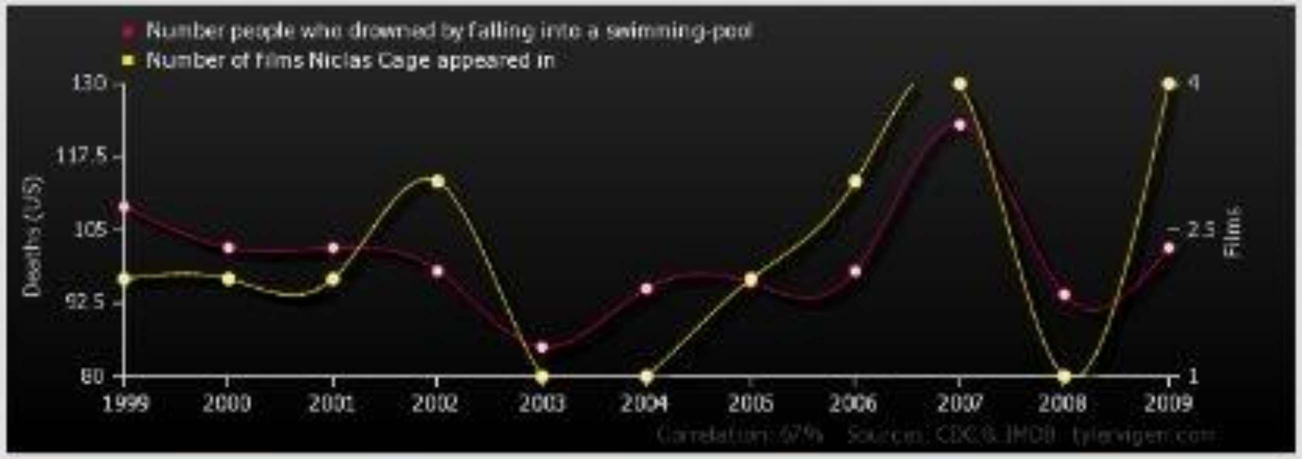
Correlation analysis

Correlation

Number people who drowned by falling into a swimming-pool

correlates with

Number of films Nicolas Cage appeared in



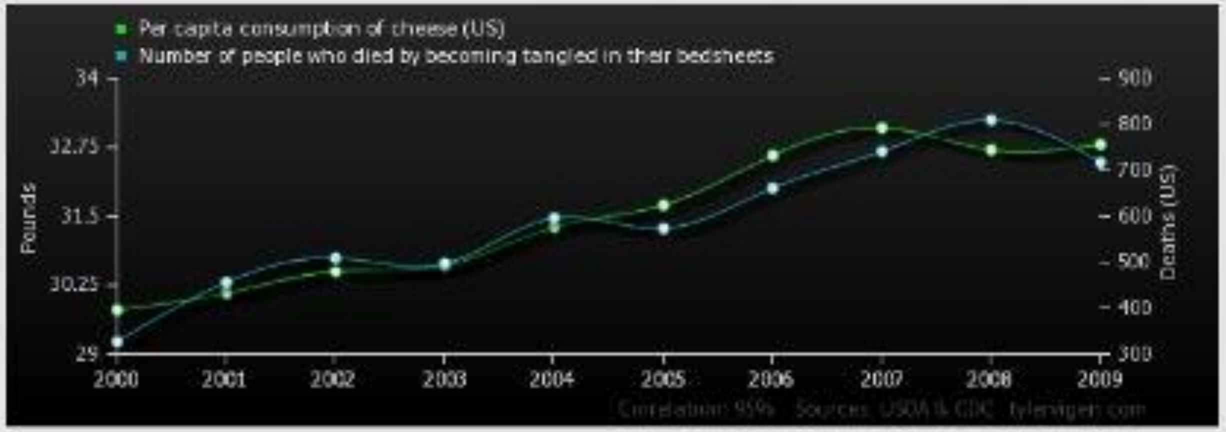
	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Number people who drowned by falling into a swimming-pool Deaths (US) (CDC)	109	102	102	98	85	95	96	98	123	94	102
Number of films Nicolas Cage appeared in Films (IMDB)	2	2	2	3	1	1	2	3	4	1	4

Correlation: 0.666004

http://dangerousminds.net/comments/spurious_correlations_between_nicolas_cage_movies_and_swimming_pool

Correlation

Per capita consumption of cheese (US)
correlates with
Number of people who died by becoming tangled in their bedsheets



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Per capita consumption of cheese (US) *Pounds (USDA)*	29.8	30.1	30.5	30.6	31.3	31.7	32.6	33.1	32.7	32.8
Number of people who died by becoming tangled in their bedsheets *Deaths (US) (CDC)*	327	456	509	497	596	573	661	741	809	717

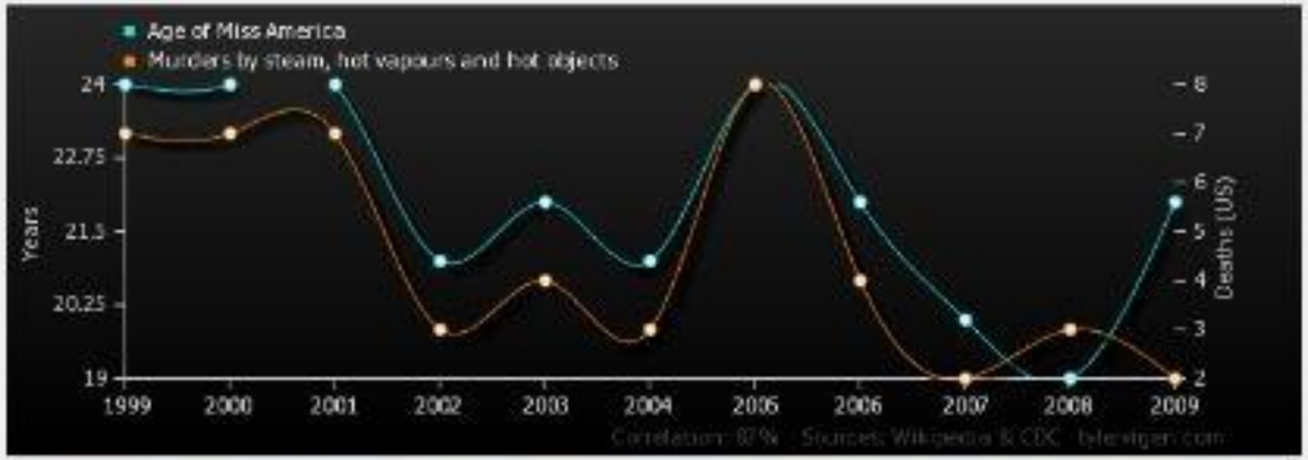
Correlation: 0.947091

Correlation

Age of Miss America

correlates with

Murders by steam, hot vapours and hot objects



	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Age of Miss America Years (Wikipedia)	24	24	24	21	22	21	24	22	20	19	22
Murders by steam, hot vapours and hot objects Deaths (US) (CDC)	7	7	7	3	4	3	8	4	2	3	2

Correlation: 0.870127

Correlation analysis - Pearson's correlation coefficient

Objective: to assess the relationship between two quantitative variables

It only evaluates the linear relationship.

$$r = \frac{cov(X,Y)}{s_x \cdot s_y}$$

where, the value of the covariance (cov) on the basis of the sample is calculated according to the following formula:

$$cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

while s_x and s_y are standard deviations for the variables: X and Y

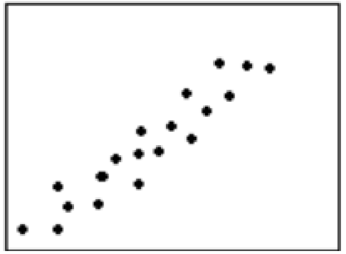
Correlation analysis - Pearson's correlation coefficient

The linear correlation coefficient always assumes values in the range [-1.1].
The greater the absolute value of the coefficient, the stronger the linear relationship between the variables.

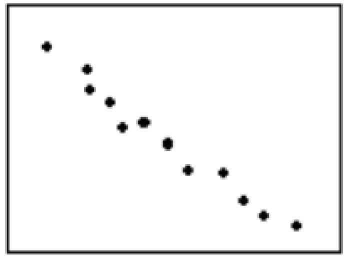
$r_{xy} = 0$ means no correlation,

$r_{xy} = 1$ means a strong positive correlation, if one variable (X) grows, the other variable (Y) also grows,
$r_{xy} = -1$ -1 means a negative correlation (if the variable X increases, then Y decreases, and vice versa).

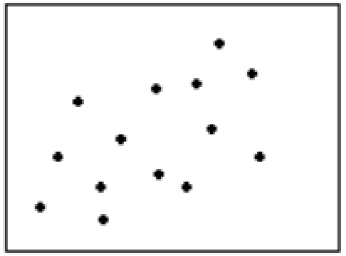
Correlation analysis - Pearson's correlation coefficient



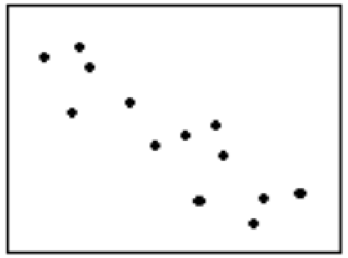
Strong positive $(r = 0{,}8)$



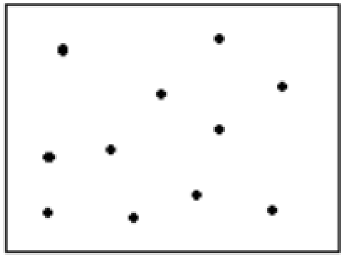
Strong negative $(r = -0{,}8)$



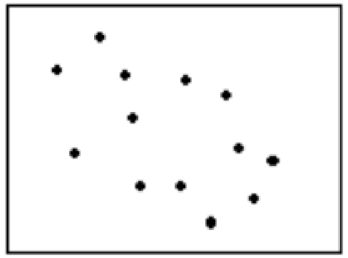
Weak positive $(r = 0{,}3)$



moderately negative $(r = -0{,}5)$

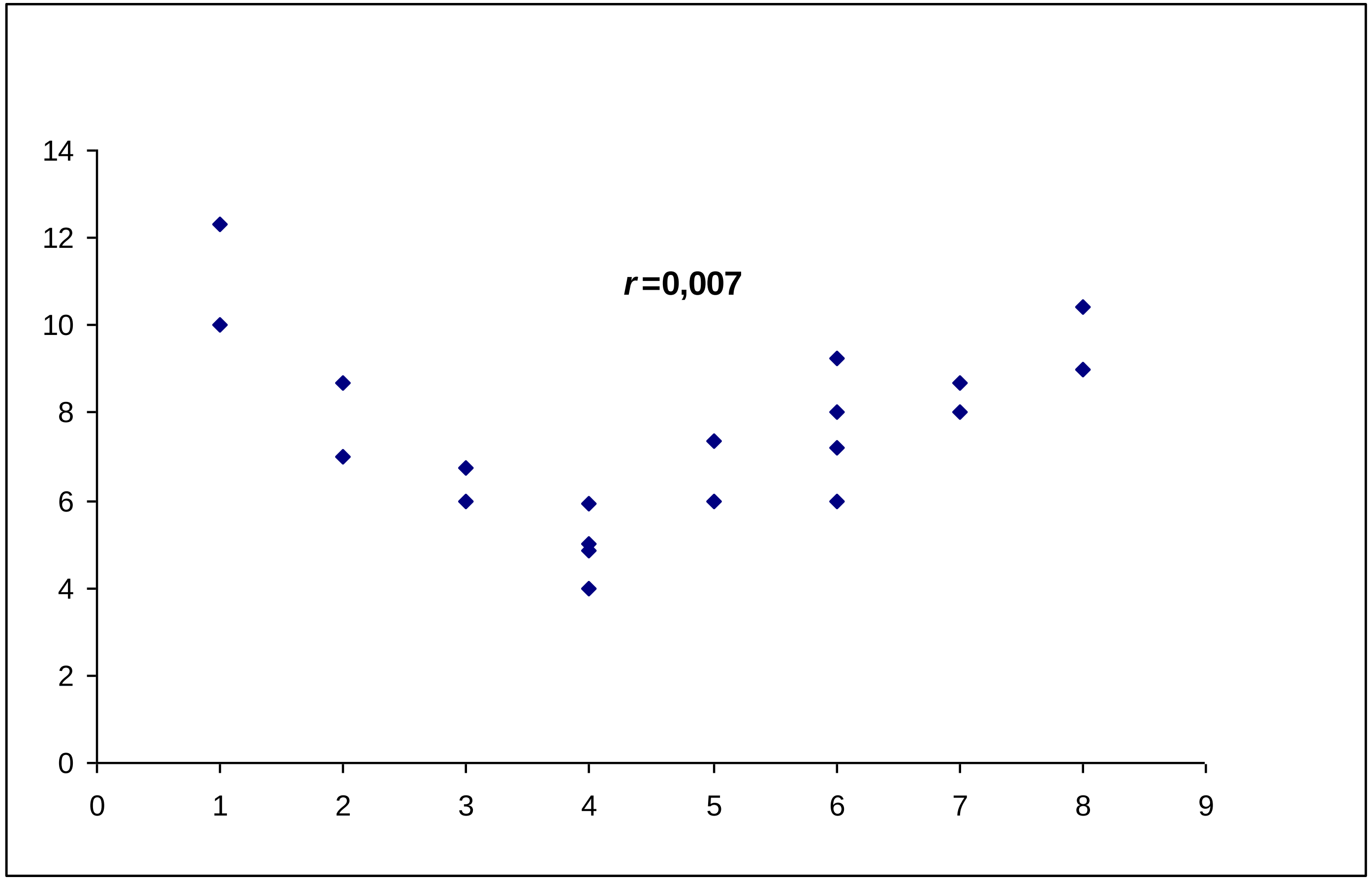


No correlation $(r = 0{,}0)$



Weak negative $(r = -0{,}3)$

Correlation analysis - Pearson's correlation coefficient vs nonlinear correlation



Correlation analysis - Pearson's correlation coefficient

Student's t-test, when variables come from the normal distribution

$$t_{emp} = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

$t_{\alpha,n-2}$ - is the critical value from the t-Student distribution

If $|t_{emp}| > t_{\alpha,n-2}$ or $p<\alpha$ then H_0 is rejected.

Correlation analysis - Pearson's correlation coefficient

Correlation significance testing

Testing is only valid when both variables are normally distributed

The null hypothesis: H_0:ϱ=0

ρ- value of the correlation coefficient for the entire population

if $|r_{emp}|>r_{\alpha,2,n-2}$ then H_0 should be rejected.

$r_{\alpha,2,n-2}$ – is the critical value of the Pearson simple correlation coefficient

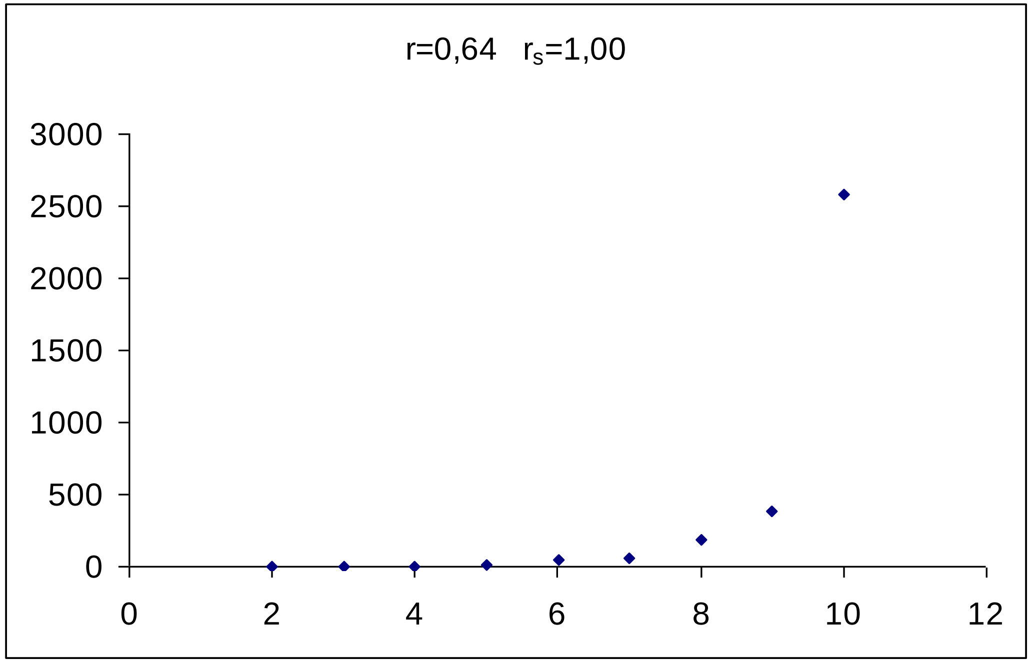
As in the case of other hypotheses in statistical programs (inference about the significance of the interdependence of two variables is based on the value of p ($p < \alpha$ means a significant correlation)

It should also be remembered that Pearson's linear correlation coefficient describes only linear relationships well. If the relationship exists but is non-linear (e.g. points are located on a parabola), the value of the correlation coefficient may be close to 0.

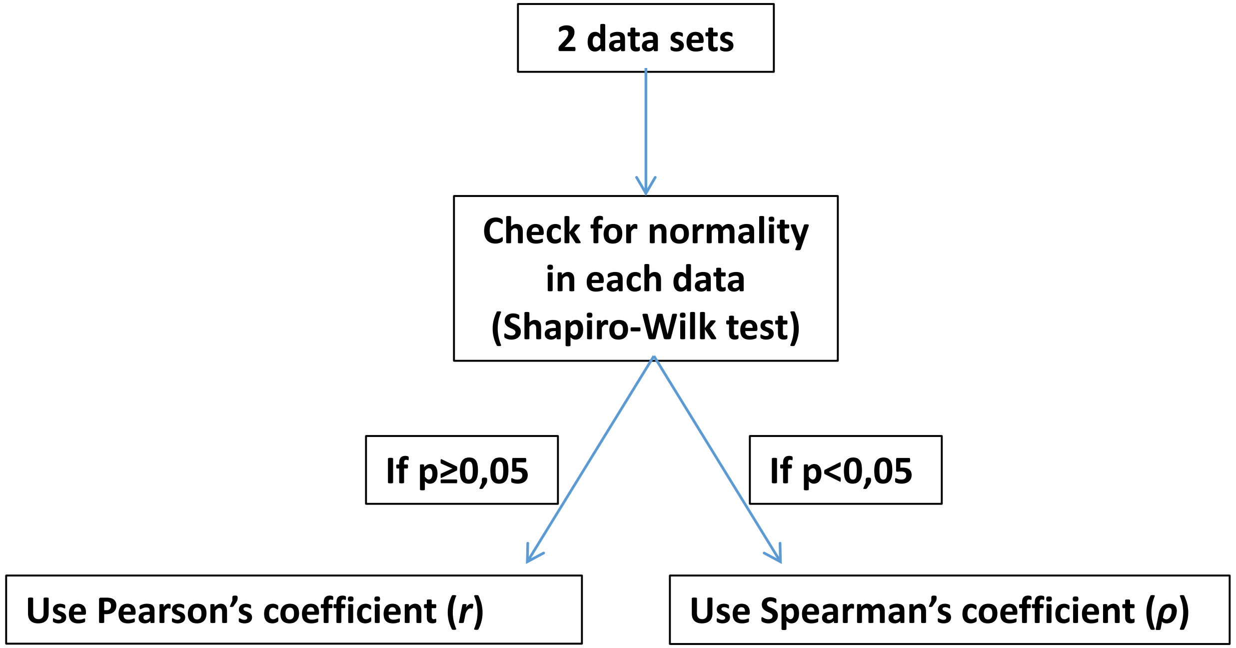
Correlation analysis - Spearman's rank correlation coefficient

Spearman's rank correlation coefficient (r_s) is used to assess the relationship between two variables. Contrary to the Pearson correlation coefficient, nonlinear dependencies can be assessed using the Spearman correlation coefficient. When testing, normality of the distribution of variables is not required, so it is possible to use this correlation coefficient when we cannot use the Persona correlation coefficient.

The values of the Spearman's rank correlation coefficient are in the range [-1, 1] and their interpretation is similar to the Pearson's correlation coefficient, i.e. the closer r_s value is to 1, the stronger the relationship is, positive, the closer it is to -1, the stronger, negative relationship , and if the r_s value is close to 0 it means no dependency or very weak dependence.



correlation



Based on:
Daniel Granato, Verônica Maria de Araújo Calado, Basil Jarvis „Observations on the use of statistical methods in Food Science and Technology" Food Research International 55 (2014) 137–149

Correlation analysis - an example

Assess with the use of correlation analysis whether there is a relationship between the features of X and Y.

online calculators
Pearson:
https://www.socscistatistics.com/pvalues/pearsondistribution.aspx

Spearman:
https://www.socscistatistics.com/tests/spearman/default.aspx

X	Y
0,5	10,3
0,7	12,3
1,2	15,6
1,4	16,8
1,6	17,5
1,55	17,9
1,4	18,5
1,8	18,2
1,7	18,6
1,9	16,2
2,3	15,8
2,4	15,4
2,5	14,6
2,1	17,1
2,8	9,6

Lecture 9

Simple regression analysis

Simple regression analysis

Simple regression is a statistical method in which we determine the dependence of one variable (Y) on another (X), i.e. the relationship is between only two variables.

Simple linear regression

Linear regression is a method of estimating the expected value of one variable (Y) by knowing the values of another variable (X) from a linear function. The searched variable Y is called the dependent variable, the variable X is called the independent variable.

Simple linear regression model

$Y = a + bX + e_i$

where:

b – regression coefficient

a – regression constant

e_i – random errors with distributionN$(0;\sigma_e^2)$

The regression constant (a) is therefore the estimated mean value of the Y variable when X = 0, while the regression coefficient (b) is the mean change in the value of Y when X is increased by one unit.

A negative value of the regression coefficient (b) indicates a negative relationship, and a positive value indicates a positive relationship

Simple linear regression model

The estimation (value estimation) of the coefficients of the regression equation is usually performed using the least squares method, which consists in minimizing the following sum of squares:

$$\sum_{i=1}^{n}(y_i - a - bx_i)^2$$

The estimators of the a and b coefficients are calculated from the formulas:

$$b = \frac{s_{xy}}{s_x^2} \qquad a = \overline{y} - b\overline{x}$$

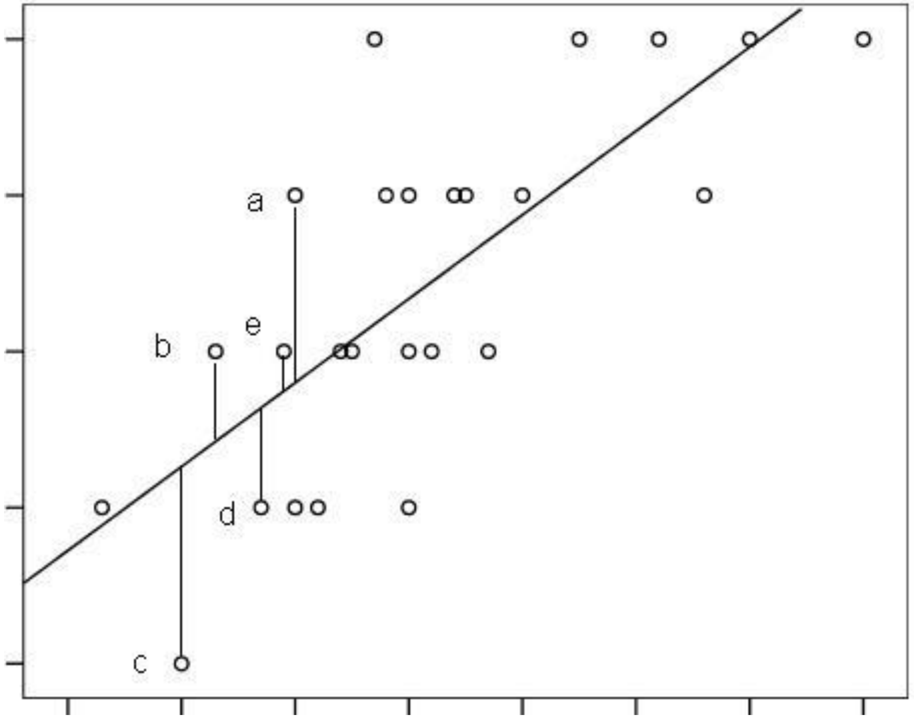
Simple linear regression model

Least squares estimators

$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$b = \bar{y} - a\bar{x}$$

Least squares method



R^2 – współczynnik determinacji

Specifies the ratio of the variability explained by the regression model to the total variation. For simple linear regression R^2=$r_{xy}{}^2$

The closer R^2 to 100% (or 1) then the depencence of Y to X is stronger, and vice versa when the value R^2 is closer to 0% (or 0) then the dependence of Y on X is weaker. The value of the coefficient of determination for linear regression is equal to the square of the Pearson correlation coefficient.

Hypothesis testing H_0: β=0 (the regression coefficient for the whole population is equal to 0) allows to assess whether there is a significant dependence of Y on X. If we reject this hypothesis, we consider that Y significantly depends on X (we reject the above hypothesis if p $<\alpha$)

Simple regression analysis example

A field experiment was carried out in which the protein content in the grain of a certain variety of winter wheat was assessed depending on the applied dose of nitrogen fertilization. Investigate the effect of nitrogen fertilization on protein content using simple linear regression.

https://www.socscistatistics.com/tests/regression/default.aspx

N (kg/ha)	protein (%)
0	11.49
10	11.55
20	11.74
30	11.88
40	11.64
50	11.62
60	11.55
70	11.64
80	12.02
90	12.08
100	12.10
110	12.27
120	12.24
130	12.26
140	12.23
150	12.10
160	12.29
170	12.44
180	12.72
190	12.45
200	12.56
210	12.54
220	12.73
230	12.82
240	12.90
250	12.86

Simple nonlinear regression

Not all relationships between two variables are linear, so sometimes it makes sense to use a non-linear regression model. Various other regression models are used for this purpose. Instead of a linear function, you can use functions such as:

square
square root
logarithmic or other.

The selection of a regression model is most often made on the basis of the value of the coefficient of determination (R2), a higher value of R2 means a better fitted regression model, and thus better describing changes in Y depending on X.
A special example of simple regression is polynomial simple regression, i.e. the use of a polynomial function in which the independent variable (X) appears in successive powers. The simplest polynomial regression model is the quadratic function (X is first and second power).

Simple nonlinear regression