Regression and Regression Analysis

ANDREW M. LINDNER
Skidmore College, USA

RYAN P. LARSON
University of Minnesota, USA

Regression analysis, a statistical technique for examining the relationship between at least one independent variable and a dependent variable, is among the most common quantitative methods used by social and natural scientists as well as professionals in politics, sports, and financial markets. While correlations are used to assess the strength and direction of an association between two variables (e.g., income and happiness), regression provides details on the substantive impact of an independent variable on a dependent variable (e.g., there is a three-point increase in happiness for every $10,000 increase in income). Multiple regression allows researchers to examine simultaneously the effect of multiple independent variables on a dependent variable, while controlling for other potentially intercorrelated independent variables.

The basic equation for linear regression (i.e., regression using a straight line of best fit) is $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. In this equation, y_i is the dependent variable and x_i is the independent variable. β_0 is the y-intercept or the value of dependent variable when the independent variable is 0. β_1 is the slope of the line of best fit between x_i and y_i and is known as the coefficient for the independent variable. For example, if the slope for the relationship between age in years and a five-point scale of political conservatism is 0.1, then people are expected to be 0.1 more conservative for each year older they are. The error term, ε_i, reflects all unobserved influences on the dependent variable. The regression coefficient is calculated by dividing the covariation of x_i and y_i (that is, the extent to which they tend to vary similarly) by the variation within x_i. Multiple regression expands on the linear model by adding further independent variables and their associated slopes (e.g., $\beta_2 X_{2i}$, $\beta_3 X_{3i}$) to the equation.

One of the key benefits to the use of multiple regression is the comparison of "like with like," which allows researchers to better estimate the causal effect of interest by ruling out alternative hypotheses. In experimental settings, random assignment of a treatment is often used to isolate the causal impact of a treatment. The randomization ensures that both treatment and control groups are, in theory, *identical in all respects apart from the treatment*, allowing the researcher to attribute the change in Y to the treatment (X). Multiple regression mimics this process by estimating independent effects "controlling for" the effects of the other variables. Regression does this by comparing the change in Y corresponding to a change an X for units that have *values identical on the other variables* included in the model. By comparing "apples to apples," the coefficient is the effect of X on Y *that is not attributable to other variables in the model*. However, unlike experimental random assignment, regression is susceptible to omitted-variables bias when important variables are left out of the model, which limits one's ability to rule out all alternative explanations to the relationship and correctly estimate coefficients. Therefore, regression estimates alone should never be mistaken for proof of a causal relationship. It is most accurate to use language like "x is significantly associated with increases in y." However, it is common practice within sociology to use verbs like "shape," "contribute to," and "affect" to describe regression results.

Regression results are often combined with tests of statistical significance to evaluate various hypotheses. Such tests tell the researcher whether the slope coefficient observed in a given sample is likely to be a real relationship between the two variables in the larger population or the result of chance sampling error. The results are usually reported as p-values or the probability of a coefficient from a sample being observed simply by chance if there was no true relationship in population. Small p-values (<.05, <.01,

The Blackwell Encyclopedia of Sociology. Edited by George Ritzer and Chris Rojek.
© 2018 John Wiley & Sons, Ltd. Published 2018 by John Wiley & Sons, Ltd.
DOI: 10.1002/9781405165518.wbeosr041

Ritzer wbeosr041.tex V1 - 01/30/2018 5:21 P.M. Page 2

and < .001 are the conventional standard for statistical significance) mean very low probabilities of finding such a result simply by chance.

OLS regression, the most common form of linear regression, produces accurate estimates of the magnitude of the relationship between the independent variables and the dependent variable only if several assumptions are met. First, the association between variables must be linear. A linear line of best fit drawn through a U-shaped scatter plot (e.g., age and number of doctor visits per year) will be flat and the slope will approach zero, failing to describe the nonlinear relationship that does exist. Second, every observation or case must be independent from other cases (e.g., children with the same teacher are not independent because they have more similar testing outcomes). Nonindependent cases can lead to estimates that over- or understate the strength of the relationship between the independent variables and the dependent variable. Third, OLS assumes that the dependent variable is continuous, which enables appropriate model fit and significance testing. Finally, the data must be homoscedastic, meaning that the data points are equally distant from the fitted line at all values of the independent variable. Using OLS with heteroscedastic data can produce estimates that describe some of the cases much better than others.

There are many extensions to linear regression that allow for analysis of more complex forms of data, as well as when assumptions of the linear model do not hold. Many dependent variables of interest are not continuous, which can be properly modeled through the use of generalized linear models (GLMs). Logistic regression can be used to model binary outcomes (e.g., yes/no on an attitude measure). Some variables have more than two outcomes, which can be analyzed using multinomial or ordinal logistic regression. Poisson or negative binomial regression is used to model count outcomes (e.g., number of crimes in a county). Finally, event history models can handle data in which the time until an event (e.g., death) is of primary concern.

Some data may have a hierarchical structure (e.g., children nested within classrooms), in which the data points are not independent. Multilevel modeling, such as random effects or mixed models, allow for nonindependence of the cases. One type of nested structure is repeated measurements over time, such as panel data (e.g., states by year). One way to address panel observations is fixed effects models, which can help with omitted-variables bias by controlling for persistent features of the units.

SEE ALSO: ANOVA (Analysis of Variance); Correlation; General Linear Model; Statistics; Variables

Further Readings

Fox, J. (2016) *Applied Regression Analysis and Generalized Linear Models*, 3rd edn, SAGE, Los Angeles, CA.

Fox, J. and Weisberg, S. (2011) *An R Companion to Applied Regression*, 2nd edn, SAGE, Los Angeles, CA.

Weisberg, S. (2005) *Applied Linear Regression*, Wiley, Hoboken, NJ.