Glossary

ADF, **method(adf)**. ADF stands for asymptotic distribution free and is a method used to obtain fitted parameters for standard linear SEMs. ADF is used by `sem` when option `method(adf)` is specified. Other available methods are ML, QML, and MLMV.

anchoring, **anchor variable**. A variable is said to be the anchor of a latent variable if the path coefficient between the latent variable and the anchor variable is constrained to be 1. `sem` and `gsem` use anchoring as a way of normalizing latent variables and thus identifying the model.

baseline model. A baseline model is a covariance model—a model of fitted means and covariances of observed variables without any other paths—with most of the covariances constrained to be 0. That is, a baseline model is a model of fitted means and variances but typically not all the covariances. Also see *saturated model*. Baseline models apply only to standard linear SEMs.

Bentler–Weeks matrices. The Bentler and Weeks (1980) formulation of standard linear SEMs places the results in a series of matrices organized around how results are calculated. See [SEM] **estat framework**.

bootstrap, **vce(bootstrap)**. The bootstrap is a replication method for obtaining variance estimates. Consider an estimation method E for estimating θ. Let $\hat{\theta}$ be the result of applying E to dataset D containing N observations. The bootstrap is a way of obtaining variance estimates for $\hat{\theta}$ from repeated estimates $\hat{\theta}_1$, $\hat{\theta}_2$, …, where each $\hat{\theta}_i$ is the result of applying E to a dataset of size N drawn with replacement from D. See [SEM] **sem option method( )** and [R] **bootstrap**.

`vce(bootstrap)` is allowed with `sem` but not `gsem`. You can obtain bootstrap results by prefixing the `gsem` command with `bootstrap:`, but remember to specify `bootstrap`'s `cluster()` and `idcluster()` options if you are fitting a multilevel model. See [SEM] **Intro 9**.

Builder. The Builder is Stata's graphical interface for building `sem` and `gsem` models. The Builder is also known as the SEM Builder. See [SEM] **Intro 2**, [SEM] **Builder**, and [SEM] **Builder, generalized**.

categorical latent variable. A categorical latent variable has levels that represent unobserved groups in the population. Latent classes are identified with the levels of the categorical latent variables and may represent healthy and unhealthy individuals, consumers with different buying preferences, or different motivations for delinquent behavior. Categorical latent variables are allowed in `gsem` but not in `sem`. See [SEM] **gsem lclass options** and [SEM] **Intro 2**.

CFA, **CFA models**. CFA stands for confirmatory factor analysis. It is a way of analyzing measurement models. CFA models is a synonym for measurement models.

cluster, **vce(cluster clustvar)**. Cluster is the name we use for the generalized Huber/White/sandwich estimator of the VCE, which is the `robust` technique generalized to relax the assumption that errors are independent across observations to be that they are independent across clusters of observations. Within cluster, errors may be correlated.

Clustered standard errors are reported when `sem` or `gsem` option `vce(cluster` *clustvar*`)` is specified. The other available techniques are OIM, OPG, robust, bootstrap, and jackknife. Also available for `sem` only is EIM.

coefficient of determination. The coefficient of determination is the fraction (or percentage) of variation (variance) explained by an equation of a model. The coefficient of determination is thus like R^2 in linear regression.

command language. Stata's sem and gsem commands provide a way to specify SEMs. The alternative is to use the Builder to draw path diagrams; see [SEM] **Intro 2**, [SEM] **Builder**, and [SEM] **Builder, generalized**.

complementary log–log regression. Complementary log–log regression is a term for generalized linear response functions that are family Bernoulli, link cloglog. It is used for binary outcome data. Complementary log–log regression is also known in Stata circles as cloglog regression or just cloglog. See *generalized linear response functions*.

conditional normality assumption. See *normality assumption, joint and conditional*.

constraints. See *parameter constraints*.

continuous latent variable. A continuous latent variable is an unobserved variable, such as mathematical ability, with values that are assumed to follow a continuous distribution. Both sem and gsem allow continuous latent variables that are assumed to follow a Gaussian distribution with a mean and variance that are either estimated or constrained to a specific value for identification. See *identification*.

correlated uniqueness model. A correlated uniqueness model is a kind of measurement model in which the errors of the measurements have a structured correlation. See [SEM] **Intro 5**.

crossed-effects models. See *multilevel models*.

curved path. See *path*.

degree-of-freedom adjustment. In estimates of variances and covariances, a finite-population degree-of-freedom adjustment is sometimes applied to make the estimates unbiased.

Let's write an estimated variance as $\hat{\sigma}_{ii}$ and write the "standard" formula for the variance as $\hat{\sigma}_{ii} = S_{ii}/N$. If $\hat{\sigma}_{ii}$ is the variance of observable variable x_i, it can readily be proven that S_{ii}/N is a biased estimate of the variances in samples of size N and that $S_{ii}/(N-1)$ is an unbiased estimate. It is usual to calculate variances with $S_{ii}/(N-1)$, which is to say the "standard" formula has a multiplicative degree-of-freedom adjustment of $N/(N-1)$ applied to it.

If $\hat{\sigma}_{ii}$ is the variance of estimated parameter β_i, a similar finite-population degree-of-freedom adjustment can sometimes be derived that will make the estimate unbiased. For instance, if β_i is a coefficient from a linear regression, an unbiased estimate of the variance of regression coefficient β_i is $S_{ii}/(N-p-1)$, where p is the total number of regression coefficients estimated excluding the intercept. In other cases, no such adjustment can be derived. Such estimators have no derivable finite-sample properties, and one is left only with the assurances provided by its provable asymptotic properties. In such cases, the variance of coefficient β_i is calculated as S_{ii}/N, which can be derived on theoretical grounds. SEM is an example of such an estimator.

SEM is a remarkably flexible estimator and can reproduce results that can sometimes be obtained by other estimators. SEM might produce asymptotically equivalent results, or it might produce identical results depending on the estimator. Linear regression is an example in which sem and gsem produce the same results as regress. The reported standard errors, however, will not look identical because the linear-regression estimates have the finite-population degree-of-freedom adjustment applied to them and the SEM estimates do not. To see the equivalence, you must undo the adjustment on the reported linear regression standard errors by multiplying them by $\sqrt{\{(N-p-1)/N\}}$.

direct, **indirect**, and **total effects**. Consider the following system of equations:

$$\begin{aligned} y_1 &= b_{10} + b_{11}y_2 + b_{12}x_1 + b_{13}x_3 + e_1 \\ y_2 &= b_{20} + b_{21}y_3 + b_{22}x_1 + b_{23}x_4 + e_2 \\ y_3 &= b_{30} + \qquad\quad b_{32}x_1 + b_{33}x_5 + e_3 \end{aligned}$$

The total effect of x_1 on y_1 is $b_{12} + b_{11}b_{22} + b_{11}b_{21}b_{32}$. It measures the full change in y_1 based on allowing x_1 to vary throughout the system.

The direct effect of x_1 on y_1 is b_{12}. It measures the change in y_1 caused by a change in x_1 holding other endogenous variables—namely, y_2 and y_3—constant.

The indirect effect of x_1 on y_1 is obtained by subtracting the total and direct effect and is thus $b_{11}b_{22} + b_{11}b_{21}b_{32}$.

EIM, **vce(eim)**. EIM stands for expected information matrix, defined as the inverse of the negative of the expected value of the matrix of second derivatives, usually of the log-likelihood function. The EIM is an estimate of the VCE. EIM standard errors are reported when `sem` option `vce(eim)` is specified. EIM is available only with `sem`. The other available techniques for `sem` are OIM, OPG, robust, cluster, bootstrap, and jackknife.

endogenous variable. A variable, observed or latent, is endogenous (determined by the system) if any path points to it. Also see *exogenous variable*.

entropy. A measure of separation between latent classes. It ranges from 0 to 1, and values closer to 1 indicate better separation between latent classes.

error, **error variable**. The error is random disturbance e in a linear equation:

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + e$$

An error variable is an unobserved exogenous variable in path diagrams corresponding to e. Mathematically, error variables are just another example of latent exogenous variables, but in `sem` and `gsem`, error variables are considered to be in a class by themselves. All (Gaussian) endogenous variables—observed and latent—have a corresponding error variable. Error variables automatically and inalterably have their path coefficients fixed to be 1. Error variables have a fixed naming convention in the software. If a variable is the error for (observed or latent) endogenous variable `y`, then the residual variable's name is `e.y`.

In `sem` and `gsem`, error variables are uncorrelated with each other unless explicitly indicated otherwise. That indication is made in path diagrams by drawing a curved path between the error variables and is indicated in command notation by including `cov(e.`*name1*`*e.`*name2*`)` among the options specified on the `sem` command. In `gsem`, errors for family Gaussian, link log responses are not allowed to be correlated.

estimation method. There are a variety of ways that one can solve for the parameters of an SEM. Different methods make different assumptions about the data-generation process, so it is important that you choose a method appropriate for your model and data; see [SEM] **Intro 4**.

exogenous variable. A variable, observed or latent, is exogenous (determined outside the system) if paths only originate from it or, equivalently, no path points to it. In this manual, we do not distinguish whether exogenous variables are strictly exogenous—that is, uncorrelated with the errors. Also see *endogenous variable*.

family distribution. See *generalized linear response functions*.

fictional data. Fictional data are data that have no basis in reality even though they might look real; they are data that are made up for use in examples.

finite mixture model. A finite mixture model (FMM) is a latent class model in which parameters of a regression model are allowed to vary across classes. The regression model may have a linear or generalized linear response function.

first- and second-order latent variables. If a latent variable is measured by other latent variables only, the latent variable that does the measuring is called first-order latent variable, and the latent variable being measured is called the second-order latent variable.

first-, second-, and higher-level (latent) variables. Consider a multilevel model of patients within doctors within hospitals. First-level variables are variables that vary at the observational (patient) level. Second-level variables vary across doctors but are constant within doctors. Third-level variables vary across hospitals but are constant within hospitals. This jargon is used whether variables are latent or not.

full joint and conditional normality assumption. See *normality assumption, joint and conditional*.

gamma regression. Gamma regression is a term for generalized linear response functions that are family gamma, link log. It is used for continuous, nonnegative, positively skewed data. Gamma regression is also known as log-gamma regression. See *generalized linear response functions*.

Gaussian regression. Gaussian regression is another term for linear regression. It is most often used when referring to generalized linear response functions. In that framework, Gaussian regression is family Gaussian, link identity. See *generalized linear response functions*.

generalized linear response functions. Generalized linear response functions include linear functions and include functions such as probit, logit, multinomial logit, ordered probit, ordered logit, Poisson, and more.

These generalized linear functions are described by a link function $g(\cdot)$ and statistical distribution F. The link function $g(\cdot)$ specifies how the response variable y_i is related to a linear equation of the explanatory variables, $\mathbf{x}_i\boldsymbol{\beta}$, and the family F specifies the distribution of y_i:

$$g\{E(y_i)\} = \mathbf{x}_i\boldsymbol{\beta}, \qquad y_i \sim F$$

If we specify that $g(\cdot)$ is the identity function and F is the Gaussian (normal) distribution, then we have linear regression. If we specify that $g(\cdot)$ is the logit function and F the Bernoulli distribution, then we have logit (logistic) regression.

In this generalized linear structure, the family may be Gaussian, gamma, Bernoulli, binomial, Poisson, negative binomial, ordinal, or multinomial. The link function may be the identity, log, logit, probit, or complementary log–log.

`gsem` fits models with generalized linear response functions.

generalized method of moments. Generalized method of moments (GMM) is a method used to obtain fitted parameters. In this documentation, GMM is referred to as ADF, which stands for asymptotic distribution free and is available for use with `sem`. Other available methods for use with `sem` are ML, QML, ADF, and MLMV.

The SEM moment conditions are cast in terms of second moments, not the first moments used in many other applications associated with GMM.

generalized SEM. Generalized SEM is a term we have coined to mean SEM optionally allowing generalized linear response functions, multilevel models, or categorical latent variables. `gsem` fits generalized SEMs.

GMM. See *generalized method of moments*.

goodness-of-fit statistic. A goodness-of-fit statistic is a value designed to measure how well the model reproduces some aspect of the data the model is intended to fit. SEM reproduces the first- and second-order moments of the data, with an emphasis on the second-order moments, and thus goodness-of-fit statistics appropriate for use after `sem` compare the predicted covariance matrix (and mean vector) with the matrix (and vector) observed in the data.

gsem. `gsem` is the Stata command that fits generalized SEMs. Also see *sem*.

GUI. See *Builder*.

identification. Identification refers to the conceptual constraints on parameters of a model that are required for the model's remaining parameters to have a unique solution. A model is said to be unidentified if these constraints are not supplied. These constraints are of two types: substantive constraints and normalization constraints.

Normalization constraints deal with the problem that one scale works as well as another for each continuous latent variable in the model. One can think, for instance, of propensity to write software as being measured on a scale of 0 to 1, 1 to 100, or any other scale. The normalization constraints are the constraints necessary to choose one particular scale. The normalization constraints are provided automatically by `sem` and `gsem` by anchoring with unit loadings.

Substantive constraints are the constraints you specify about your model so that it has substantive content. Usually, these constraints are 0 constraints implied by the paths omitted, but they can include explicit parameter constraints as well. It is easy to write a model that is not identified for substantive reasons; see [SEM] **Intro 4**.

indicator variables, **indicators**. The term "indicator variable" has two meanings. An indicator variable is a 0/1 variable that contains whether something is true. The other usage is as a synonym for measurement variables.

indirect effects. See *direct, indirect, and total effects*.

initial values. See *starting values*.

intercept. An intercept for the equation of endogenous variable y, observed or latent, is the path coefficient from `_cons` to y. `_cons` is Stata-speak for the built-in variable containing 1 in all observations. In SEM-speak, `_cons` is an observed exogenous variable.

jackknife, **vce(jackknife)**. The jackknife is a replication method for obtaining variance estimates. Consider an estimation method E for estimating θ. Let $\widehat{\theta}$ be the result of applying E to dataset D containing N observations. The jackknife is a way of obtaining variance estimates for $\widehat{\theta}$ from repeated estimates $\widehat{\theta}_1$, $\widehat{\theta}_2$, …, $\widehat{\theta}_N$, where each $\widehat{\theta}_i$ is the result of applying E to D with observation i removed. See [SEM] **sem option method()** and [R] **jackknife**.

`vce(jackknife)` is allowed with `sem` but not `gsem`. You can obtain jackknife results by prefixing the `gsem` command with `jackknife:`, but remember to specify `jackknife`'s `cluster()` and `idcluster()` options if you are fitting a multilevel model. See [SEM] **Intro 9**.

joint normality assumption. See *normality assumption, joint and conditional*.

Lagrange multiplier test. Synonym for score test.

latent class analysis. Latent class analysis is useful for identifying and understanding unobserved groups in a population. When performing a latent class analysis, we fit models that include a categorical latent variable with levels called latent classes that correspond to the unobserved groups. In latent

class analysis, we can compare models with differing numbers of latent classes and different sets of constraints on parameters to determine the best-fitting model. For a given model, we can compare parameter estimates across classes. We can estimate the proportion of the population in each latent class. And we can predict the probabilities that the individuals in our sample belong to each latent class.

latent class model. Any model with a categorical latent variable that is fit as part of a latent class analysis. In some literature, latent class models are more narrowly defined to include only categorical latent variables and the binary or categorical observed variables that are indicators of class membership, but we do not make such a restriction. See [SEM] **Intro 5**.

latent cluster model. A type of latent class model with continuous observed outcomes.

latent growth model. A latent growth model is a kind of measurement model in which the observed values are collected over time and are allowed to follow a trend. See [SEM] **Intro 5**.

latent profile model. A type of latent class model with continuous observed outcomes.

latent variable. A variable is latent if it is not observed. A variable is latent if it is not in your dataset but you wish it were. You wish you had a variable recording the propensity to commit violent crime, or socioeconomic status, or happiness, or true ability, or even income accurately recorded. Latent variables are sometimes described as imagined variables.

In the software, latent variables are usually indicated by having at least their first letter capitalized.

Also see *continuous latent variable*, *categorical latent variable*, *first- and second-order latent variables*, *first-, second-, and higher-level (latent) variables*, and *observed variables*.

linear regression. Linear regression is a kind of SEM in which there is a single equation. See [SEM] **Intro 5**.

link function. See *generalized linear response functions*.

logit regression. Logit regression is a term for generalized linear response functions that are family Bernoulli, link logit. It is used for binary outcome data. Logit regression is also known as logistic regression and also simply as logit. See *generalized linear response functions*.

manifest variables. Synonym for observed variables.

measure, **measurement**, **x a measurement of X**, **x measures X**. See *measurement variables*.

measurement models, **measurement component**. A measurement model is a particular kind of model that deals with the problem of translating observed values to values suitable for modeling. Measurement models are often combined with structural models and then the measurement model part is referred to as the measurement component. See [SEM] **Intro 5**.

measurement variables, **measure**, **measurement**, **x a measurement of X**, **x measures X**. Observed variable x is a measurement of latent variable X if there is a path connecting $x \leftarrow X$. Measurement variables are modeled by measurement models. Measurement variables are also called indicator variables.

method. Method is just an English word and should be read in context. Nonetheless, method is used here usually to refer to the method used to solve for the fitted parameters of an SEM. Those methods are ML, QML, MLMV, and ADF. Also see *technique*.

MIMIC. See *multiple indicators and multiple causes*.

mixed-effects models. See *multilevel models*.

ML, **method(ml)**. ML stands for maximum likelihood. It is a method to obtain fitted parameters. ML is the default method used by `sem` and `gsem`. Other available methods for `sem` are QML, MLMV, and ADF. Also available for `gsem` is QML.

MLMV, **method(mlmv)**. MLMV stands for maximum likelihood with missing values. It is an ML method used to obtain fitted parameters in the presence of missing values. MLMV is the method used by `sem` when the `method(mlmv)` option is specified; `method(mlmv)` is not available with `gsem`. Other available methods for use with `sem` are ML, QML, and ADF. These other methods omit from the calculation observations that contain missing values.

modification indices. Modification indices are score tests for adding paths where none appear. The paths can be for either coefficients or covariances.

moments (of a distribution). The moments of a distribution are the expected values of powers of a random variable or centralized (demeaned) powers of a random variable. The first moments are the expected or observed means, and the second moments are the expected or observed variances and covariances.

multilevel models. Multilevel models are models that include unobserved effects (latent variables) for different groups in the data. For instance, in a dataset of students, groups of students might share the same teacher. If the teacher's identity is recorded in the data, then one can introduce a latent variable that is constant within teacher and that varies across teachers. This is called a two-level model.

If teachers could in turn be grouped into schools, and school identities were recorded in the data, then one can introduce another latent variable that is constant within school and varies across schools. This is called a three-level (nested-effects) model.

In the above example, observations (students) are said to be nested within teacher nested within school. Sometimes there is no such subsequent nesting structure. Consider workers nested within occupation and industry. The same occupations appear in various industries and the same industries appear within various occupations. We can still introduce latent variables at the occupation and industry level. In such cases, the model is called a crossed-effects model.

The latent variables that we have discussed are also known as random effects. Any coefficients on observed variables in the model are known as the fixed portion of the model. Models that contain fixed and random portions are known as mixed-effects models.

multinomial logit regression. Multinomial logit regression is a term for generalized linear response functions that are family multinomial, link logit. It is used for categorical-outcome data when the outcomes cannot be ordered. Multinomial logit regression is also known as multinomial logistic regression and as mlogit in Stata circles. See *generalized linear response functions*.

multiple correlation. The multiple correlation is the correlation between endogenous variable y and its linear prediction.

multiple indicators and multiple causes. Multiple indicators and multiple causes is a kind of structural model in which observed causes determine a latent variable, which in turn determines multiple indicators. See [SEM] **Intro 5**.

multivariate regression. Multivariate regression is a kind of structural model in which each member of a set of observed endogenous variables is a function of the same set of observed exogenous variables and a unique random disturbance term. The disturbances are correlated. Multivariate regression is a special case of seemingly unrelated regression.

negative binomial regression. Negative binomial regression is a term for generalized linear response functions that are family negative binomial, link log. It is used for count data that are overdispersed relative to Poisson. Negative binomial regression is also known as nbreg in Stata circles. See *generalized linear response functions*.

nested-effects models. See *multilevel models*.

nonrecursive (structural) model (system), **recursive (structural) model (system)**. A structural model (system) is said to be nonrecursive if there are paths in both directions between one or more pairs of endogenous variables. A system is recursive if it is a system—it has endogenous variables that appear with paths from them—and it is not nonrecursive.

A nonrecursive model may be unstable. Consider, for instance,

$$y_1 = 2y_2 + 1x_1 + e_1$$
$$y_2 = 3y_1 - 2x_2 + e_2$$

This model is unstable. To see this, without loss of generality, treat $x_1 + e_1$ and $2x_2 + e_2$ as if they were both 0. Consider $y_1 = 1$ and $y_2 = 1$. Those values result in new values $y_1 = 2$ and $y_2 = 3$, and those result in new values $y_1 = 6$ and $y_2 = 6$, and those result in new values .... Continue in this manner, and you reach infinity for both endogenous variables. In the jargon of the mathematics used to check for this property, the eigenvalues of the coefficient matrix lie outside the unit circle.

On the other hand, consider these values:

$$y_1 = 0.5y_2 + 1x_1 + e_1$$
$$y_2 = 1.0y_1 - 2x_2 + e_2$$

These results are stable in that the resulting values converge to $y_1 = 0$ and $y_2 = 0$. In the jargon of the mathematics used to check for this property, the eigenvalues of the coefficient matrix lie inside the unit circle.

Finally, consider the values

$$y_1 = 0.5y_2 + 1x_1 + e_1$$
$$y_2 = 2.0y_1 - 2x_2 + e_2$$

Start with $y_1 = 1$ and $y_2 = 1$. That yields new values $y_1 = 0.5$, and $y_2 = 2$ and that yields new values $y_1 = 1$ and $y_2 = 1$, and that yields new values $y_1 = 0.5$ and $y_2 = 2$, and it will oscillate forever. In the jargon of the mathematics used to check for this property, the eigenvalues of the coefficient matrix lie on the unit circle. These coefficients are also considered to be unstable.

normality assumption, joint and conditional. The derivation of the standard, linear SEM estimator usually assumes the full joint normality of the observed and latent variables. However, full joint normality can replace the assumption of normality conditional on the values of the exogenous variables, and all that is lost is one goodness-of-fit test (the test reported by `sem` on the output) and the justification for use of optional method MLMV for dealing with missing values. This substitution of assumptions is important for researchers who cannot reasonably assume normality of the observed variables. This includes any researcher including, say, variables age and age-squared in his or her model.

Meanwhile, the generalized SEM makes only the conditional normality assumption.

Be aware that even though the full joint normality assumption is not required for the standard linear SEM, `sem` calculates the log-likelihood value under that assumption. This is irrelevant except that log-likelihood values reported by `sem` cannot be compared with log-likelihood values reported by `gsem`, which makes the lesser assumption.

See [SEM] **Intro 4**.

normalization constraints. See *identification*.

normalized residuals. See *standardized residuals*.

observed variables. A variable is observed if it is a variable in your dataset. In this documentation, we often refer to observed variables by using `x1`, `x2`, ..., `y1`, `y2`, and so on; in reality, observed variables have names such as `mpg`, `weight`, `testscore`, and so on.

In the software, observed variables are usually indicated by having names that are all lowercase.

Also see *latent variable*.

OIM, **vce(oim)**. OIM stands for observed information matrix, defined as the inverse of the negative of the matrix of second derivatives, usually of the log likelihood function. The OIM is an estimate of the VCE. OIM is the default VCE that `sem` and `gsem` report. The other available techniques are EIM, OPG, robust, cluster, bootstrap, and jackknife.

OPG, **vce(opg)**. OPG stands for outer product of the gradients, defined as the cross product of the observation-level first derivatives, usually of the log likelihood function. The OPG is an estimate of the VCE. The other available techniques are OIM, EIM, robust, cluster, bootstrap, and jackknife.

ordered complementary log–log regression. Ordered complementary log–log regression is a term for generalized linear response functions that are family ordinal, link cloglog. It is used for ordinal-outcome data. Ordered complementary log–log regression is also known as ocloglog in Stata circles. See *generalized linear response functions*.

ordered logit regression. Ordered logit regression is a term for generalized linear response functions that are family ordinal, link logit. It is used for ordinal outcome data. Ordered logit regression is also known as ordered logistic regression, as just ordered logit, and as ologit in Stata circles. See *generalized linear response functions*.

ordered probit regression. Ordered probit regression is a term for generalized linear response functions that are family ordinal, link probit. It is used for ordinal-outcome data. Ordered probit regression is also known as just ordered probit and known as oprobit in Stata circles. See *generalized linear response functions*.

parameter constraints. Parameter constraints are restrictions placed on the parameters of the model. These constraints are typically in the form of 0 constraints and equality constraints. A 0 constraint is implied, for instance, when no path is drawn connecting x with y. An equality constraint is specified when one path coefficient is forced to be equal to another or one covariance is forced to be equal to another.

Also see *identification*.

parameters, ancillary parameters. The parameters are the to-be-estimated coefficients of a model. These include all path coefficients, means, variances, and covariances. In mathematical notation, the theoretical parameters are often written as $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\alpha}$ is the vector of intercepts, $\boldsymbol{\beta}$ is the vector of path coefficients, $\boldsymbol{\mu}$ is the vector of means, and $\boldsymbol{\Sigma}$ is the matrix of variances and covariances. The resulting parameter estimates are written as $\hat{\boldsymbol{\theta}}$.

Ancillary parameters are extra parameters beyond the ones just described that concern the distribution. These include the scale parameter of gamma regression, the dispersion parameter for negative binomial regression, and the cutpoints for ordered probit, logit, and cloglog regression, and the like. These parameters are also included in θ.

path. A path, typically shown as an arrow drawn from one variable to another, states that the first variable determines the second variable, at least partially. If $x \rightarrow y$, or equivalently $y \leftarrow x$, then $y_j = \alpha + \cdots + \beta x_j + \cdots + e.y_j$, where β is said to be the $x \rightarrow y$ path coefficient. The ellipses are included to account for paths to y from other variables. α is said to be the intercept and is automatically added when the first path to y is specified.

A curved path is a curved line connecting two variables, and it specifies that the two variables are allowed to be correlated. If there is no curved path between variables, the variables are usually assumed to be uncorrelated. We say usually because correlation is assumed among observed exogenous variables and, in the command language, assumed among latent exogenous variables, and if some of the correlations are not desired, they must be suppressed. Many authors refer to covariances rather than correlations. Strictly speaking, the curved path denotes a nonzero covariance. A correlation is often called a standardized covariance.

A curved path can connect a variable to itself, and in that case, it indicates a variance. In path diagrams in this manual, we typically do not show such variance paths even though variances are assumed.

path coefficient. The path coefficient is associated with a path; see *path*. Also see *intercept*.

path diagram. A path diagram is a graphical representation that shows the relationships among a set of variables using paths. See [SEM] **Intro 2** for a description of path diagrams.

path notation. Path notation is a syntax defined by the authors of Stata's `sem` and `gsem` commands for entering path diagrams in a command language. Models to be fit may be specified in path notation or they may be drawn using path diagrams into the Builder.

Poisson regression. Poisson regression is a term for generalized linear response functions that are family Poisson, link log. It is used for count data. See *generalized linear response functions*.

probit regression. Probit regression is a term for generalized linear response functions that are family Bernoulli, link probit. It is used for binary outcome data. Probit regression is also known simply as probit. See *generalized linear response functions*.

p-value. P-value is another term for the reported significance level associated with a test. Small p-values indicate rejection of the null hypothesis.

QML, **method(ml) vce(robust)**. QML stands for quasimaximum likelihood. It is a method used to obtain fitted parameters and a technique used to obtain the corresponding VCE. QML is used by `sem` and `gsem` when options `method(ml)` and `vce(robust)` are specified. Other available methods are ML, MLMV, and ADF. Other available techniques are OIM, EIM, OPG, cluster, bootstrap, and jackknife.

quadrature. Quadrature is generic method for performing numerical integration. `gsem` uses quadrature in any model including latent variables (excluding error variables). `sem`, being limited to linear models, does not need to perform quadrature.

random-effects models. See *multilevel models*.

regression. A regression is a model in which an endogenous variable is written as a function of other variables, parameters to be estimated, and a random disturbance.

reliability. Reliability is the proportion of the variance of a variable not due to measurement error. A variable without measure error has reliability 1.

residual. In this manual, we reserve the word "residual" for the difference between the observed and fitted moments of an SEM. We use the word "error" for the disturbance associated with a (Gaussian) linear equation; see *error*. Also see *standardized residuals*.

robust, **vce(robust)**. Robust is the name we use here for the Huber/White/sandwich estimator of the VCE. This technique requires fewer assumptions than most other techniques. In particular, it merely assumes that the errors are independently distributed across observations and thus allows the errors to be heteroskedastic. Robust standard errors are reported when the `sem` (`gsem`) option `vce(robust)` is specified. The other available techniques are OIM, EIM, OPG, cluster, bootstrap, and jackknife.

saturated model. A saturated model is a full covariance model—a model of fitted means and covariances of observed variables without any restrictions on the values. Also see *baseline model*. Saturated models apply only to standard linear SEMs.

score test, **Lagrange multiplier test**. A score test is a test based on first derivatives of a likelihood function. Score tests are especially convenient for testing whether constraints on parameters should be relaxed or parameters should be added to a model. Also see *Wald test*.

scores. Scores has two unrelated meanings. First, scores are the observation-by-observation first-derivatives of the (quasi) log-likelihood function. When we use the word "scores", this is what we mean. Second, in the factor-analysis literature, scores (usually in the context of factor scores) refers to the expected value of a latent variable conditional on all the observed variables. We refer to this simply as the predicted value of the latent variable.

second-level latent variable. See *first-, second-, and higher-order latent variables*.

second-order latent variable. See *first- and second-order latent variables*.

seemingly unrelated regression. Seemingly unrelated regression is a kind of structural model in which each member of a set of observed endogenous variables is a function of a set of observed exogenous variables and a unique random disturbance term. The disturbances are correlated and the sets of exogenous variables may overlap. If the sets of exogenous variables are identical, this is referred to as multivariate regression.

SEM. SEM stands for structural equation modeling and for structural equation model. We use SEM in capital letters when writing about theoretical or conceptual issues as opposed to issues of the particular implementation of SEM in Stata with the `sem` or `gsem` commands.

sem. `sem` is the Stata command that fits standard linear SEMs. Also see *gsem*.

SSD, **ssd**. See *summary statistics data*.

standard linear SEM. An SEM without multilevel effects in which all response variables are given by a linear equation. Standard linear SEM is what most people mean when they refer to just SEM. Standard linear SEMs are fit by `sem`, although they can also be fit by `gsem`; see *generalized SEM*.

standardized coefficient. In a linear equation $y = \ldots\ bx + \ldots$, the standardized coefficient β is $(\hat{\sigma}_y/\hat{\sigma}_x)b$. Standardized coefficients are scaled to units of standard deviation change in y for a standard deviation change in x.

standardized covariance. A standardized covariance between y and x is equal to the correlation of y and x, that is, it is equal to $\sigma_{xy}/\sigma_x\sigma_y$. The covariance is equal to the correlation when variables are standardized to have variance 1.

standardized residuals, **normalized residuals**. Standardized residuals are residuals adjusted so that they follow a standard normal distribution. The difficulty is that the adjustment is not always possible. Normalized residuals are residuals adjusted according to a different formula that roughly follow a standard normal distribution. Normalized residuals can always be calculated.

starting values. The estimation methods provided by `sem` and `gsem` are iterative. The starting values are values for each of the parameters to be estimated that are used to initialize the estimation process. `sem` and `gsem` provide starting values automatically, but in some cases, these are not good enough and you must both diagnose the problem and provide better starting values. See [SEM] **Intro 12**.

structural equation model. Different authors use the term "structural equation model" in different ways, but all would agree that an SEM sometimes carries the connotation of being a structural model with a measurement component, that is, combined with a measurement model.

structural model. A structural model is a model in which the parameters are not merely a description but are believed to be of a causal nature. Obviously, SEM can fit structural models and thus so can `sem` and `gsem`. Neither SEM, `sem`, nor `gsem` are limited to fitting structural models, however.

Structural models often have multiple equations and dependencies between endogenous variables, although that is not a requirement.

See [SEM] **Intro 5**. Also see *structural equation model*.

structured (correlation or covariance). See *unstructured and structured (correlation or covariance)*.

substantive constraints. See *identification*.

summary statistics data. Data are sometimes available only in summary statistics form, as means and covariances; means, standard deviations or variances, and correlations; covariances; standard deviations or variances and correlations; or correlations. SEM can be used to fit models with such data in place of the underlying raw data. The `ssd` command creates datasets containing summary statistics.

technique. Technique is just an English word and should be read in context. Nonetheless, technique is usually used here to refer to the technique used to calculate the estimated VCE. Those techniques are OIM, EIM, OPG, robust, cluster, bootstrap, and jackknife.

Technique is also used to refer to the available techniques used with `ml`, Stata's optimizer and likelihood maximizer, to find the solution.

total effects. See *direct, indirect, and total effects*.

unstandardized coefficient. A coefficient that is not standardized. If $\texttt{mpg} = -0.006 \times \texttt{weight} + 39.44028$, then -0.006 is an unstandardized coefficient and, as a matter of fact, is measured in mpg-per-pound units.

unstructured and structured (correlation or covariance). A set of variables, typically error variables, is said to have an unstructured correlation or covariance if the covariance matrix has no particular pattern imposed by theory. If a pattern is imposed, the correlation or covariance is said to be structured.

variance–covariance matrix of the estimator. The estimator is the formula used to solve for the fitted parameters, sometimes called the fitted coefficients. The VCE is the estimated variance–covariance matrix of the parameters. The diagonal elements of the VCE are the variances of the parameters or equivalent; the square roots of those elements are the reported standard errors of the parameters.

VCE. See *variance–covariance matrix of the estimator*.

Wald test. A Wald test is a statistical test based on the estimated variance–covariance matrix of the parameters. Wald tests are especially convenient for testing possible constraints to be placed on the estimated parameters of a model. Also see *score test*.

weighted least squares. Weighted least squares (WLS) is a method used to obtain fitted parameters. In this documentation, WLS is referred to as ADF, which stands for asymptotic distribution free. Other available methods are ML, QML, and MLMV. ADF is, in fact, a specific kind of the more generic WLS.

WLS. See *weighted least squares*.

Reference

Bentler, P. M., and D. G. Weeks. 1980. Linear structural equations with latent variables. *Psychometrika* 45: 289–308. https://doi.org/10.1007/BF02293905.