

Product Moment Correlation Coefficient

1 (2 marks)

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{48}{\sqrt{40 \times 90}}$$

M1

$$= \frac{48}{60} = 0.8$$

A1

Answer: $r = 0.8$

2 marks

2 (3 marks)

$$S_{xx} = 380 - \frac{52^2}{8} = 42.0, S_{yy} = 920 - \frac{80^2}{8} = 120$$

M1

$$S_{xy} = 588 - \frac{52 \times 80}{8} = 68.0$$

M1

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \approx 0.958$$

A1

Answer: $r \approx 0.958$

3 marks

3 (3 marks)

(i) Strong negative correlation

B1

(ii) Almost no linear correlation

B1

(iii) Moderate positive correlation

B1

Answer: (i) strong negative; (ii) none; (iii) moderate positive

3 marks

4 (2 marks)

A shows y increasing as x increases: $r \approx 0.95$; B shows y decreasing as x increases: $r \approx -0.95$

B1

In both cases the points lie close to a straight line (strong linear correlation)

B1

Answer: A: 0.95; B: -0.95

2 marks

5 (2 marks)

$$-1 \leq r \leq 1$$

B1

$r = 1$: all points lie exactly on a straight line with positive gradient; $r = -1$: exactly on a straight line with negative gradient

B1

Answer: $-1 \leq r \leq 1$

2 marks

6 (3 marks)

$$\Sigma x = 10, \Sigma y = 16, \Sigma x^2 = 30, \Sigma y^2 = 74, \Sigma xy = 47; S_{xx} = 5, S_{yy} = 10, S_{xy} = 7$$

M1 A1

$$r = \frac{7}{\sqrt{50}} \approx 0.990$$

A1

Answer: $r \approx 0.990$

3 marks

7 (4 marks)

The coding is linear, so r for (x, y) equals r for (u, w)

E1

$$S_{uu} = 146 - \frac{26^2}{6} = 33.3, S_{ww} = 579 - \frac{49^2}{6} = 179$$

M1

$$S_{uw} = 289 - \frac{26 \times 49}{6} = 76.7, r = \frac{S_{uw}}{\sqrt{S_{uu}S_{ww}}} \approx 0.993$$

M1 A1

Answer: $r \approx 0.993$

4 marks

8 (3 marks)

Positive linear coding does not change r : (a) 0.8

B1

(b) Multiplying by a negative number changes the sign: -0.8

B1

(c) Both multipliers are positive: 0.8

B1

Answer: (a) 0.8; (b) -0.8; (c) 0.8

3 marks

9 (3 marks)

$$r = \frac{\text{covariance}}{s_x s_y} \quad \text{M1}$$

$$s_x = 4, s_y = 5 \quad \text{A1}$$

$$r = \frac{12}{4 \times 5} = 0.6 \quad \text{A1}$$

Answer: $r = 0.6$ 3 marks

10 (5 marks)

$$\text{(a) } \Sigma x = 37, \Sigma y = 132, \Sigma x^2 = 279, \Sigma y^2 = 3106, \Sigma xy = 713 \quad \text{B1}$$

$$S_{xx} = 50.8, S_{yy} = 202, S_{xy} = -101 \quad \text{M1 A1}$$

$$r \approx -0.997 \quad \text{A1}$$

Answer: $r \approx -0.997$ 4 marks

(b) Very strong negative linear correlation: y decreases as x increases E1

Answer: strong negative correlation 1 mark

11 (4 marks)

$$S_{xy} = r \sqrt{S_{xx} S_{yy}} \quad \text{M1}$$

$$= 0.8 \times 30 = 24 \quad \text{A1}$$

$$b = \frac{S_{xy}}{S_{xx}} \quad \text{M1}$$

$$= \frac{24}{25} = 0.96 \quad \text{A1}$$

Answer: $S_{xy} = 24$; $b = 0.96$ 4 marks

12 (4 marks)

(a) The point near (11, 25) lies far from the pattern of the others B1

It weakens the (otherwise strong, negative) correlation, pulling r towards zero E1

Answer: (11, 25); weakens correlation 2 marks

(b) Without the outlier the points lie close to a line with negative gradient E1

So r would be closer to -1 ($r \approx -0.987$ without the outlier, $r \approx 0.170$ with it) E1

Answer: yes 2 marks

13 (3 marks)

There is a fairly strong positive linear correlation: students who revise for longer tend to score higher marks B1

Correlation does not imply causation E1

Another factor, such as ability or motivation, could cause both E1

Answer: strong positive; correlation is not causation 3 marks

14 (5 marks)

$$S_{xx} = 420 - \frac{60^2}{10} = 60.0 \quad \text{M1}$$

$$S_{yy} = 1849 - \frac{129^2}{10} = 185 \quad \text{A1}$$

$$S_{xy} = 874 - \frac{60 \times 129}{10} = 100 \quad \text{M1}$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \approx 0.949 \quad \text{A1}$$

Strong positive linear correlation E1

Answer: $r \approx 0.949$; strong positive 5 marks

15 (6 marks)

$$\text{(a) } S_{xx} = 524, S_{yy} = 262, S_{xy} = 102 \quad \text{M1}$$

$$r \approx 0.274 \quad \text{A1}$$

Answer: $r \approx 0.274$ 2 marks

$$\text{(b) New sums: } \Sigma x = 83, \Sigma y = 128, \Sigma x^2 = 901, \Sigma y^2 = 2066, \Sigma xy = 1361 \quad \text{M1 A1}$$

$$S_{xx} = 136, S_{yy} = 246, S_{xy} = 181$$

M1

$$r \approx 0.990$$

A1

Answer: $r \approx 0.990$

4 marks

16 (4 marks)

$$u - \bar{u} = b(x - \bar{x}) \text{ and } w - \bar{w} = d(y - \bar{y})$$

B1

$$S_{uu} = b^2 S_{xx}, S_{ww} = d^2 S_{yy}, S_{uw} = bd S_{xy}$$

M1 A1

$$r_{uw} = \frac{bd S_{xy}}{\sqrt{b^2 d^2 S_{xx} S_{yy}}} = \frac{bd S_{xy}}{bd \sqrt{S_{xx} S_{yy}}} = r_{xy}$$

A1

Answer: shown

4 marks

17 (3 marks)

$r = 1$ means that all the points lie on a straight line

E1

Gradient from (1, 3) to (2, 5) is 2, so the line is $y = 2x + 1$

M1

$$k = 2 \times 3 + 1 = 7$$

A1

Answer: $k = 7$

3 marks

18 (5 marks)

$$\Sigma x = 0, \Sigma y = 10, \Sigma xy = -8 - 1 + 0 + 1 + 8 = 0$$

M1 A1

$$S_{xy} = 0 - \frac{0 \times 10}{5} = 0, \text{ so } r = 0$$

A1

r measures only linear correlation

E1

The points follow the exact curve $y = x^2$, which is not a straight line

E1

Answer: $r = 0$; relationship is not linear

5 marks

19 (6 marks)

Coded values: $u = 2, 4, 6, 8, 10, 12$; $w = 12, 35, 41, 73, 80, 111$

M1

$$\Sigma u = 42, \Sigma w = 352, \Sigma u^2 = 364, \Sigma w^2 = 27100, \Sigma uw = 3126$$

A1

$$S_{uu} = 70.0, S_{ww} = 6450, S_{uw} = 662$$

M1 A1

Linear coding does not alter r : $r \approx 0.985$

M1 A1

Answer: $r \approx 0.985$

6 marks

20 (5 marks)

$$\text{Gradient } b = \frac{S_{xy}}{S_{xx}}, \text{ so } S_{xy} = 1.5 \times 16 = 24$$

M1 A1

$$r^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

M1

$$0.81 = \frac{24^2}{16 S_{yy}}$$

A1

$$S_{yy} = \frac{576}{16 \times 0.81} = 44.4$$

A1

Answer: $S_{xy} = 24$; $S_{yy} = 44.4$ (= 400/9)

5 marks

21 (8 marks)

(a) Group A: $S_{xx} = 17.5, S_{yy} = 7.33, S_{xy} = 11.0, r_A \approx 0.971$

M1 A1

Group B: $S_{xx} = 17.5, S_{yy} = 7.33, S_{xy} = 11.0, r_B \approx 0.971$

A1

Answer: $r_A \approx 0.971$; $r_B \approx 0.971$

3 marks

(b) Combined: $n = 12, \Sigma x = 72, \Sigma y = 86, \Sigma x^2 = 542, \Sigma y^2 = 778, \Sigma xy = 643$

M1 A1

$$S_{xx} = 110, S_{yy} = 162, S_{xy} = 127$$

M1

$$r \approx 0.952$$

A1

Answer: $r \approx 0.952$

4 marks

(c) The combined r is higher than either separate value, because the two groups occupy different regions of the (x, y) plane and the overall trend is stronger than within each group

E1

Answer: combined r is stronger

1 mark

22 (6 marks)

(a) S_{xx} should be $190 - \frac{30^2}{6}$: the student has omitted the division by n E1

r should be $\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$: the student has not taken the square root E1

Answer: missing $\div n$ in S_{xx} ; no square root 2 marks

(b) $S_{xx} = 190 - 150 = 40$ M1

$$r = \frac{48}{\sqrt{40 \times 46}}$$

$$\approx 1.12$$

Answer: $r \approx 1.12$ A1

(c) S_{xy} is positive, so r must be positive; also S_{xx} cannot be negative 3 marks

Answer: S_{xx} cannot be negative E1

1 mark

23 (6 marks)

$$\bar{x} = 5, \bar{y} = 9$$

B1

$$S_{xx} = 155 - \frac{25^2}{5} = 30.0, S_{yy} = 469 - \frac{45^2}{5} = 64.0, S_{xy} = 267 - \frac{25 \times 45}{5} = 42.0$$

M1 A1

$$r = \frac{42.0}{\sqrt{30.0 \times 64.0}} \approx 0.959$$

A1

The new point contributes $(5 - 5)^2 = 0$, $(9 - 9)^2 = 0$ and 0 to S_{xx} , S_{yy} and S_{xy} , and the new means are unchanged M1

So r is unchanged: $r \approx 0.959$ A1

Answer: $r \approx 0.959$ both times (the new point adds nothing) 6 marks

24 (5 marks)

$$b = \frac{S_{xy}}{S_{xx}} \text{ and } r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

B1

$$\text{So } S_{xy} = r\sqrt{S_{xx}S_{yy}} \text{ and } b = r\frac{\sqrt{S_{xx}S_{yy}}}{S_{xx}} = r\sqrt{\frac{S_{yy}}{S_{xx}}} = r\frac{s_y}{s_x}$$

M1 A1

$$1.2 = r \times \frac{4}{2.5}$$

M1

$$r = \frac{1.2 \times 2.5}{4} = 0.75$$

A1

Answer: $r = 0.75$ 5 marks

25 (6 marks)

$$(a) s_x = \sqrt{\frac{S_{xx}}{n}}, s_y = \sqrt{\frac{S_{yy}}{n}}$$

B1

$$\Sigma z_x z_y = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{s_x s_y} = \frac{S_{xy}}{s_x s_y}$$

M1 A1

$$\frac{1}{n} \Sigma z_x z_y = \frac{S_{xy}}{n s_x s_y} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = r$$

A1

Answer: shown 4 marks

(b) Standardised values do not depend on the units or origin: z_x is unchanged by a change of units M1

So $r = \frac{1}{n} \Sigma z_x z_y$ is unchanged A1

Answer: shown 2 marks

26 (7 marks)

$$(a) r = \frac{-54}{\sqrt{135 \times 60}} = \frac{-54}{\sqrt{8100}}$$

M1

$$= -\frac{54}{90} = -0.6$$

A1

Answer: $r = -0.6$ 2 marks

(b) p decreases as x increases (coefficient $-\frac{1}{3}$) and q increases with y (coefficient 2) M1

So the sign of r is reversed: $r_{pq} = +0.6$

Answer: $r = 0.6$

A1

2 marks

$$(c) s_x = \sqrt{\frac{135}{15}} = 3, s_y = \sqrt{\frac{60}{15}} = 2$$

M1 A1

$$b = \frac{S_{xy}}{S_{xx}} = \frac{-54}{135} = -0.4$$

A1

Answer: $s_x = 3$; $s_y = 2$; $b = -0.4$

3 marks

27 (9 marks)

(a) Expand: $\lambda^2 S_{xx} + 2\lambda S_{xy} + S_{yy} \geq 0$ for all λ

M1 A1

A quadratic that is never negative has discriminant ≤ 0

E1

$$4S_{xy}^2 - 4S_{xx}S_{yy} \leq 0, \text{ so } S_{xy}^2 \leq S_{xx}S_{yy}$$

M1 A1

Dividing by $S_{xx}S_{yy} > 0$: $r^2 \leq 1$, so $-1 \leq r \leq 1$

A1

Answer: shown

6 marks

(b) $r^2 = 1$ means the discriminant is zero, so the quadratic has a repeated root λ_0

M1

$$\Sigma[\lambda_0(x_i - \bar{x}) + (y_i - \bar{y})]^2 = 0, \text{ a sum of squares equal to zero}$$

M1

Each term is zero: $y_i - \bar{y} = -\lambda_0(x_i - \bar{x})$, a straight line

A1

Answer: shown

3 marks

28 (8 marks)

$$(a) u - \bar{u} = (x - \bar{x}) + (y - \bar{y}), w - \bar{w} = (x - \bar{x}) - (y - \bar{y})$$

B1

$$S_{uw} = \Sigma[(x - \bar{x})^2 - (y - \bar{y})^2]$$

M1 A1

$$= S_{xx} - S_{yy} = 0$$

M1

$$\text{So } r_{uw} = \frac{S_{uw}}{\sqrt{S_{uu}S_{ww}}} = 0$$

A1

Answer: shown

5 marks

(b) y values are a rearrangement of x values, so $S_{xx} = S_{yy}$

B1

$$u = 5, 13, 5, 14, 13; w = -3, -7, 3, 0, 7; \Sigma u = 50, \Sigma w = 0, \Sigma uw = 0, \text{ so } S_{uw} = 0 - \frac{50 \times 0}{5} = 0$$

M1

$r_{xy} = -0.160$ (non-zero, so the result holds whatever r_{xy} is)

A1

Answer: $S_{uw} = 0$; $r_{xy} \approx -0.160$

3 marks

29 (7 marks)

$$S_{xy'} = \Sigma(x - \bar{x})[(y - \bar{y}) + (e - \bar{e})] = S_{xy} + S_{xe} = S_{xy}$$

M1 A1

$$S_{y'y'} = S_{yy} + 2S_{ye} + S_{ee} = S_{yy} + S_{ee}$$

M1 A1

$$r' = \frac{S_{xy}}{\sqrt{S_{xx}(S_{yy} + S_{ee})}} = r\sqrt{\frac{S_{yy}}{S_{yy} + S_{ee}}}$$

A1

$$r' = 0.8 \times \sqrt{\frac{64}{100}} = 0.8 \times 0.8 = 0.64$$

A1

Measurement error always reduces the apparent strength of the correlation (attenuation)

E1

Answer: $r' = 0.64$; error weakens correlation

7 marks

30 (9 marks)

$$(a) \Sigma x = 15, \Sigma x^2 = 55, S_{xx} = 55 - 45 = 10$$

B1

$$\Sigma y = 25 + a, \Sigma xy = 90 + 3a, \text{ so } S_{xy} = 90 + 3a - 3(25 + a) = 15$$

M1 A1

$$\Sigma y^2 = 179 + a^2, S_{yy} = 179 + a^2 - \frac{(25 + a)^2}{5} = \frac{4a^2}{5} - 10a + 54$$

A1

$$\text{Answer: } S_{xx} = 10; S_{xy} = 15; S_{yy} = \frac{4a^2}{5} - 10a + 54$$

4 marks

(b) $r > 0.9$ requires $r^2 > 0.81$ (r is positive, since $S_{xy} = 15 > 0$)

M1

$$\frac{15^2}{10(0.8a^2 - 10a + 54)} > 0.81$$

M1

$$0.8a^2 - 10a + 54 < \frac{225}{8.1}, \text{ i.e. } 0.8a^2 - 10a + 26.222 < 0$$

A1

Roots $a = 3.74$ and $a = 8.76$

M1

$3.74 < a < 8.76$

A1

Answer: $3.74 < a < 8.76$

5 marks

Rank Correlation

1 (2 marks)

Highest to lowest: 75, 69, 62, 53, 48

M1

Ranks of 62, 48, 75, 53, 69 are 3, 5, 1, 4, 2

A1

Answer: 3, 5, 1, 4, 2

2 marks

2 (2 marks)

$$r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 30}{8 \times 63}$$

M1

$$= 1 - \frac{180}{504} = \frac{9}{14} \approx 0.643$$

A1

Answer: $\frac{9}{14} \approx 0.643$

2 marks

3 (2 marks)

Ordered: 3, 3, 7, 7, 9; the two 3s share ranks 1 and 2 and the two 7s share ranks 3 and 4

M1

Ranks: 3.5, 1.5, 3.5, 5, 1.5

A1

Answer: 3.5, 1.5, 3.5, 5, 1.5

2 marks

4 (3 marks)

(i) Perfect agreement between the two rankings

B1

(ii) Strong disagreement: items ranked high in one list tend to be ranked low in the other

B1

(iii) Almost no association between the rankings

B1

Answer: perfect agreement; strong negative; none

3 marks

5 (3 marks)

$d = -1, 1, -1, 1, 0$ so $\sum d^2 = 4$

M1 A1

$$r_s = 1 - \frac{6 \times 4}{5 \times 24} = 0.8$$

A1

Answer: $r_s = 0.8$

3 marks

6 (3 marks)

Ranks 1 to 7 against 7 to 1: $d = -6, -4, -2, 0, 2, 4, 6$

M1

$$\sum d^2 = 2(36 + 16 + 4) = 112$$

A1

$$r_s = 1 - \frac{6 \times 112}{7 \times 48} = 1 - 2 = -1$$

A1

Answer: shown

3 marks

7 (5 marks)

Test 1 ranks: 7, 3, 8, 2, 5, 1, 4, 6

M1

Test 2 ranks: 6, 4, 8, 2, 7, 1, 3, 5

A1

$d = 1, -1, 0, 0, -2, 0, 1, 1$

M1

$$\sum d^2 = 8, r_s = 1 - \frac{6 \times 8}{8 \times 63}$$

M1

≈ 0.905

A1

Answer: $r_s \approx 0.905$

5 marks

8 (4 marks)

$$d = 1, -2, 2, -2, 2, -1$$

M1

$$\Sigma d^2 = 18$$

A1

$$r_s = 1 - \frac{6 \times 18}{6 \times 35} = \frac{17}{35}$$

M1

≈ 0.486 : fairly strong agreement

A1

$$\text{Answer: } r_s = \frac{17}{35} \approx 0.486$$

4 marks

9 (5 marks)

(a) The ranks of y are identical to the ranks of x, so $d = 0$ for every pair

M1

$$\Sigma d^2 = 0 \text{ and } r_s = 1$$

A1

$$\text{Answer: } r_s = 1$$

2 marks

(b) y always increases with x, so there is perfect monotonic (rank) agreement

E1

The relationship is a curve, not a straight line, so the PMCC is below 1

E1

Spearman's coefficient measures monotonic association, the PMCC measures linear association

E1

Answer: monotonic vs linear

3 marks

10 (5 marks)

x ranks: 1, 2.5, 2.5, 4, 5, 6 (the two 12s share ranks 2 and 3)

M1 A1

y ranks: 1, 3, 2, 5, 4, 6

A1

$$\Sigma d^2 = \frac{5}{2}$$

M1

$$r_s \approx 1 - \frac{6 \times \frac{5}{2}}{210} = \frac{13}{14} \approx 0.929$$

A1

$$\text{Answer: } r_s \approx 0.929$$

5 marks

11 (4 marks)

$$\frac{11}{14} = 1 - \frac{6 \times 12}{n(n^2 - 1)}$$

M1

$$\frac{72}{n(n^2 - 1)} = \frac{3}{14}, \text{ so } n(n^2 - 1) = 336$$

M1 A1

$$n = 7 \text{ (since } 7 \times 48 = 336)$$

A1

$$\text{Answer: } n = 7$$

4 marks

12 (5 marks)

The ranks 1 to 6 must each be used once; 1, 2, 3, 4, 6 are used, so E has rank 5

B1

Taster 1 ranks A to F: 3, 1, 5, 2, 6, 4 and taster 2 ranks A to F: 2, 1, 4, 3, 5, 6; $d = 1, 0, 1, -1, 1, -2$

M1 A1

$$\Sigma d^2 = 1 + 0 + 1 + 1 + 1 + 4 = 8$$

A1

$$r_s = 1 - \frac{6 \times 8}{6 \times 35} = \frac{27}{35} \approx 0.771$$

A1

$$\text{Answer: } E = 5; r_s = \frac{27}{35} \approx 0.771$$

5 marks

13 (4 marks)

(i) Spearman: the data are already ranks

B1

(ii) PMCC: the data are (approximately) bivariate normal with a linear relationship

B1 B1

(iii) Spearman: the relationship is monotonic but not linear

B1

Answer: (i) Spearman; (ii) PMCC; (iii) Spearman

4 marks

14 (7 marks)

x ranks: 2, 5, 5, 1, 8, 3, 7, 5 (three 2s share ranks 4, 5, 6 and so have rank 5)

M1 A1

y ranks: 2, 5, 3.5, 1, 8, 3.5, 7, 6 (two 3.8s share ranks 3 and 4)

A1

$$\Sigma r_x = \Sigma r_y = 36, \Sigma r_x^2 = 202, \Sigma r_y^2 = 204, \Sigma r_x r_y = 201$$

M1

$$S_{xx} = 40.0, S_{yy} = 41.5, S_{xy} = 39.0$$

M1 A1

$$r_s = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \approx 0.957$$

A1

Answer: $r_s \approx 0.957$

7 marks

15 (7 marks)

A and B: $d = -1, 1, 0, -1, 1, 0$, so $\Sigma d^2 = 4$

M1 A1

$$r_s(A, B) = 1 - \frac{6 \times 4}{210} \approx 0.886$$

A1

A and C: $\Sigma d^2 = 16$, $r_s(A, C) \approx 0.543$

M1 A1

B and C: $\Sigma d^2 = 12$, $r_s(B, C) \approx 0.657$

A1

Judges A and B agree most closely

A1

Answer: A, B: 0.886; A, C: 0.543; B, C: 0.657. A and B agree most closely

7 marks

16 (6 marks)

x ranks: 1, 2.5, 2.5, 4, 5.5, 5.5, 7

M1 A1

y ranks: 2, 1, 3, 4, 6, 5, 7

A1

$$\Sigma r_x^2 = 139, \Sigma r_y^2 = 140, \Sigma r_x r_y = 138$$

M1

$$S_{xx} = 27.0, S_{yy} = 28.0, S_{xy} = 25.5$$

M1

$$r_s = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{25.5}{\sqrt{27.0 \times 28.0}} \approx 0.927$$

A1

Answer: $r_s \approx 0.927$

6 marks

17 (4 marks)

There is a strong positive association: applicants ranked highly at interview tend to be ranked highly in the aptitude test

B1 E1

Spearman's coefficient uses ranks only, so it ignores the size of the gaps between scores

E1

Eight applicants is a small sample, so the result may not be reliable

E1

Answer: strong positive association; small sample

4 marks

18 (5 marks)

$d = -1, 1, 0, 0, 0, -1, 1$, so $\Sigma d^2 = 4$

M1 A1

$$r_s = 1 - \frac{6 \times 4}{8 \times 63} = 1 - \frac{24}{504} = \frac{20}{21} \approx 0.952$$

M1 A1

Each adjacent swap contributes $d^2 = 1 + 1 = 2$ wherever it occurs, so the two swaps always give

E1

$\Sigma d^2 = 4$ and the same r_s

$$\text{Answer: } r_s = \frac{20}{21} \approx 0.952$$

5 marks

19 (5 marks)

Σd^2 is largest when the rankings are exactly reversed

E1

Judge 2's ranks are 5, 4, 3, 2, 1 against 1, 2, 3, 4, 5: $d = -4, -2, 0, 2, 4$

M1

$$\Sigma d^2 = 16 + 4 + 0 + 4 + 16 = 40$$

A1

$$r_s = 1 - \frac{6 \times 40}{5 \times 24} = 1 - 2$$

M1

$$= -1$$

A1

Answer: $\Sigma d^2 = 40$; $r_s = -1$

5 marks

20 (5 marks)

$$0.5 = 1 - \frac{6\sum d^2}{8 \times 63}, \text{ so } \sum d^2 = \frac{0.5 \times 504}{6} = 42$$

M1 A1

Seven known values: $\sum d^2 = 17$, so the eighth $d^2 = 25$

M1 A1

 $\sum d = 0$ gives $d = -5$ (the other root, 5, does not make $\sum d$ zero)

A1

Answer: $d = -5$

5 marks

21 (8 marks)

$$(a) \sum x = 277, \sum y = 130, \sum x^2 = 48955, \sum y^2 = 2216, \sum xy = 5170$$

M1

$$S_{xx} = 39400, S_{yy} = 104, S_{xy} = 669$$

M1

$$r \approx 0.331$$

A1

Answer: $r \approx 0.331$

3 marks

(b) x ranks 1 to 8 in order of size: 1, 2, 3, 4, 5, 6, 7, 8; y ranks: 2, 1, 3, 5, 4, 7, 8, 6

M1

$$\sum d^2 = 10$$

A1

$$r_s = 1 - \frac{6 \times 10}{504} = 0.881$$

A1

Answer: $r_s \approx 0.881$

3 marks

(c) The extreme value of x dominates the PMCC (it depends on actual values), pulling the correlation down

E1

Spearman's coefficient uses only ranks so is unaffected by the size of the outlier: it is more suitable

E1

Answer: Spearman is resistant to outliers

2 marks

22 (8 marks)

$$(a) 8 \text{ occupies positions } 2, 3, 4: \text{rank } \frac{2+3+4}{3} = 3$$

M1 A1

15 occupies positions 6, 7: rank 6.5; so ranks are 1, 3, 3, 3, 5, 6.5, 6.5, 8

A1

Answer: 1, 3, 3, 3, 5, 6.5, 6.5, 8

3 marks

$$(b) \sum r = 1 + 3 + 3 + 3 + 5 + 6.5 + 6.5 + 8 = 36 = \frac{8 \times 9}{2}$$

B1

$$\sum r^2 = 1 + 27 + 25 + 84.5 + 64 = 202$$

M1 A1

Answer: $\sum r = 36$; $\sum r^2 = 202$

3 marks

$$(c) S_{rr} = 202 - 162 = 39.5$$

M1

Ties reduce S_{rr} below 42

A1

Answer: $S_{rr} = 39.5 < 42$

2 marks

23 (6 marks)

(a) The two rankings must use the same convention: rank 1 should be the highest (or the lowest) in both

E1

Using opposite conventions reverses one ranking, so r_s has the wrong sign (-0.886 instead of a positive value) and gives a false conclusion

E1

Answer: opposite ranking conventions

2 marks

(b) Reversed ranks ($7 - r$): 2, 1, 3, 5, 4, 6

M1

$$d = 1 - 2, 2 - 1, 3 - 3, 4 - 5, 5 - 4, 6 - 6 = -1, 1, 0, -1, 1, 0$$

M1

$$\sum d^2 = 4$$

A1

$$r_s = 1 - \frac{6 \times 4}{6 \times 35} = \frac{26}{35} \approx 0.743$$

A1

Answer: $r_s = \frac{26}{35} \approx 0.743$

4 marks

24 (7 marks)

$$(a) d = 0, -1, 1, 0, -1, 1, 0, 0, -1, 1$$

M1

$$\sum d^2 = 6$$

A1

$$r_s = 1 - \frac{6 \times 6}{10 \times 99} = \frac{53}{55} \approx 0.964$$

A1

Answer: $r_s = \frac{53}{55} \approx 0.964$

3 marks

(b) Perfect prediction: $d = 0$ for all runners. Swapping neighbours k and $k + 1$ gives $d = \pm 1$ for both

M1

Contribution $1^2 + 1^2 = 2$

A1

There are three separate adjacent swaps in part (a)

M1

So $\sum d^2 = 3 \times 2 = 6$

A1

Answer: shown

4 marks

25 (7 marks)

x ranks (smallest = 1): 3, 6, 6, 1.5, 10, 8, 6, 12, 9, 4, 1.5, 11

M1 A1

y ranks: 3, 6.5, 4.5, 1, 11, 8.5, 6.5, 12, 8.5, 4.5, 2, 10

M1 A1

$\sum r_x r_y = 645$, $\sum r_x^2 = 648$, $\sum r_y^2 = 648$

M1

$S_{xy} = 138$, $S_{xx} = 140$, $S_{yy} = 142$

M1

$$r_s = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \approx 0.979$$

A1

Answer: $r_s \approx 0.979$

7 marks

26 (6 marks)

(a) The differences between two sets of ranks 1 to n must sum to zero because $\sum r_x = \sum r_y$

E1

$-1 + 1 - 1 + 1 + 0 - 3 + 1 + 1 = -1 \neq 0$

B1

Answer: $\sum d \neq 0$

2 marks

(b) With -2 : $\sum d = -1 + 1 - 1 + 1 + 0 - 2 + 1 + 1 = 0$

M1

$\sum d^2 = 1 + 1 + 1 + 1 + 0 + 4 + 1 + 1 = 10$

A1

$$r_s = 1 - \frac{6 \times 10}{8 \times 63} = 1 - \frac{60}{504}$$

M1

$$= \frac{37}{42} \approx 0.881$$

A1

Answer: $r_s = \frac{37}{42} \approx 0.881$

4 marks

27 (9 marks)

$\sum r_x = \sum r_y = \frac{n(n+1)}{2}$ and $\sum r_x^2 = \sum r_y^2 = \frac{n(n+1)(2n+1)}{6}$

B1

$$S_{xx} = S_{yy} = \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4}$$

M1

$$= \frac{n(n+1)}{12} [2(2n+1) - 3(n+1)] = \frac{n(n^2-1)}{12}$$

M1 A1

$\sum d^2 = \sum r_x^2 + \sum r_y^2 - 2\sum r_x r_y$, so $\sum r_x r_y = \frac{n(n+1)(2n+1)}{6} - \frac{1}{2}\sum d^2$

M1 A1

$$S_{xy} = \sum r_x r_y - \frac{n(n+1)^2}{4} = S_{xx} - \frac{1}{2}\sum d^2$$

M1

$$r_s = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{S_{xx} - \frac{1}{2}\sum d^2}{S_{xx}}$$

A1

$$= 1 - \frac{6\sum d^2}{n(n^2-1)}$$

A1

Answer: shown

9 marks

28 (8 marks)

$$\Sigma d^2 = \Sigma r_x^2 + \Sigma r_y^2 - 2\Sigma r_x r_y = \frac{n(n+1)(2n+1)}{3} - 2\Sigma r_x r_y \quad \text{M1 A1}$$

This is largest when $\Sigma r_x r_y$ is smallest, which is when the pairing is exactly reversed, $r_y = n + 1 - r_x$ (rearrangement inequality) E1

$$\Sigma r_x(n+1-r_x) = (n+1)\frac{n(n+1)}{2} - \frac{n(n+1)(2n+1)}{6} = \frac{n(n+1)(n+2)}{6} \quad \text{M1 A1}$$

$$\Sigma d^2_{\max} = \frac{n(n+1)(2n+1)}{3} - \frac{n(n+1)(n+2)}{3} = \frac{n(n+1)(n-1)}{3} = \frac{n(n^2-1)}{3} \quad \text{A1}$$

Then $r_s = 1 - \frac{6}{n(n^2-1)} \times \frac{n(n^2-1)}{3} = -1$, so $r_s \geq -1$ M1

For $n = 6$: $\Sigma d^2 = \frac{6 \times 35}{3} = 70$; check: $d = 5, 3, 1, -1, -3, -5$ gives $25 + 9 + 1 + 1 + 9 + 25 = 70$ A1

Answer: $\Sigma d^2_{\max} = n(n^2-1)/3$; $n = 6$: 70 8 marks

29 (9 marks)

(a) $\Sigma r_x^2 = 1 + 6.25 + 6.25 + 16 + 25 + 36 = 90.5 = 91 - \frac{1}{2}$ M1 A1

$S_{xx} = 90.5 - \frac{21^2}{6} = 90.5 - 73.5 = 17$ (compared with 17.5) A1

Answer: $S_{xx} = 17$ 3 marks

(b) $\Sigma r_x r_y = 2 + 2.5 + 10 + 12 + 30 + 30 = 86.5$ M1

$S_{xy} = 86.5 - 73.5 = 13$ A1

$S_{yy} = 91 - 73.5 = 17.5$ B1

$r_s = \frac{13}{\sqrt{17 \times 17.5}} \approx 0.754$ A1

Answer: $r_s \approx 0.754$ 4 marks

(c) $d = -1, 1.5, -1.5, 1, -1, 1$, so $\Sigma d^2 = 1 + 2.25 + 2.25 + 1 + 1 + 1 = 8.5$ and $1 - \frac{6 \times 8.5}{210} \approx 0.757$ M1

The simple formula is only approximate when there are ties (here 0.757 against 0.754): use the PMCC of the ranks E1

Answer: 0.757; the formula is only approximate with ties 2 marks

30 (9 marks)

(a) $r_s = 1 - \frac{6\Sigma d^2}{4 \times 15} = 1 - \frac{\Sigma d^2}{10}$ B1

Answer: shown 1 mark

(b) $\Sigma d^2 = 0$: the identical ordering 1234 (1 ordering) B1

$\Sigma d^2 = 2$: one adjacent swap: 2134, 1324, 1243 (3 orderings) M1 A1

$\Sigma d^2 = 4$: two separate adjacent swaps 2143 (1 ordering); $\Sigma d^2 = 4$ for no other ordering M1 A1

$r_s \geq 0.8$ needs $\Sigma d^2 \leq 2$: probability $\frac{1+3}{24} = \frac{1}{6}$; $P(r_s = 1) = \frac{1}{24}$ A1

Answer: $P(r_s = 1) = \frac{1}{24}$; $P(r_s \geq 0.8) = \frac{1}{6}$ 6 marks

(c) $P(r_s \geq 0.8) = \frac{1}{6} \approx 0.167 > 0.05$, so $r_s = 0.8$ is not significant M1

$P(r_s = 1) = \frac{1}{24} \approx 0.042 < 0.05$, so only $r_s = 1$ is significant A1

Answer: only $r_s = 1$ is significant 2 marks

Linear Regression

1 (2 marks)

$$y = 2.4 + 0.7 \times 10$$

$$= 9.4$$

M1

A1

Answer: 9.4

2 marks

2 (3 marks)

$$b = \frac{S_{xy}}{S_{xx}} = \frac{60}{40} = 1.5$$

M1 A1

$$a = \bar{y} - b\bar{x} = 12 - 1.5 \times 5 = 4.5, \text{ so } y = 4.5 + 1.5x$$

A1

Answer: $y = 4.5 + 1.5x$

3 marks

3 (3 marks)

12 is the cost (£12) when no posters are printed: a fixed cost

B1 B1

3.5 is the extra cost (£3.50) of each additional poster

B1

Answer: fixed cost £12; £3.50 per poster

3 marks

4 (2 marks)

(i) Explanatory: mass of fertiliser; response: yield

B1

(ii) Explanatory: age of car; response: resale price

B1

Answer: (i) fertiliser, yield; (ii) age, price

2 marks

5 (3 marks)

 $x = 25$ lies within the range of the data (interpolation), so the estimate is likely to be reliable

B1 B1

 $x = 60$ is outside the range (extrapolation), so the estimate may be unreliable

B1

Answer: 25 reliable; 60 unreliable

3 marks

6 (2 marks)

$$\text{Predicted } y = 3 + 11 = 14$$

M1

$$\text{Residual} = 14.2 - 14 = 0.2, \text{ so the observation is above the line}$$

A1

Answer: residual 0.2; above the line

2 marks

7 (5 marks)

$$\Sigma x = 15, \Sigma y = 33, \Sigma x^2 = 55, \Sigma xy = 117$$

M1

$$S_{xx} = 55 - \frac{15^2}{5} = 10, S_{xy} = 117 - \frac{15 \times 33}{5} = 18$$

A1 A1

$$b = \frac{18}{10} = 1.8$$

M1

$$a = \frac{33}{5} - 1.8 \times 3 = 1.2, \text{ so } y = 1.2 + 1.8x$$

A1

Answer: $y = 1.2 + 1.8x$

5 marks

8 (5 marks)

$$S_{xx} = 820 - \frac{76^2}{8} = 98.0$$

M1

$$S_{xy} = 2237 - \frac{76 \times 212}{8} = 223$$

A1

$$b = \frac{S_{xy}}{S_{xx}} = 2.28$$

M1

$$a = \frac{212}{8} - b \times \frac{76}{8}$$

M1

$$y = 4.88 + 2.28x$$

A1

Answer: $y = 4.88 + 2.28x$

5 marks

9 (5 marks)

$$\text{Regression of } x \text{ on } y: \text{gradient } b' = \frac{S_{xy}}{S_{yy}} = \frac{-78.5}{122} = -0.646$$

M1 A1

$$a' = \bar{x} - b'\bar{y}$$

M1

$$x = 15.9 + (-0.646)y$$

A1

$$y = 10: x \approx 9.43$$

A1

$$\text{Answer: } x = 15.9 - 0.646y; x \approx 9.43$$

5 marks

10 (6 marks)(a) Both lines pass through $(\bar{x}, \bar{y})$

E1

$$\bar{y} = 2 + 0.5(1 + 0.8\bar{y}) = 2.5 + 0.4\bar{y}, \text{ so } \bar{y} = \frac{2.5}{0.6} = \frac{25}{6}$$

M1

$$\bar{x} = 1 + 0.8 \times \frac{25}{6} = \frac{13}{3}$$

A1

$$\text{Answer: } \bar{x} = \frac{13}{3}, \bar{y} = \frac{25}{6}$$

3 marks

$$(b) r^2 = b_{yx} \times b_{xy} = 0.5 \times 0.8 = 0.4$$

M1 A1

$$r = \sqrt{0.4} \approx 0.632 \text{ (positive since both gradients are positive)}$$

A1

$$\text{Answer: } r \approx 0.632$$

3 marks

11 (5 marks)

$$S_{uu} = 286 - \frac{36^2}{6} = 70.0, S_{uy} = 1037 - \frac{36 \times 155}{6} = 107$$

M1 A1

$$b = \frac{107}{70.0} = 1.53$$

M1

$$a = 25.8 - 1.53 \times 6.00 = 16.7, \text{ so } y = 16.7 + 1.53u$$

A1

$$\text{Replace } u \text{ by } x - 2000: y = 16.7 + 1.53(x - 2000)$$

A1

$$\text{Answer: } y = 16.7 + 1.53(x - 2000)$$

5 marks

12 (4 marks) $r = 0.95$ indicates the points lie very close to a straight line

B1

 $r = 0.4$ indicates a weak linear relationship with considerable scatter

B1

Predictions from B will have smaller errors, so B is more reliable

E1

Reliable only within the range of the data (interpolation)

E1

Answer: line B

4 marks

13 (3 marks)

$$y - 20 = 2.5(x - 6)$$

M1

$$y = 5 + 2.5x$$

A1

$$x = 8: y = 25$$

A1

$$\text{Answer: } y = 5 + 2.5x; y = 25$$

3 marks

14 (7 marks)

$$(a) \Sigma x = 49.7, \Sigma y = 171, \Sigma x^2 = 338.39, \Sigma y^2 = 3915.64, \Sigma xy = 1149.98$$

B1

$$S_{xx} = 29.63, S_{yy} = 260.5, S_{xy} = 87.64$$

M1

$$r \approx 0.9976$$

A1

$$\text{Answer: } r \approx 0.9976$$

3 marks

$$(b) b = \frac{S_{xy}}{S_{xx}} \approx 2.958$$

M1

$$a = \bar{y} - b\bar{x} \approx 2.998, \text{ so } y = 2.998 + 2.958x$$

A1

$$x = 6: y \approx 20.7$$

A1

$$\text{Answer: } y = 2.998 + 2.958x; y \approx 20.7$$

3 marks

(c) 6 is within the range of x (3.2 to 9.0) and r is close to 1, so the estimate is reliable

E1

Answer: reliable

1 mark

15 (6 marks)**(a)** Fitted values: 7.14, 10.89, 14.63, 18.37, 22.11, 25.86

M1 A1

Residuals $y - \hat{y} = -0.14, -0.89, 1.37, -1.37, 2.89, -1.86$

A1

Answer: -0.14, -0.89, 1.37, -1.37, 2.89, -1.86

3 marks

(b) Sum of residuals = $0 \approx 0$

M1

Largest residual in magnitude is for (10, 25), residual 2.89

A1

Answer: largest residual at $x = 10$

2 marks

(c) A pattern (such as a curve) means the data are not randomly scattered about the line, so a different model is needed

E1

Answer: pattern means linear model unsuitable

1 mark

16 (6 marks)**(a)** M on T

B1

T is the controlled (explanatory) variable and M is the response variable; the line minimises errors in M

E1

Answer: M on T

2 marks

(b) $S_{TT} = 1750$, $S_{TM} = 259$

M1 A1

$$b = \frac{259}{1750} = 0.1480, a = 7.767 - 0.1480 \times 35 = 2.587$$

M1

M = $2.587 + 0.1480T$, so at $T = 35$: $M \approx 7.77$ g

A1

Answer: $M = 2.587 + 0.1480T$; $M \approx 7.77$ g

4 marks

17 (5 marks) $x = 1.6\text{m}$

B1

 $y = 3 + 2(1.6\text{m}) = 3 + 3.2\text{m}$

M1 A1

 $p = 100y = 100(3 + 2x)$

M1

 $p = 300 + 200x$

A1

Answer: $y = 3 + 3.2\text{m}$; $p = 300 + 200x$

5 marks

18 (5 marks)

$$b_{yx} = r \frac{s_y}{s_x} = 0.8 \times \frac{5}{2} = 2$$

M1 A1

 $y - 30 = 2(x - 10)$, so $y = 10 + 2x$

A1

$$b_{xy} = r \frac{s_x}{s_y} = 0.8 \times \frac{2}{5} = 0.32$$

M1

 $x - 10 = 0.32(y - 30)$, so $x = 0.4 + 0.32y$

A1

Answer: $y = 10 + 2x$; $x = 0.4 + 0.32y$

5 marks

19 (7 marks)**(a)** $S_{xx} = 42.0$, $S_{xy} = 83.8$

M1

 $b = 2.00$

A1

 $y = 1.07 + 2.00x$

A1

Answer: $y = 1.07 + 2.00x$

3 marks

(b) $S_{xx} = 28.0$, $S_{xy} = 28.0$

M1

 $b = 1.00$, $a = 4.06$, $y = 4.06 + 1.00x$

A1 A1

The outlier greatly increases the gradient (1.00 to 2.00) and pulls the line away from the other points

E1

Answer: $y = 4.06 + 1.00x$

4 marks

20 (6 marks)**(a)** $S_{yy} = 213$, $S_{xy} = 177$

M1

 $x = -5.48 + 0.830y$

M1 A1

 $y = 25$: $x \approx 15.3$

A1

Answer: $x \approx 15.3$

4 marks

(b) Rearranging: $x = \frac{25 - 7.62}{1.12} \approx 15.5$

M1

The regression of x on y is the better method here, because it minimises the errors in x ; the two

E1

estimates differ (15.3 and 15.5) since $r < 1$

Answer: x on y is better

2 marks

21 (6 marks)

$$\sum e_i = \sum y_i - na - b\sum x_i$$

M1

$$= n\bar{y} - n(\bar{y} - b\bar{x}) - bn\bar{x} = 0$$

A1

$$\sum x_i e_i = \sum x_i(y_i - \bar{y}) - b\sum x_i(x_i - \bar{x})$$

M1 A1

$$(\text{using } a = \bar{y} - b\bar{x}) = S_{xy} - bS_{xx}$$

M1

$$= S_{xy} - \frac{S_{xy}}{S_{xx}}S_{xx} = 0$$

A1

Answer: shown

6 marks

22 (8 marks)

$$(a) \sum t = 45, \sum s = 553, \sum t^2 = 285, \sum ts = 2759$$

B1

$$S_{tt} = 285 - \frac{45^2}{10} = 82.5, S_{ts} = 2759 - \frac{45 \times 553}{10} = 270$$

M1 A1

$$s = 40.5 + 3.28t$$

A1

Answer: $s = 40.5 + 3.28t$

4 marks

(b) Sales are increasing by about 3.28 thousand units per year

B1

$$2017: t = 3, s = 40.5 + 3 \times 3.28 \approx 50.4$$

M1

$$\text{Residual} = 51 - 50.4 \approx 0.618$$

A1

Answer: 3.28 thousand per year; residual ≈ 0.618

3 marks

(c) $t = 18: s \approx 99.6$ thousand; it is extrapolation far beyond the data, and conditions may change

E1

Answer: 99.6 thousand; extrapolation

1 mark

23 (6 marks)

(a) The gradient of y on x is $\frac{S_{xy}}{S_{xx}}$, not $\frac{S_{xx}}{S_{xy}}$

E1

The intercept is $a = \bar{y} - b\bar{x}$ (the student has swapped x and y)

E1

Answer: inverted gradient; swapped means

2 marks

$$(b) b = \frac{94.0}{42.0} = 2.24$$

M1 A1

$$a = 18.0 - 2.24 \times 6.00 = 4.57, \text{ so } y = 4.57 + 2.24x$$

A1

Answer: $y = 4.57 + 2.24x$

3 marks

$$(c) y = 4.57 + 2.24 \times 6 \approx 18.0$$

B1

Answer: 18.0

1 mark

24 (8 marks)

(a) Both lines pass through $(\bar{x}, \bar{y})$: $\bar{y} = 0.6\bar{x} + 4$ and $\bar{x} = 0.9\bar{y} - 1$

M1

$$\bar{x} = 0.9(0.6\bar{x} + 4) - 1 = 0.54\bar{x} + 2.6, \text{ so } \bar{x} = \frac{2.6}{0.46} \approx 5.65$$

M1

$$\bar{y} = 0.6 \times 5.65 + 4 \approx 7.39$$

A1

Answer: $\bar{x} \approx 5.65, \bar{y} \approx 7.39$

3 marks

$$(b) r^2 = 0.6 \times 0.9 = 0.54$$

M1

$$r = \sqrt{0.54} \approx 0.735$$

A1

Both regression gradients are positive, so r is positive

E1

Answer: $r \approx 0.735$

3 marks

$$(c) b_{yx} = r \frac{s_y}{s_x}, \text{ so } \frac{s_y}{s_x} = \frac{0.6}{r}$$

M1

$$= \frac{0.6}{\sqrt{0.54}} \approx 0.816$$

A1

Answer: 0.816

2 marks

25 (6 marks)

The means of z_x and z_y are 0 and their standard deviations are 1

B1 B1

$$\text{Gradient} = r_z \frac{S_{z_y}}{S_{z_x}} = r_z \times 1$$

M1

$r_z = r$, since the standardisation is a positive linear change

E1

Line passes through (0, 0), so the intercept is 0

M1

$$z_y = r z_x$$

A1

Answer: shown

6 marks

26 (6 marks)

$$(a) b = r \frac{S_y}{S_x} = 0.6 \times 1 = 0.6$$

M1

$$y - 60 = 0.6(x - 60)$$

A1

$$x = 75: y = 60 + 0.6 \times 15 = 69$$

M1 A1

Answer: $y = 24 + 0.6x$; 69

4 marks

(b) Predicted 9 above the mean = 0.9 standard deviations

B1

The prediction is closer to the mean than the original score: regression towards the mean, because

E1

$r < 1$

Answer: 0.9 sd; regression to the mean

2 marks

27 (8 marks)

$$(a) b_{yx} b_{xy} = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

M1 A1

$= r^2$; S_{xx} and S_{yy} are positive, so both gradients have the sign of S_{xy} , i.e. of r

A1

Answer: shown

3 marks

$$(b) \frac{b_{yx}}{(1/b_{yx})} = b_{yx} b_{xy} = r^2 \leq 1, \text{ so } b_{yx} \leq \frac{1}{b_{xy}}$$

M1 A1

Both lines pass through $(\bar{x}, \bar{y})$; equal gradients require $r^2 = 1$

E1

Answer: shown

3 marks

$$(c) r^2 = 1.5 \times 0.8 = 1.2$$

M1

But $r^2 \leq 1$, so these cannot both be regression lines of the same data

E1

Answer: $r^2 = 1.2 > 1$

2 marks

28 (9 marks)

$$(a) A: S_{xx} = 604 - \frac{64^2}{8} = 92.0, S_{xy} = 1478 - \frac{64 \times 162}{8} = 182$$

M1 A1

$$y = 4.42 + 1.98x$$

A1

$$B: y = 10.9 + 1.76x$$

A1

Answer: A: $y = 4.42 + 1.98x$; B: $y = 10.9 + 1.76x$

4 marks

(b) Combined: $n = 15$, $\Sigma x = 152$, $\Sigma y = 393$, $\Sigma x^2 = 1848$, $\Sigma xy = 4624$

M1 A1

$$S_{xx} = 308, S_{xy} = 642$$

M1

$$b = 2.08, a = 5.07, y = 5.07 + 2.08x$$

A1

Not simply the average: the combined line depends on the spread and means of each sample (the difference between sample means also contributes)

E1

Answer: $y = 5.07 + 2.08x$

5 marks

29 (8 marks)

$$(a) y_i - \bar{y} = (\hat{y}_i - \bar{y}) + e_i \text{ and } \hat{y}_i - \bar{y} = b(x_i - \bar{x})$$

M1

$$\text{Square and sum: } S_{yy} = b^2 S_{xx} + 2b \Sigma (x_i - \bar{x}) e_i + \Sigma e_i^2$$

M1 A1

$$\text{The cross term is zero, so } S_{yy} = b^2 S_{xx} + \Sigma e_i^2$$

A1

Answer: shown

4 marks

$$(b) b^2 S_{xx} = \frac{S_{xy}^2}{S_{xx}} = r^2 S_{yy}$$

M1

$$\text{So } \Sigma e_i^2 = S_{yy} - r^2 S_{yy} = S_{yy}(1 - r^2)$$

A1

Answer: shown**2 marks****(c)** $S_{yy} = 82.8$, $S_{xy} = 64.3$, $S_{xx} = 55.3$, so $r^2 = 0.903$ **M1** $\Sigma e^2 = S_{yy}(1 - r^2) \approx 8.04$; about 90.3% of the variation is explained**A1****Answer: $r^2 \approx 0.903$; $\Sigma e^2 \approx 8.04$** **2 marks****30 (9 marks)****(a)** $b = \frac{S_{xy}}{S_{xx}} = \frac{30}{40} = 0.75$ **B1** $S_{ww} = S_{xx} + 2S_{xv} + S_{vv} = S_{xx} + S_{vv} = 50$ **M1 A1** $S_{wy} = S_{xy} + S_{vy} = S_{xy} = 30$ **M1** $b' = \frac{S_{wy}}{S_{ww}} = \frac{S_{xy}}{S_{xx} + S_{vv}} = b \frac{S_{xx}}{S_{xx} + S_{vv}}$ **A1****Answer: shown****5 marks****(b)** $b' = \frac{30}{50} = 0.6 = 0.75 \times \frac{40}{50}$ **M1**

The gradient is reduced (attenuated) towards zero, so the error makes the relationship appear weaker

E1**Answer: $b' = 0.6$; gradient underestimated****2 marks****(c)** $y = 50 + 0.75(x - 20)$ gives 57.5; $y = 50 + 0.6(x - 20)$ gives 56**M1**

The error-affected line under-predicts away from the mean because its gradient is too small

E1**Answer: 57.5 and 56****2 marks**

Least Squares Regression

1 (2 marks)

The sum of the squares of the vertical distances (residuals) between the observed y-values and the line

B1 $S = \Sigma(y_i - a - bx_i)^2$ **B1****Answer: $S = \Sigma(y - a - bx)^2$** **2 marks****2 (3 marks)**

Predicted values: 2, 3, 4, 5

M1

Residuals: 0, 0, 1, 0

A1

Sum of squares of residuals = 1

A1**Answer: 1****3 marks****3 (3 marks)** $b = \frac{S_{xy}}{S_{xx}} = 2$ **M1 A1** $a = \bar{y} - b\bar{x} = 10 - 6 = 4$, so $y = 4 + 2x$ **A1****Answer: $y = 4 + 2x$** **3 marks****4 (2 marks)**

Residuals can be positive or negative and would cancel if just added

E1

Squaring gives more weight to large residuals and leads to simple formulae (by differentiation)

E1**Answer: avoid cancelling; penalise large errors****2 marks****5 (3 marks)**Fitted values: $\frac{19}{6}$, $\frac{28}{6}$, $\frac{37}{6}$ **M1**Residuals: $\frac{-1}{6}$, $\frac{2}{6}$, $\frac{-1}{6}$ **A1**Sum = $\frac{-1 + 2 - 1}{6} = 0$ **A1****Answer: $-1/6$, $1/3$, $-1/6$; sum 0****3 marks**

6 (3 marks)Line A residuals: 0, 1, 0, 1, so $\Sigma = 2$

M1 A1

Line B residuals: $-0.5, 1, 0.5, 2$, so $\Sigma = 5.5$; line A has the smaller value and fits better

A1

Answer: A: 2; B: 5.5; line A

3 marks

7 (5 marks)

$$40 = 5a + 15b$$

M1 A1

$$146 = 15a + 55b$$

A1

From the first, $a = 8 - 3b$: $146 = 120 - 45b + 55b$

M1

$$b = 2.6, a = 0.2$$

A1

Answer: a = 0.2, b = 2.6

5 marks

8 (4 marks)

$$\frac{\partial S}{\partial a} = -2\Sigma(y_i - a - bx_i)$$

M1 A1

Setting this to zero: $\Sigma y_i - na - b\Sigma x_i = 0$

M1

Dividing by n : $\bar{y} = a + b\bar{x}$, so $(\bar{x}, \bar{y})$ lies on the line

A1

Answer: shown

4 marks

9 (4 marks)

$$\text{Gradient } b = \frac{S_{xy}}{S_{xx}} = 1.2$$

M1

$$\frac{S_{xy}^2}{S_{xx}} = \frac{576}{20} = 28.8$$

M1 A1

$$\Sigma e^2 = 45 - 28.8 = 16.2$$

A1

Answer: 16.2

4 marks

10 (5 marks)

$$\frac{dS}{db} = -2\Sigma x_i(y_i - bx_i)$$

M1 A1

Setting to zero: $\Sigma xy - b\Sigma x^2 = 0$, so $b = \frac{\Sigma xy}{\Sigma x^2}$

A1

$$\Sigma xy = 112.1, \Sigma x^2 = 55$$

M1

$$b = \frac{112.1}{55} \approx 2.04$$

A1

Answer: b ≈ 2.04

5 marks

11 (5 marks)

$$S = \Sigma(y_i - 2 - bx_i)^2, \frac{dS}{db} = -2\Sigma x_i(y_i - 2 - bx_i)$$

M1 A1

$$\text{Setting to zero: } b = \frac{\Sigma x(y - 2)}{\Sigma x^2}$$

A1

$$\Sigma x(y - 2) = 0 + 2.8 + 10.2 + 20.7 + 36.8 = 70.5, \Sigma x^2 = 30$$

M1

$$b = \frac{70.5}{30} \approx 2.35$$

A1

Answer: b ≈ 2.35

5 marks

12 (5 marks)

Fitted values: 5, 8.5, 12, 15.5

M1 A1

Residuals: 0, 0.5, -1 , 0.5

A1

$$\Sigma e = 0 = 0$$

B1

$$\Sigma xe = 0 = 0 \text{ (to rounding)}$$

B1

Answer: residuals 0, 0.5, -1 , 0.5; sums zero

5 marks

13 (4 marks)

$$\Sigma e^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}} \quad \text{M1}$$

$$= S_{yy}(1 - r^2) \quad \text{A1}$$

$$= 50(1 - 0.81) \quad \text{M1}$$

$$= 9.5 \quad \text{A1}$$

Answer: 9.5 4 marks

14 (7 marks)

$$\text{(a)} \frac{\partial S}{\partial a} = -2\Sigma(y - a - bx) = 0 \text{ gives } \Sigma y = na + b\Sigma x \quad \text{M1 A1}$$

$$\frac{\partial S}{\partial b} = -2\Sigma x(y - a - bx) = 0 \quad \text{M1}$$

$$\text{gives } \Sigma xy = a\Sigma x + b\Sigma x^2 \quad \text{A1}$$

Answer: shown 4 marks

$$\text{(b)} \Sigma x = 30, \Sigma y = 55, \Sigma x^2 = 220, \Sigma xy = 386 \quad \text{B1}$$

$$55 = 5a + 30b \text{ and } 386 = 30a + 220b \quad \text{M1}$$

$$b = 1.4, a = 2.6 \quad \text{A1}$$

Answer: a = 2.6, b = 1.4 3 marks

15 (6 marks)

$$\text{Write } y_i - a - bx_i = (y_i - \bar{y}) - b(x_i - \bar{x}) + (\bar{y} - b\bar{x} - a) \quad \text{M1}$$

$$\text{Let } c = \bar{y} - b\bar{x} - a; \text{ the cross terms with } c \text{ vanish because } \Sigma(y_i - \bar{y}) = \Sigma(x_i - \bar{x}) = 0 \quad \text{M1 A1}$$

$$S = \Sigma[(y_i - \bar{y}) - b(x_i - \bar{x})]^2 + nc^2 \quad \text{A1}$$

$$\text{The second term is smallest (zero) when } c = 0, \text{ i.e. } a = \bar{y} - b\bar{x} \quad \text{A1}$$

$$S_{\min} = S_{yy} - 2bS_{xy} + b^2S_{xx} \quad \text{A1}$$

Answer: shown 6 marks

16 (6 marks)

$$\text{By symmetry of the argument, } d = \frac{S_{xy}}{S_{yy}} \text{ and } c = \bar{x} - d\bar{y} \quad \text{B1}$$

$$S_{xy} = 46.8, S_{yy} = 48.7, S_{xx} = 45.3 \quad \text{M1 A1}$$

$$x = -1.89 + 0.960y \quad \text{A1}$$

$$y \text{ on } x: y = 2.02 + 1.03x \quad \text{A1}$$

$$\text{The lines differ unless } r = \pm 1; \text{ both pass through } (\bar{x}, \bar{y}) = (4.67, 6.83) \quad \text{E1}$$

Answer: x = -1.89 + 0.960y; y = 2.02 + 1.03x 6 marks

17 (5 marks)

$$\Sigma x_i e_i = \Sigma x_i y_i - a\Sigma x_i - b\Sigma x_i^2 \quad \text{M1 A1}$$

$$= 0 \text{ by the second normal equation} \quad \text{A1}$$

$$\text{Similarly } \Sigma e_i = \Sigma y - na - b\Sigma x = 0 \text{ by the first} \quad \text{M1}$$

$$\text{The residuals are uncorrelated with } x \text{ (no linear pattern remains in the residuals)} \quad \text{E1}$$

Answer: shown 5 marks

18 (6 marks)

$$\frac{dS}{db} = -2\Sigma w_i x_i (y_i - bx_i) \quad \text{M1 A1}$$

$$\text{Setting equal to zero gives } b = \frac{\Sigma wxy}{\Sigma wx^2} \quad \text{A1}$$

$$\Sigma wxy = 2.1 + 15.2 + 37.8 + 31.6 = 86.7 \quad \text{M1}$$

$$\Sigma wx^2 = 1 + 8 + 18 + 16 = 43 \quad \text{M1}$$

$$b = \frac{86.7}{43} \approx 2.02 \quad \text{A1}$$

Answer: b ≈ 2.02 6 marks

19 (6 marks)

$$u = -2, -1, 0, 1, 2, \text{ so } \sum u = 0 \text{ and } \sum u^2 = 10$$

B1

$$\text{With } \sum u = 0 \text{ the normal equations give } na = \sum y, \text{ so } a = \bar{y} = 18.9$$

M1 A1

$$b = \frac{\sum uy}{\sum u^2} = \frac{24.5}{10} = 2.45$$

M1 A1

$$y = 18.9 + 2.45u = 18.9 + 2.45\left(\frac{x - 2002}{2}\right)$$

A1

$$\text{Answer: } y = 18.9 + 2.45(x - 2002)/2$$

6 marks

20 (7 marks)

$$\text{Minimum over } a: S = S_{yy} - 2bS_{xy} + b^2S_{xx}$$

B1

$$\frac{dS}{db} = -2S_{xy} + 2bS_{xx} = 0$$

M1 A1

$$b = \frac{S_{xy}}{S_{xx}}; \frac{d^2S}{db^2} = 2S_{xx} > 0 \text{ so this is a minimum}$$

A1 B1

$$S_{\min} = S_{yy} - 2\frac{S_{xy}^2}{S_{xx}} + \frac{S_{xy}^2}{S_{xx}} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

M1

$$= S_{yy}\left(1 - \frac{S_{xy}^2}{S_{xx}S_{yy}}\right) = S_{yy}(1 - r^2)$$

A1

Answer: shown

7 marks

21 (7 marks)

$$(a) \frac{\partial S}{\partial a} = -2\sum(y - a - bx^2) = 0$$

M1 A1

$$\frac{\partial S}{\partial b} = -2\sum x^2(y - a - bx^2) = 0, \text{ giving the second equation}$$

A1

Answer: shown

3 marks

$$(b) \sum x^2 = 91, \sum x^4 = 2275, \sum y = 86.2, \sum x^2y = 2029.1$$

M1 A1

$$b = \frac{S_{uy}}{S_{uu}} = 0.807$$

M1

$$a = \bar{y} - b\bar{u} = 2.13$$

A1

$$\text{Answer: } a \approx 2.13, b \approx 0.807$$

4 marks

22 (7 marks)

$$\text{Line: } y - 5 = b(x - 2); \text{ minimise } S = \sum [(y - 5) - b(x - 2)]^2$$

M1

$$\frac{dS}{db} = -2\sum (x - 2)[(y - 5) - b(x - 2)] = 0$$

M1 A1

$$b = \frac{\sum (x - 2)(y - 5)}{\sum (x - 2)^2}$$

A1

$$x - 2: -2, -1, 1, 2, 4; y - 5: -4, -2, 1, 2, 6; \sum (x - 2)(y - 5) = 39, \sum (x - 2)^2 = 26$$

M1

$$b = \frac{39}{26} \approx 1.50, \text{ so } y = 5 + 1.50(x - 2)$$

A1 A1

$$\text{Answer: } y = 5 + 1.50(x - 2)$$

7 marks

23 (6 marks)

$$(a) T \text{ is linear in } a \text{ and } b, \text{ so it has no minimum (it can be made arbitrarily negative)}$$

E1

Positive and negative residuals cancel, so a poor line can give $T = 0$; the residuals must be squared

E1

Answer: T has no minimum; use squares

2 marks

$$(b) \frac{\partial S}{\partial a} = 0 \text{ and } \frac{\partial S}{\partial b} = 0 \text{ give the normal equations } \sum y = na + b\sum x, \sum xy = a\sum x + b\sum x^2$$

M1

$$\sum x = 15, \sum y = 43, \sum x^2 = 55, \sum xy = 154$$

M1

$$S_{xx} = 10, S_{xy} = 154 - \frac{15 \times 43}{5} = 25$$

A1

$$b = 2.5, a = 1.1$$

A1

$$\text{Answer: } b = 2.5, a = 1.1$$

4 marks

24 (7 marks)

$$(a) \text{ Residual} = (y_i - \bar{y}) - b(x_i - \bar{x}) \quad \text{M1}$$

$$S = \sum (y_i - \bar{y})^2 - 2b \sum (x_i - \bar{x})(y_i - \bar{y}) + b^2 \sum (x_i - \bar{x})^2 \quad \text{M1}$$

$$= S_{yy} - 2bS_{xy} + b^2S_{xx} \quad \text{A1}$$

Answer: shown 3 marks

$$(b) \bar{x} = 5, \bar{y} = 8; S_{xx} = 30, S_{yy} = 46, S_{xy} = 35 \quad \text{M1 A1}$$

$$S(1) = 46 - 2 \times 35 + 30 = 6; S(1.2) = 5.2 \quad \text{M1}$$

$$\text{Least squares } b = \frac{S_{xy}}{S_{xx}} = 1.1667 \text{ gives the minimum } S = 5.1667 \quad \text{A1}$$

Answer: S(1) = 6; S(1.2) = 5.2; minimum 5.1667 at b = 1.1667 4 marks

25 (7 marks)

$$(a) \frac{\partial S}{\partial \alpha} = 0: \sum y = n\alpha + \beta \sum u = n\alpha \quad \text{M1 A1}$$

$$\frac{\partial S}{\partial \beta} = 0: \sum uy = \alpha \sum u + \beta \sum u^2 = \beta \sum u^2 \quad \text{M1}$$

$$\text{So } \alpha = \bar{y} \text{ and } \beta = \frac{\sum uy}{\sum u^2} = \frac{S_{xy}}{S_{xx}} \quad \text{A1}$$

Answer: shown 4 marks

$$(b) \bar{x} = 7.6, u = -3.6, -1.6, -0.6, 1.4, 4.4 \quad \text{B1}$$

$$\alpha = \bar{y} = 9.6, \beta = \frac{47.2}{37.2} \approx 1.27 \quad \text{M1}$$

$$y = 9.6 + 1.27(x - 7.6) \quad \text{A1}$$

Answer: y = 9.6 + 1.27(x - 7.6) 3 marks

26 (6 marks)

$$(a) \frac{dS}{dc} = -2\sum (y_i - c) = 0 \text{ gives } c = \bar{y} \quad \text{M1 A1}$$

$$\frac{d^2S}{dc^2} = 2n > 0 \text{ (minimum); } S(\bar{y}) = \sum (y_i - \bar{y})^2 = S_{yy} \quad \text{A1}$$

Answer: shown 3 marks

$$(b) \text{ Using a line of gradient } b = 0 \text{ gives sum of squares } S_{yy}; \text{ the regression line reduces it to } \sum e^2 \quad \text{E1}$$

$$\text{The proportion of the sum of squares removed is } \frac{S_{yy} - \sum e^2}{S_{yy}} = 1 - \frac{\sum e^2}{S_{yy}} = r^2 \quad \text{M1}$$

$$r^2 = 1 - \frac{4.8}{30} \approx 0.840 \quad \text{A1}$$

Answer: r² ≈ 0.840 3 marks

27 (9 marks)

$$\text{Write } y_i - a - bx_i = e_i + (a^* - a) + (b^* - b)x_i \text{ where } e_i = y_i - a^* - b^*x_i \quad \text{M1}$$

$$\text{Square and sum: } S = \sum e_i^2 + \sum [(a^* - a) + (b^* - b)x_i]^2 + 2\sum e_i[(a^* - a) + (b^* - b)x_i] \quad \text{M1 A1}$$

$$\text{The cross term vanishes since } \sum e_i = 0 \text{ and } \sum x_i e_i = 0 \quad \text{A1}$$

$$\sum e_i^2 = S_{\min} = S_{yy} - \frac{S_{xy}^2}{S_{xx}} \quad \text{A1}$$

$$\sum [\delta + \varepsilon x_i]^2 \text{ with } \delta = a^* - a, \varepsilon = b^* - b: = n\delta^2 + 2\delta\varepsilon\sum x_i + \varepsilon^2\sum x_i^2 \quad \text{M1}$$

$$\sum [\delta + \varepsilon x_i]^2 = n\delta^2 + 2\delta\varepsilon\sum x_i + \varepsilon^2\sum x_i^2 \quad \text{M1}$$

$$\text{Both extra terms are } \geq 0, \text{ and are zero only if } \varepsilon = 0 \text{ and } \delta = 0 \quad \text{E1}$$

$$\text{So } S \geq S_{\min} \text{ with equality only at } (a^*, b^*): \text{ the unique minimiser} \quad \text{A1}$$

Answer: shown 9 marks

28 (8 marks)

$$(a) S_{\lambda}(b) = S_{yy} - 2bS_{xy} + b^2S_{xx} + \lambda b^2 \quad \text{M1}$$

$$\frac{d S_{\lambda}}{d b} = -2S_{xy} + 2b(S_{xx} + \lambda) = 0 \quad \text{M1 A1}$$

$$b_{\lambda} = \frac{S_{xy}}{S_{xx} + \lambda}; \text{ second derivative } 2(S_{xx} + \lambda) > 0 \text{ so a minimum} \quad \text{A1}$$

Answer: shown 4 marks

$$(b) S_{xx} = 10, S_{xy} = 15.0 \quad \text{M1}$$

$$b_{\lambda} = \frac{15.0}{10 + 5} = 1.00 \quad \text{A1}$$

Ordinary $b = 1.50$; b_{λ} is smaller (shrunk towards 0) A1

Answer: $b_{\lambda} = 1.00$ vs $b = 1.50$ 3 marks

$$(c) \lambda \rightarrow 0: b_{\lambda} \rightarrow \frac{S_{xy}}{S_{xx}} \text{ (ordinary least squares); } \lambda \rightarrow \infty: b_{\lambda} \rightarrow 0 \text{ (the line becomes horizontal, } y = \bar{y}) \quad \text{B1}$$

Answer: OLS gradient; 0 1 mark

29 (10 marks)

$$(a) D = \frac{S_{xx} + S_{yy}}{2} - \frac{S_{xx} - S_{yy}}{2} \cos 2\theta - S_{xy} \sin 2\theta \quad \text{M1 A1}$$

$$\frac{d D}{d \theta} = (S_{xx} - S_{yy}) \sin 2\theta - 2S_{xy} \cos 2\theta = 0, \text{ so } \tan 2\theta = \frac{2S_{xy}}{S_{xx} - S_{yy}} = \frac{4}{3} \quad \text{M1 A1}$$

$$\frac{2m}{1 - m^2} = \frac{4}{3}; 2m^2 + 3m - 2 = 0, (2m - 1)(m + 2) = 0, \text{ so } m = \frac{1}{2} \text{ (the minimising root; } m = -2 \text{ is the maximum)} \quad \text{A1}$$

$$\text{Answer: } m = \frac{1}{2} \quad \text{5 marks}$$

$$(b) \tan \theta = \frac{1}{2}; \sin^2 \theta = \frac{1}{5}, \cos^2 \theta = \frac{4}{5}, \sin \theta \cos \theta = \frac{2}{5} \quad \text{M1}$$

$$D = \frac{9}{5} - \frac{8}{5} + \frac{24}{5} = 5 \quad \text{A1}$$

$$\text{Ordinary vertical RSS} = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 6 - \frac{4}{9} = \frac{50}{9} \approx 5.56 > 5 \quad \text{A1}$$

Answer: $D_{\min} = 5 < 5.56$ 3 marks

$$(c) y \text{ on } x: \frac{S_{xy}}{S_{xx}} = \frac{2}{9} \approx 0.222; x \text{ on } y: \frac{S_{yy}}{S_{xy}} = 3 \quad \text{M1}$$

$$0.222 < 0.5 < 3 \quad \text{A1}$$

$$\text{Answer: } \frac{2}{9} < \frac{1}{2} < 3 \quad \text{2 marks}$$

30 (7 marks)

$$(a) \sum e_i = 0: \sum y - na - b \sum x = 0, \text{ so } a = \bar{y} - b \bar{x} \quad \text{M1 A1}$$

$$\sum x_i e_i = 0: \sum xy - a \sum x - b \sum x^2 = 0; \text{ substitute } a \quad \text{M1}$$

$$\sum xy - n \bar{x} \bar{y} - b(\sum x^2 - n \bar{x}^2) = 0, \text{ so } S_{xy} - b S_{xx} = 0 \quad \text{A1}$$

Answer: shown 4 marks

(b) These two conditions have exactly one solution (a, b), the least squares line E1

Answer: unique solution 1 mark

(c) $\sum e = -2$ and $\sum xe = -9$ are not both zero, so it is not the least squares line M1

$$S_{xx} = 10, S_{xy} = 12: y = 2.9 + 1.2x \quad \text{A1}$$

Answer: not LS; $y = 2.9 + 1.2x$ 2 marks

Regression Modelling

1 (3 marks)

$$\ln y = \ln 5 + 3 \ln x$$

M1

Plot $\ln y$ against $\ln x$

A1

Gradient 3, intercept $\ln 5 \approx 1.61$

A1

Answer: $\ln y$ against $\ln x$; gradient 3, intercept $\ln 5$

3 marks

2 (3 marks)

$$y = 10^{0.5 + 1.2 \log x}$$

M1

$$= 10^{0.5} \times x^{1.2}$$

A1

$$y \approx 3.16x^{1.2}$$

A1

Answer: $y \approx 3.16x^{1.2}$

3 marks

3 (2 marks)

The residuals show a clear curved (U-shaped) pattern, not a random scatter

B1

So a linear model is not suitable and a curve (for example a quadratic) should be considered

E1

Answer: not suitable

2 marks

4 (3 marks)

$\ln y = \ln a + x \ln b$, so $\ln a = 1.2$ and $\ln b = 0.3$

M1

$$a = e^{1.2} \approx 3.32$$

A1

$$b = e^{0.3} \approx 1.35$$

A1

Answer: $a \approx 3.32$; $b \approx 1.35$

3 marks

5 (2 marks)

$$P = 2.5 \times 1.4^6$$

M1

$$\approx 18.8$$

A1

Answer: 18.8

2 marks

6 (3 marks)

$\ln y = \ln 3, \ln 3 + \ln 2, \ln 3 + 2 \ln 2, \ln 3 + 3 \ln 2$

M1

Equal steps of $\ln 2$ for equal steps in x , so the graph is a straight line

A1

Gradient = $\ln 2 \approx 0.693$

A1

Answer: gradient $\ln 2 \approx 0.693$

3 marks

7 (6 marks)

$Y = \ln a + bX$, so a straight line of Y on X is expected

B1

$$\Sigma X = 6.579, \Sigma Y = 16.53, \Sigma X^2 = 9.410, \Sigma XY = 21.39$$

M1 A1

$$S_{XX} = 2.195, S_{XY} = 3.259$$

M1

$$b = \frac{S_{XY}}{S_{XX}} \approx 1.48$$

A1

$$\ln a = 2.756 - b \times 1.097 \approx 1.128, a \approx 3.09$$

A1

Answer: $a \approx 3.09$; $b \approx 1.48$

6 marks

8 (6 marks)

(a) $\ln y = \ln a + x \ln b$

B1

$$S_{xx} = 17.5, S_{xy} = 6.901$$

M1

$$\ln b = 0.3943, \ln a = 1.408$$

A1

$$a \approx 4.09, b \approx 1.48$$

A1

Answer: $a \approx 4.09$; $b \approx 1.48$

4 marks

(b) $y = 4.09 \times 1.48^6$

M1

$$\approx 43.6$$

A1

Answer: 43.6

2 marks

9 (5 marks)

$$u = 1, 0.5, 0.25, 0.2, 0.125$$

B1

$$\Sigma u = 2.075, \Sigma y = 30.20, \Sigma u^2 = 1.368, \Sigma uy = 17.12$$

M1

$$S_{uu} = 0.5070, S_{uy} = 4.587$$

A1

$$b = \frac{S_{uy}}{S_{uu}} \approx 9.05$$

M1

$$a = \bar{y} - b\bar{u} \approx 2.29$$

A1

Answer: $a \approx 2.29$; $b \approx 9.05$

5 marks

10 (5 marks)

$$w = 1, 4, 9, 16, 25; \Sigma w = 55, \Sigma y = 65.40$$

B1

$$\Sigma w^2 = 979, \Sigma wy = 1094$$

M1

$$S_{ww} = 979 - \frac{55^2}{5} = 374, S_{wy} = 375.1$$

A1

$$b = \frac{S_{wy}}{S_{ww}} \approx 1.00$$

M1

$$a \approx 2.05$$

A1

Answer: $a \approx 2.05$; $b \approx 1.00$

5 marks

11 (4 marks)Each year the population is multiplied by $e^{0.03} \approx 1.0305$

M1 A1

Percentage increase $\approx 3.05\%$

A1

$$\text{Doubling: } e^{0.03t} = 2, t = \frac{\ln 2}{0.03} \approx 23.1 \text{ years}$$

A1

Answer: 3.05% per year; 23.1 years

4 marks

12 (4 marks)

The residuals are small compared with typical y-values

B1

They alternate in sign with no trend or curvature

B1

There is no systematic pattern and the spread does not change with x

B1

So the model appears to be appropriate

E1

Answer: appropriate

4 marks

13 (4 marks)The correlation between $\ln y$ and x (0.998) is closer to 1

B1

so $\ln y = \ln a + x \ln b$ is a better linear fit than y on x

E1

Exponential model preferred: $y = ab^x$ with $a = e^{\text{intercept}}$, $b = e^{\text{gradient}}$

M1 A1

Answer: exponential model

4 marks

14 (7 marks)

$$\ln T = \ln k + n \ln L$$

B1

$$X = \ln L: -1.386, -0.693, 0, 0.405, 0.693; Y = \ln T: 0.01, 0.351, 0.693, 0.9, 1.04$$

M1 A1

$$S_{XX} = 2.855, S_{XY} = 1.417$$

M1

$$n = \frac{S_{XY}}{S_{XX}} \approx 0.496$$

A1

$$\ln k \approx 0.696 \text{ so } k \approx 2.01$$

A1

n is very close to 0.5, consistent with the theory $T \propto \sqrt{L}$

E1

Answer: $n \approx 0.496$; $k \approx 2.01$

7 marks

15 (7 marks)

$$(a) \ln N = \ln N_0 - kt$$

B1

$$S_{tt} = 1000, S_{tY} = -43.09$$

M1 A1

$$-k = \frac{S_{tY}}{S_{tt}} \approx -0.04309, \text{ so } k \approx 0.04309$$

A1

$$\ln N_0 \approx 6.887, N_0 \approx 980$$

A1

Answer: $N_0 \approx 980$; $k \approx 0.0431$

5 marks

$$(b) N = \frac{1}{2}N_0; e^{-kt} = \frac{1}{2}$$

M1

$$t = \frac{\ln 2}{k} \approx 16.1 \text{ minutes}$$

A1

Answer: 16.1 minutes

2 marks

16 (6 marks)

$$(a) S_{xx} = 28, S_{xy} = 220, b = 7.87$$

M1 A1

$$a = -11.1$$

A1

Residuals: 5.31, 0.14, -3.23, -4.5, -3.07, -0.04, 5.39

A1

Answer: residuals 5.31, 0.14, -3.23, -4.5, -3.07, -0.04, 5.39

4 marks

(b) Residuals are positive, then negative, then positive: a systematic U-shape

B1

A quadratic model $y = a + bx^2$ (or similar) would be more suitable

E1

Answer: U-shape; use a quadratic

2 marks

17 (5 marks)

$$(a) x \text{ and } y: S_{xx} = 17.5, S_{yy} = 1360, S_{xy} = 147, r \approx 0.9541$$

M1 A1

$$x \text{ and } \ln y: S_{YY} = 3.544, S_{xY} = 7.874, r \approx 0.9999$$

M1 A1

Answer: $r = 0.9541$ (linear); $r = 0.9999$ (ln y)

4 marks

$$(b) \text{ Model B has } r \text{ closer to } 1; q = e^{0.4499} \approx 1.57, p = e^{1.190} \approx 3.29$$

B1

Answer: model B; $p \approx 3.29$, $q \approx 1.57$

1 mark

18 (6 marks)

$$(a) \ln u = \ln a + t \ln b$$

B1

$$S_{tt} = 17.5, S_{tY} = 8.360$$

M1 A1

$$\ln b = 0.4777, \ln a = 0.1735; a \approx 1.19, b \approx 1.61$$

A1

Answer: $a \approx 1.19$; $b \approx 1.61$

4 marks

$$(b) t = 8: u \approx 1.19 \times 1.61^8 \approx 54.3 \text{ million}$$

M1

Extrapolation well beyond the data; exponential growth cannot continue indefinitely (market saturation)

E1

Answer: 54.3 million; extrapolation

2 marks

19 (5 marks)

$$y = 10^{0.3 + 0.8x} = 10^{0.3} \times (10^{0.8})^x$$

M1

$$a = 10^{0.3} \approx 2.00, b = 10^{0.8} \approx 6.31$$

A1 A1

$$y = 100: \log_{10} 100 = 2 = 0.3 + 0.8x$$

M1

$$x = \frac{1.7}{0.8} = 2.125$$

A1

Answer: $y = 2.00 \times 6.31^x$; $x = 2.125$

5 marks

20 (7 marks)

$$(a) z = \frac{1}{y} = c + dx$$

B1

Answer: $z = c + dx$

1 mark

$$(b) z = 2, 3.03, 3.846, 5, 5.882, 6.667$$

M1

$$S_{xx} = 17.5, S_{xz} = 16.52$$

M1 A1

$$d = 0.9441$$

A1

$$c = \bar{z} - d\bar{x} = 1.100$$

A1

$$x = 8: z \approx 8.653, \text{ so } y \approx 0.116$$

A1

Answer: $c \approx 1.10$, $d \approx 0.944$; $y \approx 0.116$

6 marks

21 (9 marks)

$$(a) \frac{\partial S}{\partial a} = -2\sum(y - a - bx - cx^2) = 0$$

M1 A1

$$\frac{\partial S}{\partial b} = -2\sum x(y - a - bx - cx^2) = 0$$

M1

$$\frac{\partial S}{\partial c} = -2\sum x^2(y - a - bx - cx^2) = 0, \text{ and rearranging gives the three equations}$$

A1

Answer: shown

4 marks

$$(b) \sum x = 0, \sum x^2 = 10, \sum x^3 = 0, \sum x^4 = 34$$

B1

$$\sum y = 23.1, \sum xy = -1.4, \sum x^2 y = 62.6$$

M1 A1

$$b = \frac{\sum xy}{\sum x^2} = -0.140$$

A1

$$5a + 10c = 23.1 \text{ and } 10a + 34c = 62.6 \text{ give } a \approx 2.28, c \approx 1.17$$

A1

Answer: a ≈ 2.28; b ≈ -0.140; c ≈ 1.17

5 marks

22 (6 marks)

(a) Taking logs of $y = ax^b$ gives $\ln y = \ln a + b \ln x$, so $\ln y$ (not y) should be plotted against $\ln x$. The student's line is $y = c + d \ln x$, a different (logarithmic) model, so the coefficients do not give a and b .

E1

E1

Answer: should plot $\ln y$ against $\ln x$

2 marks

$$(b) \ln a = 1.8, a = e^{1.8} \approx 6.05$$

M1 A1

$$b = 2.4$$

B1

$$y = 6.05 \times 3^{2.4} \approx 84.5$$

A1

Answer: a ≈ 6.05; b = 2.4; y ≈ 84.5

4 marks

23 (8 marks)

$$(a) \text{ Model A: } S_{xx} = 10, S_{xy} = 39.10, b = 3.910, a = -3.010$$

M1 A1

$$\ln y = 0.956, 1.411, 1.946, 2.434, 2.918$$

M1

$$\text{Model B: } \ln q = 0.4947, \ln p = 0.4486$$

M1

$$p \approx 1.57, q \approx 1.64$$

A1

Answer: A: $y = -3.01 + 3.91x$; B: $y = 1.57 \times 1.64^x$

5 marks

$$(b) \text{ Model A: } \sum(y - \hat{y})^2 \approx 11.7$$

M1

$$\text{Model B: } \sum(y - \hat{y})^2 \approx 0.0336$$

A1

Model B has the smaller sum, so it fits better

A1

Answer: A: 11.7; B: 0.0336; B better

3 marks

24 (8 marks)

$$(a) \ln V = 4.787, 4.927, 5.063, 5.209, 5.342, 5.485, 5.62$$

M1

$$S_{tt} = 28, S_{YY} = 0.5415, S_{tY} = 3.894$$

M1

$$r \approx 1.0000, \text{ very close to } 1$$

A1

Answer: r ≈ 1.0000

3 marks

$$(b) \ln b = \frac{S_{tY}}{S_{tt}} \approx 0.1391, b \approx 1.15$$

M1 A1

 $a \approx 120$; the investment grows by about 14.9% each year

A1

Answer: a ≈ 120; b ≈ 1.15 (14.9% growth per year)

3 marks

$$(c) \text{ Fitted } V(3) = 120 \times 1.15^3 \approx 182, \text{ residual} \approx 0.836$$

M1

$$V(10) \approx 482; \text{ this is extrapolation (} t = 10 \text{ is outside } 0 \text{ to } 6) \text{ so less reliable}$$

A1

Answer: residual ≈ 0.836; V(10) ≈ 482

2 marks

25 (6 marks)

$$(a) \ln 6 = \ln a + b \ln 2 \text{ and } \ln 48 = \ln a + b \ln 8$$

M1

$$\text{Subtracting: } \ln 8 = b \ln 4, \text{ so } b = \frac{\ln 8}{\ln 4} = \frac{3 \ln 2}{2 \ln 2} = \frac{3}{2}$$

M1 A1

$$6 = a \times 2^{3/2}, \text{ so } a = \frac{6}{2\sqrt{2}} = \frac{3}{\sqrt{2}} = \frac{3\sqrt{2}}{2}$$

A1

$$\text{Answer: } b = \frac{3}{2}; a = \frac{3\sqrt{2}}{2}$$

4 marks

$$(b) y(4) = \frac{3}{\sqrt{2}} \times 4^{3/2} = \frac{3}{\sqrt{2}} \times 8 = \frac{24}{\sqrt{2}} \approx 16.97$$

M1

17 is close to 16.97, so the data are consistent with this model to within rounding

A1

Answer: $y(4) \approx 16.97$; consistent

2 marks

26 (8 marks)

$$(a) \frac{100}{y} = 1 + e^{c - kt}$$

M1

$$\frac{100}{y} - 1 = e^{c - kt}, \text{ so } \ln\left(\frac{100}{y} - 1\right) = c - kt$$

A1

Answer: shown

2 marks

$$(b) Z = 1.992, 1.325, 0.663, -0.04, -0.754, -1.45$$

M1

$$S_{tt} = 17.5, S_{tz} = -12.08$$

M1

$$-k = -0.6900, \text{ so } k \approx 0.690$$

A1

$$c = 2.01$$

A1

Answer: $c \approx 2.01$; $k \approx 0.690$

4 marks

$$(c) y = 50: Z = \ln 1 = 0 = c - kt$$

M1

$$t = \frac{c}{k} \approx 2.92 \text{ days}$$

A1

Answer: 2.92 days

2 marks

27 (12 marks)

$$(a) \ln y = \ln a + b \ln x + \ln \varepsilon$$

M1

The error $\ln \varepsilon$ is additive and approximately constant in spread for all x , which is what linear regression assumes

E1

Answer: shown

2 marks

$$(b) X = \ln x, Y = \ln y: S_{XX} = 2.195, S_{XY} = 3.712$$

M1 A1

$$b = 1.691$$

A1

$$\ln a = 1.182, a \approx 3.26$$

A1

Answer: $a \approx 3.26$; $b \approx 1.69$

4 marks

$$(c) \text{Residuals: } 0.042, -0.031, -0.064, -0.005, 0.027, 0.03$$

M1 A1

The residuals are small and show no clear pattern, so the power-law model fits well

E1

Answer: 0.042, -0.031, -0.064, -0.005, 0.027, 0.03

3 marks

$$(d) y = 3.26 \times 8^{1.69} \approx 110$$

M1 A1

Back-transforming gives the median, not the mean: $E[\varepsilon] > 1$ when $\ln \varepsilon$ is symmetric, so the mean of y is slightly larger

E1

Answer: $y \approx 110$

3 marks

28 (9 marks)

$$(a) \text{Ratios: } 1.47, 1.48, 1.45, 1.47, 1.47, 1.48, 1.46$$

M1 A1

They are almost constant (about 1.47), so y is approximately a geometric sequence: $y = ab^x$

E1

Answer: ratios ≈ 1.47

3 marks

$$(b) S_{xx} = 42, S_{xy} = 16.11$$

M1 A1

$$\ln b = 0.3835, b \approx 1.47; \ln a = 1.097, a \approx 3.00$$

A1

Growth $\approx 46.7\%$ per unit of x

A1

Answer: $a \approx 3.00$; $b \approx 1.47$

4 marks

$$(c) \text{Linear fit: } y = -2.78 + 5.49x$$

M1

Residuals: $x = 0$: 5.78, $x = 3$: -4.29, $x = 7$: 8.34 (positive, negative, positive): a U-shape, so the data are curved (convex)

A1

Answer: residuals 5.78, -4.29, 8.34

2 marks

29 (10 marks)

$$(a) T - 20 = Ae^{-kt}, \text{ so } \ln(T - 20) = \ln A - kt$$

B1

Answer: shown

1 mark

$$(b) Z = \ln(T - 20) = 4.317, 3.942, 3.564, 3.178, 2.773, 2.351$$

M1

$$S_{tt} = 437.5, S_{tz} = -34.31$$

M1 A1

$$-k = -0.07842, k \approx 0.0784$$

A1

$$\ln A = 4.334, A \approx 76.3$$

Answer: $A \approx 76.3$; $k \approx 0.0784$

A1
5 marks

$$(c) T = 20 + 76.3 e^{-0.0784 \times 40} \\ \approx 23.3 \text{ }^{\circ}\text{C}$$

M1
A1

Answer: $23.3 \text{ }^{\circ}\text{C}$

2 marks

(d) Gradient ≈ -0.08546 (compared with -0.07842)

M1

r for the wrong assumption is -0.99883 (compared with -0.99974), so the data look less linear and the model fit is worse

A1

Answer: gradient ≈ -0.08546 ; poorer linearity

2 marks

30 (8 marks)

$$(a) \frac{y}{x} = ae^{bx}$$

M1

$$\ln\left(\frac{y}{x}\right) = \ln a + bx: \text{gradient } b, \text{intercept } \ln a$$

A1

Answer: $\ln(y/x) = \ln a + bx$

2 marks

$$(b) Z = \ln\left(\frac{y}{x}\right): 1.176, 0.99, 0.779, 0.612, 0.188, -0.255$$

M1

$$S_{xx} = 8.500, S_{xz} = -3.464$$

M1 A1

$$b \approx -0.408, \ln a \approx 1.40, a \approx 4.04$$

A1

Answer: $a \approx 4.04$; $b \approx -0.408$

4 marks

$$(c) \frac{d}{dx} \frac{y}{x} = ae^{bx}(1 + bx) = 0 \text{ at } x = -\frac{1}{b}$$

M1

$$x \approx 2.45 \text{ hours; } y = a(-1/b)e^{-1} \approx 3.65$$

A1

Answer: peak at $x \approx 2.45 \text{ h}$; $y \approx 3.65$

2 marks

Statistical Inference for Correlation

1 (2 marks)

$$H_0: \rho = 0$$

B1

$$H_1: \rho > 0$$

B1

Answer: $H_0: \rho = 0$; $H_1: \rho > 0$

2 marks

2 (3 marks)

$$0.81 > 0.6215$$

M1 A1

So reject H_0 : there is evidence of positive correlation at the 5% level

A1

Answer: reject H_0

3 marks

3 (3 marks)

$$0.62 > 0.5494, \text{ so reject } H_0$$

M1 A1

There is evidence that students who do well in mathematics tend to do well in physics (positive correlation between the marks)

A1

Answer: evidence of positive correlation

3 marks

4 (3 marks)

$$(i) \text{ Two-tailed: } H_1: \rho \neq 0$$

B1

$$(ii) \text{ One-tailed: } H_1: \rho > 0$$

B1

$$(iii) \text{ One-tailed: } H_1: \rho < 0$$

B1

Answer: (i) two-tailed; (ii), (iii) one-tailed

3 marks

5 (2 marks)

The probability of rejecting H_0 when it is actually true is 5%

B1

so there is a 5% chance of concluding there is correlation when in fact there is none

E1

Answer: $P(\text{reject } H_0 \mid H_0 \text{ true}) = 0.05$

2 marks

6 (3 marks)

$0.5 < 0.8054$, so do not reject H_0

M1 A1

With so few pairs, even a moderate r could easily occur by chance: a very large r is needed for significance

E1

Answer: do not reject H_0

3 marks

7 (4 marks)

$H_0: \rho = 0$, $H_1: \rho > 0$

B1

Critical value 0.5494

B1

$r = 0.68 > 0.5494$

M1

Reject H_0 : there is evidence of positive correlation

A1

Answer: reject H_0

4 marks

8 (4 marks)

$H_0: \rho = 0$, $H_1: \rho \neq 0$

B1

Critical values ± 0.5760

B1

$-0.58 < -0.5760$, so r lies in the critical region

M1

Reject H_0 : there is evidence of correlation (negative)

A1

Answer: reject H_0

4 marks

9 (6 marks)

$S_{xx} = 148.0$, $S_{yy} = 262.2$, $S_{xy} = 188.7$

M1 A1

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \approx 0.9577$$

A1

$H_0: \rho = 0$, $H_1: \rho > 0$

B1

$0.9577 > 0.5822$

M1

Reject H_0 : evidence of positive correlation

A1

Answer: $r \approx 0.9577$; reject H_0

6 marks

10 (4 marks)

H_0 : no association, H_1 : there is association (two-tailed)

B1

Critical value 0.7381

B1

$0.77 > 0.7381$

M1

Reject H_0 : evidence of association between the judges' rankings

A1

Answer: reject H_0

4 marks

11 (3 marks)

No: correlation does not imply causation

B1

Both are caused by a third variable, hot sunny weather

E1

More people on the beach on sunny days leads to higher sales and more sunburn

E1

Answer: no; hot weather is a confounding variable

3 marks

12 (4 marks)

$H_0: \rho = 0$, $H_1: \rho < 0$

B1

Critical value -0.5923

B1

$-0.62 > -0.5923$ so r is not in the critical region

M1

Do not reject H_0 : insufficient evidence of negative correlation at the 1% level

A1

Answer: do not reject H_0

4 marks

13 (4 marks)

$0.032 < 0.05$: reject H_0 at the 5% level

B1

$0.032 > 0.01$: do not reject H_0 at the 1% level

B1

The p-value is the probability of obtaining a sample correlation at least as extreme as that observed if H_0 is true

B1 B1

Answer: reject at 5%; not at 1%

4 marks

14 (7 marks)

$$\Sigma x = 117, \Sigma y = 251, \Sigma x^2 = 2439, \Sigma y^2 = 10811, \Sigma xy = 5100$$

M1

$$S_{xx} = 157.5, S_{yy} = 310.8, S_{xy} = 205.5$$

M1 A1

$$r \approx 0.9288$$

A1

$$H_0: \rho = 0, H_1: \rho > 0, \text{critical value } 0.7293$$

B1

$$0.9288 > 0.7293 \text{ so reject } H_0$$

M1

There is evidence that more sleep deprivation is associated with more errors

A1

Answer: $r \approx 0.9288$; reject H_0

7 marks

15 (7 marks)

x ranks: 3, 5, 1, 7, 4, 6, 2

M1 A1

y ranks: 3, 6, 1, 5, 4, 7, 2

A1

$$\Sigma d^2 = 6, r_s = 1 - \frac{6 \times 6}{7 \times 48} \approx 0.8929$$

M1 A1

H_0 : no association, H_1 : positive association; critical value 0.7143

B1

$$0.8929 > 0.7143: \text{reject } H_0$$

A1

Answer: $r_s \approx 0.8929$; reject H_0

7 marks

16 (5 marks)

(a) Diagram A

B1

The points form a roughly elliptical cloud, as expected for bivariate normal data

E1

Answer: A

2 marks

(b) Spearman's rank test

B1

The relationship is clearly curved but increasing (monotonic), so ranks measure it better and no normality is needed

E1

Answer: Spearman's

2 marks

(c) To check the data look linear and elliptical and to spot outliers

E1

Answer: check linearity/outliers

1 mark

17 (6 marks)

$H_0: \rho = 0, H_1: \rho \neq 0$; degrees of freedom 10

B1

$$t = \frac{0.62\sqrt{10}}{\sqrt{1 - 0.3844}}$$

M1

$$= \frac{1.9606}{0.7846} \approx 2.499$$

A1

Critical region $|t| > 2.2281$

B1

$$2.499 > 2.2281 \text{ so reject } H_0$$

M1

There is evidence of correlation between the variables

A1

Answer: $t \approx 2.499$; reject H_0

6 marks

18 (6 marks)

Critical $|t| = 2.1009$ with $n - 2 = 18$

B1

$$\frac{r\sqrt{18}}{\sqrt{1 - r^2}} = 2.1009$$

M1

$$r^2 \times 18 = 4.4139(1 - r^2)$$

M1

$$r^2(18 + 4.4139) = 4.4139$$

A1

$$r^2 = \frac{4.4139}{22.414} \approx 0.1969$$

M1

$$|r| \approx 0.4438$$

A1

Answer: $|r| \approx 0.4438$

6 marks

19 (5 marks)

$n = 10$: $0.40 < 0.5494$, do not reject H_0

M1 A1

$n = 30$: $0.40 > 0.3061$, reject H_0

M1 A1

The same r is significant for a larger sample because a larger sample is less likely to give such an r by chance

E1

Answer: $n = 10$: not significant; $n = 30$: significant

5 marks

20 (6 marks)

$H_0: \rho = 0$, $H_1: \rho \neq 0$ for both

B1

A: $0.55 < 0.6319$, do not reject H_0

M1 A1

B: $0.40 > 0.3961$, reject H_0

M1 A1

The larger sample size makes a smaller r less likely to occur by chance if $\rho = 0$

E1

Answer: A not significant; B significant

6 marks

21 (7 marks)

$$(a) t^2(1 - r^2) = r^2(n - 2)$$

M1

$$t^2 = r^2(n - 2 + t^2)$$

A1

$$r = \frac{t}{\sqrt{n - 2 + t^2}} \text{ (positive root for } t > 0 \text{)}$$

A1

Answer: shown

3 marks

$$(b) r = \frac{1.9432}{\sqrt{6 + 1.9432^2}}$$

M1

$$\approx 0.6215$$

A1

Answer: 0.6215

2 marks

(c) For large n , t tends to a fixed value and $n - 2$ grows

M1

so $t/\sqrt{n - 2 + t^2} \rightarrow 0$: smaller correlations become significant

E1

Answer: critical r decreases with n

2 marks

22 (6 marks)

(a) The question asks about any correlation, so H_1 should be $\rho \neq 0$ and the test is two-tailed

E1

The critical values should be ± 0.5760 (two-tailed, 5%), giving a region $|r| > 0.5760$

E1

A hypothesis test can only provide evidence, never proof

E1

Answer: one-tailed used; wrong critical value; 'proves'

3 marks

(b) $H_0: \rho = 0$, $H_1: \rho \neq 0$, critical value ± 0.5760

B1

$|-0.55| = 0.55 < 0.5760$, so r is not in the critical region

M1

Do not reject H_0 : insufficient evidence of correlation at the 5% level

A1

Answer: do not reject H_0

3 marks

23 (8 marks)

$H_0: \rho = 0$, $H_1: \rho > 0$

B1

$$\Sigma x = 62.70, \Sigma y = 396, \Sigma x^2 = 538.47, \Sigma y^2 = 20604, \Sigma xy = 3316.3$$

M1 A1

$$S_{xx} = 47.06, S_{yy} = 1002, S_{xy} = 212.7$$

M1 A1

$$r \approx 0.9793$$

A1

Critical value 0.7887; $0.9793 > 0.7887$

M1

Reject H_0 : strong evidence at the 1% level that more rainfall is associated with more hay

A1

Answer: $r \approx 0.9793$; reject H_0

8 marks

24 (6 marks)

(a) $H_0: \rho = 0$, $H_1: \rho \neq 0$

B1

$$0.45 > 0.3961$$

M1

Reject H_0 : there is evidence of correlation

A1

Answer: reject H_0

3 marks

(b) $r^2 = 0.2025$

B1

Only about 20% of the variation in y is explained by the linear relationship

E1

Predictions would have large errors (80% of the variation is unexplained)

E1

Answer: $r^2 \approx 0.20$ **3 marks****25 (7 marks)**

$$n - 2 = 7$$

B1

$$t = \frac{0.72\sqrt{7}}{\sqrt{1 - 0.5184}}$$

M1

$$= \frac{1.9050}{0.6940}$$

A1

$$t \approx 2.745$$

A12.745 > 1.8946: reject H_0 at the 5% level**M1 A1**

2.745 < 2.9980: do not reject at the 1% level, so the p-value is between 1% and 5%

A1**Answer: $t \approx 2.745$; significant at 5% but not at 1%****7 marks****26 (6 marks)****(a)** Each test: $P(\text{not significant}) = 0.95$ **M1**

$$0.95^{20} \approx 0.3585$$

A1**Answer: 0.3585****2 marks****(b)** $1 - 0.3585 \approx 0.6415$ **M1 A1**

With many tests, a “significant” result is likely to occur by chance alone

E1

Significant correlations found this way need to be confirmed with new data

E1**Answer: 0.642****4 marks****27 (9 marks)**

$$s^2 = \frac{S_{yy}(1 - r^2)}{n - 2}$$

B1

$$t = \frac{S_{xy}}{S_{xx}} \div \sqrt{\frac{S_{yy}(1 - r^2)}{(n - 2)S_{xx}}}$$

M1 A1

$$= \frac{S_{xy}}{S_{xx}} \times \sqrt{\frac{(n - 2)S_{xx}}{S_{yy}(1 - r^2)}}$$

M1

$$= \frac{S_{xy}\sqrt{n - 2}}{\sqrt{S_{xx}}\sqrt{S_{yy}}\sqrt{1 - r^2}}$$

A1

$$= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \times \frac{\sqrt{n - 2}}{\sqrt{1 - r^2}}$$

M1

$$= r \frac{\sqrt{n - 2}}{\sqrt{1 - r^2}}$$

A1The statistics are identical and have the same distribution under H_0 **E1**

so the two tests always give the same conclusion

E1**Answer: shown****9 marks****28 (7 marks)**Need $r = 0.5 >$ critical value**M1** $n = 11$: critical value $0.5214 > 0.5$, not significant**A1** $n = 12$: critical value $0.4973 < 0.5$, significant**A1**So the smallest sample size is $n = 12$ **A1**

$$n = 12: t = \frac{0.5\sqrt{10}}{\sqrt{1 - 0.25}}$$

M1

$$= 1.826$$

A1which exceeds the critical value of t_{10} , 1.8125**A1****Answer: $n = 12$** **7 marks**

29 (9 marks)

(a) $\Sigma x = 83.90$, $\Sigma y = 121.6$, $\Sigma x^2 = 949.07$, $\Sigma y^2 = 1949.6$, $\Sigma xy = 1069.2$ M1

$r \approx 0.1440$ A1

$H_0: \rho = 0$, $H_1: \rho \neq 0$; $0.1440 < 0.6319$ M1

Do not reject H_0 : no significant evidence of correlation A1

Answer: $r \approx 0.1440$; do not reject H_0 4 marks

(b) Without the pair: $r \approx 0.8157$ M1 A1

$0.8157 > 0.6664$ M1

Reject H_0 : significant positive correlation A1

Answer: $r \approx 0.8157$; reject H_0 4 marks

(c) The conclusion depends on a single point: it should be checked (and removed only if it is a genuine error), not deleted just to obtain significance E1

Answer: check before removing 1 mark

30 (9 marks)

$\Sigma x = 298$, $\Sigma y = 432$, $\Sigma x^2 = 10932$, $\Sigma y^2 = 21610$, $\Sigma xy = 14855$ M1

$S_{xx} = 2052$, $S_{yy} = 2948$, $S_{xy} = 1981$, $r \approx 0.8057$ M1 A1

PMCC test: $H_0: \rho = 0$, $H_1: \rho \neq 0$; $0.8057 > 0.6319$: reject H_0 B1 A1

Ranks of x: 2, 10, 5, 3, 1, 9, 7, 6, 4, 8; ranks of y: 1, 10, 6, 2, 5, 9, 8, 7, 3, 4 M1

$\Sigma d^2 = 38$, $r_s = 1 - \frac{6 \times 38}{990} \approx 0.7697$ A1

$r_s > 0.6485$: reject H_0 for the rank test M1

Conclusion: there is significant evidence of correlation between hours and score (r is positive, so more hours are associated with higher scores); the PMCC test assumes bivariate normality while the rank test does not E1

Answer: $r \approx 0.8057$; $r_s \approx 0.7697$; both tests significant 9 marks