A Glossary of Commonly-Seen Statistical Terms in the Social Sciences

Christopher Zorn
August 2022

Note: This glossary was written by me, and is designed to be neither comprehensive nor representative, only useful. Asterisks (*) indicate subjects covered in PLSC 502; others are covered in subsequent courses in PLSC.

Artificial Intelligence ("AI"): See *Machine Learning*.

Artificial Intelligence ("AI") (2022 version): A term used in consulting firms which can mean anything from "an HTML dashboard that shows frequency tables" to "a DeBERTaV3 / ERNIE-Doc ensemble for deep transfer learning."

ARIMA – An acronym for "autoregressive integrated moving average." A *regression model* for *time-series data*, used to study temporal relationships.

Binary Data* – see *Dichotomous Data*.

Causal Inference – In political and social science, the process of assessing the causal effect of some factor on an outcome of interest, often including the set of tools for doing so in a manner that is valid and reliable. Obtaining valid, reliable causal inferences about social and behavioral phenomena – particularly those not amenable to experimental manipulation and study – is a challenging and widely-studied area of methodology in the social sciences.

Coefficient* – A quantity *estimated* from the data that describes an aspect of the relationship(s) among the variables. In a *regression* model, the coefficients (usually denoted β) indicate the strength and direction of the relationship between one of more *independent variables* X and the *dependent variable* Y. Close synonyms include "parameter" and (in a regression context) "slope."

Confidence Interval* – A measure of uncertainty around an *estimate* or *prediction*. Similar in application to a *P-value*, confidence intervals give the reader an idea of the likelihood that the result in question might be due to (statistical) chance. They are best interpreted, however, as a range of values that are reasonably "close" to the estimate; they are thus an indication of the reliability of that estimate. As a rule, analyses based on larger amounts of data will have smaller confidence intervals.

Correlation* – In general, a measure of the direction and degree of (usually linear) association between two variables. A positive correlation means that as one variable (say, self-identification with the Republican party) increases in value, we would expect the value of the other variable (say, the likelihood of voting for McCain) to increase as well. A negative correlation means that as one variable increases, the value of the other variable would be expected to decrease (as in the relationship between GOP identification and the likelihood of voting for Obama). a commonly-used

version is "Pearson's correlation," or "Pearson's r."

Covariate* – See *Independent Variable*.

Dependent Variable* – The phenomenon of interest; the "thing" the authors are interested in explaining. Typically denoted as Y. In a medical study, the dependent variable might be the occurrence of cancer in some individuals but not others; an engineer might study why some bridges collapse and others do not. An economist might try to explain rates of inflation or economic growth.

Dichotomous Data* – A phenomenon of interest that can take on only one of two possible outcomes, e.g., a diagnosis of pregnancy (pregnant or not), whether or not someone has ever attended an NFL game, or the party of the person elected president in 2008 (Democratic or Republican).

Duration Model – See *Survival Model*.

Error-Correction Model – A form of *regression model* for *time-series data*. Used to study series where there is an expectation that, over the long run, the values of the *dependent variable* will return to some stable equilibrium value.

Estimate* – A quantity calculated from data that provides a guess as to the true value of some *coefficient*. Often denoted using a "hat," so that the estimate of the *regression* coefficient β is written as "$\hat{\beta}$."

Event Count Data* – Data where the dependent variable takes the form of a count of the number of occurrences of some event. Examples include the number of eggs laid by a boat-tailed grackle, the number of times an individual has viewed the movie *Repo Man*, or the number of bills a president vetoes in a session of Congress.

Event-History Model – see *Survival Model*.

Fixed-Effects Model – A type of *regression model* for *panel* or *time-series cross-sectional* data, in which each unit of observation has a distinct "baseline" value of the *dependent variable* from each other.

Hazard Model – See *Survival Model*.

Independent Variable* – The factor or factors that the authors believe cause, are *correlated* with, or otherwise explain the dependent variable. A medical study of lung cancer would consider each patient's history of smoking to be a key independent variable; an engineer studying bridge collapse might use bridges' age, materials of construction, and average daily load as independent variables. Economists would potentially explain inflation rates with the money supply, and so use (e.g.) in-

terest rates as a key independent variable.

Linear Regression* – A particular *regression* model in which the relationship (sometimes denoted $f(\cdot)$) between the *dependent* and *independent variables* is assumed to be linear. Linearity means that a one-unit change in X corresponds to a constant expected change in Y. Linear regression is most often used when the *dependent variable* is continuous (that is, can take on a wide range of values) and measured at the interval level of higher.

Logit – A form of *regression model* used when the *dependent variable* is *dichotomous*. Variants include "ordered logit," for dependent variables that form ordered categories (such as "Agree," "Neutral," and "Disagree" responses on a public opinion survey) and "multinomial logit" for dependent variables consisting of three or more non-ordered outcomes (e.g., $\{$Telecaster, Les Paul, Stratocaster$\}$).

Machine Learning – Computational and/or statistical tools (algorithms, estimators, etc.) that use data to improve the performance of computers at achieving some task. Machine learning is often considered a part/form of *artificial intelligence*, and encompasses a wide range of tools, including many of the entries in this document. Sometimes abbreviated "ML," it is a term that is widely used in computer and information science to refer to things that might be called "statistics" or "models" in other disciplines.

Negative Binomial Model – A type of *regression model* for *event count data*, based on the negative binomial distribution. It is generally considered to be more flexible than the *Poisson model*.

Network Analysis – Broadly, the study of relations among entities through the use of visual, conceptual, and/or mathematical/statistical graphs. In the social and behavioral sciences, network analysis is often called "social network analysis" (abbreviated "SNA"), and refers to models for studying interrelated social entities and the nature of those interrelations. SNA is a widely-used class of tools in sociology, political science, and other related fields.

Nonlinearity – The condition when an association between two variables Y and X is such that $\frac{\partial Y}{\partial X} \neq c$. Nonlinearity often refers either to a property of a *relationship* between two variables, or to a property of a *model*; for example, a *logit* regression model assumes a nonlinear functional form describes the relationship between the predictors X and the outcome $\Pr(Y = 1)$.

OLS* – "Ordinary Least Squares." A means of estimating a (typically, linear) *regression* model. See *Linear Regression*.

P-Value* – A measure of the probability of obtaining the finding observed (or a more extreme finding) given that the hypothesis being tested is not true. P-values are often mistaken for a form of *statistical significance* testing.

Panel Data – Data where we observe multiple units over multiple time points, and where the number of units (typically denoted N) is significantly greater than the number of time points (denoted T). A common example is a multi-wave public opinion survey, where (say) 1,000 respondents are interviewed at two or three separate points in time (say, once before the primary election, once between the primary and the general, and once after the general); such a study would have $N = 1000$ and $T = 3$, with $NT = 3000$.

Panel-Corrected Standard Errors – A technique used in *regression models* for *panel* and *time-series cross-sectional* data.

Poisson Model – A type of *regression model* for *event count data*, based on the Poisson distribution.

Prediction* – One or more values of the *dependent variable* Y derived by combining the *estimates* ($\hat{\beta}$) from a *regression model* and values of the *independent variables* X. Often denoted using a "hat" (as in "$\hat{Y}$"). Predictions are often used to illustrate the implications of regression model results on actual outcomes, and are often accompanied by *confidence intervals*.

Probit – A form of *regression model* used when the *dependent variable* is *dichotomous*. It is based upon the normal distribution. Variants include "ordered probit," for dependent variables that form ordered categories (such as "Agree," "Neutral," and "Disagree" responses on a public opinion survey) and "multinomial probit" for dependent variables consisting of three or more non-ordered outcomes (e.g., {Volkswagen, Subaru, Jeep}).

Random-Effects Model – A type of *regression model* for *panel* or *time-series cross-sectional* data, similar to a *fixed-effects* model.

Regression* – A form of statistical model used to relate a single *dependent variable* to one or more *independent variables*. In general, a regression takes the form:

$$Y = f(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + u)$$

where Y is the dependent variable, the Xs are the independent variables, the βs are *coefficients* describing the strength and direction of the relationship between Y and X, u is a random "error term," and $f(\cdot)$ denotes some function that describes the "shape" of the relationship between Y and the Xs.

Regression Coefficient* – A quantity that is estimated from the data that describes the strength and direction of the relationship between the *dependent* (Y) and *independent* (X) variables. In a *regression* model, these are typically denoted by the Greek letter β (beta). A positive regression coefficient (that is, $\beta > 0$) generally means that an increase in X leads to or is associated with an increase in Y. A negative regression coefficient (that is, $\beta < 0$) generally means that an increase

in X leads to or is associated with an increase in Y.

Statistical Significance* – A measure of the degree of likelihood that a statistical finding (e.g., an *estimate*) occurred by chance. High levels of statistical significance indicate that the statistical result is unlikely to have been a "fluke," and are often interpreted (somewhat incorrectly) as evidence that a finding is "real." A statistically significant result is not necessarily one that has real-world, substantive importance; for example, studies that use large amounts of data will often exhibit statistically significant results whose substantive importance is quite low.

Survival Analysis – The study of the time until some event of interest occurs; also referred to variously as "duration analysis," "reliability analysis," and "event-history analysis." Survival models are form of *regression model* where the *dependent variable* takes the form of a length of time (a "duration"). Also referred to as *event history* models, *duration* models, and *hazard* (or *hazard-rate*) models. Developed in biostatistics and medicine to model (e.g.) human lifespans, drug effectiveness, etc. In some instances, the dependent variable is the hazard (risk) of an event occurring in time.

Time-Series Cross-Sectional Data* – Data where we observe multiple units over multiple time points, and where the number of units (typically denoted N) is equal to or less than the number of time points (denoted T). An example would be a study of the annual party control of state legislatures in the thirteen original states, where $N = 13$ and $T = 233$ (that is, one observation for each state for each year, 1789-2021), so that $NT = 13 \times 233 = 3029$ total lines of data (one for each state for each year that it has existed). Panel and time-series cross-sectional data are often denoted with a double index (subscript), such as "Y_{it}," with i denoting the unit of observation (e.g., the state) and t denoting the time of observation (here, the year).

Time-Series Data* – Data consisting of long series of regular measurements on single variables for single units of observation. Examples include daily amounts of rainfall in State College, PA, monthly U.S. inflation rates, and the percentage of Democratic members of the U.S. Senate in each session of Congress. Time-series data are often denoted with an index (subscript) t, e.g. "Y_t." Combining ("pooling") multiple time series from different units of observation yields *time-series cross-sectional* data.

Time-Series Model – A term used generically to denote a *regression model* for *time-series data*. Encompasses a wide range of models.

VAR – An acronym for "vector autoregression." A *regression model* for *time-series data* in which all or nearly all *dependent variables* also serve as *independent variables* in a system of equations.