M1 Maths

S4-2 Linear Regression

- trends in data, regression lines, interpolation and extrapolation
- correlation
- causality

Summary Learn Solve Revise Answers

Summary

If we draw a scattergram of two measured variables, we sometimes find that there is a bit of a trend; in other words that, as the independent variable increases, the dependent variable tends to increase or tends to decrease. We say 'tends to' because there may be random variation superimposed on the trend.

Where this happens, a regression line can be drawn to show the most likely value of the dependent variable for each value of the independent variable.

The correlation coefficient (r or R) indicates how much of the variation in the dependent variable can be attributed to the change in the independent variable as opposed to being just random variation. In other words, it indicates how close the data points lie to the regression line. If the data points are scattered randomly with no trend, then $r = 0$; if the data points lie exactly on a straight line, then $r = 1$.

If the regression line has a negative gradient, then r will be negative also. Sometimes r^2 is quoted instead of r because it is always positive.

If there is a trend, this may mean one of three things in terms of causality:

1. The value of the dependent variable is partly determined by the value of the independent variable,
2. The value of the independent variable is partly determined by the value of the dependent variable, or
3. The value of a third variable partly determines the value of both plotted variables.

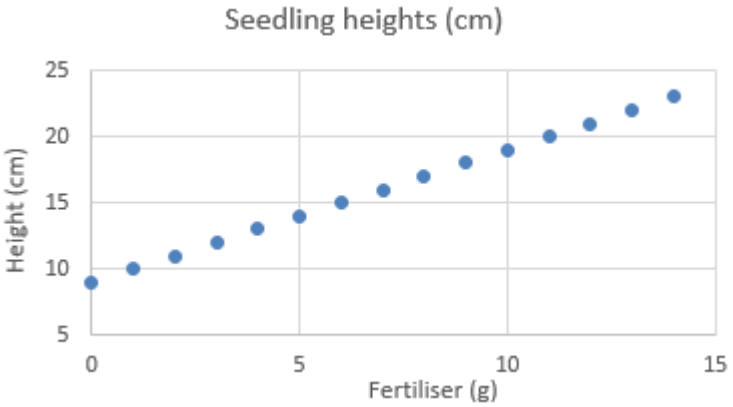
Learn

Trends in Data

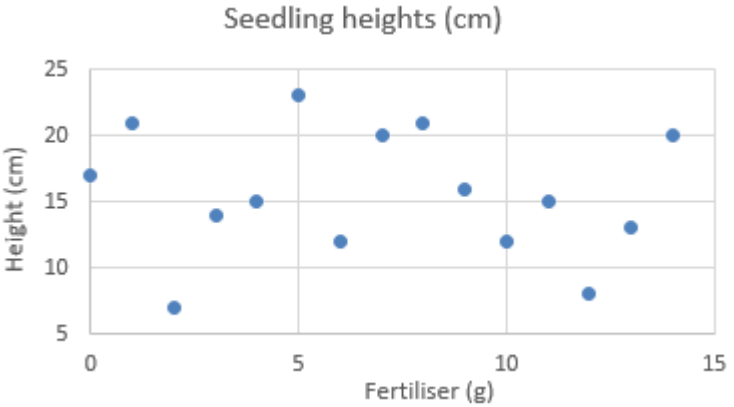
An agricultural researcher was investigating the effect of a new fertiliser on tomato plants. She prepared pots with 0, 1, 2, 3 etc. grams of the fertiliser mixed in with the soil. Then she planted one seedling in each pot. She recorded the height of each plant

21 days later. The graphs below show possible relations between the amount of fertiliser added and the height of the plant.

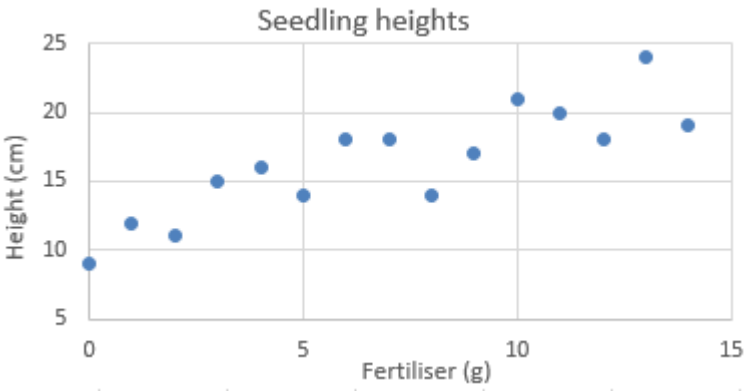
A



B



C



In graph A, the height depends on the amount of fertiliser. There is a definite pattern, so we can write the relation as a formula. It is $h = 9 + f$, there h is the height and f is the amount of fertiliser. This allows us to predict the height of another plant if we know the amount of fertiliser. For instance, a plant with 6.5 g of fertiliser will be 15.5 cm tall.

In graph B, the height varies randomly and seems not to depend at all on the amount of fertiliser. We cannot write a formula for the relation and we cannot predict the

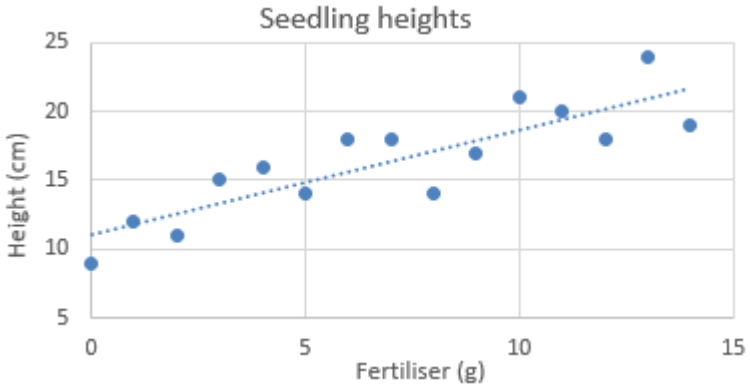
height of another plant given 6.5 g of fertiliser except to say that it will probably be somewhere in the general range from 5 cm to 25 cm.

In Graph C, there seems to be a bit of a pattern. As the amount of fertiliser increases, the height tends to increase. But there is random variation superimposed on the pattern. Instead of saying that there is a definite pattern, we say that there is a trend in the data. We say that there is some correlation between height and amount of fertiliser, but not the perfect correlation that we saw in Graph A.

Regression Lines

For the situation in Graph C, it is possible to make some sort of prediction of the height of another plant, though it won't be totally accurate. For example, another plant with 12 g of fertiliser will most likely be somewhere around 21 cm tall. It is unlikely to be 12 cm; whereas a plant with 1.5 g will most likely be somewhere around 12 cm tall. It is unlikely to be 21 cm.

We can draw a line on Graph C to show the most likely height for each amount of fertiliser. It will look like this:

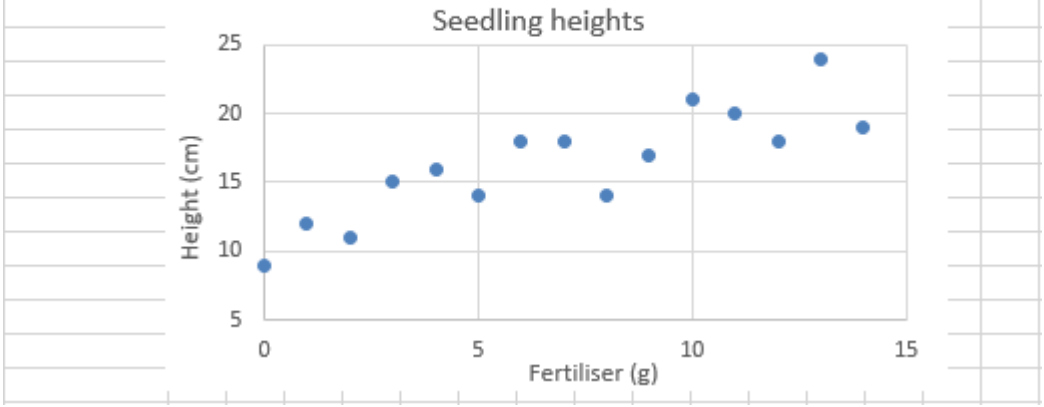


This line is called a **regression line** or a **line of best fit** or a **trend line**. Regression lines can be sketched in by eye. Or they can be calculated exactly. They are calculated by finding the line for which the sum of the squares of the deviations of the data points from the lines in minimum. There is an algorithm for this, but it is complicated and time consuming if done by hand. Calculators and spreadsheets can do it quickly, so we will use them.

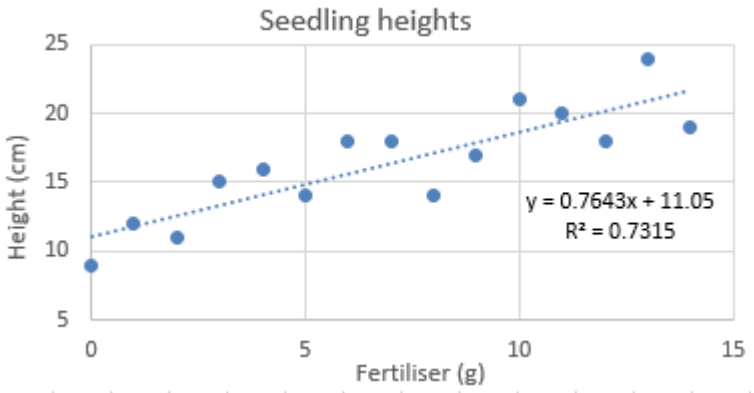
This is how to do it with Excel.

Put the data into a table, highlight the table and insert a scatter plot like this:

Fertiliser (g)	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Height (cm)	9	12	11	15	16	14	18	18	14	17	21	20	18	24	19



Then right click one of the data points and select 'Add Trendline…' in the options. Select 'Linear'. (This gives a straight regression line; the other options give curves and you will look at these in Module A5-7. The current module is named *Linear Regression* because we are dealing here only with straight regression lines.) Also tick the boxes at the bottom for 'Display Equation on chart' and 'Display R-squared value on chart'. The graph should then look like this:



The equation $y = 0.7643x + 11.05$ is the formula for the regression line, where y is the height and x is the amount of fertiliser. (Excel always uses y for the dependent variable and x for the independent variable.)

R is called the correlation coefficient, a measure of how close the data is to the regression line. We will look at this more later. R^2 is just the square of R.

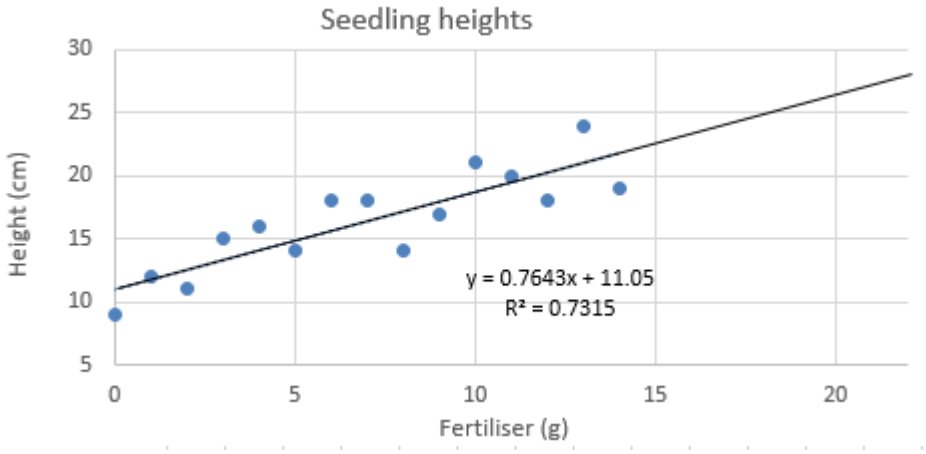
Now we can predict the most likely height for a new plant with 8.6 g of fertiliser either by reading off the graph or by subbing into the equation. Subbing, we get

$$\begin{aligned} height &= 0.7643 \times 8.6 + 11.05 \\ &= 17.6 \text{ cm} \end{aligned}$$

Interpolation and extrapolation

Finding the most likely height for a plant with 8.6 g of fertiliser is called **interpolation** (emphasis on the 'er'). Interpolation is finding most likely heights for new amounts of fertiliser (finding the value of the dependent variable for new values of the independent variable) within the range of the data (in this case between 0 and 14 g of fertiliser).

Finding the most likely heights for amounts of fertiliser outside the data range is called **extrapolation** (emphasis on the first 'a'). For instance, we can extrapolate for predict the most likely height for 20 g of fertiliser. We can extend the line like this:



Then read off that the most likely height is about 26 cm.

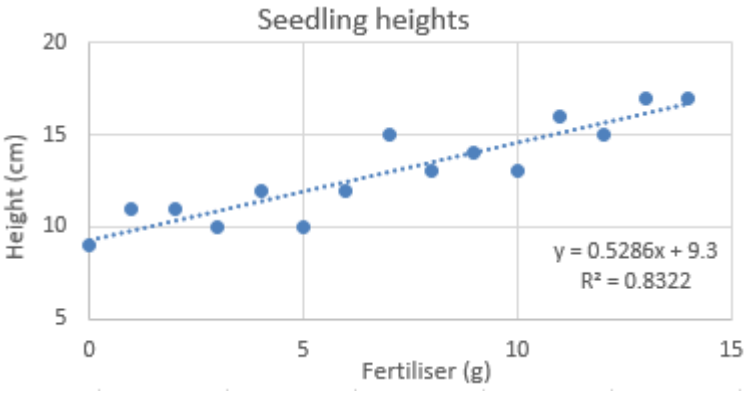
Or we can sub 20 into the formula to get *height* = 0.7643 × 20 + 11.05 = 26.3 cm.

Interpolation is fairly safe, but care is needed when extrapolation as it is not always valid, particularly when extrapolating a long way. The heights of the tomato plants would be expected to reach a maximum at the optimum amount of fertiliser, then start to decrease again after that. Our formula would tell us that a tomato plant with 500 g of fertiliser will be about 4 m tall. More likely, it would be dead from an overdose.

Correlation

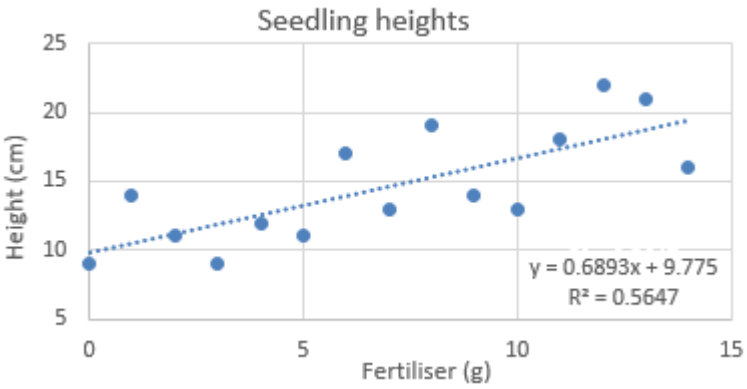
In some scatter graphs, like Graph D below, the points lie close to the regression line. In other words, most of the variation in height is related to the variation in the amount of fertiliser and not much is random variation. We then say there is a high correlation between the two variables.

D



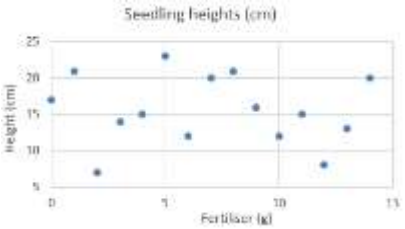
In other scatter graphs, like Graph E below, the points are more scattered. In other words, more of the variation in height is random variation and not so much is related to the variation in amount of fertiliser. We then say there is a low correlation between the two variables.

E

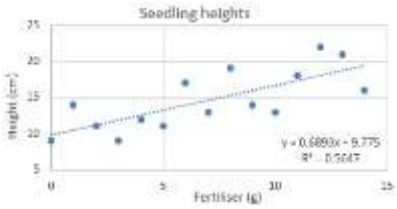


We quantify correlation by using the correlation coefficient (sometimes abbreviated to r or R). Often we use the square of the correlation coefficient (r^2 or R^2) instead. Excel uses R^2.

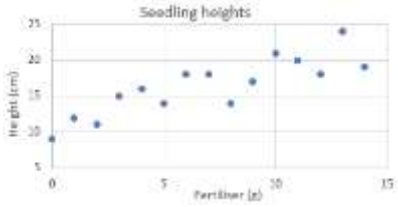
r is a number from 0 to 1 and so r^2 is also a number from 0 to 1. 0 means there is no trend at all – the data is just random. 1 means that all the data lies perfectly on the regression line and there is no random variation at all. In between means in between. Graphs A to E are reproduced below in order of increasing correlation, along with their R^2 values.



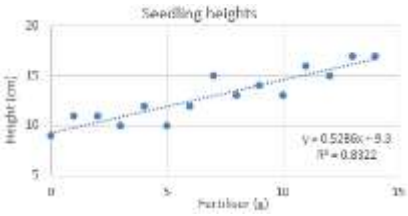
$R^2 = 0.01$



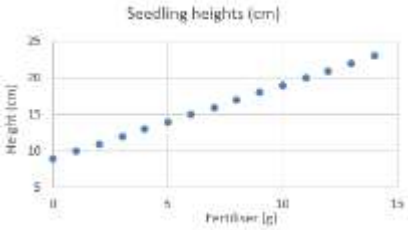
$R^2 = 0.56$



$R^2 = 0.73$



$R^2 = 0.83$

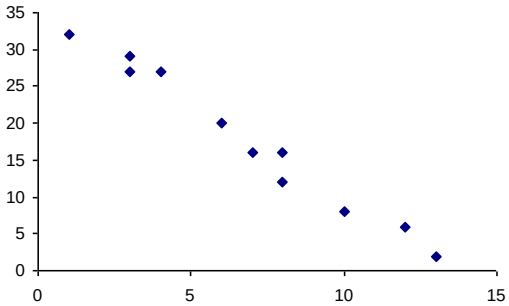


$R^2 = 1$

R^2 values of 0.5 or less look pretty random.

Positive and Negative Correlation

If a regression line has a negative gradient, then the value of r will be negative. $R = 0$ means no correlation at all (totally random); $r = -1$ means all the data lies on the line of best fit exactly. Again, in between means in between, so the graph below has a correlation coefficient of −0.90.



Of course, r^2 will always be positive and that is why it is often used.

Causality

In the experiment with the tomato plants above, there is a positive correlation between the amount of fertiliser and the height of the plant. We have assumed that this indicates that increasing the amount of fertiliser ***causes*** the plant to be taller.

A correlation between two variables generally indicates a causal relationship between them. If there is only a small number of points on the scatter graph, an apparent correlation can occur just by chance even when the data is unrelated. But where there is a reasonable amount of data, a correlation is generally a sign of a causal relationship.

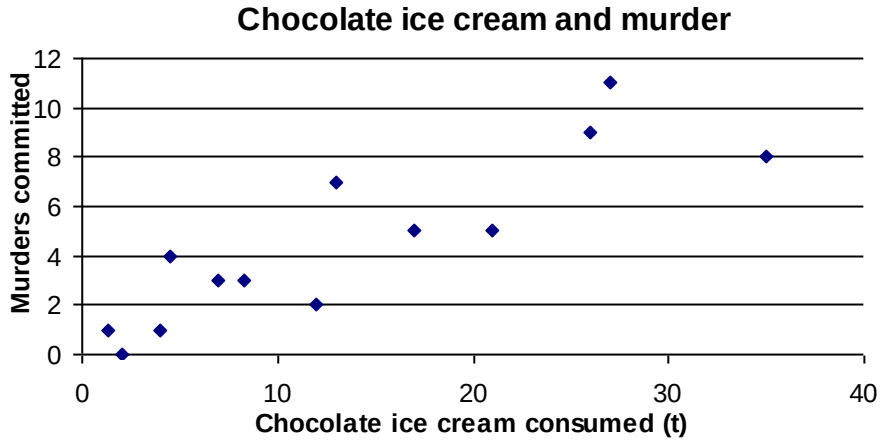
In the tomato plant example above, it would seem reasonable to assume that the amount of fertiliser affects the height – the plants given more fertiliser tended to grow taller.

But it would be possible, just from looking at the graph without thinking about the situation, that the person fertilising the plants gave more fertiliser to the taller plants. So the height affects the amount of fertiliser.

If we think about the situation, though, this might seem less likely. Also, the person collecting the data would probably know whether that was the case.

In general, if there is a correlation between two variables, then we tend to assume that a change in one variable causes a change in the other. We can usually tell which one causes the change by thinking about the situation.

But now look at the following scatter graph. It is a graph of the number of murders committed in various US towns last year plotted against the amount of chocolate ice cream eaten in those towns.



There seems to be a correlation here. This suggests a causal relation. It could be that eating more chocolate ice cream tends to cause more murders to be committed. This

doesn't seem likely. So maybe it's the other way round. If there are more murders in the town, then people tend to eat more chocolate ice cream. Somehow, that doesn't seem terribly likely either.

There is a third possibility which people tend to overlook. This is that, if a town is bigger, then more chocolate ice cream is consumed and more murders occur. This does seem reasonable. The data points in the bottom-left part of the graph are for small towns where not much ice cream is eaten and not many murders occur. The data points in the top-right part are for big towns where a lot of ice cream is eaten and a lot of murders occur.

So it doesn't mean that eating chocolate ice cream turns people into homicidal maniacs or that people go on chocolate ice cream binges when people in their town get murdered.

Thus there are three possibilities for a causal relation if a correlation is observed:

1. The first possibility is that the independent variable influences the dependent variable. In other words, if people in a town eat more chocolate ice cream, then this will lead to more murders.

2. The second possibility is that the dependent variable influences the independent variable. In other words, when more murders occur, this leads to people eating more chocolate ice cream.

3. The third possibility is that a third variable is influencing both the independent and dependent variable. In this case the third variable might be the population of the towns. Towns with a small population consume small amounts of ice cream and have small numbers of murders. Towns with a large population consume large amounts of ice cream and have large numbers of murders. This would produce the correlation observed. In this case, this third possibility seems very plausible and is the one most people would accept.

We need to be able to make an intelligent guess at which of these three types of causal relationship leads to an observed correlation.

Practice

Q1 The following table shows, for a group of students, the number of hours of sport they do per week and their score on a Physical Education test.

Hours of sport	4	1.5	18	11	10	7.5	0	6	8.5	13	7	14
Test score	16	21	36	29	31	22	16	29	34	37	12	37

(a) Plot this data as a scatter graph on Excel along with the trend line, the regression equation and the R^2 value.

(b) Interpolate by reading off the graph to find the most likely score for someone who does 12 hours of sport a week.

(c) Extrapolate using the equation to find the most likely score for someone who does 60 hours sport per week.

(d) Comment on your answer to (c).

(e) Give the values for R^2 and for R.

(f) Give the three possible causes of the correlation and say which you think is most likely.

Q2 The following table shows, for the towns in Fargo Shire, the number of activities that are available to keep youths occupied in the evenings and the number of crimes committed per year by 10 to 20 year olds.

Activities	3	7	0	6	11	2	1	6	5	12	4	8
Crimes	21	9	32	29	8	27	26	16	22	10	16	14

(a) Plot this data as a scatter graph on Excel along with the trend line, the regression equation and the R^2 value.

(b) Interpolate by reading off the graph to find the most likely number of crimes for a town with 8 activities.

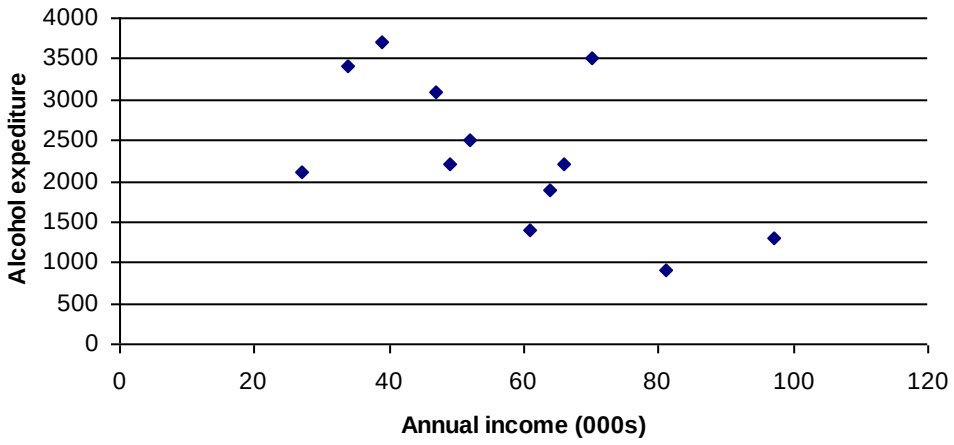
(c) Extrapolate using the equation to find the most likely number of crimes in a town with 25 activities.

(d) Comment on your answer to (c).

(e) Give the values for R^2 and for R.

(f) Give the three possible causes of the correlation and say which you think is most likely.

Q3



The graph above is for a sample of twelve 30-year old men from Caboolture. It shows the amount each spends on alcoholic drinks per year and their income per year.

(a) Copy the graph, then add a line of best fit by eye.

(b) Estimate the correlation coefficient, r.

(c) Use the line of best fit to predict the most likely amount a 30-year old man from Caboolture will spend on alcohol if he earns \$70 000 per year.

(d) Give three possible causes for the correlation between the amount spent on alcohol and income and say which you think is most likely.

Solve

Q51 Roger noted that people who drive BMWs generally had more money than people who drove cheaper cars. So he decided that the best way to get rich was to buy a BMW. Comment on his thinking.

Revise

Revision Set 1

Q61 The following table shows, for a group of policemen, their weight (mass) and the number of arrests they have made in the past year.

Mass	82	105	90	75	141	68	88	192	123	66	151	99
Arrests	16	11	21	18	7	22	19	1	9	18	3	24

(a) Plot this data as a scatter graph on Excel along with the trend line, the regression equation and the R^2 value.

(b) Interpolate by reading off the graph to find the most likely number of arrests for a policemen who weighs 160 kg.

(c) Extrapolate using the equation to find the most likely number of arrests for someone who weighs 22 kg.

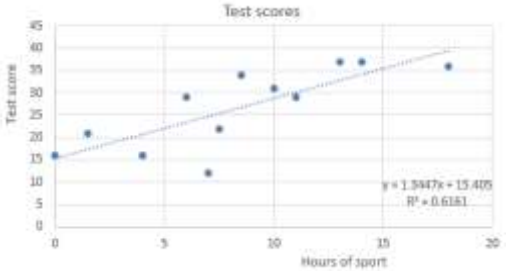
(d) Comment on your answer to (c).

(e) Give the values for R^2 and for R.

(f) Give the three possible causes of the correlation and say which you think is most likely.

Answers

Q1 (a)



(b) 32 (c) 96

(d) This result is unlikely because the student would have little of no time to spend on learning and would probably be exhausted.

(e) $R^2 = 0.62$, $R = 0.78$

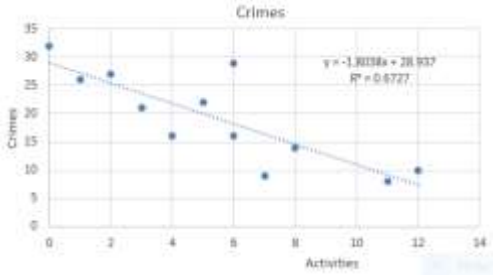
(f) 1. Doing more sport makes the students do better on the test

2. Getting good marks on the test inspires students to do more sport

3. Something else causes them to do more sport and get better test marks. That thing could be how much they are into sport.

Number 3 is probably most likely.

Q2 (a)



(b) 15 (c) −16

(d) This result is impossible as you cannot have a negative number of crimes. More likely, the number of crimes would level out at a small number as the number of activities keeps increasing.

(e) $R^2 = 0.67$, $R = -0.82$ [Note the negative.]

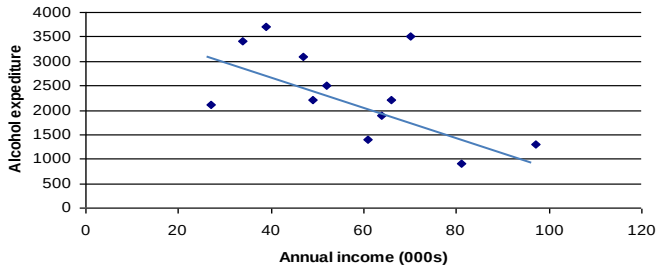
(f) 1. Providing more activities gives the youth more to occupy their time so they have less time to get up to mischief.

2. The towns with less youth crime provide more activities, possibly because they spend less money fixing crimes and so have more for activities, or possibly as a reward to the youth.

3. Something else causes more activities and less crime. That thing could be how rich the people of the town are.

Any of these are possible, but Number 1 seems the most likely.

Q3 (a)



(b) $r \approx 0.8$ (c) \$18 000

(d) 1. Earning more gives people different things to do, so they drink less or people who don't earn much drink to drown their sorrows.

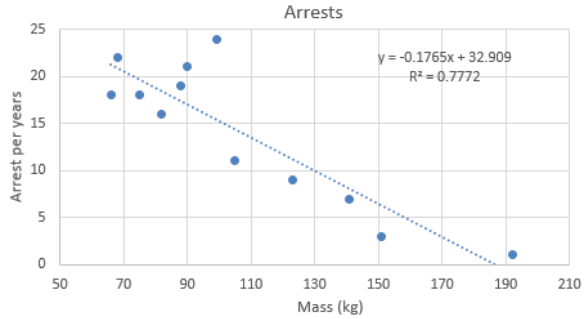
2. Drinking more causes people to be able to earn less.

3. Having a more positive attitude to life causes people to earn more and to drink less.

All of these are possible. 3 is maybe the most likely.

Q51 The correlation between owning a BMW and being rich probably has the opposite causality, i.e. people who are rich are more likely to buy a BMW rather than people who have BMWs are more likely to get rich. Buying a BMW if he can't afford it will make Roger worse off.

Q61 (a)



(b) 5 (c) 29

(d) This is unlikely as any adult weighing 22 kg would be in hospital or dead.

(e) $r^2 = 0.78$ $r = -0.88$

(f) 1. Heavy policemen can't run fast enough to catch crooks.

2. Police who don't make many arrests eat more because they feel bad.

3. Police who work mostly in the office don't make many arrests. Also, they do little exercise and so put on weight.

All are possible, but 3 might be the most likely.