcome are observed in observations from each group.

From here, we discuss a model with an ordinal endogenous treatment. The results for binary treatment models are similar.

As in *Binary and ordinal endogenous covariates*, using the results from *Likelihood for multiequation models* in [ERM] **eprobit**, the joint density of $y_i$ and $t_i$ can be written using the conditional density of the treatment error $\epsilon_{ti}$ on the outcome errors $\epsilon_{i1}, \ldots, \epsilon_{Ti}$.

Define

$$
r_i = y_i - \mathbf{x}_i\boldsymbol{\beta}_j \quad \text{if} \quad t_i = v_j
$$

The log likelihood for the model is

$$
\ln L = \sum_{i=1}^{N} w_i \ln \left\{ \Phi_1^* \left( l_{ti} - \frac{\rho_{1t}}{\sigma} r_i, u_{ti} - \frac{\rho_{1t}}{\sigma} r_i, 1 - \rho_{1t}^2 \right) \phi\left(r_i, \sigma^2\right) \right\}
$$

where $l_{ti}$ and $u_{ti}$ are the limits for the treatment probability given in *Treatment* in [ERM] **eprobit**.

The treatment effect $y_{ji} - y_{1i}$ is the difference in the outcome for individual $i$ if the individual receives the treatment $t_i = v_j$ and what the difference would have been if the individual received the control treatment $t_i = v_1$ instead.

The conditional POM for treatment group $j$ is

$$\text{POM}_j(\mathbf{x}_i) = E\left(y_{ji}|\mathbf{x}_i\right) = \mathbf{x}_i\boldsymbol{\beta}_j$$

For treatment group $j$, the treatment effect (TE) conditioned on $\mathbf{x}_i$ is

$$\text{TE}_j(\mathbf{x}_i) = E\left(y_{ji} - y_{1i}|\mathbf{x}_i\right) = \text{POM}_j(\mathbf{x}_i) - \text{POM}_1(\mathbf{x}_i)$$

For treatment group $j$, the treatment effect on the treated (TET) in group $h$ for covariates $\mathbf{x}_i$ is

$$\begin{aligned}\text{TET}_j(\mathbf{x}_i, t_i = v_h) &= E\left(y_{ji} - y_{1i}|\mathbf{x}_i, t_i = v_h\right)\\ &= \mathbf{x}_i\boldsymbol{\beta}_j - \mathbf{x}_i\boldsymbol{\beta}_1 + E\left(\epsilon_{ji}|\mathbf{x}_i, t_i = v_h\right) - E\left(\epsilon_{1i}|\mathbf{x}_i, t_i = v_h\right)\end{aligned}$$

Remembering that the outcome errors and the treatment error $\epsilon_{ti}$ are multivariate normal, for $j = 1, \ldots, T$ we can decompose $\epsilon_{ji}$ such that

$$\epsilon_{ji} = \sigma\rho_{1t}\epsilon_{ti} + \psi_{ji}$$

where $\psi_{ji}$ has mean 0.

It follows that

$$\text{TET}_j(\mathbf{x}_i, t_i = v_h) = \mathbf{x}_i\boldsymbol{\beta}_j - \mathbf{x}_i\boldsymbol{\beta}_1$$

We can take the expectation of these conditional predictions over the covariates to get population average parameters. The `estat teffects` or `margins` command is used to estimate the expectations as predictive margins once the model is estimated with `eregress`. The POM for treatment group $j$ is

$$\text{POM}_j = E\left(y_{ji}\right) = E\left\{\text{POM}_j(\mathbf{x}_i)\right\}$$

The ATE for treatment group $j$ is

$$\text{ATE}_j = E\left(y_{ji} - y_{1i}\right) = E\left\{\text{TE}_j(\mathbf{x}_i)\right\}$$

For treatment group $j$, the average treatment effect on the treated (ATET) in treatment group $h$ is

$$\text{ATET}_{jh} = E\left(y_{ji} - y_{1i}|t_i = v_h\right) = E\left\{\text{TET}_j(\mathbf{x}_i, t_i = v_h)|t_i = v_h\right\}$$

The conditional mean of $y_i$ at treatment level $v_j$ is

$$E(y_i|\mathbf{x}_i, \mathbf{z}_{ti}, t_i = v_j) = \mathbf{x}_i\boldsymbol{\beta}_j + E(\epsilon_i|\mathbf{x}_i, \mathbf{z}_{ti}, t_i = v_j)$$

In *Predictions using the full model* in [ERM] **eprobit postestimation**, we discuss how the conditional mean of $\epsilon_i$ is calculated.

If the variance and correlation parameters are potential-outcome specific, for $j = 1, \ldots, T$, $\epsilon_{ji}$ and $\epsilon_{ti}$ are bivariate normal with mean 0 and covariance

$$\mathbf{\Sigma}_j = \begin{bmatrix} \sigma_j^2 & \sigma_j\rho_{jt} \\ \sigma_j\rho_{jt} & 1 \end{bmatrix}$$

Now, define

$$\rho_i = \sum_{j=1}^{T} 1(t_i = v_j)\rho_{jt}$$

$$\sigma_i = \sum_{j=1}^{T} 1(t_i = v_j)\sigma_j$$

The log likelihood for the model is

$$\ln L = \sum_{i=1}^{N} w_i \ln \left\{ \Phi_1^* \left( l_{ti} - \frac{\rho_i}{\sigma_i} r_i, u_{ti} - \frac{\rho_i}{\sigma_i} r_i, 1 - \rho_i^2 \right) \phi\left(r_i, \sigma_i^2\right) \right\}$$

The definitions for the potential-outcome means and treatment effects are the same as in the case where the variance and correlation parameters did not vary by potential outcome. For the treatment effect on the treated (TET) of group $j$ in group $h$, we have

$$\begin{aligned} \text{TET}_j(\mathbf{x}_i, t_i = v_h) &= E\left(y_{ji} - y_{1i} | \mathbf{x}_i, t_i = v_h\right) \\ &= \mathbf{x}_i\boldsymbol{\beta}_j - \mathbf{x}_i\boldsymbol{\beta}_1 + E\left(\epsilon_{ji} | \mathbf{x}_i, t_i = v_h\right) - E\left(\epsilon_{1i} | \mathbf{x}_i, t_i = v_h\right) \end{aligned}$$

The outcome errors and the treatment error $\epsilon_{ti}$ are multivariate normal, so for $j = 1, \ldots, T$, we can decompose $\epsilon_{ji}$ such that

$$\epsilon_{ji} = \sigma_j\rho_j\epsilon_{ti} + \psi_{ji}$$

where $\psi_{ji}$ has mean 0 and is independent of $t_i$.

It follows that

$$\begin{aligned} \text{TET}_j(\mathbf{x}_i, t_i = v_h) &= E\left(y_{ji} - y_{1i} | \mathbf{x}_i, t_i = v_h\right) \\ &= \mathbf{x}_i\boldsymbol{\beta}_j - \mathbf{x}_i\boldsymbol{\beta}_1 + (\sigma_j\rho_j - \sigma_1\rho_1) E\left(\epsilon_{ti} | \mathbf{x}_i, t_i = v_h\right) \end{aligned}$$

The mean of $\epsilon_{ti}$ conditioned on $t_i$ and the exogenous covariates $\mathbf{x}_i$ can be determined using the formulas discussed in *Predictions using the full model* in [ERM] **eprobit postestimation**. It is nonzero. So the treatment effect on the treated will be equal only to the treatment effect under an exogenous treatment or when the correlation and variance parameters are identical between the potential outcomes.

As in the other case, we can take the expectation of these conditional predictions over the covariates to get population-averaged parameters. The `estat teffects` or `margins` command is used to estimate the expectations as predictive margins once the model is fit with `eregress`.

## Endogenous sample selection

### Probit endogenous sample selection

A linear regression for outcome $y_i$ with selection on $s_i$ has the form

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i > 0$$
$$s_i = 1\left(\mathbf{z}_{si}\boldsymbol{\alpha}_s + \epsilon_{si} > 0\right)$$

where $\mathbf{x}_i$ are covariates that affect the outcome and $\mathbf{z}_{si}$ are covariates that affect selection. The outcome $y_i$ is observed if $s_i = 1$ and is not observed if $s_i = 0$. The unobserved errors $\epsilon_i$ and $\epsilon_{si}$ are normal with mean 0 and covariance

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & \sigma\rho_{1s} \\ \sigma\rho_{1s} & 1 \end{bmatrix}$$

As in the previous section, using the results from *Likelihood for multiequation models* in [ERM] **eprobit**, the joint density of $y_i$ and $s_i$ can be written using the conditional density of the selection error $\epsilon_{si}$ on the outcome error $\epsilon_i$.

For the selection indicator $s_i$, we have lower and upper limits

$$l_{si} = \begin{cases} -\infty & s_i = 0 \\ -\mathbf{z}_{si}\boldsymbol{\alpha}_s - \frac{\rho_{1s}}{\sigma}\left(y_i - \mathbf{x}_i\boldsymbol{\beta}\right) & s_i = 1 \end{cases} \qquad u_{si} = \begin{cases} -\mathbf{z}_{si}\boldsymbol{\alpha}_s & s_i = 0 \\ \infty & s_i = 1 \end{cases}$$

The log likelihood for the model is

$$\ln L = \sum_{i=1}^{N} w_i \ln\Phi_1^*\left(l_{si}, u_{si}, 1 - s_i\rho_{1s}^2\right) + \sum_{i\in S} w_i \ln\phi\left(y_i - \mathbf{x}_i\boldsymbol{\beta}, \sigma^2\right)$$

where $S$ is the set of observations for which $y_i$ is observed.

The conditional mean of $y_i$ is

$$E(y_i|\mathbf{x}_i) = \mathbf{x}_i\boldsymbol{\beta}$$

### Tobit endogenous sample selection

Instead of constraining the selection indicator to be binary, tobit endogenous sample selection uses a censored continuous sample-selection indicator. We allow the selection variable to be left-censored or right-censored.

A linear regression model for outcome $y_i$ with tobit selection on $s_i$ has the form

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i > 0$$

We observe the selection indicator $s_i$, which indicates the censoring status of the latent selection variable $s_i^\star$,

$$s_i^\star = \mathbf{z}_{si}\boldsymbol{\alpha}_s + \epsilon_{si}$$

$$s_i = \begin{cases} l_i & s_i^\star \le l_i \\ s_i^\star & l_i < s_i^\star < u_i \\ u_i & s_i^\star \ge u_i \end{cases}$$

where $\mathbf{z}_{si}$ are covariates that affect selection, and $l_i$ and $u_i$ are fixed lower and upper limits.

The outcome $y_i$ is observed when $s_i^\star$ is not censored ($l_i < s_i^\star < u_i$). The outcome $y_i$ is not observed when $s_i^\star$ is left-censored ($s_i^\star \le l_i$) or $s_i^\star$ is right-censored ($s_i^\star \ge u_i$). The unobserved errors $\epsilon_i$ and $\epsilon_{si}$ are normal with mean 0 and covariance

$$\begin{bmatrix} \sigma^2 & \sigma_{1s} \\ \sigma_{1s} & \sigma_s^2 \end{bmatrix}$$

For the selected observations, we can treat $s_i$ as a continuous endogenous regressor, as in *Continuous endogenous covariates*. In fact, $s_i$ may even be used as a regressor for $y_i$ in `eregress` (specify `tobitselect(..., main)`). On the nonselected observations, we treat $s_i$ like the probit sample-selection indicator in *Probit endogenous sample selection*.

The log likelihood is

$$\begin{aligned} \ln L = &\sum_{i\in S} w_i \ln\phi_2\left(y_i - \mathbf{x}_i\boldsymbol{\beta}, s_i - \mathbf{z}_{si}\boldsymbol{\alpha}_s, \boldsymbol{\Sigma}\right) \\ &+ \sum_{i\in L} w_i \ln\Phi_1^*(l_{li}, u_{li}, 1) \\ &+ \sum_{i\in U} w_i \ln\Phi_1^*(l_{ui}, u_{ui}, 1) \end{aligned}$$

where $S$ is the set of observations for which $y_i$ is observed, $L$ is the set of observations where $s_i^\star$ is left-censored, and $U$ is the set of observations where $s_i^\star$ is right-censored. The lower and upper limits for selection—$l_{li}$, $u_{li}$, $l_{ui}$, and $u_{ui}$—are defined in *Tobit endogenous sample selection* in [ERM] **eprobit**.

When $s_i$ is not a covariate in $\mathbf{x}_i$, we use the standard conditional mean formula,

$$E(y_i|\mathbf{x}_i) = \mathbf{x}_i\boldsymbol{\beta}$$

Otherwise, we use

$$E(y_i|\mathbf{x}_i, s_i, z_{si}) = \mathbf{x}_i\boldsymbol{\beta} + \frac{\sigma_{1s}}{\sigma_s^2}(s_i - z_{si}\boldsymbol{\alpha}_s)$$

## Combinations of features

Extended linear regression models that involve multiple features can be formulated using the techniques discussed in *Likelihood for multiequation models* in [ERM] **eprobit**. Essentially, the density of the observed endogenous covariates can be written in terms of the unobserved normal errors. The observed endogenous and exogenous covariates determine the range of the errors, and the joint density can be evaluated as multivariate normal probabilities and densities.

## Confidence intervals

The estimated variances will always be nonnegative, and the estimated correlations will always fall in $(-1, 1)$. To obtain confidence intervals that accommodate these ranges, we must use transformations.

We use the log transformation to obtain the confidence intervals for variance parameters and the atanh transformation to obtain confidence intervals for correlation parameters. For details, see *Confidence intervals* in [ERM] **eprobit**.

# References


Amemiya, T. 1985. *Advanced Econometrics*. Cambridge, MA: Harvard University Press.

Elbakidze, L., L. Lu, and S. Eigenbrode. 2011. Evaluating vector-virus-yield interactions for peas and lentils under climatic variability: A limited dependent variable analysis. *Journal of Agricultural and Resource Economics* 36: 504–520.

Heckman, J. 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* 5: 475–492.

——. 1978. Dummy endogenous variables in a simultaneous equation system. *Econometrica* 46: 931–959.

——. 1979. Sample selection bias as a specification error. *Econometrica* 47: 153–161.

Keshk, O. M. G. 2003. Simultaneous equations models: What are they and how are they estimated. Program in Statistics and Methodology, Department of Political Science, Ohio State University. https://polisci.osu.edu/sites/polisci.osu.edu/files/Simultaneous Equations.pdf.

Maddala, G. S. 1983. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Maddala, G. S., and L.-F. Lee. 1976. Recursive Models with Qualitative Endogenous Variables. *Annals of Economic and Social Measurement* 5: 525–545.

Maitra, C., and P. Rao. 2014. An empirical investigation into measurement and determinants of food security in slums of Kolkata. School of Economics Discussion Paper No. 531, School of Economics, University of Queensland. espace.library.uq.edu.au/view/UQ:352184

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

White, H. L., Jr. 1996. *Estimation, Inference and Specification Analysis*. Cambridge: Cambridge University Press.

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

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

## Also see

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

[ERM] **eregress predict** — predict after eregress

[ERM] **estat teffects** — Average treatment effects for extended regression models

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

[R] **heckman** — Heckman selection model

[R] **ivregress** — Single-equation instrumental-variables regression

[R] **regress** — Linear regression

[SVY] **svy estimation** — Estimation commands for survey data

[TE] **etregress** — Linear regression with endogenous treatment effects

**[U] 20 Estimation and postestimation commands**