End of Topic Test – Chapter 7: Correlation and Regression — Mark Scheme

Total: 92 marks

1. Number of pages and book weight

Pages x: 80, 130, 100, 140, 115, 90, 160, 140, 105, 150. Weight y (g): 160, 270, 180, 290, 230, 180, 320, 270, 210, 300. ($n = 10$, $\bar{x} = 121$, $\bar{y} = 241$.)

(a) M1 correct scales chosen on both axes covering the data range. A1 at least 8 of the 10 points plotted correctly (±1 small square). A1 all 10 points plotted correctly, no line drawn. *[3]*

(b) B1 positive direction (as pages increase, weight increases). B1 fairly strong linear relationship. B1 contextual comment (heavier books tend to have more pages, which makes sense as more paper adds mass). *[3]*

(c) M1 correct method/formula for r used (by calculator or from summary statistics). M1 $S_{xx} = \sum(x - \bar{x})^2 = 6540$. M1 $S_{yy} = \sum(y - \bar{y})^2 = 28890$. M1 $S_{xy} = \sum(x - \bar{x})(y - \bar{y}) = 13590$. A1 $r = \dfrac{13590}{\sqrt{6540 \times 28890}} = 0.989$ (3 d.p.). B1 interpretation: a strong, positive correlation between number of pages and weight. *[6]*

(d) M1 regression equation used: $b = S_{xy}/S_{xx} = 13590/6540 = 2.078$, $a = \bar{y} - b\bar{x} = 241 - 2.078(121) = -10.44$, so $\hat{y} = -10.44 + 2.078x$. M1 substitute $x = 125$: $\hat{y} = -10.44 + 2.078(125)$. A1 $\hat{y} \approx 249$ g. B1 this is interpolation, since $x = 125$ lies within the range of the given data ($80 \leq x \leq 160$). *[3]*

(e) B1 reason 1, e.g. natural variation between books of the same page count (paper type/density, cover material, binding) affects weight independently of page count. B1 reason 2, e.g. measurement/recording error in pages or weight; or other factors (hardcover vs. paperback, page thickness/quality) not captured by a simple linear model. *[3]*

2. Judging consistency in an art competition

Judge X: 40,46,34,49,49,37,41,36,43,38. Judge Y: 40,39,45,34,35,38,39,43,40,42. Judge Z: 42,42,37,46,45,39,41,37,4
($n = 10$.)

(a)–(c) M1 correct scale on both axes (30–60) for each of the three scatter graphs. A1 points plotted correctly (allow ±1 small square) for each graph. *[3 each, 9 total]*

(d) M1 correct method for r (calculator or summary statistics) applied to both pairs. A1 $r(X, Y) = -0.861$ (3 d.p.). A1 $r(X, Z) = 0.920$ (3 d.p.). B1 comment: X and Y show a strong negative correlation, while X and Z show a strong positive correlation. *[4]*

(e) M1 recognise X and Z agree closely (strong positive correlation, $r \approx 0.920$). M1 recognise Y correlates negatively with both X ($r = -0.861$) and Z (which can also be shown to be strongly negative, $r(Y, Z) \approx -0.818$). A1 numerical evidence correctly cited to support the argument. B1 conclusion: Judge Y is marking inconsistently with the other two judges (tends to score paintings oppositely to X and Z). B1 possible explanation offered, e.g. Judge Y may be using different/reversed criteria, or applying a systematically different standard. B1 overall clarity and quality of the statistical argument. *[6]*

(f) B1 correlation measures the strength of a *pattern*/relative ordering of marks, not overall agreement in absolute level. B1 example: a judge could be perfectly correlated with the others but consistently score everyone, say, 10 marks higher or lower (an additive bias), which correlation would not detect. B1 conclusion: correlation alone is not sufficient evidence of fairness; absolute differences should also be considered. *[3]*

3. Area and population of Central American countries

Country data (Area km^2, Population thousands): Belize (22966, 340.84), Costa Rica (51100, 4755.23), El Salvador (21041, 6125.51), Guatemala (108889, 14647.08), Honduras (112090, 8598.56), Mexico (1964375, 120286.66), Nicaragua (130370, 5848.64), Panama (75420, 3608.43). ($n = 8$.)

(a) B1+B1+B1+B1 (one per point) the remaining four points plotted correctly, at a consistent scale with the four already shown. *[4]*

(b) M1 correct method for r used. A1 $r \approx 0.997$ (3 d.p.). B1 direction: positive. B1 strength: very strong (Mexico, with by far the largest area and population, dominates and inflates this value – see part (d)). *[4]*

(c) M1 $b = \dfrac{S_{xy}}{S_{xx}} \approx 0.0603$. M1 $a = \bar{y} - b\bar{x} \approx 1778.34$. A1 $b \approx 0.06$ (2 d.p.). A1 $a \approx 1778.34$ (2 d.p.). B1 written in the required form: $\hat{y} = 1778.34 + 0.06x$. *[5]*

(d) B1 Mexico is a high-leverage outlier (far from the other 7 points in both x and y), and dominates the sums used to calculate r, so it inflates r towards 1 – making the overall relationship look stronger than it is for typical-sized countries (removing Mexico reduces r to only around 0.59). B1 (development) so $r = 0.997$ overstates how strongly area and population are linearly related among "ordinary" countries. B1 Mexico also has a strong influence (leverage) on the slope of the regression line, pulling the fitted line strongly towards fitting this one point. B1 (development) the line may therefore not represent the area–population relationship well for the other, more typical, countries. *[4]*

(e) M1 substitute $x = 100{,}000$ into the regression equation (using the more precise, unrounded gradient for accuracy: $\hat{y} = 1778.34 + 0.0603 \times 100{,}000$). A1 $\hat{y} \approx 7{,}811$ thousand (≈ 7.8 million). B1 comment on reliability: although $x = 100{,}000$ is technically within the range of the data (interpolation), the fitted line is heavily distorted by the Mexico outlier, so this prediction should be treated with caution/may not be reliable for a "typical" country of this size. *[3]*

(f) B1+B1+B1 (any three) valid limitations, e.g. population depends on many other factors besides area (economic development, urbanisation, fertility/birth rates, history) not captured by a simple linear model; the fit is dominated by a single extreme outlier (Mexico), so it may not represent the relationship for other countries; a linear model could give unrealistic (e.g. negative) predictions for very small areas. *[3]*

4. Comparing classes: interpretation and residuals

Class A: $\hat{y} = 48 + 2.4x$; Class B: $\hat{y} = 52 + 1.9x$.

(a) B1 Class A: intercept 48 = predicted score with 0 hours' revision (baseline); slope 2.4 = each extra hour of revision is associated with a 2.4-mark increase in score. B1 Class B: intercept 52 = predicted baseline score with 0 hours' revision; slope 1.9 = each extra hour is associated with a 1.9-mark increase. B1 both interpreted in context with correct units (marks, and marks per hour). *[3]*

(b) B1 Class A has the greater slope. B1 $2.4 > 1.9$ correctly compared. B1 so each additional hour of revision is associated with a bigger increase in score in Class A than in Class B. *[3]*

(c) M1 Class A prediction: $\hat{y} = 48 + 2.4(6) = 62.4$. M1 Class B prediction: $\hat{y} = 52 + 1.9(6) = 63.4$. A1 residuals: Class A $= 65 - 62.4 = +2.6$; Class B $= 65 - 63.4 = +1.6$. A1 the student performed above expectation for *both* models, but relatively more so compared with Class A's model (larger positive residual, $+2.6 > +1.6$). *[4]*

(d) B1 reason 1 (e.g. cohort ability – Class B may have had a higher-ability intake, reflected in its higher intercept of 52). B1 explanation of its effect (raises the intercept, not necessarily the slope). B1 reason 2 (e.g. teaching method – Class A's teaching may make each hour of independent revision more productive). B1 explanation of its effect (raises the slope). B1 overall quality/coherence of

the explanation (e.g. also acknowledging possible selection effects or differing ranges of x between classes). *[5]*

5. Salary and experience: prediction and model assessment

$\hat{y} = 22.5 + 1.8x$ (salary in £000s), $r = 0.76$.

(a) B1 intercept: 22.5 (£22,500) is the predicted starting salary for someone with 0 years' experience. B1 slope: for each extra year of experience, predicted salary increases by £1,800 on average. B1 both stated clearly in context with correct units. *[3]*

(b) M1 substitute $x = 25$: $\hat{y} = 22.5 + 1.8(25)$. A1 = 67.5, i.e. £67,500. B1 caution: 25 years may be beyond (or at the edge of) the range of experience in the original sample, so this could be extrapolation, which is less reliable. B1 further caution: $r = 0.76$ shows a reasonably strong but imperfect fit, and salary growth may not stay linear indefinitely (e.g. it may plateau), so the prediction should be treated with care. *[4]*

(c) B1 variable 1 (e.g. level of education/qualifications) with brief justification (affects earning potential independently of years of experience). B1 variable 2 (e.g. industry sector, job role/seniority, or geographic location) with brief justification. B1 (development) explaining how including these in a multiple regression could improve predictive accuracy. *[3]*

(d) B1 limitation 1, e.g. omitted-variable bias – a single predictor ignores other relevant factors. B1 how to address it, e.g. use multiple regression including additional variables such as those in part (c). B1 limitation 2, e.g. assumes a constant linear rate of salary growth, which may not hold at high experience levels (diminishing returns/career plateau). B1 how to address it, e.g. use a non-linear model, or fit separate models/bands for different experience ranges. *[4]*