Theme 7. Correlation and regression.

Correlation analysis is concerned with measuring the degree of association between two variables, x and y. Initially, we assume that both x and y are numerical, e.g. height and weight. Suppose we have a pair of values (*x, y*) measured on each of the n individuals in our sample. We can mark the point corresponding to each individual's pair of values on a two-dimensional scatter diagram. Conventionally, we put the *x* variable on the horizontal axis, and the *y* variable on the vertical axis in this diagram. Plotting the points for all n individuals, we obtain a scatter of points that may suggest a relationship between the two variables.

Pearson correlation coefficient

We say that we have a linear relationship between *x* and *y* if a straight line drawn through the midst of the points provides the most appropriate approximation to the observed relationship. We measure how close the observations are to the straight line that best describes their linear relationship by calculating the ***Pearson product moment correlation coefficient***, usually simply called the correlation coefficient. Its true value in the population ρ is estimated in the sample by *r*, where

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

which is usually obtained from computer output.

Properties

• *r* ranges from -1 to +1.

• Its sign indicates whether one variable increases as the other variable increases (positive *r*) or whether one variable decreases as the other increases (negative *r*) (see Fig. 7.1).

• Its magnitude indicates how close the points are to the straight line. In particular if *r* =+ 1 or -1, then there is perfect correlation with all the points lying on the line (this is most unusual, in practice); if *r* = 0, then there is no linear correlation (although there may be a non-linear relationship). The closer *r* is to the extremes, the greater the degree of linear association (Fig. 7.1).

• It is dimensionless, i.e. it has no units of measurement.

• Its value is valid only within the range of values of *x* and *y* in the sample. Its absolute value (ignoring sign) tends to increase as the range of values of *x* and/or *y* increases and therefore you cannot infer that it will have the same value when considering values of *x* or *y* that are more extreme than the sample values.

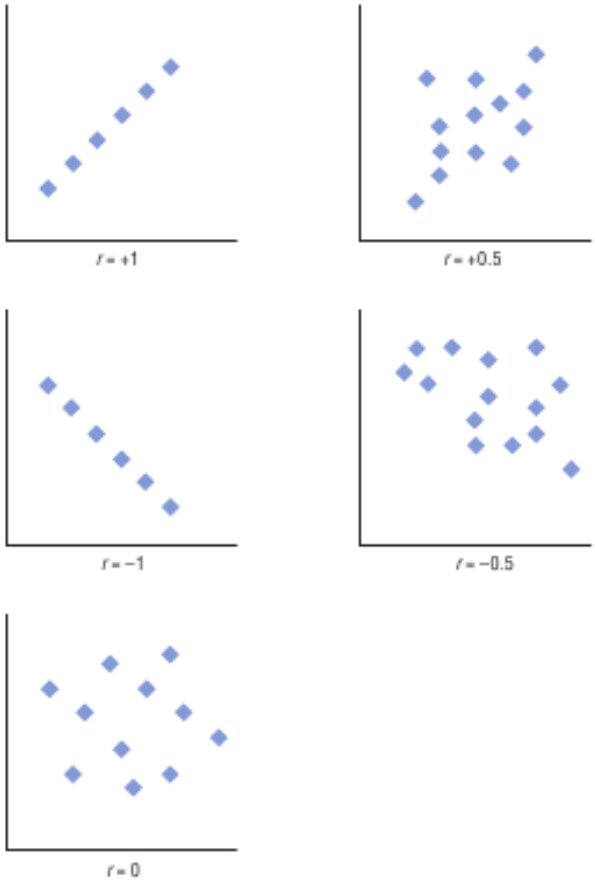

Fig. 7.1. Values of *r* in different situations

• *x* and *y* can be interchanged without affecting the value of *r*.

• A correlation between *x* and *y* does not necessarily imply a 'cause and effect' relationship.

• r^2 represents the proportion of the variability of y that can be attributed to its linear relationship with x.

Hypothesis test for the Pearson correlation coefficient

We want to know if there is any linear correlation between two numerical variables. Our sample consists of n independent pairs of values of x and y. We assume that at least one of the two variables is normally distributed.

1. Define the null and alternative hypotheses under study

H_0: $\rho = 0$

H_1: $\rho \neq 0$

2. Collect relevant data from a sample of individuals

3. Calculate the value of the test statistic specific to H_0

Calculate r.

• If $n \leq 150$, r is the test statistic

• If $n > 150$, calculate

$$T = \sqrt{\frac{(n - 2)}{(1 - r^2)}}$$

which follows a t-distribution with n - 2 degrees of freedom.

4. Compare the value of the test statistic to values from a known probability distribution

5. Interpret the P-value and results

Note that, if the sample size is large, H_0 may be rejected even if r is quite close to zero. Alternatively, even if r is large, H_0 may not be rejected if the sample size is small. For this reason, it is particularly helpful to calculate r^2, the proportion of the total variance of one variable explained by its linear relationship with the other.

Spearman's rank correlation coefficient

We calculate Spearman's rank correlation coefficient, the non-parametric equivalent to Pearson's correlation coefficient, if one or more of the following points is true:

• at least one of the variables, x or y, is measured on an ordinal scale;

• neither x nor y is Normally distributed;

• the sample size is small;

• we require a measure of the association between two variables when their relationship is non-linear.

Calculation

To estimate the population value of Spearman's rank correlation coefficient, ρ_s, by its sample value, r_s:

1. Arrange the values of x in increasing order, starting with the smallest value, and assign successive ranks (the numbers 1, 2, 3,..., n) to them. Tied values receive the mean of the ranks these values would have received had there been no ties.

2. Assign ranks to the values of y in a similar manner.

3. r_s is the Pearson correlation coefficient between the ranks of x and y.

Properties and hypothesis tests

These are the same as for Pearson's correlation coefficient, replacing r by r_s, except that:

• r_s provides a measure of association (not necessarily linear) between x and y;

• we do not calculate r_s^2.It does not represent the proportion of the total variation in one variable that can be attributed to its linear relationship with the other.

Linear regression analysis

To investigate the relationship between two numerical variables x and y we measure the values of x and y on each of the n individuals in our sample. We plot the points on a ***scatter diagram*** and say that we have a ***linear relationship*** if the data approximate a straight line. If we believe y is dependent on x, we can determine the ***linear regression line*** (the regression of y on x) that best describes the straight line relationship between the two variables.

The mathematical equation which estimates the simple linear regression line is:

$Y = a + bx$

• x is called the ***independent***, ***predictor*** or ***explanatory variable***;

• for a given value of x, Y is the value of y (called the ***dependent***, ***outcome*** or ***response*** variable), which lies on the estimated line. It is an estimate of the value we expect for y (i.e. its mean) if we know the value of x, and is called the ***fitted*** value of y;

• a is the ***intercept*** of the estimated line; it is the value of Y when x = 0 (Fig. 12);

• b is the slope or gradient of the estimated line; it represents the amount by which Y increases on average if we increase x by one unit (Fig. 7.2).

a and ***b*** are called the ***regression coefficients*** of the estimated line, although this term is often reserved only for ***b***. Simple linear regression can be extended to include more than one explanatory variable; in this case, it is known as ***multiple linear regression***.

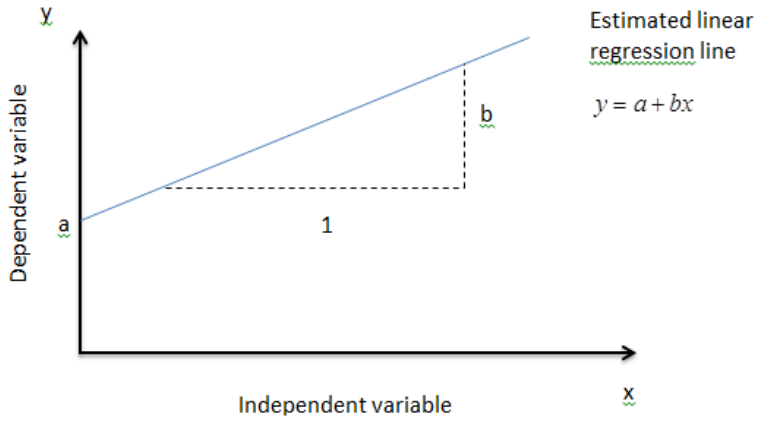

Fig. 7.2. Regression line.

To draw the line $Y = a + bx$ on the scatter diagram, we choose three values of x (i.e. x_1, x_2 and x_3) along its range. We substitute x_1 in the equation to obtain the corresponding value of Y, namely $Y_1 = a + bx_1$; Y_1 is our estimated fitted value for x_1 which corresponds to the observed value, y_1. We repeat the procedure for x_2 and x_3 to obtain the corresponding values of Y_2 and Y_3. We plot these points on the scatter diagram and join them to produce a straight line.

We can use the regression line for ***predicting*** values of y for values of x within the observed range. We predict the mean value of y for individuals who have a certain value of x by substituting that value of x into the equation of the line. So, if $x = x_0$, we predict y as $Y_0 = a + bx_0$. We use this estimated predicted value, and its standard error, to evaluate the confidence interval for the true mean value of y in the population. Repeating this procedure for various values of x allows us to construct confidence limits for the line. This is a band or region that contains the true line with, say, 95% certainty. Similarly, we can calculate a wider region within which we expect most (usually 95%) of the observations to lie.

Control questions

1. What is the purpose of correlation analysis?
2. When is Pearson correlation coefficient used?
3. When is Spearman's rank correlation coefficient used?
4. Explain properties and hypothesis tests of correlation coefficients.
5. What is linear regression analysis?

References

1. Medical statistics at a glance / Aviva Petrie, Caroline Sabin.—2nd ed., 2005. – 157 p.
2. Using and understanding medical statis tics / David E. Matthews Vernon T. Farewell. – 4th, completely rev. and enl. ed., 2007. – 322 p.
3. Handbook of statistics. Epidemiology and Medical Statistics / C.R. Rao, J.P. Miller, D.C. Rao, 2008. – 852 p.