The SAGE Dictionary of Social Research Methods

REGRESSION ANALYSIS

Definition

A body of statistical techniques in which the form of the relationship between a dependent variable and one or more independent variables is established so that knowledge of the values of the independent variables enables prediction of the value of the dependent variable or likelihood of the occurrence of an event if the dependent variable is categorical.

Distinctive Features

Regression analysis is a method by which quantitative social science seeks to establish how things are caused. The objectives are both scientific description and prediction. If we know the form of the relationship between things we have measured and know to be causal to something else, then we can predict the value of the caused thing. For example, using logistic regression, knowledge of the blood levels of various hormones and of the results of an X-ray in a patient with prostate cancer can be used to predict if the cancer has spread to the patient's lymph nodes.

The relationship between independent and dependent variables is expressed through a regression equation. In the simplest case of the linear relationship between two continuous variables this can be written as:

$$Y = a + bX$$

Here Y is the value of the dependent variable, X of the independent variable, a the value Y has when X is zero, and b the amount Y changes when X changes by one unit. We can extend this simple bivariate relationship to become multivariate by adding in more independent variables thus:

$$Y = a + b^1X^1 + b^2X^2 + b^3X^3 \ldots + b^nX^n$$

There is usually a difference between the actual value of Y and the value predicted by the regression equation – the residual. In simple linear regression – linear because

changes in *Y* are proportionate to changes in *X* – the line fitted minimizes the sum of the square of these residuals – squares because squaring eliminates negative numbers and some residuals will be positive and some negative. The simplest way to see the relationship between variables in the two or three variable case is to generate a scatter plot in two or three dimensions. Standard statistical packages, such as SPSS, fit a line to the plot for the two variable case and the degree of spread of the real values around the line can be inspected. The correlation coefficient is a measure of the degree to which the real values of *Y* correspond to those predicted by the regression equation.

Logistic regression is a robust technique in which the dependent variable can be categorical or ordinal. The independent variables can be at any level of measurement and at different levels of measurement. Logistic regression can handle nonlinear relationships and gives us an indication of the likelihood of the dependent variable having a particular value.

Evaluation

Traditional regression analysis requires that a series of demanding conditions be met by the data, especially when the data comes from a sample. However, the biggest drawback is that traditional regression procedures assume that the relationships among variables are linear. Linear relationships are not always, indeed perhaps not often, present in the social world. The detection of interaction, of the relationships among variables depending on the specific values of those variables, and changing as those values change, is an indication of nonlinearity. Logistic regression makes fewer assumptions and can cope with nonlinear relationships because it is essentially a predictive tool rather than an analytical method of establishing causal models.

Associated Concepts:
- causal model
- causality
- correlation
- discriminant function analysis
- econometrics
- forecasting

- general linear model
- log-linear analysis
- multivariate analysis
- prediction

SAGE research methods online