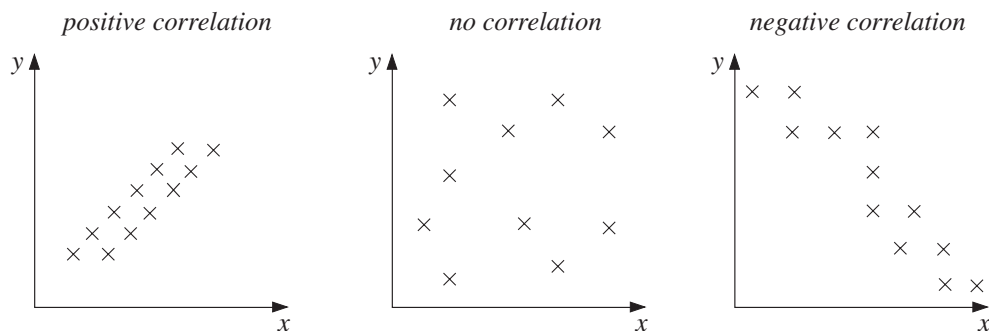


Correlation and Regression

Essential information

- **Correlation** is a measure of how two sets of data are related whereas **regression** shows how well data fits a straight line. Regression should only be used when the data is strongly correlated (positively or negatively).
- **Measures of correlation**, r , are based on $-1 \leq r \leq 1$ and $r = 1$ for perfect positive correlation and $r = 0$ for no correlation and $r = -1$ for perfect negative correlation.



- **Spearman's Rank Correlation Coefficient** is based on ranked data and is given by

$$r = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

- **Product Moment Correlation Coefficient** (sometimes known as Pearson's PMCC) is based on actual data values and is given by

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

where

$$s_x = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2}$$

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

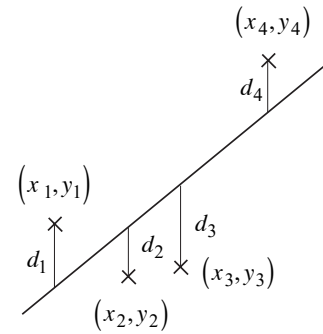
- **Regression lines** are essentially lines of best fit based on given data

Correlation and Regression

Essential information

- The **equation of the line of regression** is based on minimising the square of the distances of the data points from the regression line,

$$\text{i.e. minimising } \sum_{i=1}^n d_i^2$$



- The **y and x line of regression** is given by

$$y - \bar{y} = \frac{S_{xy}}{S_{xx}}(x - \bar{x})$$

when

$$S_{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y}$$

and

$$S_{xx} = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2$$

- Bivariate data** is data that has two variables