Correlation and regression

Scatter diagrams, explanatory and response variables, the least squares regression line, interpreting its gradient and intercept, and the product moment correlation coefficient.

Correlation and regression

Scatter diagrams

A **scatter diagram** plots pairs of values (x, y) for two variables. **Correlation** describes the pattern.

- **Positive correlation:** y tends to increase as x increases.
- **Negative correlation:** y tends to decrease as x increases.
- **No correlation:** no pattern.

The nearer the points are to a straight line, the stronger the correlation.

NOTE

Correlation is not the same as cause. Two variables may rise together because of a third variable, or by chance. Say "is associated with", not "causes".

Explanatory and response variables

The **explanatory** (independent) variable x is the one that is controlled, or is used to explain the other. The **response** (dependent) variable y is the one that is measured or predicted. On a scatter diagram the explanatory variable goes on the horizontal axis.

TRY IT

A student records the hours spent revising and the test score. Which is the explanatory variable?

Hours of revision is the explanatory variable x, and the score is the response variable y.

Summary values

For n pairs of values, the regression and correlation formulae use

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}, \qquad S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}, \qquad S_{xy} = \sum xy - \frac{\sum x \sum y}{n}$$

Note Academy

The regression line

The **least squares regression line** of y on x is the straight line $y = a + bx$ that makes the sum of the squares of the vertical distances from the points as small as possible.

$$b = \frac{S_{xy}}{S_{xx}}, \qquad a = \bar{y} - b\bar{x}$$

The line always passes through $(\bar{x}, \bar{y})$.

> **TRY IT**
>
> For 8 pairs, $\sum x = 40$, $\sum y = 96$, $\sum x^2 = 240$, $\sum xy = 520$. Find the regression line of y on x.

$$S_{xx} = 240 - \frac{40^2}{8} = 40, \qquad S_{xy} = 520 - \frac{40 \times 96}{8} = 40$$

$$b = \frac{40}{40} = 1, \qquad \bar{x} = 5, \quad \bar{y} = 12, \qquad a = 12 - 1 \times 5 = 7$$

The line is $y = 7 + x$.

Interpreting the line

- The **gradient** b is the change in y for each increase of 1 in x.
- The **intercept** a is the predicted value of y when $x = 0$. It may have no meaning if $x = 0$ is outside the data.

> **TRY IT**
>
> The line $y = 7 + x$ models a test score y against hours of revision x. Interpret the numbers 1 and 7.

The score is predicted to rise by 1 mark for each extra hour of revision. With no revision the predicted score is 7.

Using the line to predict

Substitute a value of x to predict y.

- **Interpolation** predicts inside the range of the data and is usually reliable.
- **Extrapolation** predicts outside the range and is less reliable, because the pattern may not continue.

> **Only predict y from x with the line of y on x. To predict x from y, a different line (of x on y) is needed.**

Note Academy

Product moment correlation coefficient

The **product moment correlation coefficient** r measures how close the points are to a straight line:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

It always lies between -1 and 1.

Value of r	Meaning
$r = 1$	All points on a straight line with positive gradient
r near 1	Strong positive linear correlation
r near 0	Little or no linear correlation
r near -1	Strong negative linear correlation
$r = -1$	All points on a straight line with negative gradient

TRY IT

Using the data above with $\sum y^2 = 1260$, find r.

$$S_{yy} = 1260 - \frac{96^2}{8} = 108, \qquad r = \frac{40}{\sqrt{40 \times 108}} = \frac{40}{65.7} = 0.609$$

This shows moderate positive linear correlation.

Limitations

- r only measures **linear** correlation. A curved pattern can have r near 0.
- A single unusual point (an outlier) can change r and the line a great deal.
- A high r does not show that x causes y.
- A prediction beyond the data may not hold.

NOTE

Coding the data, such as $x \to x - 10$, changes a but not b when only x is shifted, and does not change r. Check which variable was coded before using the line.

Note Academy