Correlation analysis is used to study the possible dependence of two or more columns in a n by m data matrix. For instance, consider any two columns in a n by m data matrix such as

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}$$

where we can select column j where $1 \le j \le m$, and refer to it as X, and column k where $1 \le k \le m$, and refer to it as Y, as long as $j \ne k$. As the data matrix will consist of observations subject to random variation and experimental error the following situations are possible.

1. X and Y are completely independent and there is no relationship whatsoever between them.
2. X and Y are linearly dependent, that is, components x_i and y_i are related in that $y_i \approx \alpha x_i$ for some parameter α.
3. X and Y are monotonically dependent, that is, components x_i and y_i are related in that, roughly speaking, y_i tend to be large when x_i are large, or some similar nonlinear tendency exists.
4. X and Y are nonlinearly dependent, that is, components x_i and y_i are related in that $f(x_i, y_i) \approx 0$ for some nonlinear implicit function $f(x, y) = 0$.
5. X and Y are dependent because they are separately dependent on another column or columns in the data matrix, or else some other factor not represented in the data matrix.

Actually, given any set of n nonsingular (x_i, y_i) pairs, a correlation coefficient r can always be calculated as

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

where $-1 \le r \le 1$ and, using b_{xy} for the slope of the regression of X on Y, and b_{yx} for the slope of the regression of Y on X

$$r^2 = b_{yx} b_{xy}.$$

However, only when X is normally distributed given Y, and Y is normally distributed given X can simple statistical tests be used for significant linear correlation. The most well known facts about r are as follows.

- When X and Y are linearly related with $y_i \approx \alpha x_i$ and $\alpha > 0$ then $r \to 1$.
- When X and Y are linearly related with $y_i \approx \alpha x_i$ and $\alpha < 0$ then $r \to -1$.
- When X and Y are not linearly related then $r \to 0$.
- When the (x_i, y_i) pairs are from a bivariate normal distribution with population correlation coefficient ρ_0 equal to zero, then the statistic

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

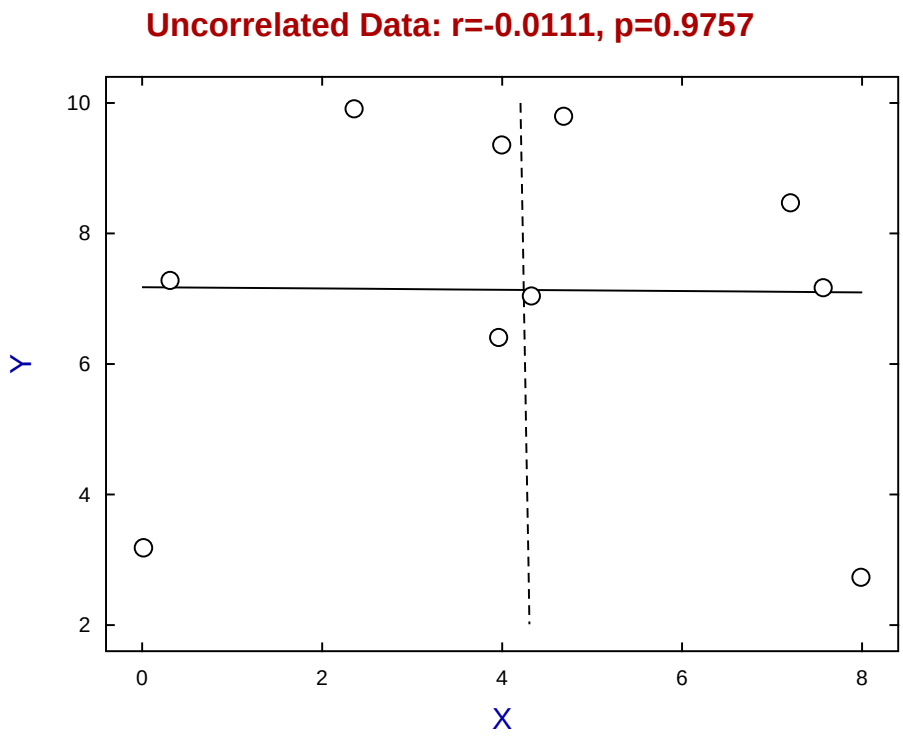
  has a Student's t-distribution with $n - 2$ degrees of freedom.

Example 1: Uncorrelated data

Consider this data set

x	y
2.3556	9.9096
0.0165	3.1851
0.3103	7.2811
3.9954	9.3582
7.9854	2.7311
4.3243	7.0423
4.6832	9.7970
7.2031	8.4710
7.5664	7.1706
3.9607	6.4083

which can be displayed as the following scattergram.



Note that, in a correlation scattergram, it is arbitrary which column of the data matrix is chosen for X, and which is chosen for Y. Hence, as it makes no sense to just show the regression line for Y as a function of X, or X as a function of Y, SimFIT allows you to plot both regression lines. If these regression lines are approximately at right angles it indicates that X and Y are not linearly correlated. Of course the visual check for perpendicularity is best when the same range and scale is used for the coordinates axes, and when a square aspect ratio is employed.

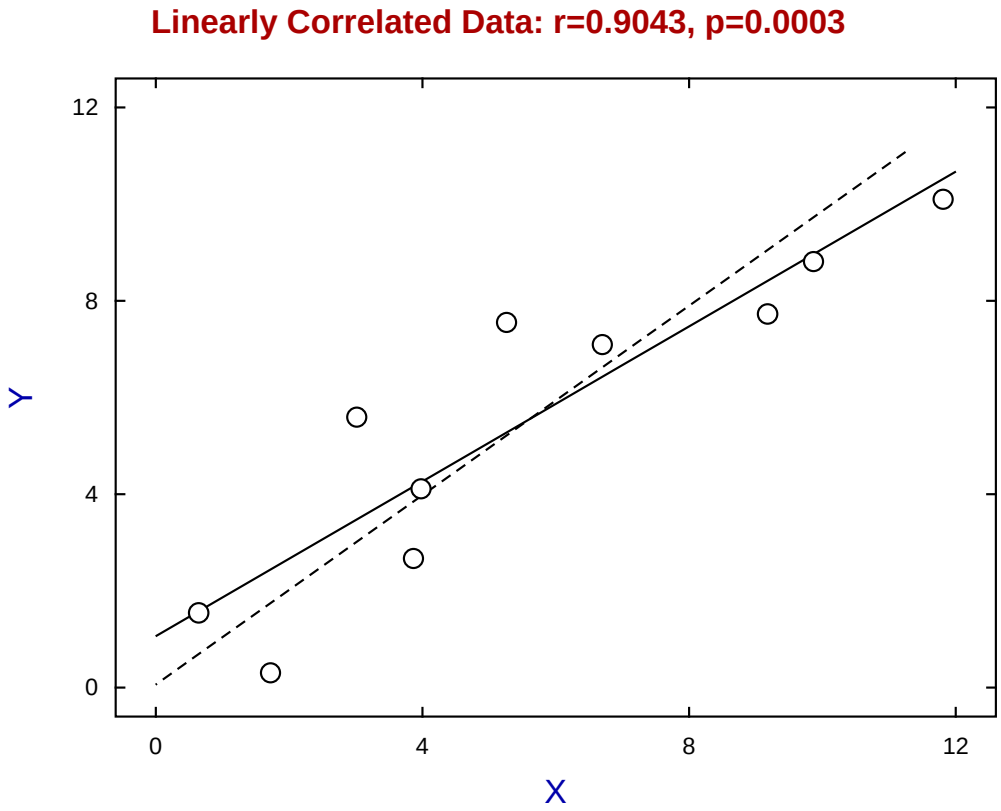
The conclusion is obvious from the scattergram, r value, p value, and almost perpendicular regression lines that these two data columns are not linearly correlated. Note that the usual type of regression line is based on the supposition that the X values are known exactly, but SimFIT also provides other techniques for plotting a best-fit single regression line when there is variation in both X and Y.

Example 2: Linearly correlated data

Consider this data set

x	y
1.7215	0.3048
0.6453	1.5455
3.8647	2.6689
3.9793	4.1100
3.0151	5.5931
5.2616	7.5528
9.1775	7.7276
6.6972	7.0932
9.8648	8.8121
11.8088	10.0993

which can be displayed as the following scattergram.



The conclusion is obvious from the scattergram, r value, p value, and almost parallel regression lines that these two data columns are linearly correlated.

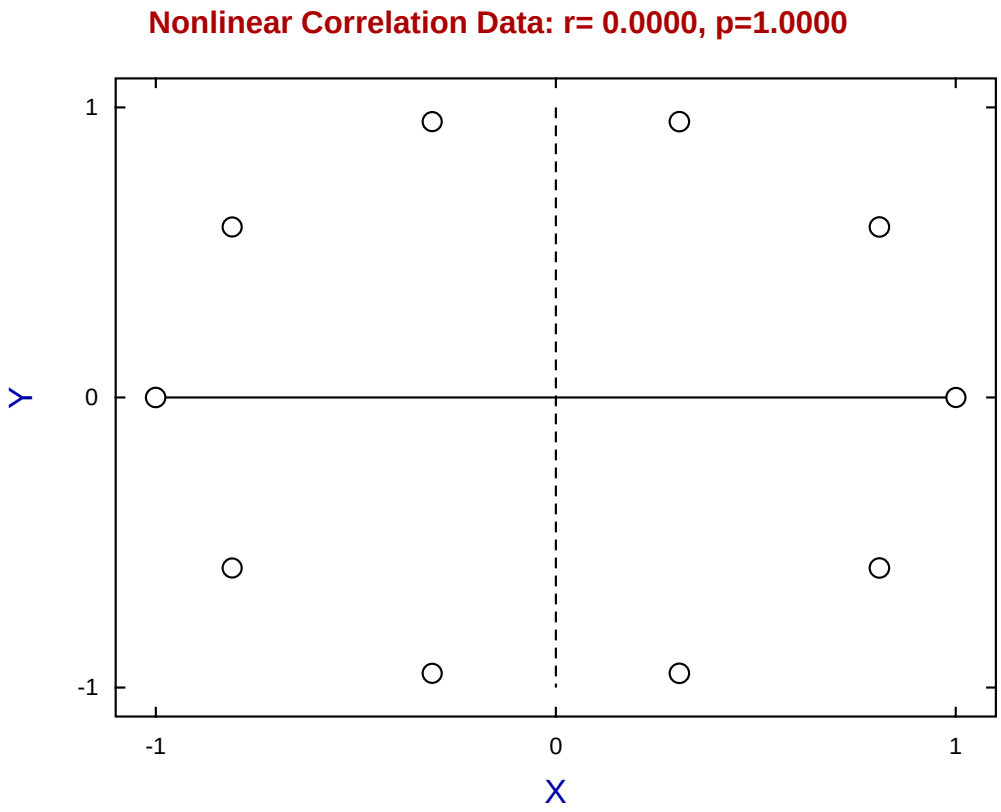
Note however, that such strong evidence for linear correlation does not imply that Y is really a linear function of X in the sense that X causes Y or vice versa. Often a strong correlation will be due to the dependence of both variables on some other factor such as time, population size, or age. For instance, one study examined that incidence of crime in several cities along with other variables such as the number of churches and reported a strong positive correlation between the incidence of crime and the number of churches. This does not, of course, mean that churches cause crime, but merely reflects the fact that large cities will tend to have more crimes but also more churches. SIMFIT provides techniques for studying these sorts of induced correlations.

Example 3: non-linearly correlated data

Consider this data set

x	y
1.0000	0.0000
0.8090	0.5878
0.3090	0.9511
-0.3090	0.9511
-0.8090	0.5878
-1.0000	0.0000
-0.8090	-0.5878
-0.3090	-0.9511
0.3090	-0.9511
0.8090	-0.5878

which can be displayed as the following scattergram.



The conclusion is obvious from the scattergram, r value, p value, and perpendicular regression lines that these two data columns are not linearly correlated.

This example emphasizes an extremely widespread misunderstanding in the application of correlation analysis. It would be harder to find a more obvious example of a data set displaying such extreme nonlinear correlation as this one. Yet the standard technique of relying on r and p values would only conclude an absence of linear correlation, and would not exclude nonlinear correlation. Scatter diagrams showing both regression lines, as in these examples, should always be inspected before making conclusions about possible correlations.