inkMaths

ANSWER BOOK

Least squares regression and residuals

Further Statistics 2 · Year FM

7 questions · 26 marks

NAVY EDITION

inkphysics.co.uk/maths

SECTION A · Warm-up

A1 A residual is the observed y minus the y the line predicts at the same x: a vertical gap, not a perpendicular distance. The least squares line minimises the sum of the squares of those residuals over all the data points. B1 for the residual as a vertical gap, B1 for the sum of squared residuals. [2]

A2 $\Sigma x = 30$, $\Sigma y = 55$, $\Sigma x^2 = 220$ and $\Sigma xy = 388$, so $S_{xx} = 220 - 900/5 = 40$ and $S_{xy} = 388 - 330 = 58$. Then $b = 58/40 = 1.45$ and $a = 11 - 1.45 \times 6 = 2.3$, giving $\mathbf{y} = 2.3 + 1.45x$. M1 for S_{xx} and S_{xy}, A1 for the gradient, A1 for the intercept. [3]

SECTION B · Standard

B1 Fitted values are 5.2, 8.1, 11.0, 13.9 and 16.8, so the residuals are **-0.2,** 0.9, -1.0, 0.1 and 0.2. They sum to zero, as they must, which is a free check on the arithmetic. Squaring and adding: 0.04 + 0.81 + 1.00 + 0.01 + 0.04 = **1.9**. [4]
By formula, $S_{yy} = 691 - 605 = 86$, so the residual sum of squares is $S_{yy} - S_{xy}^2/S_{xx} = 86 - 58^2/40 = 86 - 84.1 = \mathbf{1.9}$. M1 for the fitted values, A1 for the five residuals, M1 for squaring and adding, A1 for the residual sum of squares.

B2 The gradient says that each extra hour of practice is associated with about **1.45** more marks. The intercept, 2.3 marks with no practice at all, comes from x = 0, which lies outside the observed range of 2 to 10 hours, so it should be quoted with care. At x = 7 the prediction is 2.3 + 10.15 = **12.45** marks. [4]
At x = 20 the model is being extrapolated to double the largest observation. There is no evidence that the relationship stays linear that far out, and in this context it cannot, since scores are bounded. B1 for interpreting the gradient, B1 for interpreting the intercept, B1 for the prediction, B1 for the extrapolation comment.

B3 It should show random scatter about zero with no trend and roughly constant spread. Systematic curvature, so that residuals are negative at the ends and positive in the middle or the reverse, means the relationship is not linear. A spread that widens with x means the assumption of constant variability fails. A single very large residual instead flags a possible outlier. B1 for random scatter about zero, B1 for curvature ruling out a linear model, B1 for a widening spread. [3]

B4 Σy becomes 68 and Σxy becomes 518, so $S_{xy} = 518 - 30 \times 68/5 = 518 - 408 = 110$, while S_{xx} is unchanged at 40 because the x values have not moved. [4]
Then $b = 110/40 =$ **2.75** and $a = 13.6 - 2.75 \times 6 =$ **−2.9**, giving $y = -2.9 + 2.75x$.
One faulty reading has nearly doubled the gradient and pushed the intercept below zero. The least-squares method squares the residuals, so a single large miss counts for far more than several small ones. M1 for recomputing the sums, A1 for the new S_{xy}, A1 for the new line, B1 for the comment on the effect. It is the reason for plotting the residuals before trusting a fit.

SECTION C · Stretch

C1 $\Sigma p = 0$, $\Sigma q = 5$, $\Sigma p^2 = 10$ and $\Sigma pq = 29$. Since $\Sigma p = 0$ the correction terms vanish: $S_{pp} =$ **10** and $S_{pq} =$ **29**. [6]
So the coded gradient is $29/10 = 2.9$, and with $\bar{p} = 0$ and $\bar{q} = 1$ the coded intercept is 1. The coded line is **q** $= 1 + 2.9p$.
Now decode. Substitute $q = y - 10$ and $p = (x - 6)/2$ into that line: $y - 10 = 1 + 2.9(x - 6)/2 = 1 + 1.45(x - 6) = 1.45x - 7.7$.
So **y** $= 2.3 + 1.45x$.
M1 for the coded sums, A1 for S_{pp} and S_{pq}, M1 for the coded gradient and intercept, A1 for the coded line, M1 for decoding, A1 for the line of y on x.
Read what the coding did. Dividing x by 2 multiplied the gradient by 2, so decoding halves it back to 1.45. Subtracting 10 from y and 6 from x moved the intercept, and only substituting back recovers it. Coding is chosen to make $\Sigma p = 0$ and the numbers small, and it never changes the fitted line, only the arithmetic needed to reach it.