<div style="border:1px solid">

eintreg — Extended interval regression

</div>

Description

eintreg fits an interval regression model that accommodates any combination of endogenous co-variates, 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.

xteintreg fits a random-effects interval regression model that accommodates endogenous covari-ates, treatment, and sample selection in the same way as eintreg and also accounts for correlation of observations within panels or within groups.

Quick start

All quick start examples use an interval-measured dependent variable with the interval's lower bound recorded in variable y_l and its upper bound recorded in y_u.

Regression of [y_l, y_u] on x with continuous endogenous covariate y2 modeled by x and z
    eintreg y_l y_u x, endogenous(y2 = x z)

Same as above, but adding continuous endogenous covariate y3 modeled by x and z2
    eintreg y_l y_u x, endogenous(y2 = x z) endogenous(y3 = x z2)

Regression of [y_l, y_u] on x with binary endogenous covariate d modeled by x and z
    eintreg y_l y_u x, endogenous(d = x z, probit)

Regression of [y_l, y_u] on x with endogenous treatment recorded in trtvar and modeled by x and z
    eintreg y_l y_u x, entreat(trtvar = x z)

Regression of [y_l, y_u] on x with exogenous treatment recorded in trtvar
    eintreg y_l y_u x, extreat(trtvar)

Random-effects regression of [y_l, y_u] on x using xtset data
    xteintreg y_l y_u x

Regression of [y_l, y_u] on x with endogenous sample-selection indicator selvar modeled by x and z
    eintreg y_l y_u x, select(selvar = x z)

Same as above, but adding endogenous covariate y2 modeled by x and z2
    eintreg y_l y_u x, select(selvar = x z) endogenous(y2 = x z2)

Same as above, but adding endogenous treatment recorded in trtvar and modeled by x and z3
    eintreg y_l y_u x, select(selvar = x z) endogenous(y2 = x z2) ///
        entreat(trtvar = x z3)

Same as above, but with random effects and without endogenous treatment
    xteintreg y_l y_u x, select(selvar = x z) endogenous(y2 = x z2)

Menu

eintreg

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

xteintreg

Statistics > Longitudinal/panel data > Endogenous covariates > Models adding selection and treatment > Interval regression (RE)

Syntax

Basic interval regression with endogenous covariates

eintreg *depvar*$_1$ *depvar*$_2$ [*indepvars*] , <u>endog</u>enous(*depvars*$_{en}$ = *varlist*$_{en}$) [*options*]

Basic interval regression with endogenous treatment assignment

eintreg *depvar*$_1$ *depvar*$_2$ [*indepvars*] , <u>entr</u>eat(*depvar*$_{tr}$ [= *varlist*$_{tr}$]) [*options*]

Basic interval regression with exogenous treatment assignment

eintreg *depvar*$_1$ *depvar*$_2$ [*indepvars*] , <u>extr</u>eat(*tvar*) [*options*]

Basic interval regression with sample selection

eintreg *depvar*$_1$ *depvar*$_2$ [*indepvars*] , <u>sel</u>ect(*depvar*$_s$ = *varlist*$_s$) [*options*]

Basic interval regression with tobit sample selection

eintreg *depvar*$_1$ *depvar*$_2$ [*indepvars*] , <u>tobitsel</u>ect(*depvar*$_s$ = *varlist*$_s$) [*options*]

Basic interval regression with random effects

xteintreg *depvar*$_1$ *depvar*$_2$ [*indepvars*] [, *options*]

Interval regression combining endogenous covariates, treatment, and selection

eintreg *depvar*$_1$ *depvar*$_2$ [*indepvars*] [*if*] [*in*] [*weight*] [, *extensions options*]

Interval regression combining random effects, endogenous covariates, treatment, and selection

xteintreg *depvar*$_1$ *depvar*$_2$ [*indepvars*] [*if*] [*in*] [, *extensions options*]

depvar$_1$ and *depvar*$_2$ should have the following form:

Type of data		*depvar*$_1$	*depvar*$_2$
point data	$a = [a, a]$	a	a
interval data	$[a, b]$	a	b
left-censored data	$(-\infty, b]$	.	b
right-censored data	$[a, +\infty)$	a	.
missing		.	.

extensions	Description
Model	
<u>endog</u>enous(*enspec*)	model for endogenous covariates; may be repeated
<u>entr</u>eat(*entrspec*)	model for endogenous treatment assignment
<u>extr</u>eat(*extrspec*)	exogenous treatment
<u>sel</u>ect(*selspec*)	probit model for selection
<u>tobitsel</u>ect(*tselspec*)	tobit model for selection

options	Description
Model	
<u>nocon</u>stant	suppress constant term
<u>off</u>set(*varname_o*)	include *varname_o* in model with coefficient constrained to 1
<u>constra</u>ints(*numlist*)	apply specified linear constraints
SE/Robust	
vce(*vcetype*)	*vcetype* may be oim, <u>r</u>obust, <u>cl</u>uster *clustvar*, opg, <u>boot</u>strap,
	or <u>jack</u>knife
Reporting	
<u>l</u>evel(*#*)	set confidence level; default is level(95)
<u>nocnsr</u>eport	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	
<u>int</u>points(*#*)	set the number of integration (quadrature) points for integration over
	four or more dimensions; default is intpoints(128)
<u>triint</u>points(*#*)	set the number of integration (quadrature) points for integration over
	three dimensions; default is triintpoints(10)
<u>reint</u>points(*#*)	set the number of integration (quadrature) points for
	random-effects integration; default is reintpoints(7)
<u>reintm</u>ethod(*intmethod*)	integration method for random effects; *intmethod* may be
	<u>mva</u>ghermite (the default) or <u>g</u>hermite
Maximization	
maximize_options	control the maximization process; seldom used
<u>collinear</u>	keep collinear variables
<u>coeflegend</u>	display legend instead of statistics

enspec is *depvars*_en = *varlist*_en [, *enopts*]

    where *depvars*_en is a list of endogenous covariates. Each variable in *depvars*_en specifies an endogenous covariate model using the common *varlist*_en and options.

entrspec is *depvar*_tr[= *varlist*_tr] [, *entropts*]

    where *depvar*_tr is a variable indicating treatment assignment. *varlist*_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 [, *selopts*]

    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
nore	do not include random effects in model for endogenous covariate
noconstant	suppress constant term

nore is available only with xteintreg.

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
nore	do not include random effects in model for endogenous treatment
noconstant	suppress constant term
offset(*varname$_o$*)	include *varname$_o$* in model with coefficient constrained to 1

nore is available only with xteintreg.

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

selopts	Description
Model	
nore	do not include random effects in selection model
noconstant	suppress constant term
offset(*varname$_o$*)	include *varname$_o$* in model with coefficient constrained to 1

nore is available only with xteintreg.

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
nore	do not include random effects in tobit selection model
<u>noncons</u>tant	suppress constant term
<u>off</u>set(*varname_o*)	include *varname_o* in model with coefficient constrained to 1

* You must specify either ll() or ul().

nore is available only with xteintreg.

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

depvar$_1$, *depvar*$_2$, *indepvars*, *depvars*$_{en}$, *varlist*$_{en}$, *depvar*$_{tr}$, *varlist*$_{tr}$, *tvar*, *depvar*$_s$, and *varlist*$_s$ may contain time-series operators; see [U] **11.4.4 Time-series varlists**.

bootstrap, by, collect, jackknife, and statsby are allowed with eintreg and xteintreg. bayesboot, rolling, and svy are allowed with eintreg. 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 with eintreg; see [U] **11.1.6 weight**.

reintpoints() and reintmethod() are available only with xteintreg.

collinear and coeflegend do 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*); see [R] **Estimation options**.

> SE/Robust

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

> Reporting

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

display_options: noci, <u>nop</u>values, <u>noo</u>mitted, vsquish, noemptycells, <u>base</u>levels, <u>all</u>baselevels, <u>nofv</u>label, fvwrap(#), fvwrapon(*style*), cformat(*%fmt*), pformat(*%fmt*), sformat(*%fmt*), and nolstretch; see [R] **Estimation options**.

> Integration

intpoints(#), triintpoints(#), reintpoints(#), reintmethod(*intmethod*); see [ERM] **ERM options**.

Maximization

maximize_options: <u>difficult</u>, <u>tech</u>nique(*algorithm_spec*), <u>iter</u>ate(*#*), [no]log, <u>trace</u>,
  <u>gradient</u>, showstep, <u>hess</u>ian, <u>showtol</u>erance, <u>tol</u>erance(*#*), <u>ltol</u>erance(*#*),
  <u>nrtol</u>erance(*#*), <u>nonrtol</u>erance, and from(*init_specs*); see [R] **Maximize**.

The default technique for eintreg is technique(nr). The default technique for xteintreg is
technique(bhhh 10 nr 2).

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

The following options are available with eintreg and xteintreg but are not shown in the dialog box:

<u>collinear</u>, <u>coeflegend</u>; see [R] **Estimation options**.

Remarks and examples

eintreg and xteintreg fit models that we refer to as "extended interval regression models", mean-
ing that they accommodate endogenous covariates, nonrandom treatment assignment, endogenous sam-
ple selection, and panel data or other grouped data.

eintreg fits models for cross-sectional data (one-level models). eintreg can account for endoge-
nous covariates, treatment, and sample selection, whether these complications arise individually or in
combination.

xteintreg fits random-effects models (two-level models) for panel data or grouped data. xteintreg
accounts for endogenous covariates, treatment, and sample selection in the same way as eintreg and
also accounts for within-panel or within-group correlation among observations.

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

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

For an introduction to eintreg, xteintreg, and the other extended regression commands for con-
tinuous, binary, and ordinal outcomes, see [ERM] **Intro 1**–[ERM] **Intro 9**.

[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 re-
gression, probit regression, and ordered probit regression—that can be fit using ERM commands.
This intro also demonstrates how to fit tobit models using eintreg by transforming your depen-
dent variable into the required format. This same transformation can be used to fit random-effects
tobit models with xteintreg.

[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 eintreg
or any of the other ERM commands.

[ERM] **Intro 6** covers random-effects models for panel data and other grouped data. It discusses xteintreg and the other ERM commands for panel data.

[ERM] **Intro 7** discusses interpretation of results. You can interpret coefficients from eintreg and xteintreg 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 8** will be helpful if you are familiar with ivtobit, xtintreg, xttobit, and other commands that address endogenous covariates, sample selection, nonrandom treatment assignment, or panel data. This introduction is a Rosetta stone that maps the syntax of those commands to the syntax of eintreg and xteintreg. If your outcome is stored in a single variable, which is necessary with tobit, ivtobit, and xttobit, the Rosetta stone also demonstrates how to create the lower- and upper-bound variables necessary for eintreg and xteintreg.

[ERM] **Intro 9** 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. Although the example uses eregress, the discussion applies equally to eintreg. This intro also demonstrates how to interpret results by using margins and estat teffects.

Additional examples are presented in [ERM] **Example 1a**–[ERM] **Example 9**. For examples using eintreg, see

See *Examples* in [ERM] **Intro** for an overview of all the examples. All examples may be interesting because they handle complications in the same way. Examples using eregress and xteregress will be of particular interest because results of models fit by eintreg and xteintreg are interpreted in the same way.

eintreg and xteintreg fit many models discussed in the literature. For instance, the tobit model was originally conceived in Tobin (1958) as a model of consumption of consumer durables, where purchases were left-censored at 0. Wooldridge (2020, sec. 17.4) introduces censored and truncated regression models. Cameron and Trivedi (2022, chap. 19) discuss the tobit model using Stata examples. eintreg can also fit models like the tobit regression model with continuous endogenous regressors (Newey 1987) and the censored regression model with binary endogenous regressors (Angrist 2001). xteintreg can fit the random-effects tobit model discussed in (Wooldridge 2010, sec. 17.8). Roodman (2011) investigated interval regression models with endogenous covariates and endogenous sample selection and demonstrated how multiple observational data complications could be addressed with a triangular model structure. He and Tamás Bartus showed how random effects could be used in the triangular model structure in Bartus and Roodman (2014). Roodman's work has been used to model processes like the effect of innovation on labor productivity (Mairesse and Robin 2009) and the effect of insect-resistant crops on pesticide demand (Fernandez-Cornejo and Wechsler 2012).

Stored results

eintreg stores the following in e():

Scalars
e(N)	number of observations
e(N_selected)	number of selected observations
e(N_nonselected)	number of nonselected observations
e(N_unc)	number of uncensored observations
e(N_lc)	number of left-censored observations
e(N_rc)	number of right-censored observations
e(N_int)	number of interval-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)	χ^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)	eintreg
e(cmdline)	command as typed
e(depvar)	names of dependent variables
e(tsel_ll)	left-censoring limit for tobit selection
e(tsel_ul)	right-censoring limit for tobit selection
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 χ^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

In addition to the above, the following is stored in r():

Matrices

r(table)	matrix containing the coefficients with their standard errors, test statistics, p-values, and confidence intervals

Note that results stored in r() are updated when the command is replayed and will be replaced when any r-class command is run after the estimation command.

xteintreg stores the following in e():

Scalars

e(N)	number of observations
e(N_g)	number of groups
e(N_selected)	number of selected observations
e(N_nonselected)	number of nonselected observations
e(N_unc)	number of uncensored observations
e(N_lc)	number of left-censored observations
e(N_rc)	number of right-censored observations
e(N_int)	number of interval-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)	χ^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(n_requad)	number of integration points for random effects
e(g_min)	smallest group size
e(g_avg)	average group size
e(g_max)	largest group size
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)	xteintreg
e(cmdline)	command as typed
e(depvar)	names of dependent variables
e(tsel_ll)	left-censoring limit for tobit selection
e(tsel_ul)	right-censoring limit for tobit selection
e(ivar)	variable denoting groups
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 χ^2 test
e(vce)	*vcetype* specified in vce()
e(vcetype)	title used to label Std. err.
e(reintmethod)	integration method for random effects
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

In addition to the above, the following is stored in r():

Matrices

r(table)	matrix containing the coefficients with their standard errors, test statistics, *p*-values, and confidence intervals

Note that results stored in r() are updated when the command is replayed and will be replaced when any r-class command is run after the estimation command.

Methods and formulas

The methods and formulas presented here are for the interval model. The estimators implemented in eintreg and xteintreg are maximum likelihood estimators covered by the results in chapter 13 of Wooldridge (2010) and White (1996).

The log-likelihood functions maximized by eintreg and xteintreg are 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) and Bartus and Roodman (2014) used this result to derive the formulas discussed below.

Methods and formulas are presented under the following headings:

Introduction

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

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

where ϵ_i is normal with mean 0 and variance σ^2. Instead of observing y_i, we observe the endpoints y_{li} and y_{ui}.

If y_i is left-censored, the lower endpoint $y_{li} = -\infty$ and we know that $y_i \leq y_{ui}$. If y_i is right-censored, the upper endpoint $y_{ui} = +\infty$ and we know that $y_i \geq y_{li}$. If there is no censoring, $y_{li} = y_{ui} = y_i$. When y_{li} and y_{ui} are real valued and not equal, we know that $y_{li} \leq y_i \leq y_{ui}$.

The log likelihood is

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

$$+ \sum_{i \in L} w_i \ln \Phi\left(\frac{y_{ui} - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)$$

$$+ \sum_{i \in R} w_i \ln \Phi\left(\frac{-y_{li} + \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)$$

$$+ \sum_{i \in I} w_i \ln \left\{ \Phi\left(\frac{y_{ui} - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right) - \Phi\left(\frac{y_{li} - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right) \right\}$$

where U is the set of observations where y_i is not censored, L is the set of observations where y_i is left-censored, R is the set of observations where y_i is right-censored, I is the set of observations where y_i is interval-censored, and w_i are the weights.

The conditional mean of y_i is

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

If we wished to condition on the censoring, we could calculate an expectation on $y_i^\star =$ $\max\{y_{li}, \min(y_{ij}, y_{ui})\}$ or a constrained mean $E(y_i|y_{li} < y_i < y_{ui})$. See *Predictions using the full model* in [ERM] **eprobit postestimation** for details on how this is done.

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 Φ_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 an interval regression framework.

Endogenous covariates

Continuous endogenous covariates

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

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

As in *Introduction*, we do not observe y_i but instead observe the endpoints y_{li} and y_{ui}. 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 ϵ_i and $\boldsymbol{\epsilon}_{ci}$ are multivariate normal with mean 0 and covariance

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

Conditional on the endogenous and exogenous covariates, ϵ_i has mean and variance

$$\mu_{1|c,i} = E\left(\epsilon_i|\mathbf{w}_{ci}, \mathbf{x}_i, \mathbf{z}_{ci}\right) = \boldsymbol{\sigma}'_{1c}\boldsymbol{\Sigma}_c^{-1}\left(\mathbf{w}_{ci} - \mathbf{z}_{ci}\mathbf{A}_c\right)'$$
$$\sigma^2_{1|c} = \text{Var}\left(\epsilon_i|\mathbf{w}_{ci}, \mathbf{x}_i, \mathbf{z}_{ci}\right) = \sigma^2 - \boldsymbol{\sigma}'_{1c}\boldsymbol{\Sigma}_c^{-1}\boldsymbol{\sigma}_{1c}$$

Let

$$r_{li} = y_{li} - \mathbf{x}_i\boldsymbol{\beta} - \mathbf{w}_{ci}\boldsymbol{\beta}_c - \mu_{1|c,i}$$
$$r_{ui} = y_{ui} - \mathbf{x}_i\boldsymbol{\beta} - \mathbf{w}_{ci}\boldsymbol{\beta}_c - \mu_{1|c,i}$$

The log likelihood is

$$\ln L = \sum_{i \in U} w_i \ln \phi\left(r_{li}, \sigma^2_{1|c}\right)$$
$$+ \sum_{i \in L} w_i \ln \Phi^*_1\left(-\infty, r_{ui}, \sigma^2_{1|c}\right)$$
$$+ \sum_{i \in R} w_i \ln \Phi^*_1\left(r_{li}, \infty, \sigma^2_{1|c}\right)$$
$$+ \sum_{i \in I} w_i \ln \Phi^*_1\left(r_{li}, r_{ui}, \sigma^2_{1|c}\right)$$
$$+ \sum_{i=1}^{N} w_i \ln \phi_C(\mathbf{w}_{ci} - \mathbf{z}_{ci}\mathbf{A}_c, \mathbf{\Sigma}_c)$$

where U is the set of observations where y_i is not censored, L is the set of observations where y_i is left-censored, R is the set of observations where y_i is right-censored, and I is the set of observations where y_i is interval-censored.

The conditional mean of y_i is

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

Binary and ordinal endogenous covariates

Here we begin by formulating the interval 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**. As in *Introduction*, we do not observe y_i but instead observe the endpoints y_{li} and y_{ui}. The binary and ordinal endogenous errors $\epsilon_{b1i}, \ldots, \epsilon_{bBi}$ and outcome error ϵ_i are multivariate normal with 0 mean and covariance

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

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

As in *Binary and ordinal endogenous covariates* in [ERM] **eregress**, for the uncensored observations, we write the joint density of y_i and $\mathbf{w}_{bi}$ using the conditional density of $\epsilon_{b1i}, \ldots, \epsilon_{bBi}$ on ϵ_i. For the censored observations, we use tools discussed in *Likelihood for multiequation models* in [ERM] **eprobit** to formulate the joint density directly.

For $i \in U$, the uncensored observations, define

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

For the censored observations, define

$$r_{li} = y_{li} - \left(\mathbf{x}_i \boldsymbol{\beta} + \mathbf{wind}_{b1i} \boldsymbol{\beta}_{b1} + \cdots + \mathbf{wind}_{bBi} \boldsymbol{\beta}_{bB} \right)$$
$$r_{ui} = y_{ui} - \left(\mathbf{x}_i \boldsymbol{\beta} + \mathbf{wind}_{b1i} \boldsymbol{\beta}_{b1} + \cdots + \mathbf{wind}_{bBi} \boldsymbol{\beta}_{bB} \right)$$

Let

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

Now the log likelihood is

$$
\begin{aligned}
\ln L = & \sum_{i \in U} w_i \ln \left\{ \Phi_B^*(\mathbf{l}_i, \mathbf{u}_i, \boldsymbol{\Sigma}_{b|1}) \, \phi(r_i, \sigma^2) \right\} \\
& + \sum_{i \in L} w_i \ln \Phi_{B+1}^* ([\mathbf{l}_{bi} \quad -\infty], [\mathbf{u}_{bi} \quad r_{ui}], \boldsymbol{\Sigma}) \\
& + \sum_{i \in R} w_i \ln \Phi_{B+1}^* ([\mathbf{l}_{bi} \quad r_{li}], [\mathbf{u}_{bi} \quad \infty], \boldsymbol{\Sigma}) \\
& + \sum_{i \in I} w_i \ln \Phi_{B+1}^* ([\mathbf{l}_{bi} \quad r_{li}], [\mathbf{u}_{bi} \quad r_{ui}], \boldsymbol{\Sigma})
\end{aligned}
$$

where U is the set of observations where y_i is not censored, L is the set of observations where y_i is left-censored, R is the set of observations where y_i is right-censored, and I is the set of observations where y_i is interval-censored. The vectors $\mathbf{l}_{bi}$ and $\mathbf{u}_{bi}$ are the upper and lower limits for the binary and ordinal endogenous regressors defined in *Binary and ordinal endogenous covariates* in [ERM] **eprobit**. The vectors $\mathbf{l}_i$ and $\mathbf{u}_i$ are the upper and lower limits for the binary and ordinal endogenous regressors defined in *Binary and ordinal endogenous covariates* in [ERM] **eregress**.

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 2×12 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 $\epsilon_{1i}, \ldots, \epsilon_{Mi}$, we have

$$y_{1i} = \mathbf{x}_i\boldsymbol{\beta} + \mathbf{wind}_{b1i}\boldsymbol{\beta}_{b1} + \cdots + \mathbf{wind}_{bBi}\boldsymbol{\beta}_{bB} + \epsilon_{1i}$$

$$\vdots$$

$$y_{Mi} = \mathbf{x}_i\boldsymbol{\beta} + \mathbf{wind}_{b1i}\boldsymbol{\beta}_{b1} + \cdots + \mathbf{wind}_{bBi}\boldsymbol{\beta}_{bB} + \epsilon_{Mi}$$

$$y_i = \sum_{j=1}^{M} 1(\mathbf{w}_{bi} = \mathbf{v}_j)y_{ji}$$

For $j = 1, \ldots, M$, the endogenous errors $\epsilon_{b1i}, \ldots, \epsilon_{bBi}$ and outcome error ϵ_{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

$$\sigma_{i,b} = \sum_{j=1}^{M} 1(\mathbf{w}_{bi} = \mathbf{v}_j)\sigma_j$$

$$\boldsymbol{\Sigma}_{i,b} = \sum_{j=1}^{M} 1(\mathbf{w}_{bi} = \mathbf{v}_j)\boldsymbol{\Sigma}_j$$

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

Now the log likelihood for this model is

$$
\begin{aligned}
\ln L = &\sum_{i \in U} w_i \ln\left\{\Phi_B^*(\mathbf{l}_i, \mathbf{u}_i, \boldsymbol{\Sigma}_{i,b|1})\,\phi\!\left(r_i, \sigma_{i,b}^2\right)\right\} \\
&+ \sum_{i \in L} w_i \ln\Phi_{B+1}^*([\mathbf{l}_{bi} \quad -\infty], [\mathbf{u}_{bi} \quad r_{ui}], \boldsymbol{\Sigma}_{i,b}) \\
&+ \sum_{i \in R} w_i \ln\Phi_{B+1}^*([\mathbf{l}_{bi} \quad r_{li}], [\mathbf{u}_{bi} \quad \infty], \boldsymbol{\Sigma}_{i,b}) \\
&+ \sum_{i \in I} w_i \ln\Phi_{B+1}^*([\mathbf{l}_{bi} \quad r_{li}], [\mathbf{u}_{bi} \quad r_{ui}], \boldsymbol{\Sigma}_{i,b})
\end{aligned}
$$

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 observe only 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 [CAUSAL] **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 an ordered probit model, and we call the treatment assignment error ϵ_{ti}. An interval 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

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

As in *Introduction*, we do not observe y_i but instead observe the endpoints y_{li} and y_{ui}.

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 $j = 1, \ldots, T$, ϵ_{ji} and ϵ_{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. The likelihood is derived in a similar manner to *Binary and ordinal endogenous covariates*.

For $i \in U$, the uncensored observations, define

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

For the censored observations, define

$$r_{li} = y_{li} - \mathbf{x}_i\boldsymbol{\beta}_j \quad \text{if} \quad t_i = v_j$$
$$r_{ui} = y_{ui} - \mathbf{x}_i\boldsymbol{\beta}_j \quad \text{if} \quad t_i = v_j$$

Now the log likelihood is

$$\ln L = \sum_{i \in U} w_i \ln \left\{ \Phi_1^* \left(l_{ti} - \frac{\rho_{1t}}{\sigma} r_i, u_{ti} - \frac{\rho_{1t}}{\sigma} r_i, 1 - \rho_{1t}^2 \right) \phi(r_i, \sigma^2) \right\}$$
$$+ \sum_{i \in L} w_i \ln \Phi_2^*([l_{ti} \quad -\infty], [u_{ti} \quad r_{ui}], \boldsymbol{\Sigma})$$
$$+ \sum_{i \in R} w_i \ln \Phi_2^*([l_{ti} \quad r_{li}], [u_{ti} \quad \infty], \boldsymbol{\Sigma})$$
$$+ \sum_{i \in I} w_i \ln \Phi_2^*([l_{ti} \quad r_{li}], [u_{ti} \quad r_{ui}], \boldsymbol{\Sigma})$$

where U is the set of observations where y_i is not censored, L is the set of observations where y_i is left-censored, R is the set of observations where y_i is right-censored, and I is the set of observations where y_i is interval-censored. 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 is

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

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

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

where ψ_{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 `eintreg`. 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 ϵ_i is calculated.

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

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

Define

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

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

$$\boldsymbol{\Sigma}_i = \sum_{j=1}^{T} 1(t_i = v_j)\boldsymbol{\Sigma}_j$$

Now the log likelihood for the model is

$$\begin{aligned}
\ln L = &\sum_{i \in U} w_i \ln\left\{\Phi_1^*\left(l_{ti} - \frac{\rho_i}{\sigma_i}r_i, u_{ti} - \frac{\rho_i}{\sigma_i}r_i, 1 - \rho_i^2\right)\phi(r_i, \sigma_i^2)\right\} \\
&+ \sum_{i \in L} w_i \ln\Phi_2^*([l_{ti} \quad -\infty], [u_{ti} \quad r_{ui}], \boldsymbol{\Sigma}_i) \\
&+ \sum_{i \in R} w_i \ln\Phi_2^*([l_{ti} \quad r_{li}], [u_{ti} \quad \infty], \boldsymbol{\Sigma}_i) \\
&+ \sum_{i \in I} w_i \ln\Phi_2^*([l_{ti} \quad r_{li}], [u_{ti} \quad r_{ui}], \boldsymbol{\Sigma}_i)
\end{aligned}$$

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 ϵ_{ti} are multivariate normal, so for $j = 1, \ldots, T$, we can decompose ϵ_{ji} such that

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

where ψ_{ji} has mean 0 and is independent of t_i.

It follows that

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

The mean of ϵ_{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 `eintreg`.

Endogenous sample selection

Probit endogenous sample selection

The regression for outcome y_i with selection on s_i has the form

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

where $\mathbf{x}_i$ are covariates that affect the outcome and $\mathbf{z}_{si}$ are covariates that affect selection. As in the *Introduction* above, we do not observe y_i but instead observe the endpoints y_{li} and y_{ui}. If $s_i = 1$, then the observation is selected, and there is an interval regression contribution to the likelihood. If $s_i = 0$, then the observation is not selected, and there is no interval regression contribution to the likelihood.

The unobserved errors ϵ_i and ϵ_{si} are normal with mean 0 and covariance

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

The likelihood is derived in a similar manner to that in *Treatment*.

For $i \in U$, the uncensored and selected observations, define

$$
r_i = y_i - \mathbf{x}_i\boldsymbol{\beta}
$$

Let

$$
\begin{aligned}
\mu_{s|1,i} &= \frac{\rho_{1s}}{\sigma} r_i \\
\sigma_{s|1} &= 1 - \rho_{1s}^2
\end{aligned}
$$

For the selection indicator s_i, the lower and upper limits on ϵ_{si} are

$$
l_{si} = \begin{cases} -\infty & s_i = 0 \\ -\mathbf{z}_{si}\boldsymbol{\alpha}_s & s_i = 1 \end{cases}
\qquad
u_{si} = \begin{cases} -\mathbf{z}_{si}\boldsymbol{\alpha}_s & s_i = 0 \\ \infty & s_i = 1 \end{cases}
$$

For the censored but selected observations, $i \notin U$, define

$$r_{li} = y_{li} - \mathbf{x}_i \boldsymbol{\beta}_j$$
$$r_{ui} = y_{ui} - \mathbf{x}_i \boldsymbol{\beta}_j$$

Now the log likelihood is

$$
\begin{aligned}
\ln L = &\sum_{i \in U} w_i \ln \left\{ \Phi_1^*(l_{si} - \mu_{s|1,i}, u_{si} - \mu_{s|1,i}, \sigma_{s|1}^2) \, \phi(r_i, \sigma^2) \right\} \\
&+ \sum_{i \in L} w_i \ln \Phi_2^*([l_{si} \quad -\infty], [u_{si} \quad r_{ui}], \boldsymbol{\Sigma}) \\
&+ \sum_{i \in R} w_i \ln \Phi_2^*([l_{si} \quad r_{li}], [u_{si} \quad \infty], \boldsymbol{\Sigma}) \\
&+ \sum_{i \in I} w_i \ln \Phi_2^*([l_{si} \quad r_{li}], [u_{si} \quad r_{ui}], \boldsymbol{\Sigma}) \\
&\sum_{i \notin S} w_i \ln \Phi_1^*(l_{si}, u_{si}, 1)
\end{aligned}
$$

where U is the set of observations where y_i is not censored, L is the set of observations where y_i is left-censored, R is the set of observations where y_i is right-censored, I is the set of observations where y_i is interval-censored, and S is the set of selected observations.

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 endogenous sample-selection indicator. We allow the selection variable to be left-censored or right-censored.

The underlying regression model for y_i with tobit selection on s_i has the form

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

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

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

$$
s_i = \begin{cases}
l_i & s_i^\star \leq l_i \\
s_i^\star & l_i < s_i^\star < u_i \\
u_i & s_i^\star \geq u_i
\end{cases}
$$

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

As in *Introduction*, y_i is observed via the endpoints y_{li} and y_{ui}. If $s_i^\star$ is not censored ($l_i < s_i^\star < u_i$), then the observation is selected, and there is an interval regression contribution to the likelihood. Otherwise, if $s_i^\star$ is left-censored ($s_i^\star < l_i$) or right-censored ($s_i^\star > l_i$), then the observation is not selected, and there is no interval regression contribution to the likelihood. The unobserved errors ϵ_i and ϵ_{si} are normal with mean 0 and covariance

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

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 eintreg (specify tobitselect(... main)). On the nonselected observations, we treat s_i like the probit endogenous sample-selection indicator in *Probit endogenous sample selection*.

Conditional on $s_i^\star$ and the exogenous covariates, ϵ_i has mean and variance

$$\mu_{1|s,i} = E\left(\epsilon_i | s_i^\star, \mathbf{x}_i, \mathbf{z}_{si}\right) = \sigma_{1s}\sigma_s^{-2}\left(\mathbf{s}_i^\star - \mathbf{z}_{si}\boldsymbol{\alpha}_s\right)$$
$$\sigma_{1|s}^2 = \mathrm{Var}\left(\epsilon_i | s_i^\star, \mathbf{x}_i, \mathbf{z}_{si}\right) = \sigma^2 - \sigma_{1s}\sigma_s^{-2}\sigma_{1s}$$

Let

$$r_{li} = y_{li} - \mathbf{x}_i\boldsymbol{\beta} - \mu_{1|s,i}$$
$$r_{ui} = y_{ui} - \mathbf{x}_i\boldsymbol{\beta} - \mu_{1|s,i}$$

The log likelihood is

$$\begin{aligned} \ln L = &\sum_{i \in U} w_i \ln \phi\left(r_{li}, \sigma_{1|s}^2\right) \\ &+ \sum_{i \in L} w_i \ln \Phi_1^*\left(-\infty, r_{ui}, \sigma_{1|s}^2\right) \\ &+ \sum_{i \in R} w_i \ln \Phi_1^*\left(r_{li}, \infty, \sigma_{1|s}^2\right) \\ &+ \sum_{i \in I} w_i \ln \Phi_1^*\left(r_{li}, r_{ui}, \sigma_{1|s}^2\right) \\ &+ \sum_{i \in S} w_i \ln \phi(s_i - \mathbf{z}_{si}\boldsymbol{\alpha}_s, \sigma_s^2) \\ &+ \sum_{i \in L_n} w_i \ln\Phi_1^*(l_{li}, u_{li}, 1) \\ &+ \sum_{i \in R_n} w_i \ln\Phi_1^*(l_{ui}, u_{ui}, 1) \end{aligned}$$

where S is the set of observations for which y_{li} and y_{ui} are observed, $U \subset S$ is the set of observations where y_i is not censored, $L \subset S$ is the set of observations where y_i is left-censored, $R \subset S$ is the set of observations where y_i is right-censored, $I \subset S$ is the set of observations where y_i is interval-censored, L_n is the set of observations for which $s_i^\star$ is left-censored, and R_n is the set of observations for which $s_i^\star$ 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

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

Random effects

For an interval regression with random effects, we observe panel data. For panel $i = 1, \ldots, N$ and observation $j = 1, \ldots, N_i$, an interval regression of y_{ij} on exogenous covariates $\mathbf{x}_{ij}$ with random effect u_i has the form

$$y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \epsilon_{ij} + u_i$$

As in *Introduction*, we do not observe y_{ij} but instead observe endpoints y_{lij} and y_{uij}. The random effect u_i is normal with mean 0 and variance σ_u^2. It is independent of the observation-level error ϵ_{ij}, which is normal with mean 0 and variance σ^2.

We derive the likelihood by using the conditional density of y_{lij} and y_{uij} on the random effect u_i and the marginal density of u_i. Multiplying them together we have the joint density, which is integrated over u_i.

Let

$$l_{ij}(u) = \sum_{j\in U_i} \phi\big(y_{ij} - \mathbf{x}_{ij}\boldsymbol{\beta} - u, \sigma^2\big)$$
$$+ \sum_{j\in L_i} \Phi\left(\frac{y_{uij} - \mathbf{x}_{ij}\boldsymbol{\beta} - u}{\sigma}\right)$$
$$+ \sum_{i\in R_i} \Phi\left(\frac{-y_{lij} + \mathbf{x}_{ij}\boldsymbol{\beta} - u}{\sigma}\right)$$
$$+ \sum_{i\in I_i} \left\{\Phi\left(\frac{y_{uij} - \mathbf{x}_{ij}\boldsymbol{\beta} - u}{\sigma}\right) - \Phi\left(\frac{y_{lij} - \mathbf{x}_{ij}\boldsymbol{\beta} - u}{\sigma}\right)\right\}$$

where U_i is the set of observations where y_{ij} is not censored, L_i is the set of observations where y_{ij} is left-censored, R_i is the set of observations where y_{ij} is right-censored, and I_i is the set of observations where y_{ij} is interval-censored.

The likelihood for panel i is

$$L_i = \int_{-\infty}^{\infty} \phi\left(\frac{u_i}{\sigma_u}\right) \prod_{j=1}^{N_i} l_{ij}(u_i) du_i$$

We can approximate this integral using Gauss–Hermite quadrature. For q-point Gauss–Hermite quadrature, let the abscissa and weight pairs be denoted by (a_{ki}, w_{ki}), $k = 1, \ldots, q$. The Gauss–Hermite quadrature approximation is then

$$\int_{-\infty}^{\infty} f(x)\exp(-x^2)\,dx \approx \sum_{k=1}^{q} w_{ki} f(a_{ki})$$

The default approximation used by xteintreg is mean–variance adaptive Gauss–Hermite quadrature. This chooses optimal abscissa and weights for each panel. See *Likelihood for multiequation models* in [ERM] **eprobit** for more information on the use of mean–variance adaptive Gauss–Hermite quadrature.

Using the quadrature approximation, the log likelihood is

$$\ln L = \sum_{i=1}^{N} \ln \left\{ \sum_{k=1}^{q} w_{ki} \prod_{j=1}^{N_i} l_{ij}(\sigma_u a_{ki}) \right\}$$

The conditional mean of y_{ij} is

$$E(y_{ij}|\mathbf{x}_{ij}) = \mathbf{x}_{ij}\boldsymbol{\beta}$$

Combinations of features

Extended interval 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.

Angrist, J. D. 2001. Estimation of limited dependent variable models with dummy endogenous regressors: Simple strategies for empirical practice. *Journal of Business and Economic Statistics* 19: 2–28. https://doi.org/10.1198/07350010152472571.

Bartus, T., and D. Roodman. 2014. Estimation of multiprocess survival models with cmp. *Stata Journal* 14: 756–777.

Cameron, A. C., and P. K. Trivedi. 2022. *Microeconometrics Using Stata*. 2nd ed. College Station, TX: Stata Press.

Fernandez-Cornejo, J., and S. Wechsler. 2012. Revisiting the impact of Bt corn adoption by U.S. farmers. *Agricultural and Resource Economics Review* 41: 377–390. https://doi.org/10.1017/S1068280500001337.

Heckman, J. J. 1976. "The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models". In *Annals of Economic and Social Measurement*, edited by S. V. Berg, vol. 5: 475–492. Cambridge, MA: National Bureau of Economic Research.

———. 1979. Sample selection bias as a specification error. *Econometrica* 47: 153–161. https://doi.org/10.2307/1912352.

Maddala, G. S. 1983. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511810176.

Maddala, G. S., and L.-F. Lee. 1976. "Recursive models with qualitative endogenous variables". In *Annals of Economic and Social Measurement*, edited by S. V. Berg, vol. 5: 525–545. Cambridge, MA: National Bureau of Economic Research.

Mairesse, J., and S. Robin. 2009. Innovation and productivity: A firm-level analysis for French manufacturing and services using CIS3 and CIS4 data (1998–2000 and 2002–2004). CREST–ENSAE.

Newey, W. K. 1987. Efficient estimation of limited dependent variable models with endogenous explanatory variables. *Journal of Econometrics* 36: 231–250. https://doi.org/10.1016/0304-4076(87)90001-7.

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

Tobin, J. 1958. Estimation of relationships for limited dependent variables. *Econometrica* 26: 24–36. https://doi.org/10.2307/1907382.

White, H. L., Jr. 1996. *Estimation, Inference and Specification Analysis*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CCOL0521252806.

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. https://doi.org/10.1016/j.jeconom.2014.04.020.

———. 2020. *Introductory Econometrics: A Modern Approach*. 7th ed. Boston: Cengage.

Also see

[ERM] **eintreg postestimation** — Postestimation tools for eintreg and xteintreg

[ERM] **eintreg predict** — predict after eintreg and xteintreg

[ERM] **predict advanced** — predict's advanced features

[ERM] **predict treatment** — predict for treatment statistics

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

[ERM] **Intro 9** — Conceptual introduction via worked example

[R] **intreg** — Interval regression

[R] **ivtobit** — Tobit model with continuous endogenous covariates

[R] **tobit** — Tobit regression

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

[XT] **xtintreg** — Random-effects interval-data regression model

[XT] **xttobit** — Random-effects tobit model

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

For suggested citations, see the FAQ on citing Stata documentation.