

Section 3.1.14 The Coefficient of Determination (r^2)

VCAA "Dot Points"

Investigating data distributions, including:

- *cause and effect; the difference between observation and experimentation when collecting data and the need for experimentation to definitively determine cause and effect*
- *non-causal explanations for an observed association including common response, confounding, and coincidence; discussion and communication of these explanations in a particular situation in a systematic and concise manner.*

The Coefficient of Determination (r^2)

The **coefficient of determination** (r^2) is useful when we have two variables, which have a linear relationship. It tells us the **proportion of variation** in one variable that can be explained by the **variation in the other variable**.

The coefficient of determination provides a measure of **how well** the linear rule linking the two variables (x and y) **predicts** the value of y when we are given the value of x .

Coefficient of determination = r^2

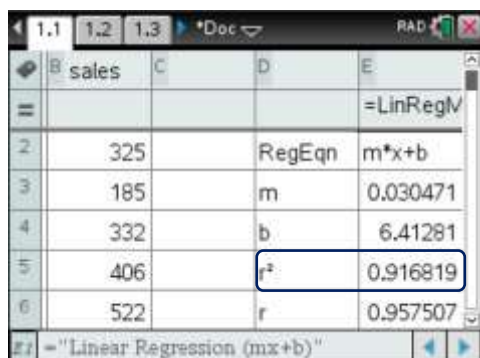
Example 1

The ice cream sales v daily temperature example from Notes 3.1.12 had a Pearson product-moment correlation coefficient (r) of 0.958. What is the coefficient of determination (r^2) for this set of data?

$$r^2 = ? \qquad r^2 = (0.958)^2$$

$$r = 0.958 \qquad = 0.918 \text{ (3 decimal places)}$$

NB: The coefficient of determination can also be found using the TI-Nspire CAS Calculator via the linear regression function.



	B sales	C	D	E
=				=LinRegM
2	325		RegEqn	m*x+b
3	185		m	0.030471
4	332		b	6.41281
5	406		r^2	0.916819
6	522		r	0.957507
#1	="Linear Regression (mx+b)"			

NB: There is a small discrepancy in the value of r^2 , due to the rounding to 3 decimal places in the above example.

Example 2

The driver accidents v driver's age example from Notes 3.1.12 had a Pearson product-moment correlation coefficient (r) of -0.948. What is the coefficient of determination (r^2) for this set of data?

$$r^2 = ? \qquad r^2 = (-0.948)^2$$

$$r = -0.948 \qquad = 0.899 \text{ (3 decimal places)}$$

	B	C	D	E
=				=LinRegV
2	21		RegEqn	m*x+b
3	19		m	-1.52121
4	15		b	46.6848
5	15		r^2	0.89841
6	14		r	-0.9478...

NB: There is a small discrepancy in the value of r^2 , due to the rounding to 3 decimal places in the above example.

NB: Because the coefficient of determination (r^2) is calculated by squaring the Pearson product-moment correlation coefficient (r), it is always a positive value ranging from 0 to 1.

$$1 \geq r^2 \geq 0$$

Interpreting the coefficient of Determination (r^2)

The **coefficient of determination (r^2)** can be used to make a statement about how **much the variation in the response variable** can be **explained** by the **variation in the explanatory variable**.

Please use the following template response when interpreting the coefficient of determination:

"We can conclude from this that *(the coefficient of determination as a percentage)* of the variation in the *(response variable)* can be explained by the variation in the *(explanatory variable)*."

Example 3

Interpretation of $r^2=0.918$ for the ice cream sales v daily temperature example

"We can conclude from this that **91.8%** of the variation in the **ice cream sales** can be explained by the variation in the **daily temperature**."

Example 4

Interpretation of $r^2=0.899$ for the driver accidents v driver's age example

"We can conclude from this that **89.9%** of the variation in the **driver accidents** can be explained by the variation in the **driver's age**."

Calculating the correlation coefficient (r), from the coefficient of determination (r^2)

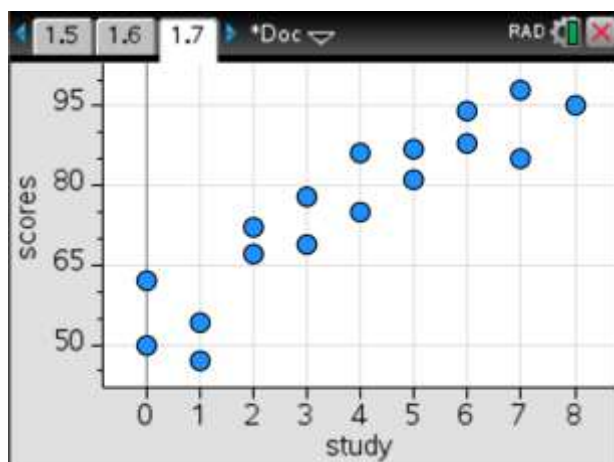
The correlation coefficient (r) can be found by taking the **square root** of the coefficient of determination (r^2).

$$r = \pm \sqrt{\text{coefficient of determination}}$$

Example 5

The scatterplot for test score v hours studied example from Notes 3.1.12 is shown below. This set of data has coefficient of determination (r^2) of 0.842

Calculate the correlation coefficient (r).



$$\begin{aligned} r &= \pm \sqrt{\text{coefficient of determination}} \\ &= \pm \sqrt{0.842} \\ &= \pm 0.918 \end{aligned}$$

To determine whether the correlation coefficient is a positive (+) or negative (-) value, check the trend of the graph!

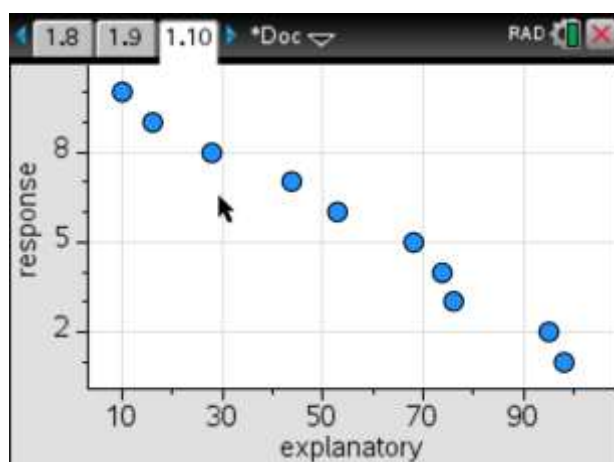
Trend is POSITIVE (upwards)

$$\therefore r = +0.918$$

Example 6

The scatterplot displayed below has a coefficient of determination (r^2) of 0.983.

Calculate the correlation coefficient (r).



$$\begin{aligned} r &= \pm \sqrt{\text{coefficient of determination}} \\ &= \pm \sqrt{0.983} \\ &= \pm 0.991 \end{aligned}$$

To determine whether the correlation coefficient is a positive (+) or negative (-) value, check the trend of the graph!

Trend is NEGATIVE (downwards)

$$\therefore r = -0.991$$

Correlation and causality

Whilst a high correlation coefficient (r^2) indicates the strength to which the response variable can be predicted by the explanatory variable, it does not mean that one variable causes the other to change.

Or in other words;

“Correlation does not mean causation”

There are many examples of variables that have extremely high correlation coefficients, but clearly no causality. Consider the following examples:

1. Tipping and corruption (2012 Harvard Study)
Countries with greater tipping had more problems with political corruption.
2. Finger Length and SAT scores (2007 British Journal of Psychology)
Boys with a higher ring to index finger ratio had higher maths SAT scores
Girls with a lower ring to index finger ratio had higher verbal SAT scores
3. Studying ethics and stealing books (2009 study)
Modern texts on ethics and morality were 50% more likely to go missing from libraries than all other books

Other factors that may contribute towards an association between two variables:

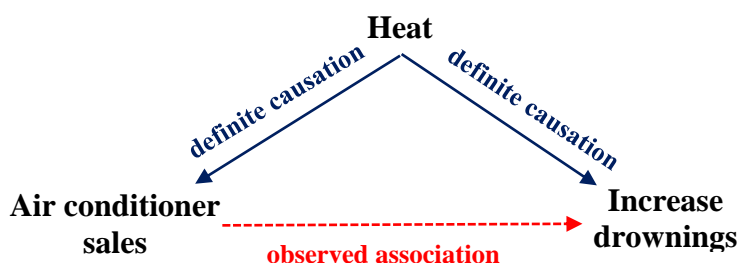
The following are examples of other factors that may contribute towards an association between two variables:

1. Common response to a third party

Scenario:

There is a strong positive correlation between the sale of air conditioners and the number of drownings.

The unintended third variable in this case would be the increase in heat.



2. Confounding variables

Scenario:

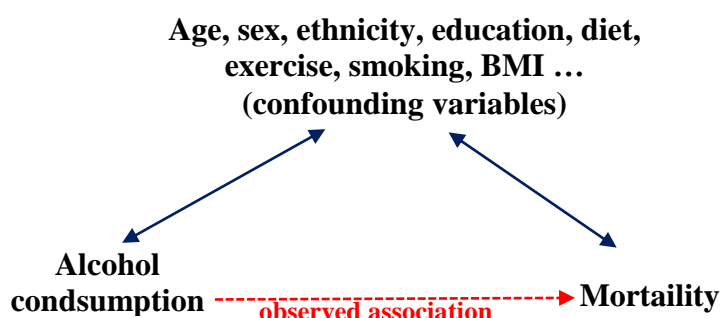
Statistics show a strong correlation between alcohol consumption and mortality.

On the surface, this would suggest that people who consume more alcohol are more likely to die.

However, numerous **confounding variables** may influence this association.

For example:

- Age
- Sex
- Diet
- Smoker
- BMI



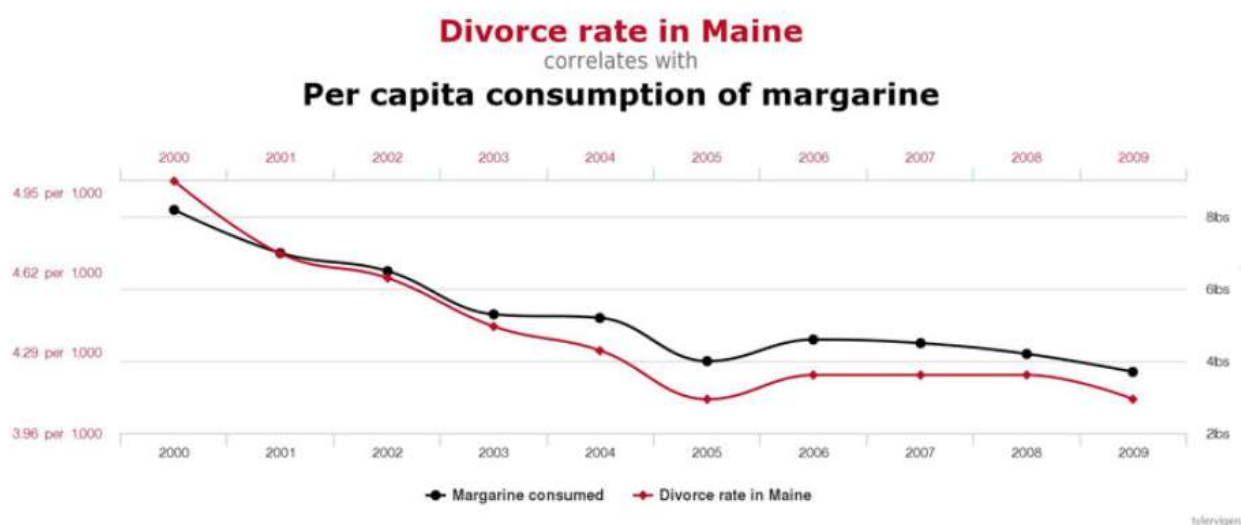
3. Coincidence

Sometimes, a correlation between two variables may be purely coincidental.

Scenario:

Over a 10-year period, Americans' fondness for margarine correlated strongly with the divorce rate in Maine. Yet there is no reason to think one caused the other.

It is an instance of two unrelated data sets showing a **coincidental pattern**.



Source: Science News for Students