

inkMaths

ANSWER BOOK

Least squares regression and residuals

Further Statistics 2 · Year FM

6 questions · 20 marks

NAVY EDITION

SECTION A · Warm-up

- A1** A residual is the observed y minus the y the line predicts at the same x : a vertical gap^[2] not a perpendicular distance. The least squares line minimises the sum of the squares of those residuals over all the data points.
- A2** $\Sigma x = 30$, $\Sigma y = 55$, $\Sigma x^2 = 220$ and $\Sigma xy = 388$, so $S_{xx} = 220 - 900/5 = 40$ and $S_{xy} = 388 - 330/5 = 58$. Then $b = 58/40 = 1.45$ and $a = 11 - 1.45 \times 6 = 2.3$, giving $\hat{y} = 2.3 + 1.45x$.

SECTION B · Standard

- B1** Predicted 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. Squaring and adding: $0.04 + 0.81 + 1.00 + 0.01 + 0.04 = 1.9$. By formula, $S_{yy} = 691 - 605 = 86$, so $RSS = 86 - 58^2/40 = 86 - 84.1 = 1.9$.
- B2** The gradient says that each extra hour of practice is associated with about 1.45 more^[4] marks. The intercept, 2.3 marks with no practice at all, is a value the line gives outside the observed range and should be treated with care. At $x = 7$ the prediction is $2.3 + 10.15 = 12.45$ marks^[3]. At $x = 20$ the model is being extrapolated to double the largest observation, where no evidence of linearity exists.
- B3** It should show random scatter about zero with no trend and roughly constant spread^[3]. 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 model's assumption of constant variability fails. A single very large residual instead flags a possible outlier.

SECTION C · Stretch

- C1** Σy becomes 68 and Σxy becomes 518, so $S_{xy} = 518 - 408 = 110$ while S_{xx} is unchanged at^[4] 40. Then $b = 110/40 = 2.75$ and $a = 13.6 - 16.5 = -2.9$. One faulty reading has nearly doubled the gradient and pushed the intercept negative, because least squares squares the residuals and so weights a large miss heavily. It is the reason for plotting residuals before trusting a fit.