

SUMMARY: Correlation & Linear Regression

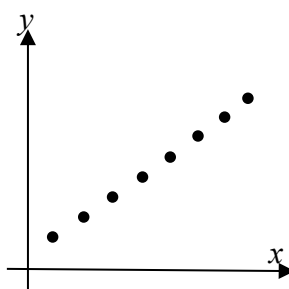
Linear correlation

Linear correlation is the study of the linear relationship between two random variables X and Y .

The **product moment correlation coefficient**, r , is a numerical measure of the strength of the linear relationship and is given by

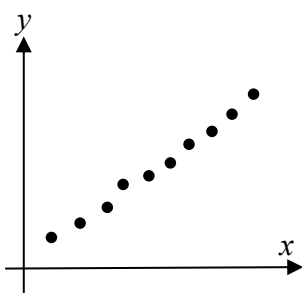
$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\left\{ \Sigma(x - \bar{x})^2 \right\} \left\{ \Sigma(y - \bar{y})^2 \right\}}} = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\sqrt{\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n} \right) \left(\Sigma y^2 - \frac{(\Sigma y)^2}{n} \right)}} \quad (\text{in MF27})$$

Relationship between the value of r and the scatter diagram



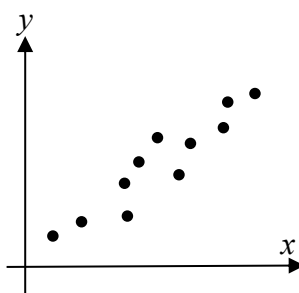
$r = 1$

Perfect positive linear relationship between x and y



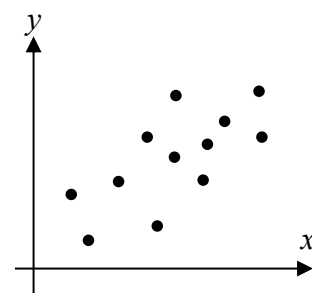
$r \approx 0.9$

Strong positive linear relationship between x and y



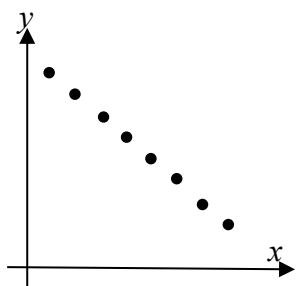
$r \approx 0.7$

Moderate positive linear relationship between x and y



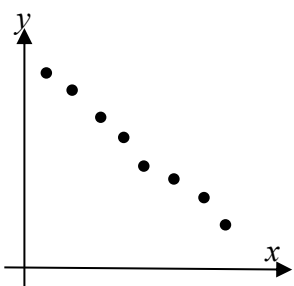
$r \approx 0.3$

Weak positive linear relationship between x and y



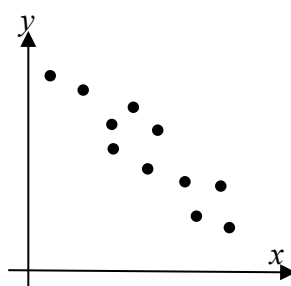
$r = -1$

Perfect negative linear relationship between x and y



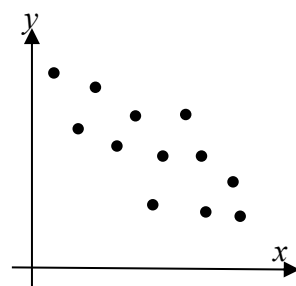
$r \approx -0.9$

Strong negative linear relationship between x and y



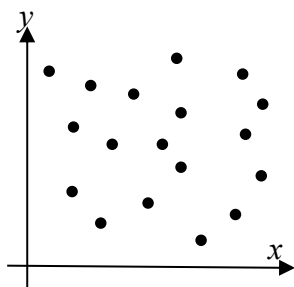
$r \approx -0.7$

Moderate negative linear relationship between x and y



$r \approx -0.3$

Weak negative linear relationship between x and y



$r \approx 0$

No linear relationship between x and y

- (i) If $r = \pm 1$, this is called a perfect linear correlation. All data points lie exactly on a straight line.
- (ii) The closer the value of r to ± 1 , the stronger the linear correlation between the two variables i.e. the closer the points on the scatter diagram are to a straight line.

Note:

- A scatter diagram should be drawn to establish the linear relationship between the variables before the correlation coefficient is calculated.
- The value of r is then calculated (either using formula or GC) to determine the degree of linear relationship between the two variables

Example

$r = 0.95$ between number of hours of sleep (x) and math marks (y) seems to indicate a strong positive linear correlation between number of hours of sleep and math marks.

- The existence of correlation between two variables does not imply cause and effect. It is possible that the two variables are each related in turn to another variable which produces the changes in both.

Linear Regression

Linear regression is a way to model the relationship between two variables by fitting a linear equation to a set of observed data.

Given a data set, the least square regression lines are lines of best fit that are ‘as close as possible’ to the data points. The equations of the least square regression lines can be found using the GC:

	y on x	x on y
Formula	Least Squares Regression line: $y = a + bx$ where $b = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2} \quad \leftarrow \text{(in MF27)}$ $= \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}$	Least Squares Regression line: $x = c + dy$ where $d = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(y - \bar{y})^2} \quad \leftarrow \text{(NOT in MF27)}$ $= \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma y^2 - \frac{(\Sigma y)^2}{n}}$
Properties	1. The line minimises the sum of squares of y-errors. 2. $(\bar{x}, \bar{y})$ is on the regression line. 3. The value of b gives the <u>gradient</u> of the regression line i.e. the amount of change in y for a unit change in x. (e.g. for regression line $y = 2.48 + 0.608x$, we can say that when rainfall, x increases by 1 unit, crop yield, y increases by about 0.608 units per hectare.) 4. The value of a gives the value of y when $x = 0$. (e.g. for regression line $y = 2.48 + 0.608x$, when there is no rainfall, the predicted crop yield = 2.48 units per hectare ($x = 0, y = 2.48$). <u>Caution:</u> To check whether the value of $x = 0$ is within or outside the data range of values of x.)	1. The line minimises the sum of squares of x-errors. 2. $(\bar{x}, \bar{y})$ is on the regression line. 3. The value of $\frac{1}{d}$ gives the gradient of the regression line when plotting y against x.

(A) To determine the Appropriate Regression Line for Estimation

Case 1

One of the variables is an independent variable.

- When x is an independent variable, **regression line of y on x** is used for any estimation.
- When y is an independent variable, **regression line of x on y** is used for any estimation.

Case 2

Neither variable can be considered independent.

- **Regression line of y on x** is used to estimate the value of y for a given value of x .
- **Regression line of x on y** is used to estimate the value of x for a given value of y .

(B) To determine the Reliability of an Estimate

- **Interpolation** involves making a prediction within the ranges of values for the data.
 - **Extrapolation** involves making a prediction outside the range of values for the data.
- (i) An estimate is **reliable** if **BOTH** the following conditions are satisfied:
- a) We are estimating **within** the data range, and
 - b) The **value of the product moment correlation coefficient** is close to 1 or -1 . i.e. there is a strong linear correlation between the variables (This may also be reflected in the scatter diagram when the data points indicate a strong linear correlation).
- (ii) An estimate is NOT reliable as long as either one of the above conditions is not satisfied. For example, the estimated value of y will not be reliable if the x values used are outside the given range of the data i.e. the extrapolated values of y may not be reliable as the relationship may not continue to be linear.