CORRELATION

CORRELATION

- Correlation is used to describe the strength and direction of the relationship between two variables (usually continuous – but can be used when one of the variables is dichotomous i.e. has only two values)
- The most common statistic obtained is Pearson's product-moment correlation (r)
- Partial correlation is used when you wish to explore the relationship between two variables while statistically controlling for a third variable
- Correlation techniques are used to: Explore the association between pairs of variables (correlation)
- Use in regression models to predict other variables

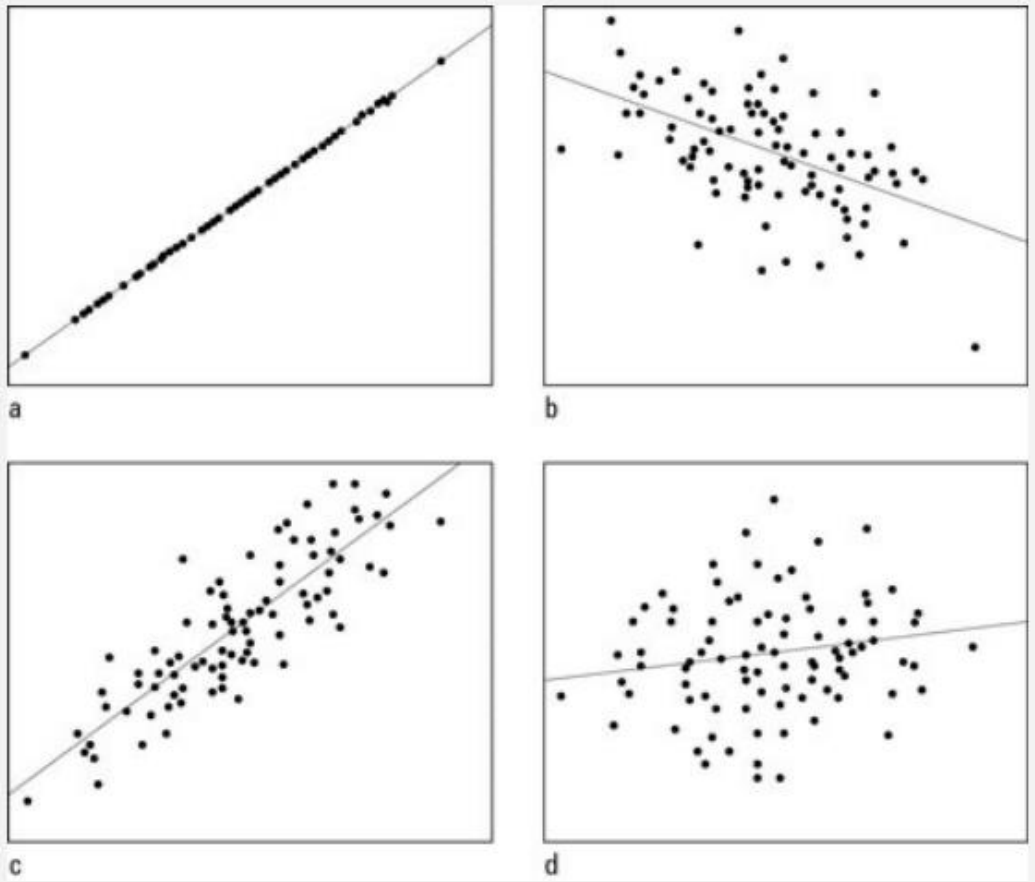
CORRELATION VS CAUSALITY

- Correlation provides an indication of a relationship between variables
- It does not indicate that one variable causes the other
- Strong correlation between variables A & B
  - A causes B?
  - B causes A?
  - C causes both A & B ?
- Ice cream sales and homicides in New York
- Smoking and lung cancer

CORRELATION & SCATTERPLOT

- The relationship between variables can be inspected visually by generating a scatter plot/diagram
- A scatterplot will provide information on both the direction of the relationship(positive or negative) and the strength of the relationship
- Scatterplots with correlations of

a) +1.00 b) –0.50 c) +0.85 d) +0.15



ASSUMPTION

- Outliers can have a dramatic effect on the correlation coefficient (especially with small samples) – check scatterplot

- The scale of measurement for the variables should be interval or ratio (continuous)

• Independence of Observations: Each observation or measurement must not be influenced by any other observation or measurement.

Linearity: Relationship between two variables should be linear.

CORRELATION COFFICENT

- The correlation coefficient (r) provides an indication of the linear (straight line) relationship between variables.

– Pearson's r will underestimate the strength of relationship when the variables are related in non-linear form.

• Be careful of a restricted range of scores – there should be as wide a range of scores on each of the two variables as possible.

THE COEFFICIENT OF CORRELATION, R

- The value of r ranges between (-1) and (+1)
- The value of r denotes the strength of the association as illustrated by the following diagram.
- What represents a strong correlation really depends on the field in which you are using the correlation.

- Correlation coefficient $r = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{(n-1)s_x s_y}$

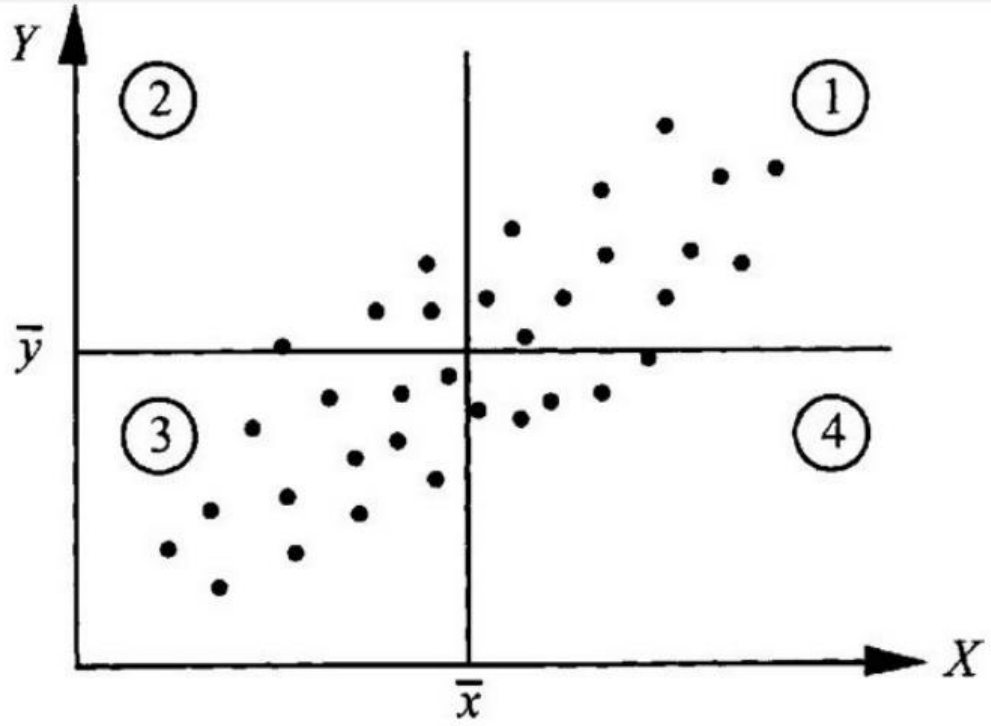


Table 2.2 Algebraic Signs of the Quantities $(y_i - \bar{y})$ and $(x_i - \bar{x})$

Quadrant	$y_i - \bar{y}$	$x_i - \bar{x}$	$(y_i - \bar{y})(x_i - \bar{x})$
1	+	+	+
2	+	−	−
3	−	−	+
4	−	+	−

EXERCISE

- Calculate the correlation coefficient between

Customer visits and the sales counts?

	Customer Visits	Sales
1	20	30
2	40	60
3	20	40
4	30	60
5	10	30
6	10	40
7	20	40
8	20	50
9	20	30
10	30	70