Correlation and Regression

Correlation and linear regression are the most commonly used techniques for investigating the relationship between two quantitative variables.

The goal of a correlation analysis is to see whether two measurement variables co vary, and to quantify the strength of the relationship between the variables, whereas regression expresses the relationship in the form of an equation.

For example, in students taking a Maths and English test, we could use correlation to determine whether students who are good at Maths tend to be good at English as well, and regression to determine whether the marks in English can be predicted for given marks in Maths.

What a Scatter Diagram Tells Us

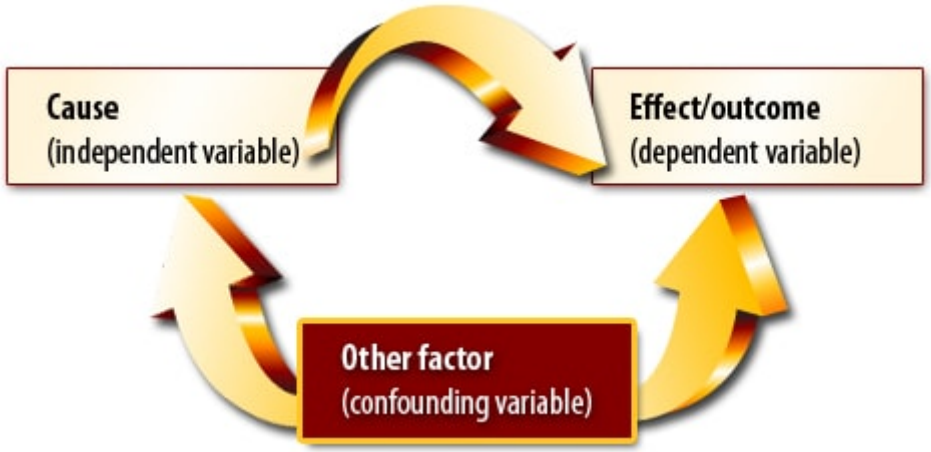
The starting point is to draw a scatter of points on a graph, with one variable on the X-axis and the other variable on the Y-axis, to get a feel of the relationship (if any) between the variables as suggested by the data. The closer the points are to a straight line, the stronger the linear relationship [1] between two variables.

Why Use Correlation?

We can use the correlation coefficient, such as the Pearson Product Moment Correlation Coefficient [2], to test if there is a linear relationship between the variables. To quantify the strength of the relationship, we can calculate the correlation coefficient (r). Its numerical value ranges from +1.0 to -1.0. r > 0 indicates positive linear relationship, r < 0 indicates negative linear relationship while r = 0 indicates no linear relationship.

A Caveat

It must, however, be considered that there may be a third variable [3] related to both of the variables being investigated, which is responsible for the apparent correlation. Correlation does not imply causation [4]. Also, a nonlinear relationship [5] may exist between two variables that would be inadequately described, or possibly even undetected, by the correlation coefficient.



Why Use Regression

In regression analysis [6], the problem of interest is the nature of the relationship itself between the dependent variable [7] (response) and the (explanatory) independent variable [8].

The analysis consists of choosing and fitting an appropriate model, done by the method of least squares, with a view to exploiting the relationship between the variables [9] to help estimate the expected response for a given value of the independent variable. For example, if we are interested in the effect of age on height, then by fitting a regression line, we can predict [10] the height for a given age.

Assumptions

Some underlying assumptions governing the uses of correlation and regression [11] are as follows.

The observations are assumed to be independent. For correlation [12], both variables should be random variables, but for regression [6] only the dependent variable Y must be random. In carrying out hypothesis tests [13], the response variable should follow Normal distribution [14] and the variability of Y should be the same for each value of the predictor variable. A scatter diagram of the data provides an initial check of the assumptions for regression.

Uses of Correlation and Regression

There are three main uses for correlation and regression.

- One is to test hypotheses [13] about cause-and-effect [15] relationships. In this case, the experimenter determines the values of the X-variable and sees whether variation in X causes variation in Y. For example, giving people different amounts of a drug and measuring their blood pressure.
- The second main use for correlation and regression is to see whether two variables are associated, without necessarily inferring a cause-and-effect [15] relationship. In this case, neither variable is determined by the experimenter; both are naturally variable. If an association is found, the inference is that variation in X may cause variation in Y, or variation in Y may cause variation in X, or variation in some other factor may affect both X and Y.
- The third common use of linear regression is estimating the value of one variable corresponding to a particular value of the other variable.

Links
[1] https://explorable.com/linear-relationship
[2] https://explorable.com/pearson-product-moment-correlation
[3] https://explorable.com/confounding-variables
[4] https://explorable.com/correlation-and-causation
[5] https://explorable.com/non-linear-relationship
[6] https://explorable.com/linear-regression-analysis
[7] https://explorable.com/dependent-variable
[8] https://explorable.com/independent-variable
[9] https://explorable.com/research-variables
[10] https://explorable.com/prediction-in-research
[11] http://www.biology.ed.ac.uk/archive/jdeacon/statistics/tress11
[12] https://explorable.com/statistical-correlation
[13] https://explorable.com/hypothesis-testing
[14] https://explorable.com/normal-probability-distribution
[15] https://explorable.com/cause-and-effect