

YEAR 1 STATISTICS · CHAPTER 4

Correlation

Specification: Edexcel Paper 3 §2.2 — Bivariate data · AQA section L2

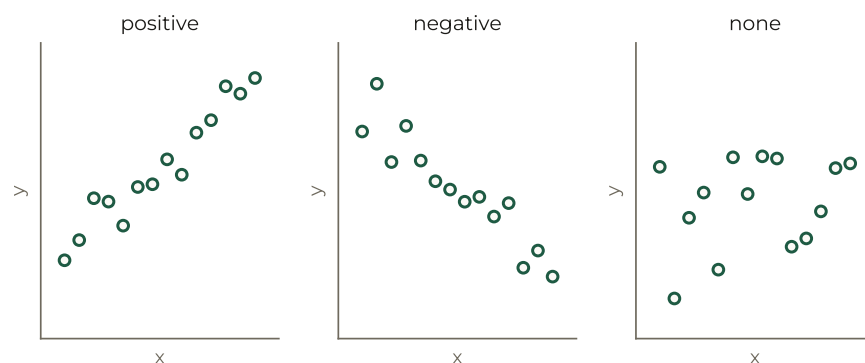
KEY FACTS — THIS CHAPTER

- **Bivariate** data pair two measurements for each item. Plot them on a **scatter diagram**, with the **explanatory** variable on the x -axis and the **response** variable on the y -axis.
- **Correlation** describes a linear pattern: **positive** (as one rises, so does the other), **negative** (one rises as the other falls), or **none**.
- A **regression line** $y = a + bx$ is the line of best fit. In Year 1 it is always **given** to you. The gradient b is the change in y per unit of x ; the intercept a is the value of y when $x = 0$.
- Predicting **inside** the range of the data is **interpolation** (reasonably reliable); predicting **outside** it is **extrapolation** (unreliable, because you have no evidence the pattern continues).
- **Correlation does not imply causation**. A third variable may explain both, or the link may be coincidence.
- Watch for **distinct subgroups**: a single line through two different groups can point the opposite way to each group.

So far each chapter has handled one variable at a time. Bivariate data come in pairs — a temperature and a sales figure, hours of revision and a mark — and the question becomes whether the two move together, and what you can safely predict from that.

1. Scatter diagrams and correlation

Plot each pair as a point. By convention the **explanatory** (independent) variable goes on the horizontal axis, and the **response** (dependent) variable, the one you think is being affected, on the vertical axis.



Positive, negative and no correlation. The closer the points lie to a straight line, the stronger the correlation.

KEY RESULT

Describe correlation with **two** words: its **direction** (positive or negative) and its **strength** (strong, moderate, weak) — then say what it means in context.

IN CONTEXT

“The scatter shows strong positive correlation: students who revised for longer generally scored higher marks.”

COMMON MISTAKE

“There is a correlation” on its own earns nothing. Name the direction, and describe it in the words of the question. And correlation is about a **linear** pattern: a clear curved relationship can have almost no linear correlation at all.

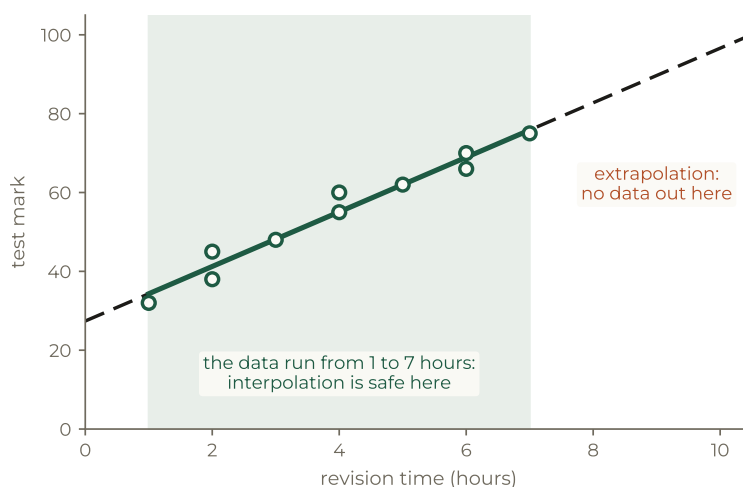
2. The regression line

The **regression line of y on x** is the straight line that best fits the data, chosen to make the vertical gaps between the points and the line as small as possible. You will always be **given** its equation — calculating it is Year 2 work.

INTERPRETING $Y = A + BX$

- b (the gradient) is the change in y for each **1 unit** increase in x .
- a (the intercept) is the value of y predicted when $x = 0$. It's only meaningful if $x = 0$ is near or inside the data.

Always interpret both **with units, in context**.



The solid line spans the data. Beyond it the line is a guess: nothing in the data says the pattern continues.

WORKED EXAMPLE 4.1 • USING A GIVEN LINE

For ten students, the regression line of test mark y on revision time x hours is $y = 27.4 + 6.92x$. The revision times in the sample ranged from 1 to 7 hours.

(a) Interpret the gradient. (b) Predict the mark of a student who revises for 3.5 hours. (c) Predict the mark for 10 hours and comment.

(a) Each extra hour of revision is associated with about **6.9 more marks**.

(b) $y = 27.4 + 6.92(3.5) = 51.6$, so about **52 marks**. This is **interpolation** — 3.5 hours is inside the range 1 to 7 — so it should be reasonably reliable.

(c) $y = 27.4 + 6.92(10) = 96.6$, so about 97 marks. But 10 hours is well outside the data, so this is **extrapolation** and unreliable: the improvement may level off with tiredness, and the model would happily predict over 100 marks for 11 hours.

IN CONTEXT

Use the line for students revising between 1 and 7 hours, and treat anything outside that range with caution.

COMMON MISTAKE

The intercept is often nonsense in context. Here $a = 27.4$ claims a student who revises for 0 hours scores 27 marks — but nobody in the sample revised for less than an hour, so that's extrapolation too. Say so when you interpret it.

3. Correlation and causation

A strong correlation does not mean one variable **causes** the other. There are three explanations to weigh every time.

KEY RESULT

1. One variable really does affect the other.
2. A **third variable** affects both. (Ice cream sales and sunburn cases rise together, but the weather drives both.)
3. **Coincidence**, particularly in small samples.

Only a controlled experiment, not a scatter diagram, can establish cause.

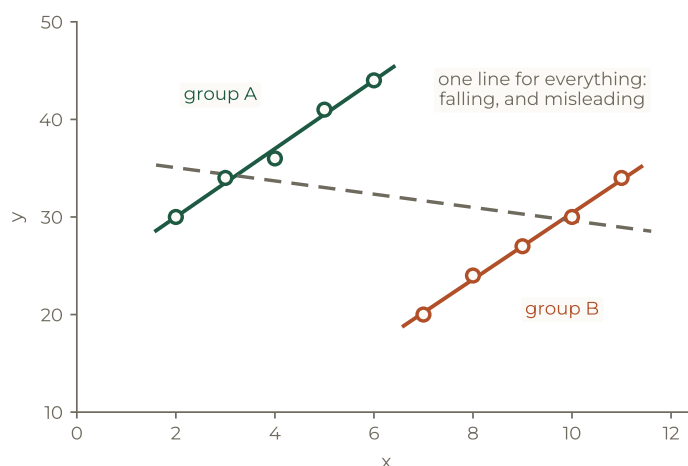
WORKED EXAMPLE 4.2 · JUDGING A CLAIM

A newspaper reports strong positive correlation between the number of hours children spend on homework and their height, and concludes that homework makes children taller. Comment.

The conclusion doesn't follow. **Age** explains both: older children are taller *and* are set more homework. Age is a third variable driving the two together, so the correlation is real but the causal claim is not.

4. Subgroups in one scatter

If a data set mixes two different groups, a single line can be badly misleading.



Both groups show positive correlation, yet one line through everything slopes down. Always look at the scatter, not just the line.

WHY THE LINE FLIPS

Group B sits to the right of group A but lower down. Fitting one line to the combined data mostly measures the *gap between the groups*, not the pattern inside them. It's a reminder that a regression line summarises whatever you feed it, sensibly or not.

IN CONTEXT

“The scatter shows two distinct groups. Within each group the correlation is positive, so a single regression line for all the data would be misleading. It would be better to fit a separate line to each group.”

Practice

Twelve questions in three tiers. Try each one properly before opening the solution. In this chapter every regression line is given: the marks are for using and criticising it, in context.

Warm-up

Question 1

[3 marks]

For each pair of variables, state the correlation you would expect and which variable you would plot on the x -axis.

- The age of a car and its value.
- A person's height and their shoe size.

- (c) A person's house number and their exam mark.

SOLUTION

- (a) Negative correlation: older cars are worth less. Age is the explanatory variable, so it goes on the x -axis.
- (b) Positive correlation: taller people tend to have larger feet. Height on the x -axis.
- (c) No correlation: the two are unrelated. (Either could go on the x -axis, since neither explains the other.)

Question 2

[3 marks]

The regression line of y on x is $y = 12.5 + 3.4x$.

- (a) Predict y when $x = 6$.
- (b) Interpret the gradient.

SOLUTION

- (a) $y = 12.5 + 3.4(6) = 32.9$.
- (b) For each increase of 1 in x , y increases by about 3.4.

Question 3

[3 marks]

A regression line is based on data with x values from 2 to 15. State whether using the line to predict y at each of these values is interpolation or extrapolation, and which prediction you would trust.

- (a) $x = 5$
- (b) $x = 20$

SOLUTION

- (a) Interpolation, since 5 lies inside the range 2 to 15. This prediction can be trusted (as far as the model fits).
- (b) Extrapolation, since 20 is outside the range. There is no evidence the pattern continues that far, so it is unreliable.

Question 4

[3 marks]

In a seaside town, weekly ice cream sales and the number of people treated for sunburn show strong positive correlation. A student concludes that eating ice cream causes sunburn. Explain why this conclusion is wrong.

SOLUTION

Correlation does not imply causation. A third variable — **hot, sunny weather** — increases both ice cream sales and the time people spend outside getting sunburnt. The two rise together without either causing the other.

Standard**Question 5**

[6 marks]

For ten days, a café records the midday temperature t °C and the number of cold drinks sold, s . The temperatures ranged from 12 °C to 26 °C, and the regression line is $s = 26.2 + 4.88t$.

- (a) Interpret the gradient in context.
- (b) Estimate the sales when the temperature is 19 °C.
- (c) Estimate the sales at 35 °C and comment on your answer.

SOLUTION

(a) For each 1 °C rise in midday temperature, the café sells about 4.9 more cold drinks.

(b) $s = 26.2 + 4.88(19) = 118.9$, so about 119 drinks. This is interpolation, so it should be reasonably reliable.

(c) $s = 26.2 + 4.88(35) = 197$, so about 197 drinks. But 35 °C is far outside the recorded range of 12–26 °C, so this is extrapolation. Sales might level off once the café runs out of stock or seats, so the estimate is unreliable.

Question 6

[4 marks]

The regression line of a car's value v (£) on its age a years is $v = 14\,500 - 1750a$, based on cars aged 1 to 9 years.

- (a) Interpret the gradient and the intercept in context.
- (b) Explain why the model cannot be used for a 12-year-old car.

SOLUTION

(a) The value falls by about £1750 for each extra year of age. The intercept suggests a new car (age 0) would be worth £14 500, though no car in the sample was new, so that is an extrapolation.

(b) 12 years is outside the range 1 to 9 years. The model would predict $14\,500 - 1750(12) = -£6500$, a negative value, which is impossible: the linear pattern clearly does not continue.

Question 7

[5 marks]

A scatter diagram of x against y shows two clearly separated clusters of points. Within each cluster the points rise from left to right, but a single regression line fitted to all the data has a negative gradient.

- (a) Explain how this can happen.
- (b) State what should be done instead.

SOLUTION

(a) The second cluster lies to the right of the first but lower down. A single line mostly measures the difference *between* the two groups rather than the pattern *within* them, so it slopes the opposite way.

(b) The data should be separated, with a regression line fitted to each group, and the two groups described separately. The scatter should always be looked at, not just the equation of the line.

Question 8

[5 marks]

The table shows the mass, m kg, and the length, ℓ cm, of eight fish.

Length, ℓ (cm)	18	22	25	27	30	33	36	40
Mass, m (kg)	0.9	1.3	1.6	1.9	2.4	2.8	3.3	4.1

Mass and length of eight fish

- (a) Describe the correlation and what it means here.
- (b) Which variable is the explanatory variable? Justify your choice.
- (c) A ninth fish is 24 cm long with mass 3.9 kg. Describe where it would lie on the scatter, and what you would do about it.

SOLUTION

- (a) Strong positive correlation: longer fish tend to be heavier.
- (b) Length. A fish's mass depends on its size, rather than the other way round, so length goes on the x -axis as the explanatory variable.
- (c) It would sit well above the pattern: at 24 cm the other fish weigh around 1.5 kg, so 3.9 kg is far heavier than expected. It should be investigated as a possible recording error before being included. If it is genuine, it should be kept and commented on, since it may be a different species.

Exam style**Question 9**

[8 marks]

Ten students' weekly study hours x and end-of-term scores y are recorded. The study times range from 5 to 28 hours, and the regression line is $y = 37.4 + 1.06x$.

- (a) Interpret the gradient in context.
- (b) Estimate the score of a student who studies 14 hours a week.
- (c) Estimate the score for 40 hours a week, and give two reasons to distrust it.
- (d) A student says the line proves that studying causes higher scores. Comment.

SOLUTION

- (a) Each extra hour of study per week is associated with about 1.06 more marks.
- (b) $y = 37.4 + 1.06(14) = 52.2$, so about 52 marks (interpolation, so reasonably reliable).
- (c) $y = 37.4 + 1.06(40) = 79.8$, so about 80 marks. Two reasons to distrust it: 40 hours is far outside the range 5–28 hours (extrapolation), and the relationship is unlikely to stay linear — extra hours probably have less and less effect, and tiredness may reduce performance.
- (d) The data show association, not cause. More able or better-motivated students may both study more *and* score higher, so a third factor could explain the link. Only a controlled experiment could establish cause.

Question 10

[5 marks]

A study finds that towns with more libraries per person have higher average exam results, and recommends building more libraries to raise results.

- (a) Give a possible third variable that could explain the correlation.

- (b) Explain the danger of acting on this recommendation.
- (c) Describe what evidence would be needed to support the claim.

SOLUTION

- (a) The wealth of the town. Wealthier areas can fund more libraries *and* tend to have better-resourced schools, which raises results.
- (b) If wealth is the real driver, building libraries alone may not improve results at all, and the money might be better spent directly on schools.
- (c) A controlled study: compare similar towns where libraries are added with ones where they are not, over time, keeping other factors as alike as possible.

Question 11

[5 marks]

The regression line of y on x for some data is $y = 5 + 2x$. A student wants to estimate the value of x when $y = 31$, so writes $31 = 5 + 2x$ and gets $x = 13$.

- (a) Explain why this is not the recommended way to estimate x from y .
- (b) State the circumstances in which the estimate would still be reasonable.

SOLUTION

- (a) The regression line of y on x is chosen to make the **vertical** distances from the points small, so it is designed for predicting y **from** x . Rearranging it to predict x from y is not the same as the line of x on y , which would be fitted the other way round, and it generally gives a different answer.
- (b) If the correlation is very strong, the two lines are close together, so the estimate is reasonable — provided $y = 31$ also lies inside the range of the observed y values.

Question 12

[6 marks]

A biologist measures the height of a plant each week for 12 weeks. The regression line of height h cm on week w is $h = 2.1 + 3.4w$, and the scatter shows the points lying close to a straight line for the first 8 weeks, then curving and flattening out.

- (a) Comment on the suitability of a linear model for these data.
- (b) The biologist predicts the height in week 20. Give two criticisms of this prediction.
- (c) Suggest how the model could be improved.

SOLUTION

(a) A straight line fits the first 8 weeks well, but the later points curve away from it. Over the whole 12 weeks, a linear model is not suitable: it will overestimate the later heights.

(b) Week 20 is outside the data (extrapolation), and the pattern was already bending away from the line by week 12, so the linear trend has clearly stopped. The prediction of $2.1 + 3.4(20) = 70.1$ cm is almost certainly far too high.

(c) Use a model that levels off, since plants approach a maximum height — for example fitting a curve, or a straight line only to the first 8 weeks while stating that it applies to that period alone.