

---

## Glossary

**Arellano–Bond estimator.** The Arellano–Bond estimator is a generalized method of moments (GMM) estimator for linear dynamic panel-data models that uses lagged levels of the endogenous variables as well as first differences of the exogenous variables as instruments. The Arellano–Bond estimator removes the panel-specific heterogeneity by first-differencing the regression equation.

**autoregressive process.** In autoregressive processes, the current value of a variable is a linear function of its own past values and a white-noise error term. For panel data, a first-order autoregressive process, denoted as an AR(1) process, is  $y_{it} = \rho y_{i,t-1} + \epsilon_{it}$ , where  $i$  denotes panels,  $t$  denotes time, and  $\epsilon_{it}$  is white noise.

**balanced data.** A longitudinal or panel dataset is said to be balanced if each panel has the same number of observations. See also *weakly balanced* and *strongly balanced*.

**between estimator.** The between estimator is a panel-data estimator that obtains its estimates by running OLS on the panel-level means of the variables. This estimator uses only the between-panel variation in the data to identify the parameters, ignoring any within-panel variation. For it to be consistent, the between estimator requires that the panel-level means of the regressors be uncorrelated with the panel-specific heterogeneity terms.

**BLUPs.** BLUPs are best linear unbiased predictions of either random effects or linear combinations of random effects. In linear models containing random effects, these effects are not estimated directly but instead are integrated out of the estimation. Once the fixed effects and variance components have been estimated, you can use these estimates to predict group-specific random effects. These predictions are called BLUPs because they are unbiased and have minimal mean squared error among all linear functions of the response.

**canonical link.** Corresponding to each family of distributions in a generalized linear model is a canonical link function for which there is a sufficient statistic with the same dimension as the number of parameters in the linear predictor. The use of canonical link functions provides the GLM with desirable statistical properties, especially when the sample size is small.

**conditional fixed-effects model.** In general, including panel-specific dummies to control for fixed effects in nonlinear models results in inconsistent estimates. For some nonlinear models, the fixed-effect term can be removed from the likelihood function by conditioning on a sufficient statistic. For example, the conditional fixed-effect logit model conditions on the number of positive outcomes within each panel.

**correlation structure.** A correlation structure is a set of assumptions imposed on the within-panel variance–covariance matrix of the errors in a panel-data model. See [XT] **xtgee** for examples of different correlation structures.

**crossed-effects model.** A crossed-effects model is a mixed model in which the levels of random effects are not nested. A simple crossed-effects model for cross-sectional time-series data would contain a random effect to control for panel-specific variation and a second random effect to control for time-specific random variation. Rather than being nested within panel, in this model a random effect due to a given time is the same for all panels.

**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.

**cross-sectional time-series data.** Cross-sectional time-series data is another name for panel data. The term *cross-sectional time-series data* is sometimes reserved for datasets in which a relatively small number of panels were observed over many periods. See also *panel data*.

**disturbance term.** The disturbance term encompasses any shocks that occur to the dependent variable that cannot be explained by the conditional (or deterministic) portion of the model.

**dynamic model.** A dynamic model is one in which prior values of the dependent variable or disturbance term affect the current value of the dependent variable.

**endogenous variable.** An endogenous variable is a regressor that is correlated with the unobservable error term. Equivalently, an endogenous variable is one whose values are determined by the equilibrium or outcome of a structural model.

**error-components model.** The error-components model is another name for the random-effects model. See also *random-effects model*.

**exogenous variable.** An exogenous variable is a regressor that is not correlated with any of the error terms in the model. Equivalently, an exogenous variable is one whose values change independently of the other variables in a structural model.

**fixed-effects model.** The fixed-effects model is a model for panel data in which the panel-specific errors are treated as fixed parameters. These parameters are panel-specific intercepts and therefore allow the conditional mean of the dependent variable to vary across panels. The linear fixed-effects estimator is consistent, even if the regressors are correlated with the fixed effects. See also *random-effects model*.

**generalized estimating equations (GEE).** The method of generalized estimating equations is used to fit population-averaged panel-data models. GEE extends the GLM method by allowing the user to specify a variety of different within-panel correlation structures.

**generalized linear model (GLM).** The generalized linear model is an estimation framework in which the user specifies a distributional family for the dependent variable and a link function that relates the dependent variable to a linear combination of the regressors. The distribution must be a member of the exponential family of distributions. GLM encompasses many common models, including linear, probit, and Poisson regression.

**hierarchical model.** A hierarchical model is one in which successively more narrowly defined groups are nested within larger groups. For example, in a hierarchical model, patients may be nested within doctors who are in turn nested within the hospital at which they practice.

**idiosyncratic error term.** In longitudinal or panel-data models, the idiosyncratic error term refers to the observation-specific zero-mean random-error term. It is analogous to the random-error term of cross-sectional regression analysis.

**instrumental variables.** Instrumental variables are exogenous variables that are correlated with one or more of the endogenous variables in a structural model. The term *instrumental variable* is often reserved for those exogenous variables that are not included as regressors in the model.

**instrumental-variables (IV) estimator.** An instrumental variables estimator uses instrumental variables to produce consistent parameter estimates in models that contain endogenous variables. IV estimators can also be used to control for measurement error.

**interval data.** Interval data are data in which the true value of the dependent variable is not observed. Instead, all that is known is that the value lies within a given interval.

**link function.** In a GLM, the link function relates a linear combination of predictors to the expected value of the dependent variable. In a linear regression model, the link function is simply the identity function.

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

- mixed model.** A mixed model contains both fixed and random effects. The fixed effects are estimated directly, whereas the random effects are summarized according to their (co)variances. Mixed models are used primarily to perform estimation and inference on the regression coefficients in the presence of complicated within-panel correlation structures induced by multiple levels of grouping.
- negative binomial regression model.** The negative binomial regression model is for applications in which the dependent variable represents the number of times an event occurs. The negative binomial regression model is an alternative to the Poisson model for use when the dependent variable is overdispersed, meaning that the variance of the dependent variable is greater than its mean.
- one-level model.** A one-level mixed model is a mixed model with one level of random variation. Suppose that you have a panel dataset consisting of patients at hospitals; a one-level model would contain a set of random effects “at the hospital level” to control for hospital-specific random variation.
- overidentifying restrictions.** The order condition for model identification requires that the number of exogenous variables excluded from the model be at least as great as the number of endogenous regressors. When the number of excluded exogenous variables exceeds the number of endogenous regressors, the model is overidentified, and the validity of the instruments can then be checked via a test of overidentifying restrictions.
- panel-corrected standard errors (PCSEs).** The term *panel-corrected standard errors* refers to a class of estimators for the variance–covariance matrix of the OLS estimator when there are relatively few panels with many observations per panel. PCSEs account for heteroskedasticity, autocorrelation, or cross-sectional correlation.
- 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.
- Poisson regression model.** The Poisson regression model is used when the dependent variable represents the number of times an event occurs. In the Poisson model, the variance of the dependent variable is equal to the conditional mean.
- pooled estimator.** A pooled estimator ignores the longitudinal or panel aspect of a dataset and treats the observations as if they were cross-sectional.
- population-averaged model.** A population-averaged model is used for panel data in which the parameters measure the effects of the regressors on the outcome for the average individual in the population. The panel-specific errors are treated as uncorrelated random variables drawn from a population with zero mean and constant variance, and the parameters measure the effects of the regressors on the dependent variable after integrating over the distribution of the random effects.
- predetermined variable.** A predetermined variable is a regressor in which its contemporaneous and future values are not correlated with the unobservable error term but past values are correlated with the error term.
- prewhiten.** To prewhiten is to apply a transformation to a time series so that it becomes white noise.
- production function.** A production function describes the maximum amount of a good that can be produced, given specified levels of the inputs.
- quadrature.** Quadrature is a set of numerical methods to evaluate an integral. Two types of quadrature commonly used in fitting panel-data models are Gaussian and Gauss–Hermite quadrature.

**random-coefficients model.** A random-coefficients model is a panel-data model in which group-specific heterogeneity is introduced by assuming that each group has its own parameter vector, which is drawn from a population common to all panels.

**random-effects model.** A random-effects model for panel data treats the panel-specific errors as uncorrelated random variables drawn from a population with zero mean and constant variance. The regressors must be uncorrelated with the random effects for the estimates to be consistent.

**REML (restricted maximum likelihood).** REML is a method of fitting linear mixed models that involves transforming out the fixed effects so as to focus solely on variance-component estimation.

**restricted maximum likelihood.** See *REML*.

**robust standard errors.** Robust standard errors, also known as Huber/White or Taylor linearization standard errors, are based on the sandwich estimator of variance. Robust standard errors can be interpreted as representing the sample-to-sample variability of the parameter estimates, even when the model is misspecified. See also *semirobust standard errors*.

**semirobust standard errors.** Semirobust standard errors are closely related to robust standard errors and can be interpreted as representing the sample-to-sample variability of the parameter estimates, even when the model is misspecified, as long as the mean structure of the model is specified correctly. See also *robust standard errors*.

**sequential limit theory.** The sequential limit theory is a method of determining asymptotic properties of a panel-data statistic in which one index, say,  $N$ , the number of panels, is held fixed, while  $T$ , the number of time periods, goes to infinity, providing an intermediate limit. Then one obtains a final limit by studying the behavior of this intermediate limit as the other index ( $N$  here) goes to infinity.

**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.

**two-level model.** A two-level mixed model is a mixed model with two levels of random variation. Suppose that you have a dataset consisting of patients overseen by doctors at hospitals, and each doctor practices at one hospital. Then a two-level model would contain a set of random effects to control for hospital-specific variation and a second set of random effects to control for doctor-specific random variation.

**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*.

**variance components.** In a mixed model, the variance components refer to the variances and covariances of the various random effects.

**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.

**white noise.** A variable,  $u_t$ , represents a white-noise process if the mean of  $u_t$  is zero, the variance of  $u_t$  is  $\sigma^2$ , and the covariance between  $u_t$  and  $u_s$  is zero for all  $s \neq t$ .

**within estimator.** The within estimator is a panel-data estimator that removes the panel-specific heterogeneity by subtracting the panel-level means from each variable and then performing ordinary least squares on the demeaned data. The within estimator is used in fitting the linear fixed-effects model.