Glossary

average structural function. The average structural function (ASF) is used to calculate predicted values of ERMs.

The ASF averages out the heterogeneity caused by the endogeneity from a conditional mean or a conditional probability in a model with endogenous covariates. Applying the ASF to a conditional mean produces an average structural mean (ASM). Applying the ASF to a conditional probability produces an average structural probability (ASP). Contrasts of ASMs or ASPs produced by a covariate change define a causal structural effect. Blundell and Powell (2003, 2004) and Wooldridge (2005, 2014) are seminal papers that define and extend the ASF. See Wooldridge (2010, 22–24) for a textbook introduction.

average structural mean. The average structural mean (ASM) is the result of applying the average structural function to a conditional mean.

average structural probability. The average structural probability (ASP) is the result of applying the average structural function to a conditional probability.

average treatment effect. See *treatment effects*.

average treatment effect on the treated. See *treatment effects*.

average treatment effect on the untreated. See *treatment effects*.

binary variable. A binary variable is any variable that records two values, the two values representing false and true, such as whether a person is sick. We usually speak of the two values as being 0 and 1 with 1 meaning true, but Stata requires merely that 0 means false and nonzero and nonmissing mean true. Also see *continuous variable*, *categorical variable*, and *interval variable*.

categorical variable. A categorical variable is a variable that records the category number for, say, lives in the United States, lives in Europe, and lives in Asia. Categorical variables play no special role in this manual, but ordered categorical variables do. The example given is unordered. The categories United States, Europe, and Asia have no natural ordering. We listed the United States first only because the author of this manual happens to live in the United States.

The way we use the term, categorical variables usually record two or more categories, and the term binary variable is used for categorical variables having two categories.

We usually speak of categorical variables as if they take on the values 1, 2, ... Stata does not require that. However, the values do need to be integers.

censored, **left-censored**, **right-censored**, and **interval-censored**. Censoring involves not observing something but knowing when and where you do not observe it.

For instance, sometimes patients, subjects, or units being studied—observations in your dataset—have values equal to missing. Such observations are said to be censored when there is a reason they are missing. A variable is missing because a potential worker chooses not to work, because a potential patient chooses not to be a patient, because a potential subject was not prescribed the treatment, etc. Such censored outcomes cause difficulty when there is an unobserved component to the reason they are censored that is correlated with the outcome being studied. ERM option `select()` addresses these issues.

Another type of censoring—interval-censoring—involves not observing a value precisely but knowing its range. You do not observe blood pressure, but you know it is in the range 120 to 140. Or you know it is less than 120 or greater than 160. ERM command `eintreg` fits models in which the dependent variable is interval-censored.

Left-censoring is open-ended interval-censoring in which measurements below a certain value are unobserved. Blood pressure is less than 120.

Right-censoring is open-ended interval-censoring in which measurements above a certain value are unobserved. Blood pressure is above 160.

conditional mean. The conditional mean of a variable is the expected value based on a function of other variables. If y is a linear function of x_1 and x_2—$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \text{noise}$—then the conditional mean of y for $x_1 = 2$ and $x_2 = 4$ is $\beta_0 + 2\beta_1 + 4\beta_2$.

confounding variable, confounder. A confounding variable is an omitted explanatory variable in a model that is correlated with variables included in the model. The fitted coefficients on the observed variables will include the effect of the variables, as intended, plus the effect of being correlated with the omitted variable.

Confounders are often omitted from the model because they are unobserved. See [ERM] **Intro 3**.

continuous variable. A continuous variable is a variable taking on any value on the number line. In this manual, however, we use the term to mean the variable is not a binary variable, not a categorical variable, and not an interval variable.

counterfactual. The result that would be expected from a thought experiment that assumes things counter to what are currently true. What would be the average income if everyone had one more year of schooling? What would be the effect of an experimental medical treatment if the treatment were made widely available? Stata's `margins` command produces statistical answers to these kinds of thought experiments and reports standard errors as well.

counterfactual predictions. Counterfactual predictions are used when you have endogenous covariates in your main equation and you wish to estimate either counterfactuals or the effect on the outcome of changing the values of covariates. They are obtained using `predict` options `base()` and `fix()`.

covariate. A covariate is a variable appearing on the right-hand side (RHS) of a model. Covariates can be exogenous or endogenous, but when the term is used without qualification, it usually means exogenous covariate. Covariates are also known as explanatory variables. Also see *endogenous covariate* and *exogenous covariate*.

cross-sectional data. Cross-sectional data refers to data collected over a set of individuals, such as households, firms, or countries sampled from a population at a given point in time.

dependent variable. A dependent variable is a variable appearing on the left-hand side of an equation in a model. It is the variable to be explained. Every equation of a model has a dependent variable. The term "the dependent variable" is often used in this manual to refer to the dependent variable of the main equation. Also see [ERM] **Intro 3**.

endogenous and exogenous treatment assignment. See *treatment assignment*.

endogenous covariate. An endogenous covariate is a covariate appearing in a model 1) that is correlated with omitted variables that also affect the outcome; 2) that is measured with error; 3) that is affected by the dependent variable; or 4) that is correlated with the model's error. See [ERM] **Intro 3**.

endogenous sample selection. Endogenous sample selection refers to situations in which the subset of the data used to fit a model has been selected in a way correlated with the model's outcome.

Mechanically, the subset used is the subset containing nonmissing values of variables used by the model. A variable is unobserved—contains missing values—because a potential worker chooses not to work, because a potential patient chooses not to be a patient, because a potential subject was not prescribed the treatment, etc. Such censored outcomes cause difficulty when there is an

unobserved component to the reason they are censored that is correlated with the outcome being studied.

ERM option `select()` can address these issues when the dataset contains observations for which the dependent variable was missing.

error. Error is the random component (residual) appearing at the end of the equations in a model. These errors account for the unobserved information explaining the outcome variable. Errors in this manual are written as $e.depvarname$, such as $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e.y$.

exogenous covariate. An exogenous covariate is a covariate that is uncorrelated with the error term in the model. See [ERM] **Intro 3**.

explanatory variable. Explanatory variable is another word for covariate.

extended regression models. Extended regression models (ERMs) are generalized structural equation models that allow identity and probit links and Gaussian, binomial, and ordinal families for the main outcome. They extend interval regression, ordered probit, probit, and linear regression models by accommodating endogenous covariates, nonrandom and endogenous treatment assignment, endogenous sample selection, and random effects.

individual-level treatment effect. An individual-level treatment effect is the difference in the individuals outcome that would occur when given one treatment instead of another. It is the difference between two potential outcomes for the individual. The blood pressure after taking a pill minus the blood pressure were the pill not taken is the individual-level treatment effect of the pill on blood pressure.

informative missingness. See *missingness*.

instrument. Instrument is an informal word for instrumental variable.

instrumental variable. An instrumental variable is a variable that affects an endogenous covariate but does not affect the dependent variable. See [ERM] **Intro 3**.

interval measurement. Interval measurement is a synonym for interval-censored. See *censored*.

interval variable. An interval variable is actually a pair of variables that record the lower and upper bounds for a variable whose precise values are unobserved. `ylb` and `yub` might record such values for a variable y. Then it is known that, for each observation i, $\texttt{ylb}_i \leq y \leq \texttt{yub}_i$. ERM estimation command `eintreg` fits such models. Also see *censored*.

interval-censored. See *censored*.

left-hand-side (LHS) variable. A left-hand-side variable is another word for dependent variable.

longitudinal data. Longitudinal data is another term for panel data. See also *panel data*.

loss to follow-up. Subjects are lost to follow-up if they do not complete the course of the study for reasons unrelated to the event of interest. For example, loss to follow-up occurs if subjects move to a different area or decide to no longer participate in a study. Loss to follow-up should not be confused with administrative censoring. If subjects are lost to follow-up, the information about the outcome these subjects would have experienced at the end of the study, had they completed the study, is unavailable.

main equation. The main equation in an ERM is the first equation specified, the equation appearing directly after the `eregress`, `eintreg`, `eprobit`, `eoprobit`, `xteregress`, `xteintreg`, `xteprobit`, or `xteoprobit` command. The purpose of ERMs is to produce valid estimates of the coefficients in the main equation, meaning the structural coefficients, in the presence of complications such as endogeneity, selection, treatment assignment, or random effects.

measurement error, measured with error. A variable measured with error has recorded value equal to $x + \epsilon$, where x is the true value. The error is presumably uncorrelated with all other errors in the model. In that case, fitted coefficients will be biased toward zero. See [ERM] **Intro 3**.

missing at random (MAR). See *missingness*.

missing completely at random (MCAR). See *missingness*.

missing not at random (MNAR). See *missingness*.

missingness. Missingness refers to how missing observations in data occur. The categories are 1) missing not at random (MNAR), 2) missing at random (MAR), and 3) missing completely at random (MCAR).

In what follows we will refer to missing observations to mean not only observations entirely missing from a dataset but also the omitted observations because of missing values when fitting models.

MNAR observations refer to cases in which the missingness depends on the outcome under study. The solution in this case is to model that dependency. When observations are missing because of missing values, ERM option `select()` can be used to model the missingness.

MAR observation refer to cases in which the missingness does not depend on the outcome under study but does depend on other variables correlated with the outcome. The solution for some of the problems raised is to include those other variables as covariates in your model. Importantly, you do not need to model the reason for missingness.

MCAR observations are just that and obviously not a problem other than to cause loss of efficiency.

The MNAR and MAR cases are known jointly as informative missingness.

multivalued treatment. A multivalued treatment is a treatment with more than two arms. See *treatment arms*.

observational data. Observational data are data collected over which the researcher had no control. The opposite of observational data is experimental data. Use of observational data often introduces statistical issues that experimental data would not. For instance, in a treatment study based on observational data, researchers had no control over treatment assignment; thus the treatment assignment needs to be modeled.

omitted variables. Omitted variables is an informal term for covariates that should appear in the model but do not. They do not because they are unmeasured, because of ignorance or other reasons. Problems arise when the variables that are not omitted are correlated with the omitted variables.

ordered categorical variable. An ordered categorical variable is a categorical variable in which the categories can be ordered, such as healthy, sick, and very sick. Actually recorded in the variable are integers such as 1, 2, and 3. The integers need not be sequential, but they must reflect the ordering. Also see *binary variable* and *continuous variable*.

outcome variable. See *dependent variable*.

panel data. Panel data are data in which the same units were observed over multiple periods. The units, called panels, are often firms, households, or patients who were observed at several points in time. In a typical panel dataset, the number of panels is large, and the number of observations per panel is relatively small.

potential outcome. Potential outcome is a term used in the treatment-effects literature. It is the outcome an individual would have had if given a specific treatment. Individual in this case means conditional on the individual's covariates, which are in the main equation in models fit by ERMs. It is the outcome that would have been observed for that individual. For instance, each patient in

a study has one potential blood pressure after taking a pill and another had he or she not taken it. Also see *treatment effects*.

potential-outcome means. Potential-outcome means (POMs) is a term used in the treatment-effects literature. They are the means (averages) of potential outcomes. The average treatment effect (see *treatment effects*) is the difference between the potential-outcome mean for treated and untreated over the population.

random-effects model. A random-effects model for panel data treats the panel-specific errors for each equation as random variables drawn from a population with zero mean and constant variance. The regressors not distinctly specified as endogenous must be uncorrelated with the random effects for the estimates to be consistent.

recursive (structural) model. ERMs fit recursive models. A model is not recursive when one endogenous variable depends (includes its equation) on another endogenous variable that depends on the first. Said in symbols, when A depends on B, which depends on A. A model is also not recursive when A depends on B depends on C, which depends on A, and so on. See [ERM] **Triangularize**.

reverse causation and simultaneous causation. We use the term reverse causation in this manual when the dependent variable in the main equation of an ERM affects a covariate as well as when the covariate affects the dependent variable. Stressed persons may be physically unhealthy because they are stressed and further stressed because they are unhealthy. When a covariate suffers from reverse causation, the solution is to make it endogenous and find instruments for it.

Our use of the term reverse causation is typical of how it is used elsewhere. Reverse causation is a reason to make a variable endogenous. Reverse causation is discussed in [ERM] **Intro 3**.

The term simultaneous causation is sometimes used as a synonym for reverse causation elsewhere, but we draw a distinction. We use the term when two already endogenous variables affect each other. Simultaneous causation is discussed in [ERM] **Triangularize**.

right-hand-side (RHS) variable. A right-hand-side variable is another word for covariate.

sample selection. Sample selection is another term for endogenous sample selection.

selection. Selection is another term for endogenous sample selection.

selection on unobservables. Selection on unobservables is another term for endogenous sample selection.

simultaneous causation. See *recursive (structural) model*.

simultaneous system. A simultaneous system is a multiple-equation model in which dependent variables can affect each other freely. The equation for `y1` could include `y2`, and the equation for `y2` include `y1`. ERMs cannot fit simultaneous systems. Because the focus of ERMs is on one equation in particular—the main equation—you can substitute the covariates for `y1` into the `y2` equation to form the reduced-form result and still obtain estimates of the structural parameters of the `y1` equation. In this manual, we discuss this issue using the terms reverse causation and recursive (structural) model. In the manual, it is discussed in [ERM] **Triangularize**.

strongly balanced. A longitudinal or panel dataset is said to be strongly balanced if each panel has the same number of observations and the observations for different panels were all made at the same times.

TE. See *treatment effect*.

tobit estimator. Tobit is an estimation technique for dealing with dependent variables that are censored. The classic tobit model dealt with left-censoring, in which the outcome variable was recorded as zero if it would have been zero or below. The estimator has since been generalized to dealing

with models in which observations can be left-censored, right-censored, or interval-censored. See *censored*.

treatment. A treatment is a drug, government program, or anything else administered to a patient, job seeker, etc., in hopes of improving an outcome.

treatment arms. Sometimes, experiments are run on more than one treatment simultaneously. Each different treatment is called an arm of the treatment. The controls (those not treated) are also an arm of the treatment.

treatment assignment. Treatment assignment is the process by which subjects are assigned to a treatment arm. That process can be endogenous or exogenous, meaning that the random component (error) in the assignment is correlated or is not correlated with the outcomes of the treatments. It is often endogenous because doctors assign subjects or subjects choose based in part on unobserved factors correlated with the treatment's outcome.

treatment effects. A treatment effect (TE) is the effect of a treatment in terms of a measured outcome such as blood pressure, ability to walk, likelihood of finding employment, etc. The statistical problem is to measure the effect of a treatment in the presence of complications such as censoring, treatment assignment, and so on.

ERMs fit treatment-effect models when one of the options `entreat()` or `extreat()` is specified for endogenous or exogenous treatment assignment. Meanwhile, the outcome model is specified in the main equation.

The TE is, for each person, the difference in the predicted outcomes based on the covariates in the main equation given that treatment is locked at treated or untreated.

The treatment effect on the treated (TET) is, for each person who was treated, the difference in the predicted outcomes based on the covariates in the main equation and the fact that they were assigned to or choose to be treated.

The treatment effect on the untreated (TEU) is, for each person who was not treated, the difference in predicted outcomes based on the covariates in the main equation and the fact that they were assigned to or choose not to be treated.

The average treatment effect (ATE) is an estimate of the average effect in a population after accounting for statistical issues.

The average effect on the treated (ATET) is an estimate of the average effect that would have been observed for those who were in fact treated in the data.

The average effect on the untreated (ATEU) is an estimate of the average effect that would have been observed for those who were in fact not treated in the data.

triangular system. See *recursive (structural) model*.

unbalanced data. A longitudinal or panel dataset is said to be unbalanced if each panel does not have the same number of observations. See also *weakly balanced* and *strongly balanced*.

weakly balanced. A longitudinal or panel dataset is said to be weakly balanced if each panel has the same number of observations but the observations for different panels were not all made at the same times.

References

Blundell, R. W., and J. L. Powell. 2003. Endogeity in nonparametric and semiparametric regression models. In *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, ed. M. Dewatripont, L. P. Hansen, and S. J. Turnovsky, vol. 2, 312–357. Cambridge: Cambridge University Press.

——. 2004. Endogeneity in semiparametric binary response models. *Review of Economic Studies* 71: 655–679. https://doi.org/10.1111/j.1467-937X.2004.00299.x.

Wooldridge, J. M. 2005. Unobserved heterogeneity and estimation of average partial effects. In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, ed. D. W. K. Andrews and J. H. Stock, 27–55. New York: Cambridge University Press.

——. 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.