GLOSSARY

A

Adjusted *R*-Squared: A goodness-of-fit measure in multiple regression analysis that penalizes additional explanatory variables by using a degrees of freedom adjustment in estimating the error variance.

Alternative Hypothesis: The hypothesis against which the null hypothesis is tested.

AR(1) Serial Correlation: The errors in a time series regression model follow an AR(1) model.

Asymptotic Bias: *See* inconsistency.

Asymptotic Confidence Interval: A confidence interval that is approximately valid in large sample sizes.

Asymptotic Normality: The sampling distribution of a properly normalized estimator converges to the standard normal distribution.

Asymptotic Properties: Properties of estimators and test statistics that apply when the sample size grows without bound.

Asymptotic Standard Error: A standard error that is valid in large samples.

Asymptotic *t* Statistic: A *t* statistic that has an approximate standard normal distribution in large samples.

Asymptotic Variance: The square of the value we must divide an estimator by in order to obtain an asymptotic standard normal distribution.

Asymptotically Efficient: For consistent estimators with asymptotically normal distributions, the estimator with the smallest asymptotic variance.

Asymptotically Uncorrelated: A time series process in which the correlation between random variables at two points in time tends to zero as the time interval between them increases. (*See also* weakly dependent.)

Attenuation Bias: Bias in an estimator that is always toward zero; thus, the expected value of an estimator with attenuation bias is less in magnitude than the absolute value of the parameter.

Augmented Dickey-Fuller Test: A test for a unit root that includes lagged changes of the variable as regressors.

Autocorrelation: *See* serial correlation.

Autoregressive Conditional Heteroskedasticity (ARCH): A model of dynamic heteroskedasticity where the variance of the error term, given past information, depends linearly on the past squared errors.

Autoregressive Process of Order One [AR(1)]: A time series model whose current value depends linearly on its most recent value plus an unpredictable disturbance.

Auxiliary Regression: A regression used to compute a test statistic—such as the test statistics for heteroskedasticity and serial correlation—or any other regression that does not estimate the model of primary interest.

Average: The sum of n numbers divided by n.

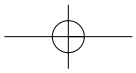
B

Balanced Panel: A panel data set where all years (or periods) of data are available for all cross-sectional units.

Base Group: The group represented by the overall intercept in a multiple regression model that includes dummy explanatory variables.

Base Period: For index numbers, such as price or production indices, the period against which all other time periods are measured.

Base Value: The value assigned to the base period for constructing an index number; usually the base value is one or 100.



Benchmark Group: *See* base group.

Bernoulli Random Variable: A random variable that takes on the values zero or one.

Best Linear Unbiased Estimator (BLUE): Among all linear, unbiased estimators, the estimator with the smallest variance. OLS is BLUE, conditional on the sample values of the explanatory variables, under the Gauss-Markov assumptions.

Beta Coefficients: *See* standardized coefficients.

Bias: The difference between the expected and the population parameter values of an estimator.

Biased Estimator: An estimator whose expectation, or sampling mean, is different from the population value it is supposed to be estimating.

Biased Towards Zero: A description of an estimator whose expectation in absolute value is less than the absolute value of the population parameter.

Binary Response Model: A model for a binary (dummy) dependent variable.

Binary Variable: *See* dummy variable.

Binomial Distribution: The probability distribution of the number of successes out of n independent Bernoulli trials, where each trial has the same probability of success.

Bivariate Regression Model: *See* simple linear regression model.

BLUE: *See* best linear unbiased estimator.

Breusch-Godfrey Test: An asymptotically justified test for AR(p) serial correlation, with AR(1) being the most popular; the test allows for lagged dependent variables as well as other regressors that are not strictly exogenous.

Breusch-Pagan Test: A test for heteroskedasticity where the squared OLS residuals are regressed on the explanatory variables in the model.

C

Causal Effect: A ceteris paribus change in one variable has an effect on another variable.

Censored Regression Model: A multiple regression model where the dependent variable has been censored above or below some known threshold.

Central Limit Theorem: A key result from probability theory which implies that the sum of independent random variables, or even weakly dependent random variables, when standardized by its standard deviation, has a distribution that tends to standard normal as the sample size grows.

Ceteris Paribus: All other relevant factors are held fixed.

Chi-Square Distribution: A probability distribution obtained by adding the squares of independent standard normal random variables. The number of terms in the sum equals the degrees of freedom in the distribution.

Chow Statistic: An F statistic for testing the equality of regression parameters across different groups (say, men and women) or time periods (say, before and after a policy change).

Classical Errors-in-Variables (CEV): A measurement error model where the observed measure equals the actual variable plus an independent, or at least an uncorrelated, measurement error.

Classical Linear Model: The multiple linear regression model under the full set of classical linear model assumptions.

Classical Linear Model (CLM) Assumptions: The ideal set of assumptions for multiple regression analysis: for cross-sectional analysis, Assumptions MLR.1 through MLR.6 and for time series analysis, Assumptions TS.1 through TS.6. The assumptions include linearity in the parameters, no perfect collinearity, the zero conditional mean assumption, homoskedasticity, no serial correlation, and normality of the errors.

Cluster Effect: An unobserved effect that is common to all units, usually people, in the cluster.

Cluster Sample: A sample of natural clusters or groups which usually consist of people.

Cochrane-Orcutt (CO) Estimation: A method of estimating a multiple linear regression model with AR(1) errors and strictly exogenous explanatory variables; unlike Prais-Winsten, Cochrane-Orcutt does not use the equation for the first time period.

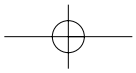
Coefficient of Determination: *See* R-squared.

Cointegration: The notion that a linear combination of two series, each of which is integrated of order one, is integrated of order zero.

Composite Error: In a panel data model, the sum of the time constant unobserved effect and the idiosyncratic error.

Conditional Distribution: The probability distribution of one random variable, given the values of one or more other random variables.

Conditional Expectation: The expected or average value of one random variable, called the dependent or explained variable, that depends on the values of one or more other variables, called the independent or explanatory variables.



Conditional Forecast: A forecast that assumes the future values of some explanatory variables are known with certainty.

Conditional Variance: The variance of one random variable, given one or more other random variables.

Confidence Interval (CI): A rule used to construct a random interval so that a certain percentage of all data sets, determined by the confidence level, yields an interval that contains the population value.

Confidence Level: The percentage of samples in which we want our confidence interval to contain the population value; 95% is the most common confidence level, but 90% and 99% are also used.

Consistent Estimator: An estimator that converges in probability to the population parameter as the sample size grows without bound.

Consistent Test: A test where, under the alternative hypothesis, the probability of rejecting the null hypothesis converges to one as the sample size grows without bound.

Constant Elasticity Model: A model where the elasticity of the dependent variable, with respect to an explanatory variable, is constant; in multiple regression, both variables appear in logarithmic form.

Contemporaneously Exogenous Regressor: In time series or panel data applications, a regressor that is uncorrelated with the error term in the same time period, but not necessarily in other time periods.

Continuous Random Variable: A random variable that takes on any particular value with probability zero.

Control Group: In program evaluation, the group that does not participate in the program.

Control Variable: *See* explanatory variable.

Corner Solution: A nonnegative dependent variable that is roughly continuous over strictly positive values but takes on the value zero with some regularity.

Correlation Coefficient: A measure of linear dependence between two random variables that does not depend on units of measurement and is bounded between -1 and 1.

Count Variable: A variable that takes on nonnegative integer values.

Covariance: A measure of linear dependence between two random variables.

Covariance Stationary: A time series process with constant mean and variance where the covariance between any two random variables in the sequence depends only on the distance between them.

Covariate: *See* explanatory variable.

Critical Value: In hypothesis testing, the value against which a test statistic is compared to determine whether or not the null hypothesis is rejected.

Cross-Sectional Data Set: A data set collected from a population at a given point in time.

Cumulative Distribution Function (cdf): A function that gives the probability of a random variable being less than or equal to any specified real number.

D

Data Censoring: A situation that arises when we do not always observe the outcome on the dependent variable because at an upper (or lower) threshold we only know that the outcome was above (or below) the threshold. (*See also* censored regression model.)

Data Frequency: The interval at which time series data are collected. Yearly, quarterly, and monthly are the most common data frequencies.

Data Mining: The practice of using the same data set to estimate numerous models in a search to find the "best" model.

Davidson-MacKinnon Test: A test that is used for testing a model against a nonnested alternative; it can be implemented as a t test on the fitted values from the competing model.

Degrees of Freedom (*df*): In multiple regression analysis, the number of observations minus the number of estimated parameters.

Denominator Degrees of Freedom: In an F test, the degrees of freedom in the unrestricted model.

Dependent Variable: The variable to be explained in a multiple regression model (and a variety of other models).

Descriptive Statistic: A statistic used to summarize a set of numbers; the sample average, sample median, and sample standard deviation are the most common.

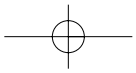
Deseasonalizing: The removing of the seasonal components from a monthly or quarterly time series.

Detrending: The practice of removing the trend from a time series.

Dickey-Fuller Distribution: The limiting distribution of the t statistic in testing the null hypothesis of a unit root.

Dickey-Fuller (DF) Test: A t test of the unit root null hypothesis in an AR(1) model. (*See also* augmented Dickey-Fuller test.)

Difference in Slopes: A description of a model where some slope parameters may differ by group or time period.



Difference-in-Differences Estimator: An estimator that arises in policy analysis with data for two time periods. One version of the estimator applies to independently pooled cross sections and another to panel data sets.

Diminishing Marginal Effect: The marginal effect of an explanatory variable becomes smaller as the value of the explanatory variable increases.

Discrete Random Variable: A random variable that takes on at most a finite or countably infinite number of values.

Distributed Lag Model: A time series model that relates the dependent variable to current and past values of an explanatory variable.

Disturbance: *See* error term.

Downward Bias: The expected value of an estimator is below the population value of the parameter.

Dummy Dependent Variable: *See* binary response model.

Dummy Variable: A variable that takes on the value zero or one.

Dummy Variable Regression: In a panel data setting, the regression that includes a dummy variable for each cross-sectional unit, along with the remaining explanatory variables. It produces the fixed effects estimator.

Dummy Variable Trap: The mistake of including too many dummy variables among the independent variables; it occurs when an overall intercept is in the model and a dummy variable is included for each group.

Duration Analysis: An application of the censored regression model, where the dependent variable is time elapsed until a certain event occurs, such as the time before an unemployed person becomes reemployed.

Durbin-Watson (DW) Statistic: A statistic used to test for first order serial correlation in the errors of a time series regression model under the classical linear model assumptions.

Dynamically Complete Model: A time series model where no further lags of either the dependent variable or the explanatory variables help to explain the mean of the dependent variable.

E

Econometric Model: An equation relating the dependent variable to a set of explanatory variables and unobserved disturbances, where unknown population parameters determine the ceteris paribus effect of each explanatory variable.

Economic Model: A relationship derived from economic theory or less formal economic reasoning.

Economic Significance: *See* practical significance.

Elasticity: The percent change in one variable given a 1% ceteris paribus increase in another variable.

Empirical Analysis: A study that uses data in a formal econometric analysis to test a theory, estimate a relationship, or determine the effectiveness of a policy.

Endogeneity: A term used to describe the presence of an endogenous explanatory variable.

Endogenous Explanatory Variable: An explanatory variable in a multiple regression model that is correlated with the error term, either because of an omitted variable, measurement error, or simultaneity.

Endogenous Sample Selection: Nonrandom sample selection where the selection is related to the dependent variable, either directly or through the error term in the equation.

Endogenous Variables: In simultaneous equations models, variables that are determined by the equations in the system.

Engle-Granger Two-Step Procedure: A two-step method for estimating error correction models whereby the cointegrating parameter is estimated in the first stage, and the error correction parameters are estimated in the second.

Error Correction Model: A time series model in first differences that also contains an error correction term, which works to bring two I(1) series back into long-run equilibrium.

Error Term: The variable in a simple or multiple regression equation that contains unobserved factors that affect the dependent variable. The error term may also include measurement errors in the observed dependent or independent variables.

Error Variance: The variance of the error term in a multiple regression model.

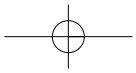
Errors-in-Variables: A situation where either the dependent variable or some independent variables are measured with error.

Estimate: The numerical value taken on by an estimator for a particular sample of data.

Estimator: A rule for combining data to produce a numerical value for a population parameter; the form of the rule does not depend on the particular sample obtained.

Event Study: An econometric analysis of the effects of an event, such as a change in government regulation or economic policy, on an outcome variable.

Excluding a Relevant Variable: In multiple regression analysis, leaving out a variable that has a nonzero partial effect on the dependent variable.



Exclusion Restrictions: Restrictions which state that certain variables are excluded from the model (or have zero population coefficients).

Exogenous Explanatory Variable: An explanatory variable that is uncorrelated with the error term.

Exogenous Sample Selection: Sample selection that either depends on exogenous explanatory variables or is independent of the error term in the equation of interest.

Exogenous Variable: Any variable that is uncorrelated with the error term in the model of interest.

Expected Value: A measure of central tendency in the distribution of a random variable, including an estimator.

Experiment: In probability, a general term used to denote an event whose outcome is uncertain. In econometric analysis, it denotes a situation where data are collected by randomly assigning individuals to control and treatment groups.

Experimental Data: Data that have been obtained by running a controlled experiment.

Experimental Group: *See* treatment group.

Explained Sum of Squares (SSE): The total sample variation of the fitted values in a multiple regression model.

Explained Variable: *See* dependent variable.

Explanatory Variable: In regression analysis, a variable that is used to explain variation in the dependent variable.

Exponential Function: A mathematical function defined for all values that have an increasing slope but a constant proportionate change.

Exponential Smoothing: A simple method of forecasting a variable that involves a weighting of all previous outcomes on that variable.

Exponential Trend: A trend with a constant growth rate.

F

***F* Distribution:** The probability distribution obtained by forming the ratio of two independent chi-square random variables, where each has been divided by its degrees of freedom.

***F* Statistic:** A statistic used to test multiple hypotheses about the parameters in a multiple regression model.

Feasible GLS (FGLS) Estimator: A GLS procedure where variance or correlation parameters are unknown and therefore must first be estimated. (*See also* generalized least squares estimator.)

Finite Distributed Lag (FDL) Model: A dynamic model where one or more explanatory variables are allowed to have lagged effects on the dependent variable.

First Difference: A transformation on a time series constructed by taking the difference of adjacent time periods, where the earlier time period is subtracted from the later time period.

First-Differenced Equation: In time series or panel data models, an equation where the dependent and independent variables have all been first-differenced.

First-Differenced Estimator: In a panel data setting, the pooled OLS estimator applied to first differences of the data across time.

First Order Conditions: The set of linear equations used to solve for the OLS estimates.

Fitted Values: The estimated values of the dependent variable when the values of the independent variables for each observation are plugged into the OLS regression line.

Fixed Effect: *See* unobserved effect.

Fixed Effects Estimator: For the unobserved effects panel data model, the estimator obtained by applying pooled OLS to a time-demeaned equation.

Fixed Effects Transformation: For panel data, the time-demeaned data.

Forecast Error: The difference between the actual outcome and the forecast of the outcome.

Forecast Interval: In forecasting, a confidence interval for a yet unrealized future value of a time series variable. (*See also* prediction interval.)

Functional Form Misspecification: A problem that occurs when a model has omitted functions of the explanatory variables (such as quadratics) or uses the wrong functions of either the dependent variable or some explanatory variables.

G

Gauss-Markov Assumptions: The set of assumptions (Assumptions MLR.1 through MLR.5 or TS.1 through TS.5) under which OLS is BLUE.

Gauss-Markov Theorem: The theorem which states that, under the five Gauss-Markov assumptions (for cross-sectional or time series models), the OLS estimator is BLUE (conditional on the sample values of the explanatory variables).

Generalized Least Squares (GLS) Estimator: An estimator that accounts for a known structure of the error variance (heteroskedasticity), serial correlation pattern in the errors, or both, via a transformation of the original model.

Geometric (or Koyck) Distributed Lag: An infinite distributed lag model where the lag coefficients decline at a geometric rate.

Goodness-of-Fit Measure: A statistic that summarizes how well a set of explanatory variables explains a dependent or response variable.

Granger Causality: A limited notion of causality where past values of one series (x_t) are useful for predicting future values of another series (y_t), after past values of y_t have been controlled for.

Growth Rate: The proportionate change in a time series from the previous period. It may be approximated as the difference in logs or reported in percentage form.

H

Heckit Method: An econometric procedure used to correct for sample selection bias due to incidental truncation or some other form of nonrandomly missing data.

Heterogeneity Bias: The bias in OLS due to omitted heterogeneity (or omitted variables).

Heteroskedasticity: The variance of the error term, given the explanatory variables, is not constant.

Heteroskedasticity of Unknown Form: Heteroskedasticity that may depend on the explanatory variables in an unknown, arbitrary fashion.

Heteroskedasticity-Robust *F* Statistic: An *F*-type statistic that is (asymptotically) robust to heteroskedasticity of unknown form.

Heteroskedasticity-Robust *LM* Statistic: An *LM* statistic that is robust to heteroskedasticity of unknown form.

Heteroskedasticity-Robust Standard Error: A standard error that is (asymptotically) robust to heteroskedasticity of unknown form.

Heteroskedasticity-Robust *t* Statistic: A *t* statistic that is (asymptotically) robust to heteroskedasticity of unknown form.

Highly Persistent Process: A time series process where outcomes in the distant future are highly correlated with current outcomes.

Homoskedasticity: The errors in a regression model have constant variance, conditional on the explanatory variables.

Hypothesis Test: A statistical test of the null, or maintained, hypothesis against an alternative hypothesis.

I

Identified Equation: An equation whose parameters can be consistently estimated, especially in models with endogenous explanatory variables.

Idiosyncratic Error: In panel data models, the error that changes over time as well as across units (say, individuals, firms, or cities).

Impact Elasticity: In a distributed lag model, the immediate percentage change in the dependent variable given a 1% increase in the independent variable.

Impact Multiplier: *See* impact propensity.

Impact Propensity: In a distributed lag model, the immediate change in the dependent variable given a one-unit increase in the independent variable.

Incidental Truncation: A sample selection problem whereby one variable, usually the dependent variable, is only observed for certain outcomes of another variable.

Inclusion of an Irrelevant Variable: The including of an explanatory variable in a regression model that has a zero population parameter in estimating an equation by OLS.

Inconsistency: The difference between the probability limit of an estimator and the parameter value.

Independent Random Variables: Random variables whose joint distribution is the product of the marginal distributions.

Independent Variable: *See* explanatory variable.

Independently Pooled Cross Section: A data set obtained by pooling independent random samples from different points in time.

Index Number: A statistic that aggregates information on economic activity, such as production or prices.

Infinite Distributed Lag (IDL) Model: A distributed lag model where a change in the explanatory variable can have an impact on the dependent variable into the indefinite future.

Influential Observations: *See* outliers.

Information Set: In forecasting, the set of variables that we can observe prior to forming our forecast.

In-Sample Criteria: Criteria for choosing forecasting models that are based on goodness-of-fit within the sample used to obtain the parameter estimates.

Instrumental Variable (IV): In an equation with an endogenous explanatory variable, an IV is a variable that does not appear in the equation, is uncorrelated with the error in the equation, and is (partially) correlated with the endogenous explanatory variable.

Instrumental Variables (IV) Estimator: An estimator in a linear model used when instrumental variables are available for one or more endogenous explanatory variables.

Integrated of Order One [I(1)]: A time series process that needs to be first-differenced in order to produce an I(0) process.

Integrated of Order Zero [I(0)]: A stationary, weakly dependent time series process that, when used in regression analysis, satisfies the law of large numbers and the central limit theorem.

Interaction Effect: In multiple regression, the partial effect of one explanatory variable depends on the value of a different explanatory variable.

Interaction Term: An independent variable in a regression model that is the product of two explanatory variables.

Intercept Parameter: The parameter in a multiple linear regression model that gives the expected value of the dependent variable when all the independent variables equal zero.

Intercept Shift: The intercept in a regression model differs by group or time period.

Internet: A global computer network that can be used to access information and download data bases.

Interval Estimator: A rule that uses data to obtain lower and upper bounds for a population parameter. (*See also* confidence interval.)

Inverse Mills Ratio: A term that can be added to a multiple regression model to remove sample selection bias.

J

Joint Distribution: The probability distribution determining the probabilities of outcomes involving two or more random variables.

Joint Hypothesis Test: A test involving more than one restriction on the parameters in a model.

Jointly Statistically Significant: The null hypothesis that two or more explanatory variables have zero population coefficients is rejected at the chosen significance level.

Just Identified Equation: For models with endogenous explanatory variables, an equation that is identified but would not be identfied with one fewer instrumental variable.

L

Lag Distribution: In a finite or infinite distributed lag model, the lag coefficients graphed as a function of the lag length.

Lagged Dependent Variable: An explanatory variable that is equal to the dependent variable from an earlier time period.

Lagged Endogenous Variable: In a simultaneous equations model, a lagged value of one of the endogenous variables.

Lagrange Multiplier Statistic: A test statistic with large sample justification that can be used to test for omitted variables, heteroskedasticity, and serial correlation, among other model specification problems.

Large Sample Properties: *See* asymptotic properties.

Latent Variable Model: A model where the observed dependent variable is assumed to be a function of an underlying latent, or unobserved, variable.

Law of Iterated Expectations: A result from probability that relates unconditional and conditional expectations.

Law of Large Numbers (LLN): A theorem which says that the average from a random sample converges in probability to the population average; the LLN also holds for stationary and weakly dependent time series.

Leads and Lags Estimator: An estimator of a cointegrating parameter in a regression with I(1) variables, where the current, some past, and some future first differences in the explanatory variable are included as regressors.

Level-Level Model: A regression model where the dependent variable and the independent variables are in level (or original) form.

Level-Log Model: A regression model where the dependent variable is in level form and (at least some of) the independent variables are in logarithmic form.

Likelihood Ratio Statistic: A statistic that can be used to test single or multiple hypotheses when the constrained and unconstrained models have been estimated by maximum likelihood. The statistic is twice the difference in the unconstrained and constrained log-likelihoods.

Limited Dependent Variable: A dependent or response variable whose range is restricted in some important way.

Linear Function: A function where the change in the dependent variable, given a one-unit change in an independent variable, is constant.

Linear Probability Model (LPM): A binary response model where the response probability is linear in its parameters.

Linear Time Trend: A trend that is a linear function of time.

Linear Unbiased Estimator: In multiple regression analysis, an unbiased estimator that is a linear function of the outcomes on the dependent variable.

Logarithmic Function: A mathematical function defined for positive arguments that has a positive, but diminishing, slope.

Log-Level Model: A regression model where the dependent variable is in logarithmic form and the independent variables are in level (or original) form.

Log-Log Model: A regression model where the dependent variable and (at least some of) the explanatory variables are in logarithmic form.

Logit Model: A model for binary response where the response probability is the logit function evaluated at a linear function of the explanatory variables.

Log-Likelihood Function: The sum of the log-likelihoods, where the log-likelihood for each observation is the log of the density of the dependent variable given the explanatory variables; the log-likelihood function is viewed as a function of the parameters to be estimated.

Long-Run Elasticity: The long-run propensity in a distributed lag model with the dependent and independent variables in logarithmic form; thus, the long-run elasticity is the eventual percentage increase in the explained variable, given a permanent 1% increase in the explanatory variable.

Long-Run Multiplier: *See* long-run propensity.

Long-Run Propensity: In a distributed lag model, the eventual change in the dependent variable given a permanent, one-unit increase in the independent variable.

Longitudinal Data: *See* panel data.

Loss Function: A function that measures the loss when a forecast differs from the actual outcome; the most common examples are absolute value loss and squared loss.

M

Marginal Effect: The effect on the dependent variable that results from changing an independent variable by a small amount.

Martingale: A time series process whose expected value, given all past outcomes on the series, simply equals the most recent value.

Martingale Difference Sequence: The first difference of a martingale. It is unpredictable (or has a zero mean), given past values of the sequence.

Matched Pairs Sample: A sample where each observation is matched with another, as in a sample consisting of a husband and wife or a set of two siblings.

Matrix: An array of numbers.

Matrix Notation: A convenient mathematical notation, grounded in matrix algebra, for expressing and manipulating the multiple regression model.

Maximum Likelihood Estimation (MLE): A broadly applicable estimation method where the parameter estimates are chosen to maximize the log-likelihood function.

Mean: *See* expected value.

Mean Absolute Error (MAE): A performance measure in forecasting, computed as the average of the absolute values of the forecast errors.

Mean Squared Error: The expected squared distance that an estimator is from the population value; it equals the variance plus the square of any bias.

Measurement Error: The difference between an observed variable and the variable that belongs in a multiple regression equation.

Median: In a probability distribution, it is the value where there is a 50% chance of being below the value and a 50% chance of being above it. In a sample of numbers, it is the middle value after the numbers have been ordered.

Method of Moments Estimator: An estimator obtained by using the sample analog of population moments; ordinary least squares and two stage least squares are both method of moments estimators.

Micronumerosity: A term introduced by Arthur Goldberger to describe properties of econometric estimators with small sample sizes.

Minimum Variance Unbiased Estimator: An estimator with the smallest variance in the class of all unbiased estimators.

Missing Data: A data problem that occurs when we do not observe values on some variables for certain observations (individuals, cities, time periods, and so on) in the sample.

Moving Average Process of Order One [MA(1)]: A time series process generated as a linear function of the current value and one lagged value of a zero-mean, constant variance, uncorrelated stochastic process.

Multicollinearity: A term that refers to correlation among the independent variables in a multiple regression model; it is usually invoked when some correlations are "large," but an actual magnitude is not well-defined.

Multiple Hypothesis Test: A test of a null hypothesis involving more than one restriction on the parameters.

Multiple Linear Regression (MLR) Model: A model linear in its parameters, where the dependent variable is a function of independent variables plus an error term.

Multiple Regression Analysis: A type of analysis that is used to describe estimation of and inference in the multiple linear regression model.

Multiple Restrictions: More than one restriction on the parameters in an econometric model.

Multiple Step-Ahead Forecast: A time series forecast of more than one period into the future.

Multiplicative Measurement Error: Measurement error where the observed variable is the product of the true unobserved variable and a positive measurement error.

N

***n*-*R*-Squared Statistic:** *See* Lagrange multiplier statistic.

Natural Experiment: A situation where the economic environment—sometimes summarized by an explanatory variable—exogenously changes, perhaps inadvertently, due to a policy or institutional change.

Natural Logarithm: *See* logarithmic function.

Nominal Variable: A variable measured in nominal or current dollars.

Nonexperimental Data: Data that have not been obtained through a controlled experiment.

Nonlinear Function: A function whose slope is not constant.

Nonnested Models: Two (or more) models where no model can be written as a special case of the other by imposing restrictions on the parameters.

Nonrandom Sample Selection: A sample selection process that cannot be characterized as drawing randomly from the population of interest.

Nonstationary Process: A time series process whose joint distributions are not constant across different epochs.

Normal Distribution: A probability distribution commonly used in statistics and econometrics for modeling a population. Its probability distribution function has a bell shape.

Normality Assumption: The classical linear model assumption which states that the error (or dependent variable) has a normal distribution, conditional on the explanatory variables.

Null Hypothesis: In classical hypothesis testing, we take this hypothesis as true and require the data to provide substantial evidence against it.

Numerator Degrees of Freedom: In an F test, the number of restrictions being tested.

O

Observational Data: *See* nonexperimental data.

OLS: *See* ordinary least squares.

OLS Intercept Estimate: The intercept in an OLS regression line.

OLS Regression Line: The equation relating the predicted value of the dependent variable to the independent variables, where the parameter estimates have been obtained by OLS.

OLS Slope Estimate: A slope in an OLS regression line.

Omitted Variable Bias: The bias that arises in the OLS estimators when a relevant variable is omitted from the regression.

Omitted Variables: One or more variables, which we would like to control for, have been omitted in estimating a regression model.

One-Sided Alternative: An alternative hypothesis which states that the parameter is greater than (or less than) the value hypothesized under the null.

One-Step-Ahead Forecast: A time series forecast one period into the future.

One-Tailed Test: A hypothesis test against a one-sided alternative.

On-Line Data Bases: Data bases that can be accessed via a computer network.

On-Line Search Services: Computer software that allows the Internet or data bases on the Internet to be searched by topic, name, title, or key words.

Order Condition: A necessary condition for identifying the parameters in a model with one or more endogenous explanatory variables: the total number of exogenous variables must be at least as great as the total number of explanatory variables.

Ordinal Variable: A variable where the ordering of the values conveys information but the magnitude of the values does not.

Ordinary Least Squares (OLS): A method for estimating the parameters of a multiple linear regression model. The ordinary least squares estimates are obtained by minimizing the sum of squared residuals.

Outliers: Observations in a data set that are substantially different from the bulk of the data, perhaps because of errors or because some data are generated by a different model than most of the other data.

Out-of-Sample Criteria: Criteria used for choosing forecasting models that are based on a part of the sample that was not used in obtaining parameter estimates.

Overall Significance of a Regression: A test of the joint significance of all explanatory variables appearing in a multiple regression equation.

Overdispersion: In modeling a count variable, the variance is larger than the mean.

Overidentified Equation: In models with endogenous explanatory variables, an equation where the number of instrumental variables is strictly greater than the number of endogenous explanatory variables.

Overidentifying Restrictions: The extra moment conditions that come from having more instrumental variables than endogenous explanatory variables in a linear model.

Overspecifying a Model: *See* inclusion of an irrelevant variable.

P

***p*-value:** The smallest significance level at which the null hypothesis can be rejected. Equivalently, the largest significance level at which the null hypothesis cannot be rejected.

Panel Data: A data set constructed from repeated cross sections over time. With a *balanced* panel, the same units appear in each time period. With an *unbalanced* panel, some units do not appear in each time period, often due to attrition.

Pairwise Uncorrelated Random Variables: A set of two or more random variables where each pair is uncorrelated.

Parameter: An unknown value that describes a population relationship.

Parsimonious Model: A model with as few parameters as possible for capturing any desired features.

Partial Effect: The effect of an explanatory variable on the dependent variable, holding other factors in the regression model fixed.

Percent Correctly Predicted: In a binary response model, the percentage of times the prediction of zero or one coincides with the actual outcome.

Percentage Change: The porportionate change in a variable, multiplied by 100.

Percentage Point Change: The change in a variable that is measured as a percent.

Perfect Collinearity: In multiple regression, one independent variable is an exact linear function of one or more other independent variables.

Plug-In Solution to the Omitted Variables Problem: A proxy variable is substituted for an unobserved omitted variable in an OLS regression.

Point Forecast: The forecasted value of a future outcome.

Poisson Distribution: A probability distribution for count variables.

Poisson Regression Model: A model for a count dependent variable where the dependent variable, conditional on the explanatory variables, is nominally assumed to have a Poisson distribution.

Policy Analysis: An empirical analysis that uses econometric methods to evaluate the effects of a certain policy.

Pooled Cross Section: A data configuration where independent cross sections, usually collected at different points in time, are combined to produce a single data set.

Pooled OLS Estimation: OLS estimation with independently pooled cross sections, panel data, or cluster samples, where the observations are pooled across time (or group) as well as across the cross-sectional units.

Population: A well-defined group (of people, firms, cities, and so on) that is the focus of a statistical or econometric analysis.

Population Model: A model, especially a multiple linear regression model, that describes a population.

Population *R*-Squared: In the population, the fraction of the variation in the dependent variable that is explained by the explanatory variables.

Population Regression Function: *See* conditional expectation.

Power of a Test: The probability of rejecting the null hypothesis when it is false; the power depends on the values of the population parameters under the alternative.

Practical Significance: The practical or economic importance of an estimate, which is measured by its sign and magnitude, as opposed to its statistical significance.

Prais-Winsten (PW) Estimation: A method of estimating a multiple linear regression model with AR(1) errors and strictly exogenous explanatory variables; unlike Cochrane-Orcutt, Prais-Winsten uses the equation for the first time period in estimation.

Predetermined Variable: In a simultaneous equations model, either a lagged endogenous variable or a lagged exogenous variable.

Predicted Variable: *See* dependent variable.

Prediction: The estimate of an outcome obtained by plugging specific values of the explanatory variables into an estimated model, usually a multiple regression model.

Prediction Error: The difference between the actual outcome and a prediction of that outcome.

Prediction Interval: A confidence interval for an

unknown outcome on a dependent variable in a multiple regression model.

Predictor Variable: *See* explanatory variable.

Probability Density Function (pdf): A function that, for discrete random variables, gives the probability that the random variable takes on each value; for continuous random variables, the area under the pdf gives the probability of various events.

Probability Limit: The value to which an estimator converges as the sample size grows without bound.

Probit Model: A model for binary responses where the response probability is the standard normal cdf evaluated at a linear function of the explanatory variables.

Program Evaluation: An analysis of a particular private or public program using econometric methods to obtain the causal effect of the program.

Proportionate Change: The change in a variable relative to its initial value; mathematically, the change divided by the initial value.

Proxy Variable: An observed variable that is related but not identical to an unobserved explanatory variable in multiple regression analysis.

Q

Quadratic Functions: Functions that contain squares of one or more explanatory variables; they capture diminishing or increasing effects on the dependent variable.

Qualitative Variable: A variable describing a nonquantitative feature of an individual, a firm, a city, and so on.

Quasi-Demeaned Data: In random effects estimation for panel data, it is the original data in each time period minus a fraction of the time average; these calculations are done for each cross-sectional observation.

Quasi-Differenced Data: In estimating a regression model with AR(1) serial correlation, it is the difference between the current time period and a multiple of the previous time period, where the multiple is the parameter in the AR(1) model.

Quasi-Experiment: *See* natural experiment.

Quasi-Likelihood Ratio Statistic: A modification of the likelihood ratio statistic that accounts for possible distributional misspecification, as in a Poisson regression model.

Quasi-Maximum Likelihood Estimation: Maximum likelihood estimation but where the log-likelihood function may not correspond to the actual conditional distribution of the dependent variable.

R

***R*-Bar Squared:** *See* adjusted *R*-squared.

***R*-Squared:** In a multiple regression model, the proportion of the total sample variation in the dependent variable that is explained by the independent variable.

***R*-Squared Form of the *F* Statistic:** The *F* statistic for testing exclusion restrictions expressed in terms of the *R*-squareds from the restricted and unrestricted models.

Random Effects Estimator: A feasible GLS estimator in the unobserved effects model where the unobserved effect is assumed to be uncorrelated with the explanatory variables in each time period.

Random Effects Model: The unobserved effects panel data model where the unobserved effect is assumed to be uncorrelated with the explanatory variables in each time period.

Random Sampling: A sampling scheme whereby each observation is drawn at random from the population. In particular, no unit is more likely to be selected than any other unit, and each draw is independent of all other draws.

Random Variable: A variable whose outcome is uncertain.

Random Walk: A time series process where next period's value is obtained as this period's value, plus an independent (or at least an uncorrelated) error term.

Random Walk with Drift: A random walk that has a constant (or drift) added in each period.

Rank Condition: A sufficient condition for identification of a model with one or more endogenous explanatory variables.

Rational Distributed Lag (RDL) Model: A type of infinite distributed lag model where the lag distribution depends on relatively few parameters.

Real Variable: A monetary value measured in terms of a base period.

Reduced Form Equation: A linear equation where an endogenous variable is a function of exogenous variables and unobserved errors.

Reduced Form Error: The error term appearing in a reduced form equation.

Reduced Form Parameters: The parameters appearing in a reduced form equation.

Regressand: *See* dependent variable.

Regression Error Specification Test (RESET): A general test for functional form in a multiple regression model; it is an *F* test of joint significance of the

squares, cubes, and perhaps higher powers of the fitted values from the initial OLS estimation.

Regression Through the Origin: Regression analysis where the intercept is set to zero; the slopes are obtained by minimizing the sum of squared residuals, as usual.

Regressor: *See* explanatory variable.

Rejection Region: The set of values of a test statistic that leads to rejecting the null hypothesis.

Rejection Rule: In hypothesis testing, the rule that determines when the null hypothesis is rejected in favor of the alternative hypothesis.

Residual: The difference between the actual value and the fitted (or predicted) value; there is a residual for each observation in the sample used to obtain an OLS regression line.

Residual Analysis: A type of analysis that studies the sign and size of residuals for particular observations after a multiple regression model has been estimated.

Residual Sum of Squares: *See* sum of squared residuals.

Response Probability: In a binary response model, the probability that the dependent variable takes on the value one, conditional on explanatory variables.

Response Variable: *See* dependent variable.

Restricted Model: In hypothesis testing, the model obtained after imposing all of the restrictions required under the null.

Root Mean Squared Error (RMSE): Another name for the standard error of the regression in multiple regression analysis.

S

Sample Average: The sum of n numbers divided by n; a measure of central tendency.

Sample Correlation: For outcomes on two random variables, the sample covariance divided by the product of the sample standard deviations.

Sample Covariance: An unbiased estimator of the population covariance between two random variables.

Sample Regression Function: *See* OLS regression line.

Sample Selection Bias: Bias in the OLS estimator which is induced by using data that arise from endogenous sample selection.

Sample Standard Deviation: A consistent estimator of the population standard deviation.

Sample Variance: An unbiased, consistent estimator of the population variance.

Sampling Distribution: The probability distribution of an estimator over all possible sample outcomes.

Sampling Variance: The variance in the sampling distribution of an estimator; it measures the spread in the sampling distribution.

Score Statistic: *See* Lagrange multiplier statistic.

Seasonal Dummy Variables: A set of dummy variables used to denote the quarters or months of the year.

Seasonality: A feature of monthly or quarterly time series where the average value differs systematically by season of the year.

Seasonally Adjusted: Monthly or quarterly time series data where some statistical procedure—possibly regression on seasonal dummy variables—has been used to remove the seasonal component.

Selected Sample: A sample of data obtained not by random sampling but by selecting on the basis of some observed or unobserved characteristic.

Semi-Elasticity: The percentage change in the dependent variable given a one-unit increase in an independent variable.

Sensitivity Analysis: The process of checking whether the estimated effects and statistical significance of key explanatory variables are sensitive to inclusion of other explanatory variables, functional form, dropping of potentially outlying observations, or different methods of estimation.

Serial Correlation: In a time series or panel data model, correlation between the errors in different time periods.

Serial Correlation-Robust Standard Error: A standard error for an estimator that is (asymptotically) valid whether or not the errors in the model are serially correlated.

Serially Uncorrelated: The errors in a time series or panel data model are pairwise uncorrelated across time.

Short-Run Elasticity: The impact propensity in a distributed lag model when the dependent and independent variables are in logarithmic form.

Significance Level: The probability of Type I error in hypothesis testing.

Simple Linear Regression Model: A model where the dependent variable is a linear function of a single independent variable, plus an error term.

Simultaneity: A term that means at least one explanatory variable in a multiple linear regression model is determined jointly with the dependent variable.

Simultaneity Bias: The bias that arises from using OLS to estimate an equation in a simultaneous equations model.

Simultaneous Equations Model (SEM): A model that jointly determines two or more endogenous variables,

where each endogenous variable can be a function of other endogenous variables as well as of exogenous variables and an error term.

Slope Parameter: The coefficient on an independent variable in a multiple regression model.

Spreadsheet: Computer software used for entering and manipulating data.

Spurious Correlation: A correlation between two variables that is not due to causality, but perhaps to the dependence of the two variables on another unobserved factor.

Spurious Regression Problem: A problem that arises when regression analysis indicates a relationship between two or more unrelated time series processes simply because each has a trend, is an integrated time series (such as a random walk), or both.

Stable AR(1) Process: An AR(1) process where the parameter on the lag is less than one in absolute value. The correlation between two random variables in the sequence declines to zero at a geometric rate as the distance between the random variables increases, and so a stable AR(1) process is weakly dependent.

Standard Deviation: A common measure of spread in the distribution of a random variable.

Standard Deviation of $\hat{\beta}_j$: A common measure of spread in the sampling distribution of $\hat{\beta}_j$.

Standard Error of $\hat{\beta}_j$: An estimate of the standard deviation in the sampling distribution of $\hat{\beta}_j$.

Standard Error of the Estimate: *See* standard error of the regression.

Standard Error of the Regression (SER): In multiple regression analysis, the estimate of the standard deviation of the population error, obtained as the square root of the sum of squared residuals over the degrees of freedom.

Standard Normal Distribution: The normal distribution with mean zero and variance one.

Standardized Coefficient: A regression coefficient that measures the standard deviation change in the dependent variable given a one standard deviation increase in an independent variable.

Standardized Random Variable: A random variable transformed by subtracting off its expected value and dividing the result by its standard deviation; the new random variable has mean zero and standard deviation one.

Static Model: A time series model where only contemporaneous explanatory variables affect the dependent variable.

Stationary Process: A time series process where the marginal and all joint distributions are invariant across time.

Statistical Inference: The act of testing hypotheses about population parameters.

Statistically Different from Zero: *See* statistically significant.

Statistically Insignificant: Failure to reject the null hypothesis that a population parameter is equal to zero, at the chosen significance level.

Statistically Significant: Rejecting the null hypothesis that a parameter is equal to zero against the specified alternative, at the chosen significance level.

Stochastic Process: A sequence of random variables indexed by time.

Strict Exogeneity: An assumption that holds in a time series or panel data model when the explanatory variables are strictly exogenous.

Strictly Exogenous: A feature of explanatory variables in a time series or panel data model where the error term at any time period has zero expectation, conditional on the explanatory variables in all time periods; a less restrictive version is stated in terms of zero correlations.

Strongly Dependent: *See* highly persistent process.

Structural Equation: An equation derived from economic theory or from less formal economic reasoning.

Structural Error: The error term in a structural equation, which could be one equation in a simultaneous equations model.

Structural Parameters: The parameters appearing in a structural equation.

Sum of Squared Residuals: In multiple regression analysis, the sum of the squared OLS residuals across all observations.

Summation Operator: A notation, denoted by Σ, used to define the summing of a set of numbers.

T

***t* Distribution:** The distribution of the ratio of a standard normal random variable and the square root of an independent chi-square random variable, where the chi-square random variable is first divided by its *df*.

***t* Ratio:** *See t* statistic.

***t* Statistic:** The statistic used to test a single hypothesis about the parameters in an econometric model.

Test Statistic: A rule used for testing hypotheses where each sample outcome produces a numerical value.

Text Editor: Computer software that can be used to edit text files.

Text (ASCII) File: A universal file format that can be transported across numerous computer platforms.

Time-Demeaned Data: Panel data where, for each cross-sectional unit, the average over time is subtracted from the data in each time period.

Time Series Data: Data collected over time on one or more variables.

Time Series Process: *See* stochastic process.

Time Trend: A function of time that is the expected value of a trending time series process.

Tobit Model: A model for a dependent variable that takes on the value zero with positive probability but is roughly continuously distributed over strictly positive values. (*See also* corner solution.)

Top Coding: A form of data censoring where the value of a variable is not reported when it is above a given threshold; we only know that it is at least as large as the threshold.

Total Sum of Squares (SST): The total sample variation in a dependent variable about its sample average.

Treatment Group: In program evaluation, the group that participates in the program. (*See also* experimental group.)

Trending Process: A time series process whose expected value is an increasing or decreasing function of time.

Trend-Stationary Process: A process that is stationary once a time trend has been removed; it is usually implicit that the detrended series is weakly dependent.

Truncated Regression Model: A classical linear regression model for cross-sectional data in which the sampling scheme entirely excludes, on the basis of outcomes on the dependent variable, part of the population.

True Model: The actual population model relating the dependent variable to the relevant independent variables, plus a disturbance, where the zero conditional mean assumption holds.

Two Stage Least Squares (2SLS) Estimator: An instrumental variables estimator where the IV for an endogenous explanatory variable is obtained as the fitted value from regressing the endogenous explanatory variable on all exogenous variables.

Two-Sided Alternative: An alternative where the population parameter can be either less than or greater than the value stated under the null hypothesis.

Two-Tailed Test: A test against a two-sided alternative.

Type I Error: A rejection of the null hypothesis when it is true.

Type II Error: The failure to reject the null hypothesis when it is false.

U

Unbalanced Panel: A panel data set where certain years (or periods) of data are missing for some cross-sectional units.

Unbiased Estimator: An estimator whose expected value (or mean of its sampling distribution) equals the population value (regardless of the population value).

Unconditional Forecast: A forecast that does not rely on knowing, or assuming values for, future explanatory variables.

Uncorrelated Random Variables: Random variables that are not linearly related.

Underspecifying a Model: *See* excluding a relevant variable.

Unidentified Equation: An equation with one or more endogenous explanatory variables where sufficient instrumental variables do not exist to identify the parameters.

Unit Root Process: A highly persistent time series process where the current value equals last period's value, plus a weakly dependent disturbance.

Unobserved Effect: In a panel data model, an unobserved variable in the error term that does not change over time. For cluster samples, an unobserved variable that is common to all units in the cluster.

Unobserved Effects Model: A model for panel data or cluster samples where the error term contains an unobserved effect.

Unobserved Heterogeneity: *See* unobserved effect.

Unrestricted Model: In hypothesis testing, the model that has no restrictions placed on its parameters.

Upward Bias: The expected value of an estimator is greater than the population parameter value.

V

Variance: A measure of spread in the distribution of a random variable.

Variance of the Prediction Error: The variance in the error that arises when predicting a future value of the dependent variable based on an estimated multiple regression equation.

Vector Autoregressive (VAR) Model: A model for two or more time series where each variable is modeled as a linear function of past values of all variables, plus disturbances that have zero means given all past values of the observed variables.

W

Weakly Dependent: A term that describes a time series process where some measure of dependence between random variables at two points in time—such as correlation—diminishes as the interval between the two points in time increases.

Weighted Least Squares (WLS) Estimator: An estimator used to adjust for a known form of heteroskedasticity, where each squared residual is weighted by the inverse of the (estimated) variance of the error.

White Test: A test for heteroskedasticity that involves regressing the squared OLS residuals on the OLS fitted values and on the squares of the fitted values; in its most general form, the squared OLS residuals are regressed on the explanatory variables, the squares of the explanatory variables, and all the nonredundant cross products of the explanatory variables.

Within Estimator: *See* fixed effects estimator.

Within Transformation: *See* fixed effects transformation.

Y

Year Dummy Variables: For data sets with a time series component, dummy (binary) variables equal to one in the relevant year and zero in all other years.

Z

Zero Conditional Mean Assumption: A key assumption used in multiple regression analysis which states that, given any values of the explanatory variables, the expected value of the error equals zero. (*See* Assumptions MLR.3, TS.2, and TS.2′.)