

The Correlation Coefficient

What It Does Mean and What It Doesn't Mean

The correlation coefficient is usually represented by the symbol r ; it measures the strength of the linear relationship in bivariate data.

The following data set shows that an r -value near 1 does not indicate that data exhibits a linear association.

x	3	4	5	6	7	8	9
y	27	64	125	216	343	512	729

Even though $r = 0.9649$, the residuals exhibit a dramatic pattern and the data is not linear. In fact, the cubic function $f(x) = x^3$ is a good model for this data and produces residuals that are all zero.

The value of the correlation coefficient does **not** tell you whether or not the relationship between two variables is linear. Rather, if there is a linear relationship, then the value of r indicates how strong that relationship is. The decision about whether or not there is a linear relationship should be based on residuals, not on the correlation coefficient.

It is also possible for the value of r to be zero when there is a strong (non-linear) association between x and y .

x	20	30	40	50	60
y	24	28	30	28	24

The value of r is usually defined as follows:

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right).$$

Within this summation,

- n represents the number of data points in the data set,
- (x_i, y_i) are the data points,
- \bar{x} and \bar{y} are the means of x_i and y_i , respectively, so

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \text{ and } \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

- s_x and s_y are the standard deviations of the x-values and the y-values, respectively, so $s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$ and

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

When the ordered pairs (x_i, y_i) do not represent "real data", it is possible that their relationship is perfectly linear. i.e. that y is actually a linear function of x . This would mean that each y_i is equal to $m \cdot x_i + b$; the same m and b would describe the relationship between (x_1, y_1) (x_2, y_2) (x_3, y_3) all the way up to (x_n, y_n) .

We can verify that if y_i is equal to $m \cdot x_i + b$ for all i , then the following are true:

$$\bar{y} = m \cdot \bar{x} + b,$$

$$s_y = m \cdot s_x \text{ if } m > 0$$

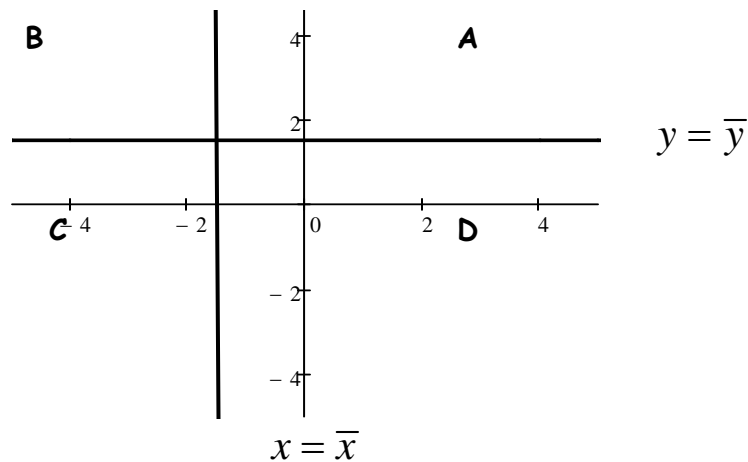
$$s_y = -m \cdot s_x \text{ if } m < 0$$

We can use algebra to show that $r = 1$ when $m > 0$ and $r = -1$ when $m < 0$.

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) = \dots = \pm 1$$

If x and y have a perfect linear relationship, then $r = \pm 1$. If x and y have a strong (but not perfect) linear relationship, then r is near 1 or near -1. But the converse is NOT true: an r value near 1 or -1 does not indicate that x and y have a strong linear relationship.

It is useful to understand how a scatter plot of ordered pairs (x_i, y_i) can reveal the sign of r . Think of the coordinate plane as being divided into quadrants A, B, C and D determined by the vertical line through \bar{x} and the horizontal line through \bar{y} .



For any point (x_i, y_i) in quadrant A, $x_i > \bar{x}$ and $y_i > \bar{y}$. This

means that in the summation $r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$, the

term $\left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$ is positive. Similarly, $\left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$ is

positive when (x_i, y_i) is in quadrant C, since $x_i < \bar{x}$ and $y_i < \bar{y}$.

By similar reasoning, any point (x_i, y_i) in quadrant B or quadrant D makes a negative contribution to the value of r .

If a scatter plot of ordered pairs (x_i, y_i) contains points only in quadrants A and C, then r will be positive. If the points are only in quadrant B and D, then r will be negative. If points are uniformly scattered through the four quadrants, then r will be zero.

If the relationship between x and y is linear, then the value of r^2 represents the proportion of the total variation in the y -values that can be accounted for by the least squares line and differences in x -values.

The table gives the number of minutes played in a season and the number of points scored for 8 NBA basketball players.

Minutes played	2770	690	780	2570	1808	1502	2030	1371
Points scored	1212	198	235	1057	814	528	650	542

For this data, the value of r^2 is 0.9523. This means that about 95% of the variation in points scored can be accounted for by the number of minutes played and the least squares line. The remaining 5% of the variation represents the extent to which the regression line fails to perfectly model the data. This variation shows up as the "scatter" of the data about the least squares line.

The table below gives the number of hours that 6 students spent studying and their test scores.

Hours studying	0.25	0.5	1	1.25	2	2.5
Test score	72	60	85	88	75	90

For this data, r^2 is about 0.375. This means that less than 40% of the variation in test scores can be accounted for by the number of hours spent studying.

