



ANSWER BOOK

Correlation and regression

Statistics · Year 12–13

7 questions · 19 marks

NAVY EDITION

SECTION A · Warm-up

- A1** Strong negative correlation. As one variable rises the other falls, and the points sit close to a downhill straight line. B1 for the direction, B1 for the strength. The sign gives the direction and the closeness to ± 1 gives the strength, so both words are needed. [2]
- A2** The gradient 0.6 says that each extra hour of revision is associated with about 0.6 more marks. The intercept 3.5 says that a student who did no revision is predicted about 3.5 marks. B1 for the gradient in context, B1 for the intercept in context. Phrasing earns these marks. Write 'associated with' rather than 'causes', and name hours and marks rather than x and y. [2]

SECTION B · Standard

- B1** $y = 3.5 + 0.6 \times 10 = 9.5$ marks. Since 10 hours lies inside the observed range of 2 to 15 hours this is interpolation, so the prediction is not exposed to the extra risk of extrapolation and the data do at least speak to that part of the line. M1 for the substitution, A1 for 9.5 marks identified as interpolation. That is all the marks claim. Interpolation is not a guarantee. How much the number is worth also depends on how well a straight line fits, how tightly the points hug it, and how the data were collected. A weak r or an obvious curve would undermine 9.5 even at 10 hours. Note also that this line predicts y from x only. To go the other way and estimate the hours behind a known mark you would need the regression line of x on y, which is a different line. [2]
- B2** The arithmetic gives $3.5 + 24 = 27.5$ marks, and the number deserves no trust. 40 hours lies far outside the 2 to 15 hour window the line was fitted to, so this is extrapolation. Nothing in the data shows the straight-line pattern continues that far out, and in practice the returns to revision would flatten. B1 for 27.5, B1 for naming extrapolation, B1 for quoting the fitted range. The marks are for the refusal together with the reason, so name extrapolation and quote the range. [3]
- B3** It shows that the two rise and fall together. It does not show that either one causes the other. Hot weather drives both, bringing more swimming and more ice cream, so a lurking third variable has manufactured a correlation between two things that never touch. B1 for the association, B1 for ruling out causation, B1 for naming a plausible third variable. Correlation is not causation, and an exam answer has to name a plausible third variable to score in full. [3]

- B4** There is almost no linear relationship. The points form a shapeless cloud about any straight line. A relationship is not ruled out, though, because a strong curve can sit behind a near-zero r . B1 for almost no linear relationship, B1 for a relationship not being ruled out. The coefficient measures straightness and nothing else. [2]

SECTION C · Stretch

- C1** Take logs of the model first: $\log_{10}v = \log_{10}a + t \log_{10}b$. Comparing with the fitted line, $\log_{10}a = 0.8$ and $\log_{10}b = 0.12$, so $a = 10^{0.8} = 6.31$ and $b = 10^{0.12} = 1.32$. At $t = 5$, $\log_{10}v = 0.8 + 0.6 = 1.4$, so $v = 10^{1.4} = 25.1$ bacteria. M1 for taking logs of the model, A1 for $a = 6.31$, A1 for $b = 1.32$, M1 for substituting $t = 5$, A1 for 25.1. Substituting into $v = 6.31 \times 1.32^5$ gives the same 25.1, which is a sound check. Marks are lost by matching the coefficients to a and b directly instead of to their logarithms. The straight line lives in log space, so every constant read off it has to be raised as a power of 10 before it means anything about v . In context, $b = 1.32$ says the culture grows by about 32% each hour. [5]