

Measurement of dependence, correlation and regression

Statistical methods for determining the type and strength of dependence between two quantities. In medicine, this method is most often applied when investigating the relationship between a disease and its possible causes.

The type and strength of the dependence for the random selection of the range "n" can be roughly assessed from the dot chart, in which each pair of data (x, y) is graphically represented by one point.

Type of dependence determines the shape of the curve, which we can interpolate the points - **linear, exponential, logarithmic** etc.

Regression

When choosing a regression function, we follow the **method of least squares**' (see linear regression), i.e. we look for the function that is closest to the values of the data we entered and then analyze the statistical properties of the line selected by this method.

File:Regrese.png

A model example of choosing a regression function by the method of least squares – the values (red dots) are best matched by a linear function (blue line)

Linear Regression

It can be used if the dependence of the quantity y on x is linear.

In practice: fitting the points in the graph with a regression **line** $y = a + bx$ so that the sum of the squares of the deviations of individual points from the line is minimal (method of least squares) .

a, b = regression coefficients.

- a – shift on the y axis (the place where the regression line crosses the vertical axis),
- b – slope of the regression line.

note squared = squared

Quadratic regression

A special case of linear regression, when we fit the data set with a quadratic function (*parabola*) ' $y = ax^2 + bx + c$ '.

a, b, c are regression coefficients that can be estimated in practice again by the method of least squares.

Logarithmic regression

A special case of linear regression, when we fit the data set with the **logarithmic** function $y = a + b \cdot \ln(x)$ '.

Strength of statistical dependence = correlation

We express it by various appropriate measures, which include, for example, correlation coefficients. The requirement that the *absolute value* of the measure of statistical dependence lie *within a closed interval from 0 to 1*. 'However, statistical dependence does not necessarily mean causality!'

Pearson's correlation coefficient ρ is used to measure the strength of dependence. According to general rules, it takes values from -1 to $+1$. If the type of dependence is linear, then:

- **zero value** ' ρ - expresses the linear independence of quantities (Correlation does not say anything about functional dependence, but only about the linear one! Only in the case of a normal distribution, if the quantities are linearly independent (zero correlation), they are at the same time functionally independent.),
- $\rho > 0$ – as the values of one quantity increase, the values of the other also increase (or both decrease),
- $\rho < 0$ – as the values of one quantity increase, the values of the other decrease and vice versa,
- The **extreme values $+1$ and -1** indicate a **functional linear dependence** of both quantities.

A high degree of dependence (correlation) often reflects causation, but this may not always be the case.

Sometimes we do not clearly determine which quantity is independent and which is dependent. A linear regression of X on Y does not give the same regression line as a regression of Y on X. The squared correlation coefficient is called the *coefficient of determination* and its value measures the magnitude of the linear relationship between X and Y regardless of which variable is dependent and which independent – this coefficient obtained from both regressions is the same.

From the linear regression graph, it can be inferred that the value of ρ – the smaller the angle formed by the two regression lines (expressing the dependence of x on y and y on x), the greater the absolute value of ρ .

Correlation study

 For more information see *Descriptive study#Correlational study*.

To assess the influence of third factors, the calculation of **partial correlation coefficients** is used, which are determined for individual pairs of characters whose association is being investigated (e.g. in a set where age, blood pressure and blood cholesterol level are recorded correlation coefficients for relationships: r_1 – for the relationship age and pressure, r_2 – for the relationship age and chol., r_3 – for the relationship chol. and pressure). Thus, a partial coefficient can be calculated, for example, for the relationship between cholesterol level and BP when age is eliminated as a third factor, and after testing statistical significance, the association between these characteristics can be confirmed or not.

Links

Related Articles

- Four-field and pivot table

External links

- Lineární regrese
- Metoda nejmenších čtverců

used literature

- MACHEK, Josef – LIKEŠ, Jiří. *Matematická statistika*. 2. edition. SNTL, 1988. ISBN 1. Jiří Likeš, Josef Machek, Matematická statistika, SNTL Praha 1988, s. 165-169.
- ZVÁROVÁ, Jana. *Biomedicínská statistika I. : Základy statistiky pro biomedicínské obory* [online] . dotisk 1 edition. Karolinum, 1998. 218 pp. Available from <<http://www.euromise.cz/education/textbooks.html>>. ISBN 80-7184-786-0.
- BENCKO, Vladimír. *Epidemiologie, výukové texty pro studenty 1. LFUK, Praha*. 2. edition. Univerzita Karlova v Praze – Nakladatelství Karolinum, 2002. 168 pp. pp. 78-80. ISBN 80-246-0383-7.