Title

stata.com

eoprobit — Extended ordered probit regression

Description

`eoprobit` fits an ordered probit 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

Ordered probit regression of `y` on `x` with continuous endogenous covariate `y2` modeled by `x` and `z`

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

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

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

Ordered probit regression of `y` on `x` with binary endogenous covariate `d` modeled by `x` and `z`

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

Ordered probit regression of `y` on `x` with endogenous treatment recorded in `trtvar` and modeled by `x` and `z`

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

Ordered probit regression of `y` on `x` with exogenous treatment recorded in `trtvar`

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

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

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

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

```
eoprobit 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`

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

Menu

Statistics > Endogenous covariates > Models adding selection and treatment > Ordered probit regression

Syntax

Basic ordered probit regression with endogenous covariates

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

Basic ordered probit regression with endogenous treatment assignment

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

Basic ordered probit regression with exogenous treatment assignment

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

Basic ordered probit regression with sample selection

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

Basic ordered probit regression with tobit sample selection

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

Ordered probit regression combining endogenous covariates, treatment, and selection

```
eoprobit 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	
`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*$_{\text{tr}}$] [`,` *tropts*]

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

extrspec is *tvar* [`,` `nomain` `nocutsinteract` `nointeract`]

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
nomain	do not add endogenous covariate to main equation
noconstant	suppress constant term

tropts	Description
Model	
nomain	do not add treatment indicator to main equation
nocutsinteract	do not interact treatment with cutpoints
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

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**.

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 eoprobit but is not shown in the dialog box:

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

Remarks and examples

stata.com

eoprobit fits models that we refer to as "extended ordered probit regression models", meaning that they accommodate endogenous covariates, nonrandom treatment assignment, and endogenous sample selection. eoprobit can account for these complications whether they arise individually or in combination.

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

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

For an introduction to eoprobit and the other extended regression commands (eregress, eintreg, and eprobit), 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 when fitting a linear, interval, probit, or ordered probit model.

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

[ERM] **intro 6** discusses interpretation of results. You can interpret coefficients from eoprobit 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 particularly helpful if you are familiar with heckoprobit 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 eoprobit.

[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. Although the example uses eregress, the discussion applies equally to eoprobit. 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 eoprobit, see

[ERM] **example 6a**	Ordered probit regression with endogenous treatment
[ERM] **example 6b**	Ordered probit regression with endogenous covariate and 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 eoprobit can fit. For instance, eoprobit can be used to fit models like the ordered probit model with endogenous sample selection discussed in De Luca and Perotti (2011) and the ordered probit models with continuous or binary endogenous covariates discussed in Wooldridge (2010, sec. 16.3.3). Roodman (2011) investigated ordered probit 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 living with a child on the happiness of the elderly (Chyi and Mao 2012) and the effect of parental migration on child education (Botezat and Pfeiffer 2014).

Stored results

eoprobit 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)	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)	eoprobit
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 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(marginsdefault)	default predict() specification for 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 ordered probit model. The estimator implemented in `eoprobit` is a maximum likelihood estimator covered by the results in chapter 13 of Wooldridge (2010) and White (1996).

The log-likelihood function maximized by `eoprobit` 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

An ordered probit regression of outcome y_i on covariates $\mathbf{x}_i$ may be written as

$$y_i = v_h \quad \text{iff} \quad \kappa_{h-1} < \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i \leq \kappa_h$$

The values $v_1, \ldots, v_H$ are real numbers such that $v_h < v_m$ for $h < m$. κ_0 is taken as $-\infty$ and κ_H is taken as $+\infty$. The unobserved error ϵ_i is standard normal.

The log likelihood for this model is

$$\ln L = \sum_{i=1}^{N} w_i \ln \left[\begin{array}{l} 1(y_i = v_1)\Phi\left(-\mathbf{x}_i\boldsymbol{\beta}\right) \\ \quad + \displaystyle\sum_{h=2}^{H-1} 1(y_i = v_h)\left\{\Phi\left(\kappa_h - \mathbf{x}_i\boldsymbol{\beta}\right) - \Phi\left(\kappa_{h-1} - \mathbf{x}_i\boldsymbol{\beta}\right)\right\} \\ \quad + 1(y_i = v_H)\Phi\left(\mathbf{x}_i\boldsymbol{\beta}\right) \end{array} \right]$$

where w_i are the weights.

For $h = 0, \ldots, H$, define

$$c_{ih} = \begin{cases} -\infty & h = 0 \\ \kappa_h - \mathbf{x}_i\boldsymbol{\beta} & h = 1, \ldots, H-1 \\ \infty & h = H \end{cases} \tag{1}$$

This leads to the limits

$$l_{1i} = c_{i(h-1)} \quad \text{if} \quad y_i = v_h \tag{2}$$

and

$$u_{1i} = c_{ih} \quad \text{if} \quad y_i = v_h \tag{3}$$

These are limits on the unobserved ϵ_i based on the observed values of y_i and $\mathbf{x}_i$. They let us rewrite the log likelihood concisely as

$$\ln L = \sum_{i=1}^{N} w_i \ln \Phi_1^*(l_{1i}, u_{1i}, 1)$$

The conditional probabilities of success can be written using similar notation. For $h = 1, \ldots, H$,

$$\Pr(y_i = v_h | \mathbf{x}_i) = \Phi_1^*(c_{i(h-1)}, c_{ih}, 1) \tag{4}$$

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 ordered probit framework.

Endogenous covariates

Continuous endogenous covariates

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

$$y_i = v_h \quad \text{iff} \quad \kappa_{h-1} < \mathbf{x}_i\boldsymbol{\beta} + \mathbf{w}_{ci}\boldsymbol{\beta}_c + \epsilon_i \leq \kappa_h$$

$$\mathbf{w}_{ci} = \mathbf{z}_{ci}\mathbf{A}_c + \boldsymbol{\epsilon}_{ci}$$

The values $v_1, \ldots, v_H$ are real numbers such that $v_h < v_m$ for $h < m$. κ_0 is taken as $-\infty$ and κ_H is taken as $+\infty$. The vector $\mathbf{z}_{ci}$ contains variables from $\mathbf{x}_i$ and other covariates that affect $\mathbf{w}_{ci}$. The unobserved errors ϵ_i and $\boldsymbol{\epsilon}_{ci}$ are multivariate normal with mean 0 and covariance

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

As in *Continuous endogenous covariates* in [ERM] **eprobit**, the likelihood can be written using the conditional density of ϵ_i on $\mathbf{w}_{ci}$.

Now, for $h = 0, \ldots, H$, define

$$c_{ih} = \begin{cases} -\infty & h = 0 \\ \kappa_h - \mathbf{x}_i\boldsymbol{\beta} - \boldsymbol{\sigma}_{1c}'\boldsymbol{\Sigma}_c^{-1}(\mathbf{w}_{ci} - \mathbf{z}_{ci}\mathbf{A}_c)' & h = 1, \ldots, H-1 \\ \infty & h = H \end{cases}$$

These expressions used the conditional mean of ϵ_i. The lower and upper limits for the y_i probability are

$$l_{1i} = c_{i(h-1)} \quad \text{if} \quad y_i = v_h$$

and

$$u_{1i} = c_{ih} \quad \text{if} \quad y_i = v_h$$

Using these limits, the conditional variance, and the conditional density of $\mathbf{w}_{ci}$, we obtain the log likelihood

$$\ln L = \sum_{i=1}^{N} w_i \left\{ \ln \Phi_1^* \left(l_{1i}, u_{1i}, 1 - \boldsymbol{\sigma}_{1c}' \boldsymbol{\Sigma}_c^{-1} \boldsymbol{\sigma}_{1c} \right) + \ln \phi_C (\mathbf{w}_{ci} - \mathbf{z}_{ci}\mathbf{A}_c, \boldsymbol{\Sigma}_c) \right\}$$

The conditional probabilities of success can be written using similar notation. For $h = 1, \ldots, H$,

$$\Pr(y_i = v_h | \mathbf{x}_i) = \Phi_1^*(c_{i(h-1)}, c_{ih}, 1 - \boldsymbol{\sigma}_{1c}' \boldsymbol{\Sigma}_c^{-1} \boldsymbol{\sigma}_{1c})$$

Binary and ordinal endogenous covariates

Here, we begin by formulating the ordered probit 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**. So we have

$$y_i = v_h \quad \text{iff} \quad \kappa_{h-1} < \mathbf{x}_i\boldsymbol{\beta} + \mathbf{wind}_{b1i}\boldsymbol{\beta}_{b1} + \cdots + \mathbf{wind}_{bBi}\boldsymbol{\beta}_{bB} + \epsilon_i \leq \kappa_h$$

where the values $v_1, \ldots, v_H$ are real numbers such that $v_h < v_m$ for $h < m$. κ_0 is taken as $-\infty$ and κ_H is taken as $+\infty$. The $\mathbf{wind}_{bji}$ vectors are defined in *Binary and ordinal endogenous covariates* in [ERM] **eprobit**. The outcome error ϵ_i and binary and ordinal endogenous errors $\epsilon_{b1i}, \ldots, \epsilon_{bBi}$ are multivariate normal with mean 0 and covariance

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

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

Now, for $h = 0, \ldots, H$, define

$$c_{ih} = \begin{cases} -\infty & h = 0 \\ \kappa_h - \mathbf{x}_i\boldsymbol{\beta} - \mathbf{wind}_{b1i}\boldsymbol{\beta}_{b1} - \cdots - \mathbf{wind}_{bBi}\boldsymbol{\beta}_{bB} & h = 1, \ldots, H-1 \\ \infty & h = H \end{cases}$$

The lower and upper limits for the y_i probability are

$$l_{1i} = c_{i(h-1)} \quad \text{if} \quad y_i = v_h$$

and

$$u_{1i} = c_{ih} \quad \text{if} \quad y_i = v_h$$

Let

$$\mathbf{l}_i = [\, l_{1i} \quad l_{b1i} \quad \ldots \quad l_{bBi} \,]$$
$$\mathbf{u}_i = [\, u_{1i} \quad u_{b1i} \quad \ldots \quad u_{bBi} \,]$$

where the l_{bji} and u_{bji} are the lower and upper limits for the binary and ordinal endogenous covariate probabilities. They are defined in *Binary and ordinal endogenous covariates* in [ERM] **eprobit**.

So the log likelihood for this model is

$$\ln L = \sum_{i=1}^{N} w_i \ln \Phi^*_{B+1}(\mathbf{l}_i, \mathbf{u}_i, \mathbf{\Sigma})$$

Now, let

$$\mathbf{l}_{bi} = [\, l_{b1i} \quad \ldots \quad l_{bBi} \,]$$
$$\mathbf{u}_{bi} = [\, u_{b1i} \quad \ldots \quad u_{bBi} \,]$$
$$\mathbf{l}_{ih1} = [\, c_{i(h-1)} \quad \mathbf{l}_{bi} \,]$$
$$\mathbf{u}_{ih1} = [\, c_{ih} \quad \mathbf{u}_{bi} \,]$$

The conditional probabilities are

$$\Pr(y_i = v_h | \mathbf{x}_i, \mathbf{z}_{b1i}, \ldots, \mathbf{z}_{bBi}, \mathbf{w}_{bi}) = \frac{\Phi^*_{B+1}(\mathbf{l}_{ih1}, \mathbf{u}_{ih1}, \mathbf{\Sigma})}{\Phi^*_{B}(\mathbf{l}_{bi}, \mathbf{u}_{bi}, \mathbf{\Sigma}_b)}$$

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}, …, y_{Ti}.

When we observe treatment t_i with levels v_1, …, v_T, we have

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

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.

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 ϵ_{ti}. An ordered probit regression of y_i on treatment t_i with levels $v_{t1}, \ldots, v_{tT}$ has the form

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

where for $j = 1, \ldots, T$ and exogenous covariates $\mathbf{x}_i$

$$y_{ji} = v_h \quad \text{iff} \quad \kappa_{(h-1)j} < \mathbf{x}_i\boldsymbol{\beta}_j + \epsilon_{ji} \leq \kappa_{hj}$$

The values $v_1, \ldots, v_H$ are real numbers such that $v_h < v_m$ for $h < m$. For $j = 1, \ldots, T$, κ_{0j} is taken as $-\infty$ and κ_{Hj} is taken as $+\infty$. For $j = 1, \ldots, T$, ϵ_{ji} and ϵ_{ti} are bivariate normal with mean 0 and covariance

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

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. Because the unobserved errors are bivariate normal, we can express the log likelihood in terms of the Φ_2^* function.

For $j = 1, \ldots, T$ and $h = 0, \ldots, H$, let

$$c_{1ihj} = \begin{cases} -\infty & h = 0 \\ \kappa_{hj} - \mathbf{x}_i\boldsymbol{\beta}_j & h = 1, \ldots, H-1 \\ \infty & h = H \end{cases}$$

The lower and upper limits for the y_i probability are

$$l_{1i} = c_{i(h-1)j} \quad \text{if} \quad y_i = v_h, t_i = v_{tj}$$

and

$$u_{1i} = c_{ihj} \quad \text{if} \quad y_i = v_h, t_i = v_{tj}$$

The log likelihood for the model is

$$\ln L = \sum_{i=1}^{N} w_i \ln \Phi_2^*([\, l_{1i} \quad l_{ti} \,], [\, u_{1i} \quad u_{ti} \,], \boldsymbol{\Sigma})$$

where the lower and upper limits for the treatment probability, l_{ti} and u_{ti}, are defined in *Treatment* in [ERM] **eprobit**.

The conditional probability of obtaining treatment level v_{th} is

$$\Pr(t_i = v_{th} | \mathbf{z}_{ti}) = \Phi_1^*(c_{ti(h-1)}, c_{tih}, 1)$$

where the cutpoints for the treatment probabilities c_{tij} are defined in *Treatment* in [ERM] **eprobit**.

For $h = 1, \ldots, H$, the conditional probabilities for outcome level v_h at treatment level v_{tj} are

$$\Pr(y_i = v_h|\mathbf{x}_i, \mathbf{z}_{ti}, t_i = v_{tj}) = \frac{\Phi_2^*([\, c_{1i(h-1)j} \quad c_{ti(j-1)} \,], [\, c_{1ihj} \quad c_{tij} \,], \boldsymbol{\Sigma})}{\Phi_1^*(c_{ti(j-1)}, c_{tij}, 1)}$$

The conditional POM for treatment group j and outcome category h is

$$\text{POM}_{hj}(\mathbf{x}_i) = E\{1(y_{ji} = v_h)|\mathbf{x}_i\} = \Phi_1^*(c_{1i(h-1)j}, c_{1i(h-1)j}, 1)$$

Conditional on the covariates $\mathbf{x}_i$ and $\mathbf{z}_{ti}$ and the treatment $t_i = v_m$, the POM for treatment group j and outcome category h is

$$\begin{aligned}\text{POM}_{hj}(\mathbf{x}_i, \mathbf{z}_{ti}, t_i = v_m) &= E\{1(y_{ji} = v_h)|\mathbf{x}_i, \mathbf{z}_{ti}, t_i = v_{tm}\} \\ &= \frac{\Phi_2^*([\, c_{1i(h-1)j} \quad c_{ti(m-1)} \,], [\, c_{1ihj} \quad c_{tim} \,], \boldsymbol{\Sigma})}{\Phi_1^*(c_{ti(m-1)}, c_{tim}, 1)}\end{aligned}$$

Without loss of generality, $t_i = v_{t1}$ corresponds to the control or base level of the treatment. Treatment effects are the differences between the potential outcomes $y_{2i}, \ldots, y_{Ti}$ and the control y_{1i}. When the potential outcomes are ordered probit, the treatment effect on a particular category is of interest.

The treatment effect of treatment group j on category h is $1(y_{ji} = v_h) - 1(y_{1i} = v_h)$, the difference in the outcome for individual i on being in category h if the individual receives the treatment $t_i = v_{tj}$ instead of the control $t_i = v_{t1}$. Evaluating this treatment effect lets us see how the treatment affects the probability of belonging to outcome category h.

For treatment group j, the treatment effect on category h conditioned on $\mathbf{x}_i$ is

$$\begin{aligned}\text{TE}_{hj}(\mathbf{x}_i) &= E\{1(y_{ji} = v_h) - 1(y_{1i} = v_h)|\mathbf{x}_i\} \\ &= \text{POM}_{hj}(\mathbf{x}_i) - \text{POM}_{h1}(\mathbf{x}_i)\end{aligned}$$

For treatment group j, the treatment effect on the treated (TET) on category h in treatment group m conditioned on $\mathbf{x}_i$ and $\mathbf{z}_{ti}$ is

$$\begin{aligned}\text{TET}_{hj}(\mathbf{x}_i, \mathbf{z}_{ti}, t_i = v_m) &= E\{1(y_{ji} = v_h) - 1(y_{1i} = v_h)|\mathbf{x}_i, t_i = v_{t,m}\} \\ &= \text{POM}_{hj}(\mathbf{x}_i, \mathbf{z}_{ti}, t_i = v_m) - \text{POM}_{h1}(\mathbf{x}_i, \mathbf{z}_{ti}, t_i = v_m)\end{aligned}$$

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 fit with `eoprobit`. The POM for treatment group j and outcome category h is

$$\text{POM}_{hj} = E\{1(y_{ji} = v_h)\} = E\{\text{POM}_{hj}(\mathbf{x}_i)\}$$

The ATE for treatment group j and outcome category h is

$$\text{ATE}_{hj} = E\{1(y_{ji} = v_h) - 1(y_{1i} = v_h)\} = E\{\text{TE}_{hj}(\mathbf{x}_i)\}$$

For treatment group j, the average treatment effect on the treated (ATET) for outcome category h in treatment group m is

$$\begin{aligned}\text{ATET}_{hjm} &= E\{1(y_{ji} = v_h) - 1(y_{1i} = v_h)|t_i = v_m\} \\ &= E\{\text{TET}_{hj}(\mathbf{x}_i, \mathbf{z}_{ti}, t_i = v_m)|t_i = v_m\}\end{aligned}$$

Endogenous sample selection

Probit endogenous sample selection

An ordered probit model for outcome y_i with selection on s_i has the form

$$y_i = v_h \quad \text{iff} \quad \kappa_{h-1} < \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i \leq \kappa_h$$

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

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 values $v_1, \ldots, v_H$ are real numbers such that $v_h < v_m$ for $h < m$. κ_0 is taken as $-\infty$ and κ_H is taken as $+\infty$.

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

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

The lower and upper limits for the y_i probability, l_{1i} and u_{1i}, are as defined in (1)–(3). For the selection indicator, the lower and upper limits l_{si} and u_{si} are defined in *Probit endogenous sample selection* in [ERM] **eprobit**.

The log likelihood for the model is

$$\ln L = \sum_{i\in S} w_i \ln\Phi_2^*([\,l_{1i} \quad l_{si}\,],[\,u_{1i} \quad u_{si}\,],\boldsymbol{\Sigma}) + \sum_{i\notin S} w_i \ln\Phi_1^*(l_{si}, u_{si}, 1)$$

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

In this model, the probability of success is usually predicted conditional on the covariates $\mathbf{x}_i$ and not on the selection status s_i. The formulas for the conditional probability are thus the same as in (4).

The conditional probability of selection is

$$\Pr\left(s_i = 1|\mathbf{z}_{si}\right) = \Phi_1^*(-\mathbf{z}_{si}\boldsymbol{\alpha}_s, \infty, 1)$$

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- or right-censored.

An ordered probit model for outcome y_i with tobit selection on s_i has the form

$$y_i = v_h \quad \text{iff} \quad \kappa_{h-1} < \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i \leq \kappa_h$$

where the values $v_1, \ldots, v_H$ are real numbers such that $v_h < v_m$ for $h < m$. κ_0 is taken as $-\infty$ and κ_H is taken as $+\infty$.

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. If $l_i < s_i^\star < u_i$, then y_i is observed. y_i is not observed if $s_i^\star \le l_i$, that is, if $s_i^\star$ is left-censored. y_i is also not observed if $s_i^\star$ is right-censored, $s_i^\star \ge u_i$. The unobserved errors ϵ_i and ϵ_{si} are normal with mean 0 and covariance

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

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

The conditional mean of ϵ_i is used in the lower and upper limits for the y_i probability for selected observations. Let

$$c_{i,h} = \begin{cases} -\infty & h = 0 \\ \kappa_h - \mathbf{x}_i\boldsymbol{\beta} - \rho_{1s}\sigma_s^{-1}\left(s_i - \mathbf{z}_{si}\alpha_c\right) & h = 1, \ldots, H-1 \\ \infty & h = H \end{cases}$$

The limits for the y_i probability for selected observations are

$$l_{1i} = c_{i(h-1)} \quad \text{if} \quad y_i = v_h$$

and

$$u_{1i} = c_{ih} \quad \text{if} \quad y_i = v_h$$

It follows that the log likelihood is

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

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**.

The conditional probabilities on $s_i = S_i$ are

$$\Pr(y_i = v_h|\mathbf{x}_i) = \Phi_1^*(c_{i(h-1)}, c_{ih}, 1 - \rho_{1s}^2)$$

If we do not include s_i in the main outcome equation, the probability of success is calculated as (4) again.

Combinations of features

Extended ordered probit 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 confidence intervals for variance parameters, and we use 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.

Botezat, A., and F. Pfeiffer. 2014. The impact of parents' migration on the well-being of children left behind: Initial evidence from Romania. IZA Discussion Paper No. 8225, Institute for the Study of Labor (IZA). http://ftp.iza.org/dp8225.pdf

Chyi, H., and S. Mao. 2012. The determinants of happiness of China's elderly population. *Journal of Happiness Studies* 13: 167–185.

De Luca, G., and V. Perotti. 2011. Estimation of ordered response models with sample selection. *Stata Journal* 11: 213–239.

Gregory, C. A. 2015. Estimating treatment effects for ordered outcomes using maximum simulated likelihood. *Stata Journal* 15: 756–774.

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.

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

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.

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] **eoprobit postestimation** — Postestimation tools for eoprobit

[ERM] **eoprobit predict** — predict after eoprobit

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

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

[R] **heckoprobit** — Ordered probit model with sample selection

[R] **oprobit** — Ordered probit regression

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

[U] 20 Estimation and postestimation commands