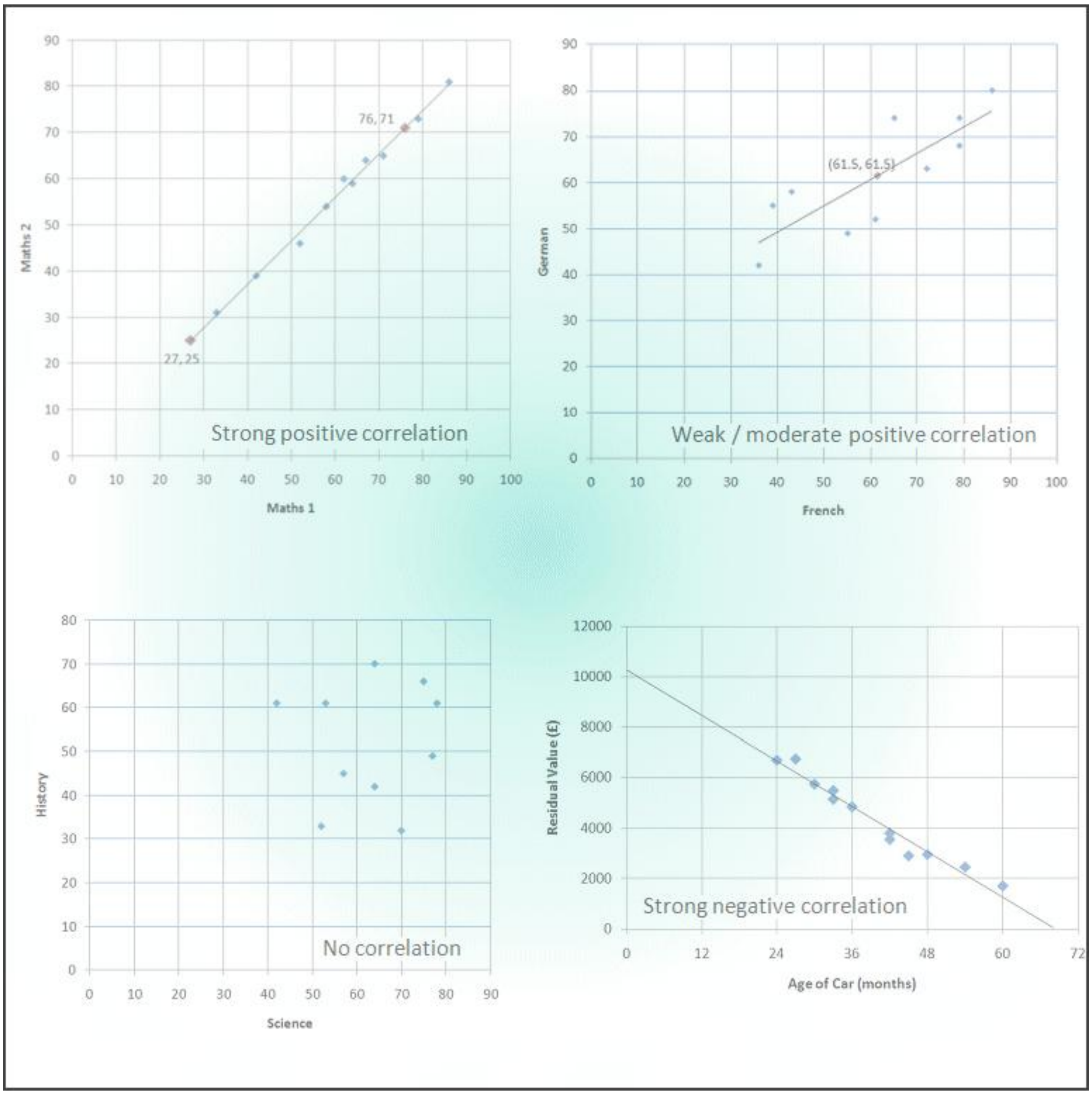
M.K. HOME TUITION

Mathematics Revision Guides
Level: GCSE Higher Tier

SCATTER DIAGRAMS



Version: 4.2 Date: 23-07-2016

Scatter Diagrams.

Scatter diagrams, or scatter plots, are used to investigate connections between two features or variables, such as maths exam results against science exam results, or height against weight.

Scatter plots are two-dimensional, and from them it is possible to infer whether two sets of data are connected - in other words, correlated.

If there is strong correlation, then the data will tend to fall in a linear pattern; if there is little or no correlation, the data will generally fall in a random pattern.

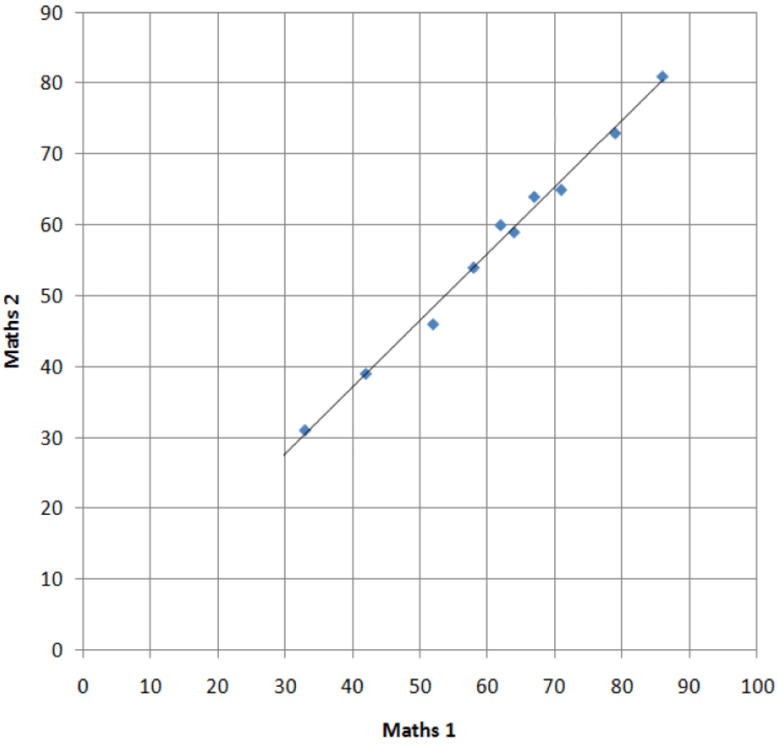
Example (1): Ten pupils were chosen at random from Year 11 and the percentage scores for the mock Maths 1 and Maths 2 Higher Tier exams were obtained as follows (Maths 1 quoted first).

64, 59	79, 73	86, 81	42, 39	58, 54
67, 64	33, 31	52, 46	71, 65	62, 60

Plot the pairs of results on a scatter diagram and attempt to draw a best-fit line. Is there a correlation between the result sets ?

i) There is a strong positive correlation between the results of the two maths exams, and hence it is easy to plot a line of best fit.

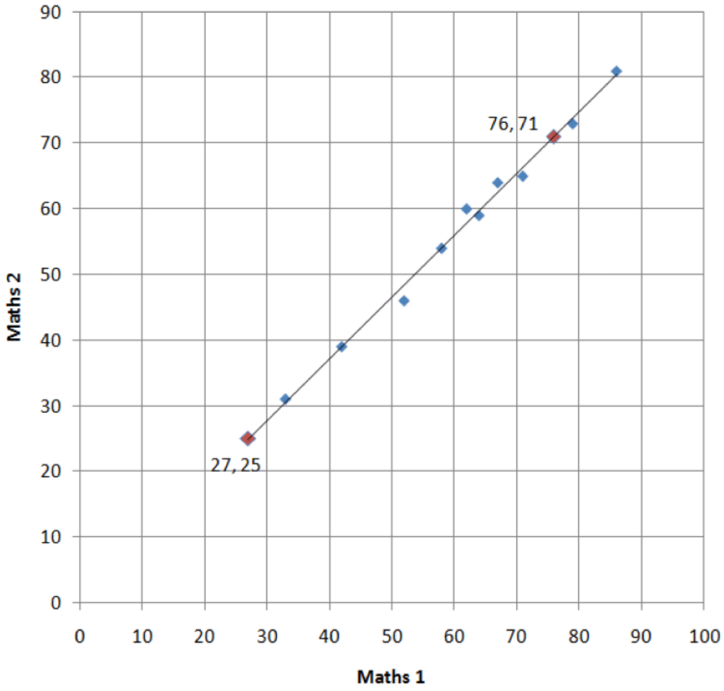
Scatter plots can also be used for the purposes of estimation, especially if the correlation between sets of data is strong.



Example (1a): i) Emily received 76% in her Maths 1 exam. Estimate her Maths 2 exam result.

ii) Fred took the Higher Tier exams in error when he should have taken the Foundation Tier exams. As a result, he obtained 25% in his Maths 2 exam. Estimate his Higher Tier Maths 1 exam score.

iii) Comment on the reliability of the estimated results from parts i) and ii).



i) Reading off the line of best fit, a mark of 76% in the Maths 1 exam corresponds to 71% in the Maths 2 exam.

∴ Emily's Maths 2 estimated score is 71%.

We are inferring an estimate from **within** the main set of marks, and such an inference is known as **interpolation.**

ii) Again, reading off the best fit line, a mark of 25% in Maths 2 corresponds to 27% in Maths 1.

∴ Fred's Maths 1 estimated score is 27%.

This time, we are trying to make an estimate from **outside** the set of marks - known as **extrapolation.**

iii) The result from part (i), namely Emily's score of 71 for Maths 2, is likely to be more reliable because it lies within the main set of marks.

The result from part (ii) is outside the range of the rest of the marks, which is less 'safe'.

In general, interpolation is more accurate than extrapolation.

Example (2) :
Ten pupils were chosen at random from Year 11 and the percentage results for the mock French and German exams were obtained as follows (French quoted first).

86,80	79,68	79,74	72,63	65,74
61,52	55,49	43,58	39,55	36,42.

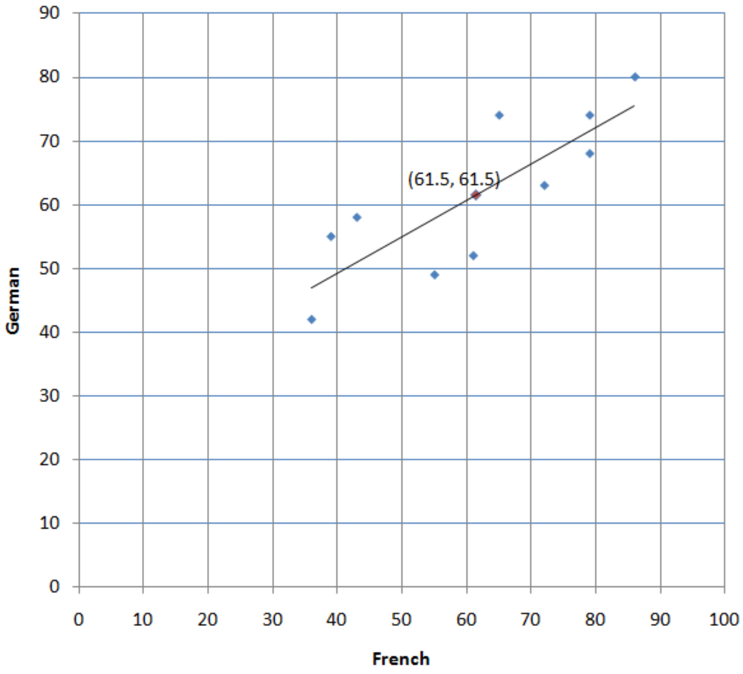
i) Produce a scatter diagram, plotting the French results on the *x*-axis.
ii) Show that the mean score for each exam is equal, and plot a best fit line passing through that mean point.

The French and German result sets both sum to 615, giving mean scores of 61.5% for each language.

The line of best fit has been plotted through the mean point (61.5, 61.5).

There is a positive correlation between the results for French and German here, but weaker than in Example 1.

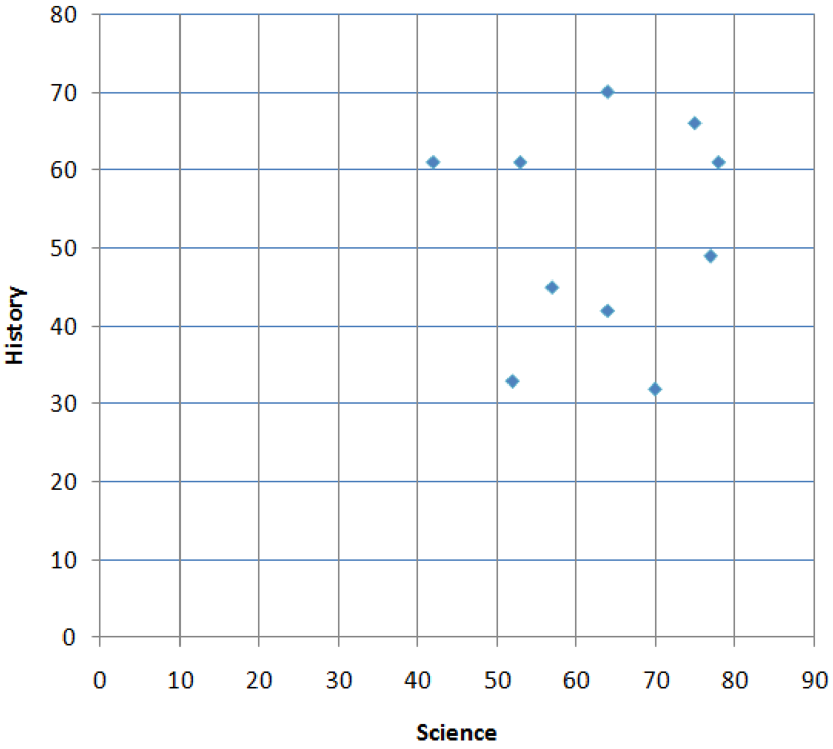
Thus, the line of best fit does not pass through any of the original points on the scatter graph.



Example (3) :
Eleven pupils were chosen at random from Year 11 and the results for the mock Science and History exams were obtained as follows (Science quoted first).

Plot the resulting scatter diagram. Is there any correlation at all between the data sets ?

78,61	77,49	75,66	70,32	64,42
64,70	57,45	53,61	52,33	42,61
35,51				



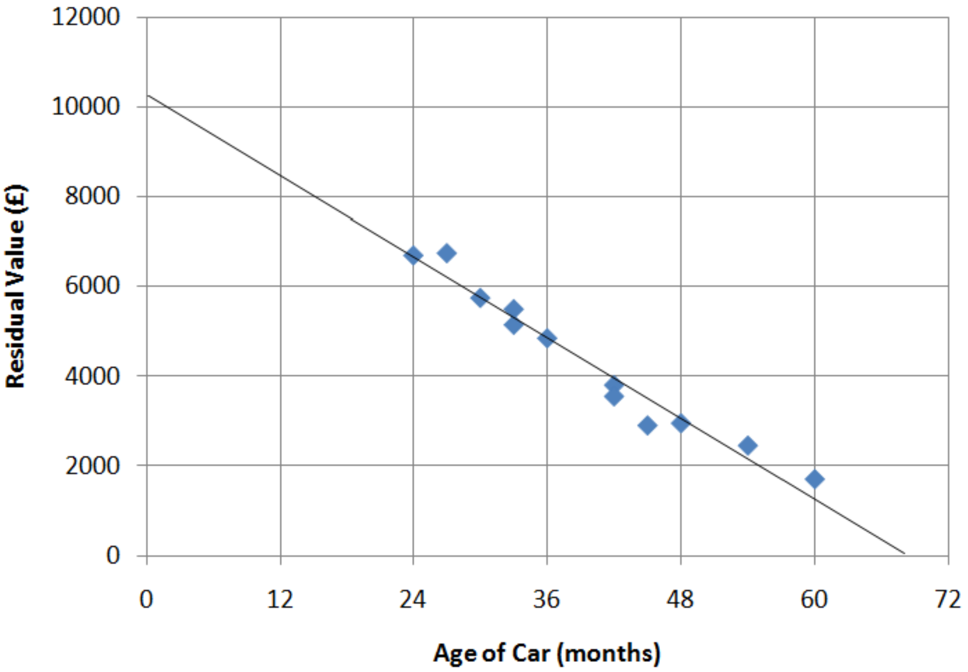
The results appear to be randomly scattered here, with no correlation at all apparent.

Example (4) : An insurance company has recorded the residual values of a sample of twelve cars of a particular model. The ages of the cars (in months) and the residual values (in £) are given in the list below.

24, 6700	27, 6900	30, 5750	33, 5500	33, 5400	36, 4850
42, 3800	42, 3550	45, 2900	48, 2950	54, 2450	60, 1700

i) Plot a scatter graph and draw a best-fit straight line through the points. Is the correlation strong or weak ? Is it positive or negative ?

ii) Explain why the best-fit straight line is unsuitable when estimating the residual value of a 6-year-old car, and suggest why it is unsuitable in the case of a brand new car.



i) The best-fit straight line has a negative gradient and passes close to all the points, therefore there is a strong negative correlation.

ii) The straight line meets the *x*-axis when the car is less than 72 months, or 6 years, old. This would imply a negative residual value for a 6-year-old car, which is nonsense. In addition, all of the cars in the sample are between two and five years old, so a brand new car falls outside the category. It is a known fact that a new car loses much of its value the instant it leaves the showroom !
This example illustrates the dangers of extrapolation very well.

Correlation and Causation.

Looking back at Example (4), there is a clear negative correlation between the ages of the cars and their residual values.

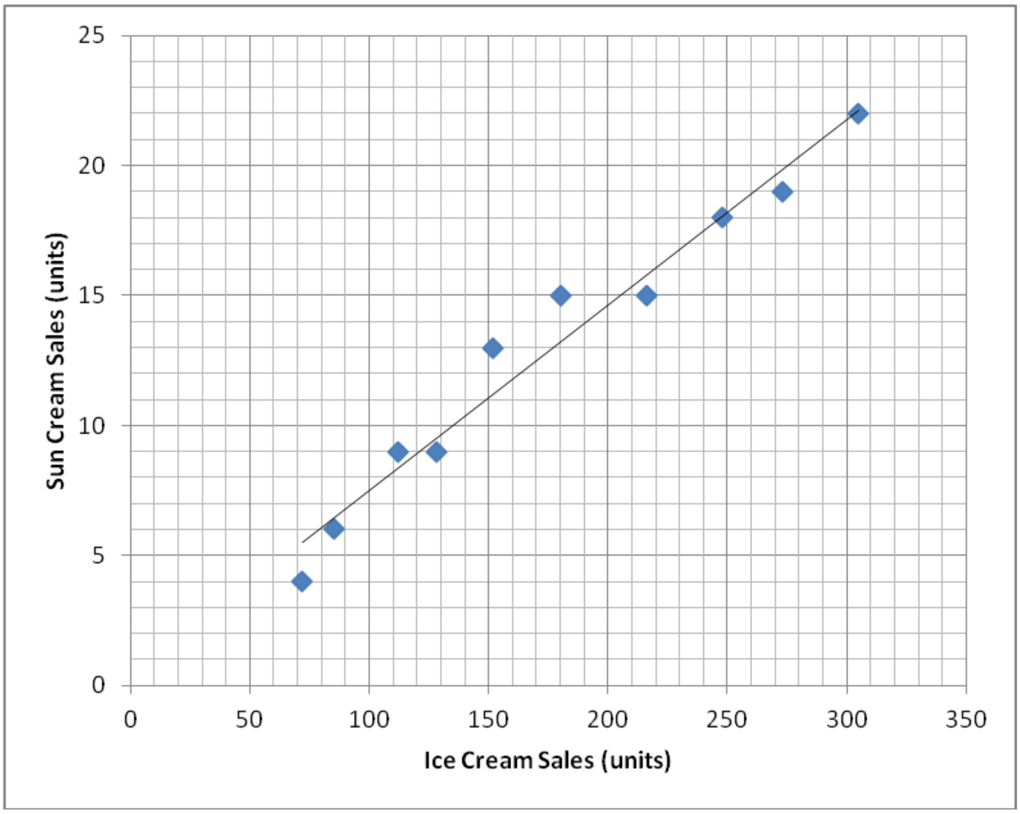
It can also be said that there is a **causal relation** between the cars' ages and values. The older the car, the lower would be the expected resale value, due to depreciation of any manufactured item with age.

So in this example, we have both a correlation and causation.

Example (5): Tony is a shopkeeper in a small seaside resort, and has recorded the sales (in units) of ice cream and sun-tan cream in his shop over a ten-week period.

i) Produce a scatter graph from the information and plot the best-fit line.
ii) Comment on the results, including a statement on causation.

Week	1	2	3	4	5	6	7	8	9	10
Ice cream sales (units)	180	128	216	72	112	85	152	305	273	248
Sun cream sales (units)	15	9	15	4	9	6	13	22	19	18



At first sight, there is a strong positive correlation between ice cream sales and sun cream sales, but there is **no causal relation** between the two.

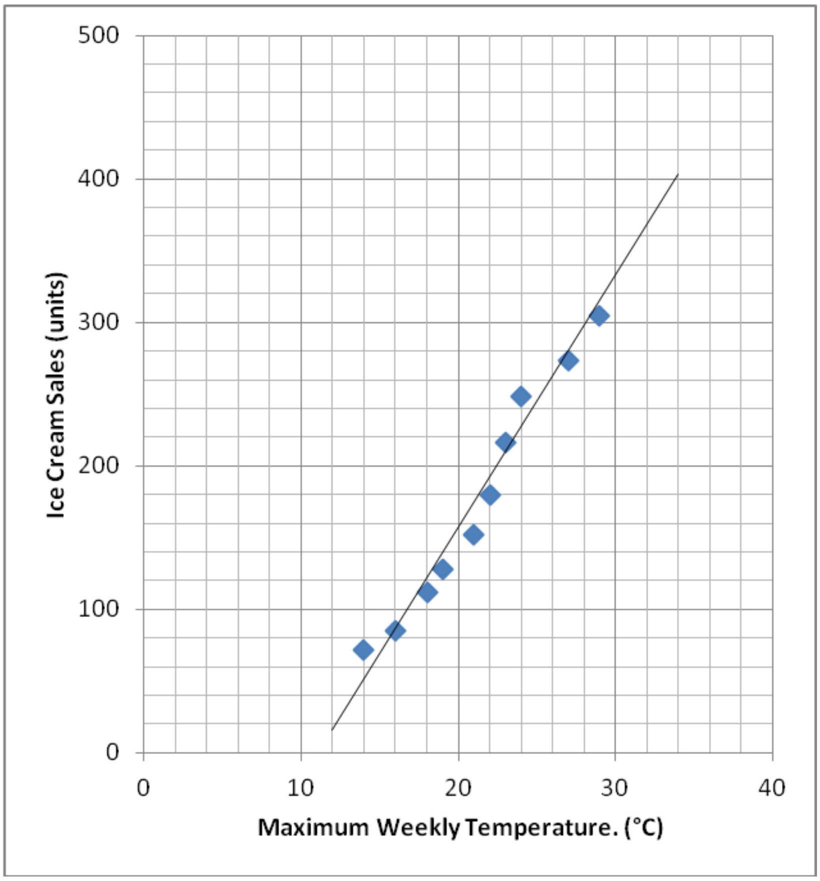
In other words, buying an ice cream does not cause a person to buy sun cream as if the second was dependent on the first. There are other causal factors at work, such as outside temperature, and the amount of sunshine, influencing the sales of both ice cream and sun cream.

Another case concerns the French and German examination results in Example (2). There is a rather weak correlation between the two, and also a debatable causation. Pupils *may* have an equal aptitude for different foreign languages, but then again, a pupil could have French as a mother tongue but have difficulty in other languages such as English and German.

Example 5(a): Tony’s brother, Louis, has decided to produce an improved scatter graph by also recording the maximum weekly temperature over the same ten-week period.

Week	1	2	3	4	5	6	7	8	9	10
Max. weekly temperature (°C)	22	19	23	14	18	16	21	29	27	24
Ice cream sales (units)	180	128	216	72	112	85	152	305	273	248

i) Produce a scatter graph from the information and plot the best-fit line.
ii) The maximum temperature for the week after Tony’s survey was 20°C. Estimate the ice cream sales for that week.
iii) The previous year, the resort experienced a heatwave where, for one week, the maximum temperature reached 33°C. Estimate the ice cream sales for that week.
iv) Comment on the reliability of the results from ii) and iii).
v) Is there a causal relationship between the maximum weekly temperature and the sales of ice cream ?



ii) The best-fit line passes through the point (20, 160), so the estimated weekly sales of ice cream for that week would be 160 units.

iii) The line passes near the point (33, 380), so the estimated weekly sales of ice cream for that heatwave week would be 380 units.

iv) The estimate of 160 units for a maximum temperature of 20°C was interpolated as it was within the main range of data. The estimate of 380 units for a maximum temperature of 33°C, on the other hand, was extrapolated as it fell outside that range.

The estimate in part ii) is therefore more reliable.

v) Visitors to a seaside resort usually eat ice cream to cool down when the weather is hot, so there is a correlation **with causation** between outside temperature and ice cream sales.