

Scatter Diagrams

1 (3 marks)

- (a) Positive: taller people tend to have longer arm spans
 (b) Negative: warmer weather means a lower heating bill
 (c) No correlation: unrelated

B1

B1

B1

Answer: (a) positive (b) negative (c) none

3 marks

2 (2 marks)

The points lie close to an upward sloping straight line
 Strong positive correlation: more revision goes with higher marks

B1

B1

Answer: Strong positive correlation

2 marks

3 (2 marks)

Explanatory: mass of fertiliser (the variable that is controlled or changed)
 Response: yield of tomatoes

B1

B1

Answer: Explanatory: fertiliser; response: yield

2 marks

4 (2 marks)

$$\bar{x} = \frac{2 + 4 + 6 + 8 + 10}{5} = 6$$

B1

$$\bar{y} = \frac{5 + 9 + 10 + 14 + 17}{5} = 11$$

B1

Answer: (6, 11)

2 marks

5 (2 marks)

Strong negative correlation
 Older cars tend to be worth less

B1

E1

Answer: Strong negative: value falls with age

2 marks

6 (2 marks)

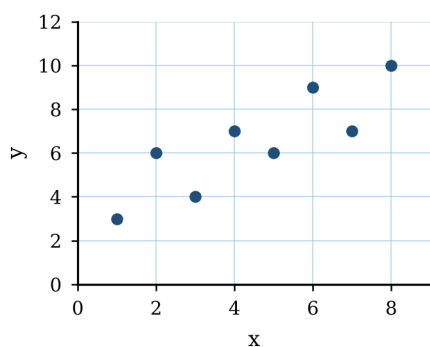
Points that scatter widely
 but with a general upward trend

B1

B1

Answer: See diagram

2 marks



7 (3 marks)

(a) Reading from the graph: (7, 3)

B1

Answer: (7, 3)

1 mark

(b) Possible reason: recording error or a patient who was discharged early for another reason
 Check the original record before deciding whether to remove it

E1

B1

Answer: Check data; possible error

2 marks

8 (5 marks)

$$(a) \bar{x} = \frac{1 + 1.4 + 1.6 + 2 + 2.4 + 3}{6} = 1.9$$

B1

$$\bar{y} = \frac{5.1 + 6 + 6.4 + 7.2 + 8.1 + 9.2}{6} = 7$$

B1

Answer: (1.9, 7)

2 marks

(b) All six points plotted accurately

B1

Mean point (1.9, 7) plotted

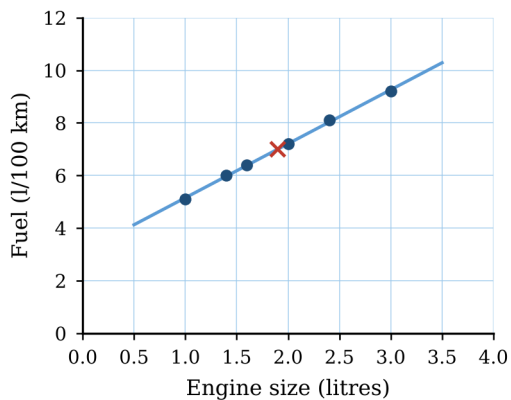
B1

Straight line through the mean point following the trend

B1

Answer: See diagram

3 marks



9 (3 marks)

(a) A: strong positive correlation

B1

B: weaker positive correlation

B1

Answer: A strong, B weak (both positive)

2 marks

(b) A, since its points lie closer to a straight line

E1

Answer: A

1 mark

10 (2 marks)

Both variables are linked to a third factor, the weather (hot weather)

E1

More people eat ice cream and more people swim in hot weather, so correlation does not imply causation

E1

Answer: Hot weather is a common cause

2 marks

11 (2 marks)

The relationship is clearly curved (y decreases then increases)

B1

A straight line cannot fit data that change direction, so the correlation coefficient would be misleading

E1

Answer: Curved: not linear

2 marks

12 (2 marks)

The points show no pattern, so there is no (linear) correlation

B1

Shoe size does not help to predict spelling mark

E1

Answer: No correlation

2 marks

13 (4 marks)

$$\text{Gradient} = \frac{60 - 20}{50 - 10} = 1$$

M1 A1

$$y - 20 = 1(x - 10), \text{ so } y = 10 + x$$

A1

When $x = 35$, $y = 45$

A1

Answer: $y = 10 + x$; 45

4 marks

14 (5 marks)

$$(a) \Sigma x = 83, \text{ so } \bar{x} = 9.2222$$

B1

$$\Sigma y = 199.7, \text{ so } \bar{y} = 22.1889$$

B1

Answer: (9.2222, 22.1889)

2 marks

$$(b) y - 22.1889 = 1.8(x - 9.2222)$$

M1

$$y = 5.5889 + 1.8x$$

A1

When $x = 20$, $y = 41.5889$

A1

Answer: $y = 5.5889 + 1.8x$; 41.5889

3 marks

15 (4 marks)**(a)** Strong positive correlation overall

B1

Answer: Strong positive

1 mark

(b) Within each group the points show a weaker, still positive, trend

B1

Group 2 has higher values of both variables than Group 1 (the groups are separated)

E1

The overall correlation is mostly caused by the difference between the groups, so the groups should be analysed separately

E1

Answer: Group differences inflate correlation

3 marks

16 (5 marks)

$$\Sigma x = 60, \Sigma y = 176, \Sigma x^2 = 548, \Sigma y^2 = 4124, \Sigma xy = 1474$$

M1

$$S_{xx} = 548 - \frac{60^2}{8} = 98$$

M1 A1

$$S_{xy} = 1474 - \frac{60 \times 176}{8} = 154, S_{yy} = 252$$

A1

 $S_{xy} > 0$ so positive correlation

E1

Answer: $S_{xx} = 98, S_{xy} = 154, S_{yy} = 252$; positive

5 marks

17 (5 marks)

$$(a) \bar{x} = \frac{36}{8} = 4.5$$

B1

$$\bar{y} = \frac{84}{8} = 10.5$$

B1

Answer: (4.5, 10.5)

2 marks

$$(b) \bar{x} = \frac{45}{9} = 5$$

B1

$$\bar{y} = \frac{87}{9} = 9.6667$$

B1

P pulls the mean of y down by 0.833, because it lies far below the trend

E1

Answer: (5, 9.6667)

3 marks

18 (3 marks)

The two coefficients are almost the same

E1

The strength of correlation depends on how closely the points follow a line, not on the gradient

E1

So the student is wrong: both sets show equally strong positive correlation

E1

Answer: Student is wrong

3 marks

19 (5 marks)

$$(a) \Sigma x = 0, \Sigma xy = \Sigma x^3 = 0$$

M1

$$S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 0 - 0$$

M1

$$= 0$$

A1

Answer: Shown

3 marks

(b) $r = 0$ means no linear correlation

E1

Here y is exactly determined by x through a curved relationship

E1

Answer: r only measures linear association

2 marks

20 (2 marks)

The pattern is seasonal (it rises then falls), so a straight line trend does not apply

E1

Only one year of data is shown, so the pattern cannot be extended

E1

Answer: Seasonal pattern; one year only

2 marks

21 (7 marks)**(a)** (27, 20)

B1

y should be about 50 for this x, as other pairs show y increasing steadily with x

E1

Answer: (27, 20)

2 marks

(b) Using a calculator: $r = 0.722$

M1 A1

Answer: 0.722

2 marks

(c) Without the pair: $r = 0.994$

M1 A1

The correlation is much stronger once the outlier is removed, showing that one point can greatly reduce r	E1
Answer: 0.994	3 marks
22 (6 marks)	
(a) Group A: $r = 0.9976$	B1
Group B: $r = 0.9935$	B1
All pairs: $r = 0.9859$	B1
Answer: 0.9976, 0.9935, 0.9859	3 marks
(b) The groups differ systematically (B has much higher values of both variables), producing a big jump between the groups	E1
The overall figure measures this separation as well as the trends within groups	E1
Answer: Group separation	2 marks
(c) Analyse each group separately	E1
Answer: Separately	1 mark
23 (6 marks)	
(a) By calculator: $r = 0.9608$	M1 A1
Answer: 0.9608	2 marks
(b) Same value 0.9608	B1
Linear coding with positive multipliers does not change r	E1
Answer: 0.9608	2 marks
(c) The multiplier -3 is negative, so the sign changes	M1
$r = -0.9608$	A1
Answer: -0.9608	2 marks
24 (6 marks)	
(a) $r = 0.944$	B1
Answer: 0.944	1 mark
(b) $\ln y = 1.569, 1.808, 2.303, 2.573, 3.04, 3.318, 3.766, 4.078$	M1
$r = 0.9983$	A1
$\ln y$ is more strongly (linearly) correlated with x than y is	E1
Answer: 0.9983	3 marks
(c) Exponential model $y = ab^x$	B1
Since $\ln y = \ln a + x \ln b$ is linear in x	E1
Answer: $y = ab^x$	2 marks
25 (4 marks)	
(a) Positive correlation	B1
The points are tightly grouped for small x and become more scattered as x increases	B1
Answer: Positive; spread increases with x	2 marks
(b) The variability of y about the trend is larger for large x	E1
So the true value could be much further from the line	E1
Answer: More variability at large x	2 marks
26 (6 marks)	
(a) Increasing over time	B1
$t = 5$ (2020) is unusually low	B1
Answer: Upward trend; 2020 low	2 marks
(b) All years: $r = 0.687$	B1
Without 2020: $r = 0.99$	B1
Answer: 0.687 and 0.99	2 marks
(c) A closure or reduced opening (for example pandemic restrictions)	E1
It is a genuine observation with an explanation, so should be kept but noted, or analysed separately	E1
Answer: Genuine but special	2 marks

27 (7 marks)**(a)** By calculator $r = 0.955$

M1 A1

Answer: 0.955

2 marks

(b) Pairs (5, 14), (6, 20), (7, 17), (8, 24)

M1

 $r = 0.816$

A1

Answer: 0.816

2 marks

(c) The correlation coefficient is smaller for the restricted range

E1

A narrow range of x hides the overall trend and a few points have a large influence

E1

A sample of only four points gives unreliable evidence

E1

Answer: Restricted range lowers r

3 marks

28 (9 marks)**(a)** Expand: $\Sigma(xy - \bar{x}\bar{y} - \bar{x}\bar{y} + \bar{x}\bar{y})$

M1

$$= \Sigma xy - \bar{y}\Sigma x - \bar{x}\Sigma y + n\bar{x}\bar{y}$$

M1

Using $\Sigma x = n\bar{x}$ and $\Sigma y = n\bar{y}$: $\Sigma xy - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y} = \Sigma xy - n\bar{x}\bar{y}$

A1

Answer: Shown

3 marks

(b) $u - \bar{u} = a(x - \bar{x})$, $v - \bar{v} = c(y - \bar{y})$

M1

$$S_{uv} = \Sigma a(x - \bar{x}) c(y - \bar{y})$$

M1

$$= ac S_{xy}$$

A1

Answer: Shown

3 marks

(c) $S_{uu} = a^2 S_{xx}$ and $S_{vv} = c^2 S_{yy}$

M1

$$r_{uv} = \frac{ac S_{xy}}{\sqrt{a^2 S_{xx} \times c^2 S_{yy}}}$$

M1

$$= \frac{ac S_{xy}}{ac \sqrt{S_{xx} S_{yy}}} = r_{xy}$$

A1

Answer: Shown

3 marks

29 (8 marks)**(a)** Fitted $y = 6.1133 + 1.6606x$

M1

Residuals: $-0.33, 0.44, -1.28, 0.30, -1.72, 5.46, -1.16, -1.58, 0.00, -0.13$

M1 A1

Answer: $-0.33, 0.44, -1.28, 0.30, -1.72, 5.46, -1.16, -1.58, 0.00, -0.13$

3 marks

(b) $2s = 4.397$

B1

Flagged: (12, 31.5)

A1

Answer: (12, 31.5)

2 marks

(c) With all pairs $r = 0.9794$

B1

Without: $r = 0.9969$

M1

Only a small change, so no single point dominates

E1

Answer: 0.9969

3 marks

30 (10 marks)**(a)** $\ln y = \ln a + n \ln x$

M1

Compare with $Y = c + mX$: $Y = \ln y$, $X = \ln x$

M1

Gradient n , intercept $\ln a$

A1

Answer: Gradient n , intercept $\ln a$

3 marks

(b) $r(x, y) = 0.9911$

B1

 $r(x, \ln y) = 0.9639$

B1

 $r(\ln x, \ln y) = 0.9965$

B1

Answer: 0.9911, 0.9639, 0.9965

3 marks

(c) The coefficient closest to 1 is for $(\ln x, \ln y)$, so $y = ax^n$ is the better model

E1

Gradient $n = 1.445$

M1 A1

$$a = e^{0.825} = 2.282$$

A1

Answer: $n = 1.445$, $a = 2.282$

4 marks

Correlation

1 (4 marks)

- (a) strong positive B1
 (b) very weak negative (almost none) B1
 (c) strong negative B1
 (d) moderate positive B1

Answer: (a) strong positive (b) very weak negative (c) strong negative (d) moderate positive 4 marks

2 (3 marks)

- $-1 \leq r \leq 1$ B1
 $r = 1$: perfect positive linear correlation (all points on a line with positive gradient) B1
 $r = -1$: perfect negative linear correlation B1

Answer: $-1 \leq r \leq 1$ 3 marks

3 (2 marks)

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{48}{\sqrt{40 \times 90}} \quad \text{M1}$$

$$= \frac{48}{60} = 0.8 \quad \text{A1}$$

Answer: 0.8 2 marks

4 (2 marks)

- $H_0: \rho = 0$ B1
 $H_1: \rho > 0$ B1

Answer: $H_0: \rho = 0$, $H_1: \rho > 0$ 2 marks

5 (2 marks)

- $0.71 > 0.5494$ M1
 Reject H_0 : there is evidence of positive correlation A1

Answer: Reject H_0 2 marks

6 (3 marks)

- (a) 0.6 B1
 (b) 0.6 B1
 (c) -0.6 B1

Answer: 0.6, 0.6, -0.6 3 marks

7 (4 marks)

$$S_{xx} = 1064 - \frac{84^2}{8} = 182 \quad \text{M1}$$

$$S_{yy} = 1857 - \frac{113^2}{8} = 260.875, \quad S_{xy} = 1397 - \frac{84 \times 113}{8} = 210.5 \quad \text{A1}$$

$$r = \frac{210.5}{\sqrt{182 \times 260.875}} \quad \text{M1}$$

$$= 0.966 \quad \text{A1}$$

Answer: 0.966 4 marks

8 (4 marks)

- $r = 0.9709$ (calculator) B1
 $H_0: \rho = 0$, $H_1: \rho > 0$ B1
 $0.9709 > 0.6215$ M1

Reject H_0 : there is evidence of positive correlation A1

Answer: $r = 0.971$; reject H_0 4 marks

9 (3 marks)

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

B1

$$|r| = 0.62 > 0.576$$

M1

Reject H_0 : there is evidence of (negative) correlation

A1

Answer: Reject H_0

3 marks

10 (2 marks)

Strong positive linear correlation

E1

Taller students tend to have a greater mass

E1

Answer: Taller students tend to be heavier

2 marks

11 (2 marks)

0.78

B1

A linear transformation with a positive multiplier does not change r

E1

Answer: 0.78

2 marks

12 (3 marks)

$$0.58 < 0.5822$$

M1

So the result is not significant at 5%; there is insufficient evidence of correlation

A1

Significance depends on the sample size as well as the size of r

E1

Answer: Not significant

3 marks

13 (3 marks)

$$r = -0.9817 \text{ (calculator)}$$

B1 B1

Strong negative correlation: students with more absences tend to have lower marks

E1

Answer: $r = -0.982$, strong negative

3 marks

14 (6 marks)

$$(a) r = -0.9738$$

B1

Answer: -0.9738

1 mark

$$(b) H_0: \rho = 0, H_1: \rho < 0$$

B1

Critical region: $r < -0.7155$

B1

$$-0.9738 < -0.7155, \text{ so reject } H_0$$

M1

There is evidence that more television is associated with lower marks

A1

Answer: Reject H_0

4 marks

(c) No, correlation does not prove causation; another factor such as time spent revising might explain both

E1

Answer: No

1 mark

15 (5 marks)

$$r = 0.8845 \text{ (calculator)}$$

B1

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

B1

Two-tailed 10% means 5% in each tail: critical value 0.4973

M1

$$0.8845 > 0.4973$$

M1

Reject H_0 : evidence of correlation

A1

Answer: $r = 0.884$; reject H_0

5 marks

16 (6 marks)

$$S_{xx} = 1930 - \frac{144^2}{12} = 202$$

M1 A1

$$S_{yy} = 3000 - \frac{180^2}{12} = 300, S_{xy} = 2450 - \frac{144 \times 180}{12} = 150$$

M1

$$r = \frac{150}{\sqrt{202 \times 300}} = 0.6093$$

A1

$$0.6093 < 0.6581$$

M1

Do not reject H_0 : insufficient evidence of positive correlation

A1

Answer: $r = 0.609$; do not reject H_0

6 marks

17 (5 marks)**(a)** All seven: $r = 0.8163$

B1

First six: $r = 0.8908$

B1

Answer: 0.8163 and 0.8908

2 marks

(b) Seven: $0.8163 > 0.6694$ so reject H_0

M1

Six: $0.8908 > 0.7293$ so reject H_0

M1

Both are significant; the point (7, 20) lowers the correlation, because it lies far above the trend

E1

Answer: Both significant; (7, 20) lowers r

3 marks

18 (5 marks)**(a)** Covariance = $\frac{\sum xy}{n} - \bar{x}\bar{y} = \frac{3450}{15} - 12 \times 18 = 14$

M1 A1

$$r = \frac{14}{\sqrt{16 \times 25}} = \frac{14}{20} = 0.7$$

A1

Answer: 0.7

3 marks

(b) $H_0: \rho = 0$, $H_1: \rho > 0$

B1

 $0.7 > 0.4409$, so reject H_0 : significant evidence of positive correlation

E1

Answer: Reject H_0

2 marks

19 (5 marks)**(a)** With more pairs, a given value of r is less likely to occur by chance if there is no correlation

E1

So a smaller r is enough to give evidence of correlation

E1

Answer: Larger samples need smaller r

2 marks

(b) $n = 10$: $0.5 < 0.5494$, not significant

B1

 $n = 20$: $0.5 > 0.3783$, significant

B1

The same r can be significant or not, depending on n

E1

Answer: No for $n = 10$; yes for $n = 20$

3 marks

20 (3 marks)

The points follow a curve, not a straight line

E1

 r measures linear association; a high value does not prove the relationship is linear

E1

Plot the data first; a different model (quadratic) is more suitable

E1

Answer: Relationship is curved; r only measures linear correlation

3 marks

21 (8 marks)

$$(a) S_{uu} = 150 - \frac{(-12)^2}{10} = 135.6$$

M1 A1

$$S_{vv} = 420 - \frac{55^2}{10} = 117.5, S_{uv} = -168 - \frac{(-12)(55)}{10} = -102$$

M1

$$r = -0.8081$$

A1

Answer: -0.8081

4 marks

(b) Same, -0.8081

B1

 u and v are positive linear transformations of x and y

E1

Answer: -0.8081

2 marks

(c) $|r| = 0.8081 > 0.5494$

M1

Reject H_0 : evidence of negative correlation

A1

Answer: Reject H_0

2 marks

22 (6 marks)**(a)** 0.05 (the significance level)

B1

Answer: 0.05

1 mark

(b) $P(\text{none significant}) = 0.95^{20} = 0.3585$

M1 A1

 $P(\text{at least one}) = 1 - 0.3585 = 0.642$

A1

Answer: 0.642

3 marks

(c) Some "significant" results will occur by chance alone

E1

So any such result needs further testing or a pre-stated hypothesis

E1

Answer: Chance findings

2 marks

23 (5 marks)

$$r = 0.9746$$

B1

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

B1

$$0.9746 > 0.7155 \text{ so reject } H_0$$

M1 A1

The test assumes a bivariate normal distribution with linear association; here y is an exact curved function of x , so the test is not appropriate

E1

Answer: Reject H_0 but the test is not valid for a curved relationship

5 marks

24 (11 marks)

$$(a) r = \frac{135}{\sqrt{90 \times 250}}$$

M1

$$r = 0.9$$

A1

Answer: 0.9

2 marks

$$(b) \Sigma x = 10 \times 8 = 80, \Sigma y = 10 \times 20 = 200$$

B1

$$\Sigma x^2 = S_{xx} + n\bar{x}^2 = 90 + 10 \times 64 = 730, \Sigma y^2 = 250 + 10 \times 400 = 4250$$

M1

$$\Sigma xy = S_{xy} + n\bar{x}\bar{y} = 135 + 10 \times 8 \times 20 = 1735$$

A1

Answer: Shown

3 marks

$$(c) \text{ New sums: } \Sigma x = 94, \Sigma y = 232, \Sigma x^2 = 926, \Sigma y^2 = 5274, \Sigma xy = 2183$$

M1 A1

$$S_{xx} = 926 - \frac{94^2}{11} = 122.7273, S_{yy} = 5274 - \frac{232^2}{11} = 380.9091$$

M1

$$S_{xy} = 2183 - \frac{94 \times 232}{11} = 200.4545$$

A1

$$r = 0.9271$$

A1

Answer: $r = 0.927$

5 marks

(d) r rises slightly, from 0.9 to 0.927, because the new pair lies close to the line through the other points and widens the range of x

E1

Answer: r rises slightly

1 mark

25 (6 marks)

$$(a) r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \text{ so } S_{xy} = 0.8 \times \sqrt{3600}$$

M1

$$= 0.8 \times 60 = 48$$

A1

Answer: 48

2 marks

$$(b) y \text{ on } x: \frac{S_{xy}}{S_{xx}} = \frac{48}{36} = \frac{4}{3}$$

B1

$$x \text{ on } y: \frac{S_{xy}}{S_{yy}} = \frac{48}{100} = 0.48$$

B1

Answer: $4/3$ and 0.48

2 marks

$$(c) \frac{4}{3} \times 0.48 = 0.64$$

M1

$$r^2 = 0.64$$

A1

Answer: Shown

2 marks

26 (5 marks)

$$(a) \bar{x} = \frac{43}{6} = 7.1667, \bar{y} = \frac{76}{6} = 12.6667$$

B1 B1

Answer: (7.1667, 12.6667)

2 marks

(b) The mean is unchanged and the new point has deviations (0, 0), so adds 0 to each sum of products

M1 A1

$$S_{xx} = 50.8333, S_{xy} = 65.3333, S_{yy} = 89.3333 \text{ are unchanged, so } r = 0.9695$$

A1

Answer: r unchanged: 0.9695

3 marks

27 (9 marks)

(a) Expand the square: $a^2(x - \bar{x})^2 - 2a(x - \bar{x})(y - \bar{y}) + (y - \bar{y})^2$ M1

Sum each term M1

$$a^2 S_{xx} - 2a S_{xy} + S_{yy}$$
A1

Answer: Shown 3 marks

(b) It is a sum of squares, so ≥ 0 for every a E1

The quadratic in a therefore has at most one root, so discriminant ≤ 0 M1

$$(-2S_{xy})^2 - 4S_{xx}S_{yy} \leq 0$$
M1

$$S_{xy}^2 \leq S_{xx}S_{yy}$$
A1

Answer: Shown 4 marks

(c) $r^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} \leq 1$ M1

So $|r| \leq 1$ A1

Answer: Shown 2 marks

28 (8 marks)

(a) Gradient y on x = $\frac{S_{xy}}{S_{xx}}$; gradient x on y = $\frac{S_{xy}}{S_{yy}}$ B1

$$\text{Product} = \frac{S_{xy}^2}{S_{xx}S_{yy}} = r^2 = 0.9 \times 0.5 = 0.45$$
M1

$$r = \sqrt{0.45} = 0.671$$
A1

Answer: 0.671 3 marks

(b) gradient y on x = $r \frac{S_y}{S_x}$ M1

$$\frac{S_y}{S_x} = \frac{0.9}{0.6708}$$
M1

$$= 1.342$$
A1

Answer: 1.342 3 marks

(c) Both gradients multiply to 1, so the two lines coincide E1

All points lie exactly on one straight line E1

Answer: Lines coincide 2 marks

29 (8 marks)

(a) $r^2(n - 2) = t^2(1 - r^2)$ M1

$$r^2(n - 2 + t^2) = t^2$$
M1

$$r = \frac{t}{\sqrt{n - 2 + t^2}} = \frac{1.8125}{\sqrt{10 + 3.28516}}$$
M1

$$= 0.4973$$
A1

Answer: 0.4973 4 marks

(b) $t = \frac{0.55 \times \sqrt{10}}{\sqrt{1 - 0.55^2}} = 2.0825$ M1 A1

$$2.0825 > 1.8125, \text{ so reject } H_0$$
A1

Answer: Reject H0 3 marks

(c) $0.55 > 0.4973$ B1

Answer: Consistent 1 mark

30 (7 marks)

(a) A: $0.45 < 0.5494$, do not reject H_0 B1

B: $0.40 > 0.3783$, reject H_0 B1

C: $0.32 > 0.2638$, reject H_0 B1

Answer: A not significant; B and C significant 3 marks

(b) Significance depends on n as well as r E1

A has only 10 pairs, so a correlation of 0.45 could easily occur by chance E1

Answer: Small n 2 marks

(c) $r = 0.32$ is a weak correlation even though it is statistically significant
Significance shows evidence that some correlation exists, not that it is strong

Answer: Significant but weak

E1

E1

2 marks

Regression Lines

1 (3 marks)

$$y = 12.4 + 3.5 \times 8 = 40.4$$

M1 A1

Gradient 3.5, so y increases by $3.5 \times 2 = 7$

B1

Answer: 40.4; increases by 7

3 marks

2 (3 marks)

$$b = \frac{S_{xy}}{S_{xx}} = \frac{120}{40} = 3$$

M1 A1

$$a = \bar{y} - b\bar{x} = 20 - 15 = 5, \text{ so } y = 5 + 3x$$

A1

Answer: $y = 5 + 3x$

3 marks

3 (2 marks)

Mass on height (mass is y , the response variable)

B1

The line of y on x minimises the vertical distances in the variable being predicted

E1

Answer: Mass on height

2 marks

4 (2 marks)

$$b = 2.34$$

B1

$$a = 1.80, \text{ so } y = 1.8 + 2.343x$$

B1

Answer: $y = 1.80 + 2.34x$

2 marks

5 (2 marks)

$$\text{Fitted value} = 2 + 3 \times 7 = 23$$

M1

$$\text{Residual} = 25 - 23 = 2$$

A1

Answer: 2

2 marks

6 (2 marks)

$$\text{When } x = 6, y = 4 + 2 \times 6 = 16$$

M1

This equals $\bar{y}$, so the line passes through $(\bar{x}, \bar{y})$

A1

Answer: Shown

2 marks

7 (5 marks)

$$S_{xx} = 1085 - \frac{95^2}{10} = 182.5$$

M1 A1

$$S_{xy} = 2285 - \frac{95 \times 210}{10} = 290$$

M1

$$b = \frac{290}{182.5} = 1.589, a = 21 - 1.589 \times 9.5 = 5.9041$$

A1

$$y = 5.904 + 1.589x$$

A1

Answer: $y = 5.904 + 1.589x$

5 marks

8 (4 marks)

Using a calculator: $b = 1.8095$

M1 A1

$$a = 11.5238, \text{ so } y = 11.524 + 1.81x$$

A1

$$x = 21: y = 49.5238 \approx 50$$

A1

Answer: $y = 11.524 + 1.81x$; about 50

4 marks

9 (3 marks)

$$y - 30 = 2.5(x - 8)$$

M1

$$y = 30 + 2.5x - 20$$

M1

$$y = 10 + 2.5x$$

A1

Answer: $y = 10 + 2.5x$

3 marks

10 (2 marks)

$$b = r \times \frac{s_y}{s_x}$$

M1

$$= 0.8 \times \frac{12}{5} = 1.92$$

A1

Answer: 1.92

2 marks

11 (3 marks)Substitute $u = x - 2020$: $y = 3.2 + 0.8(x - 2020)$

M1

$$y = 0.8x - 1612.8$$

A1

For 2026: $u = 6$, $y = 3.2 + 4.8 = 8$

B1

Answer: $y = 0.8x - 1612.8$; 8

3 marks

12 (2 marks)

The residuals sum to zero for a least squares line

E1

$$\text{Fifth} = -(1.2 - 0.8 + 0.5 - 2.1) = 1.2$$

B1

Answer: 1.2

2 marks

13 (3 marks)

35 is the fixed charge (the cost for 0 days)

E1

48 is the extra cost per day

E1

$$C = 35 + 48 \times 3 = £179$$

B1

Answer: £179

3 marks

14 (7 marks)

$$(a) S_{xx} = 1518 - \frac{120^2}{12} = 318$$

M1 A1

$$S_{xy} = 5020 - \frac{120 \times 426}{12} = 760$$

M1

$$b = 2.3899, a = 11.6006, \text{ so } y = 11.6 + 2.39x$$

A1

Answer: $y = 11.6 + 2.39x$

4 marks

$$(b) y = 11.6006 + 2.3899 \times 10 = 35.5$$

M1 A1

$$\text{Residual} = 36 - 35.5 = 0.5$$

A1

Answer: 35.5; residual 0.5

3 marks

15 (6 marks)

$$S_{uu} = 105 - \frac{25^2}{10} = 42.5$$

M1

$$S_{uv} = -150 - \frac{25 \times (-30)}{10} = -75$$

M1

$$b = \frac{-75}{42.5} = -1.7647$$

A1

$$a = \bar{v} - b\bar{u} = -3 - (-1.7647 \times 2.5) = 1.4118, \text{ so } v = 1.4118 + -1.7647u$$

M1 A1

$$y - 100 = 1.4118 + -1.7647(x/10), \text{ so } y = 101.4118 - 0.1765x$$

A1

Answer: $y = 101.4118 - 0.1765x$

6 marks

16 (6 marks)

$$(a) y \text{ on } x: b = 1.581, a = 2.5448$$

M1 A1

$$x \text{ on } y: b = \frac{S_{xy}}{S_{yy}} = 0.5955, a = -0.9799$$

M1 A1

Answer: $y = 2.545 + 1.581x$; $x = -0.98 + 0.595y$

4 marks

$$(b) y \text{ on } x \text{ at } x = 9.1429 \text{ gives } y = 17$$

M1

 $x \text{ on } y \text{ at } y = 17 \text{ gives } x = 9.1429, \text{ so both pass through } (9.1429, 17)$

A1

Answer: Shown**2 marks****17 (3 marks)**

$$h = 100H, \text{ so } M = -48 + 0.52 \times 100H$$

M1

$$M = -48 + 52H$$

A1

$$H = 1.40: M = -48 + 72.8 = 24.8 \text{ kg}$$

A1

$$\text{Answer: } M = -48 + 52H; 24.8 \text{ kg}$$

3 marks**18 (5 marks)**

$$\text{Gradient (y on x)} = r \frac{s_y}{s_x} = 0.9 \times 3 = 2.7$$

M1 A1

$$y - 34 = 2.7(x - 10), \text{ so } y = 7 + 2.7x$$

A1

$$\text{Gradient (x on y)} = r \frac{s_x}{s_y} = 0.9 \times \frac{1}{3} = 0.3$$

M1

$$x - 10 = 0.3(y - 34), \text{ so } x = -0.2 + 0.3y$$

A1

$$\text{Answer: } y = 7 + 2.7x; x = -0.2 + 0.3y$$

5 marks**19 (5 marks)**

$$\bar{x} = 5.2, \bar{y} = 10.4$$

B1

$$\text{Deviations } x - \bar{x}: -3.20, -1.20, -0.20, 1.80, 2.80; y - \bar{y}: -4.40, -3.40, 0.60, 1.60, 5.60$$

M1

$$S_{xy} = \sum(x - \bar{x})(y - \bar{y}) = 36.6, S_{xx} = 22.8$$

M1 A1

$$b = 1.605, y = 2.053 + 1.605x$$

A1

$$\text{Answer: } y = 2.053 + 1.605x$$

5 marks**20 (6 marks)**

$$\text{(a) Fitted } y = 8 + 2x$$

M1

$$\text{Residuals: } 0.00, -1.00, 2.00, -1.00, 1.00, -2.00, 1.00, 0.00$$

M1 A1

$$\text{Answer: } 0.00, -1.00, 2.00, -1.00, 1.00, -2.00, 1.00, 0.00$$

3 marks

$$\text{(b) Largest in size: } (6, 22) \text{ with residual } 2.00$$

B1

$$\Sigma e^2 = 0 + 1 + 4 + 1 + 1 + 4 + 1 + 0$$

M1

$$= 12$$

A1

$$\text{Answer: } (6, 22); \Sigma e^2 = 12$$

3 marks**21 (6 marks)**

$$\text{(a) Both lines pass through the mean point}$$

B1

$$\text{Substitute } y = 2 + 0.8x: x = 1 + 0.5(2 + 0.8x) = 2 + 0.4x$$

M1

$$x = 3.3333 \text{ and } y = 4.6667$$

A1

$$\text{Answer: } (3.3333, 4.6667)$$

3 marks

$$\text{(b) } r^2 = 0.8 \times 0.5 = 0.4$$

M1 A1

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

A1

$$\text{Answer: } 0.632$$

3 marks**22 (6 marks)**

$$\text{(a) } \frac{dD}{db} = -2\Sigma x(y - bx)$$

M1 A1

$$\text{Set to zero: } \Sigma xy = b\Sigma x^2$$

M1

$$b = \frac{\Sigma xy}{\Sigma x^2}; \text{ second derivative } 2\Sigma x^2 > 0 \text{ so a minimum}$$

A1

$$\text{Answer: Shown}$$

4 marks

$$\text{(b) } \Sigma xy = 138, \Sigma x^2 = 55$$

M1

$$b = \frac{138}{55} = 2.5091$$

A1

$$\text{Answer: } 2.5091$$

2 marks

23 (6 marks)

(a) $b = 0.9571$ M1

$y = 5 + 0.957x$ A1

Answer: $y = 5 + 0.957x$ 2 marks

(b) New: $y = 15.393 + 0.611x$ M1 A1

Old prediction at 55: 57.6429; new: 48.9821 M1

The added point lowers the line, so predictions are lower: it has high influence A1

Answer: $y = 15.393 + 0.611x$; at $x = 55$: old 57.643, new 48.982 4 marks

24 (6 marks)

(a) $b = 0.6574$ M1

$a = 5.6571$ A1

$y = 5.657 + 0.657x$ A1

Answer: $y = 5.657 + 0.657x$ 3 marks

(b) $a \approx 5.657$: the fixed charge (a journey of 0 km), about £5.657 E1 $b \approx 0.657$: the extra cost per kilometre, about £0.657 E1**Answer: Fixed charge and cost per km** 2 marks

(c) $5.657 + 0.657 \times 10 = £12.231$ B1

Answer: £12.231 1 mark

25 (7 marks)

(a) $\ln y = \ln a + \ln b^x = \ln a + x \ln b$ M1

Plot or regress $\ln y$ against x : the intercept is $\ln a$ and the gradient is $\ln b$ E1**Answer: Shown** 2 marks

(b) $\ln y$ values: 1.435, 1.808, 2.251, 2.639, 3.02, 3.431 M1

$\ln y = 1.0306 + 0.4001x$ M1 A1

$a = e^{1.031} = 2.80$, $b = e^{0.4} = 1.49$ A1

Answer: $a = 2.80$, $b = 1.49$ 4 marks

(c) $y = 2.80 \times 1.49^8 = 68.8$ B1

Answer: 68.8 1 mark

26 (6 marks)

(a) $\ln y = \ln a + n \ln x$ M1

Gradient estimates n and intercept estimates $\ln a$ A1**Answer: Shown** 2 marks(b) Using $\ln x$ and $\ln y$ values M1

$\ln y = 1.1128 + 1.5156 \ln x$ M1 A1

$n = 1.516$, $a = e^{1.113} = 3.043$ A1

Answer: $n = 1.516$, $a = 3.043$ 4 marks

27 (9 marks)

(a) $\frac{\partial D}{\partial a} = -2\sum(y - a - bx)$ M1

Set to 0: $\sum y - na - b\sum x = 0$ M1

$a = \bar{y} - b\bar{x}$ A1

Answer: Shown 3 marks

(b) $\frac{\partial D}{\partial b} = -2\sum x(y - a - bx) = 0$ M1

Substitute a : $\sum xy - (\bar{y} - b\bar{x})\sum x - b\sum x^2 = 0$ M1

$\sum xy - n\bar{y}\bar{x} = b(\sum x^2 - n\bar{x}^2)$ M1

$S_{xy} = bS_{xx}$, so $b = \frac{S_{xy}}{S_{xx}}$ A1

Answer: Shown 4 marks

(c) $\sum(y - a - bx) = \sum y - na - b\sum x$ M1

 $= 0$ by the first equation in (a) A1**Answer: Shown** 2 marks

28 (8 marks)**(a)** Fitted values $0.7333 + 1.7429x$

M1

Residuals: 0.524, 0.781, -1.962 , 0.295, -0.448 , 0.81

M1 A1

Sum = $0.00 \approx 0$

A1

Answer: sum is 0

4 marks

(b) $\sum x e = 0.00 \approx 0$

M1 A1

It is the condition $\frac{\partial D}{\partial b} = 0$ for the minimum of the sum of squares

E1

Otherwise the line could be tilted to reduce D

E1

Answer: $\sum x e = 0$

4 marks

29 (9 marks)**(a)** The regression line passes through $(\bar{x}, \bar{y})$; $\bar{x} = 3$

M1

 $\bar{y} = 4 + 3 \times 3 = 13$, so $\sum y = 65$

M1

 $k = 65 - (8 + 9 + 15 + 20) = 13$

A1

Answer: 13

3 marks

(b) Fitted values: 7, 10, 13, 16, 19

M1

Residuals: 1, -1 , 0, -1 , 1

A1

Sum = 0

B1

Answer: 1, -1 , 0, -1 , 1; sum 0

3 marks

(c) The residuals of a least squares line sum to zero

E1

 $1.2 - 0.8 + 0.5 - 2.1 + 1.0 = -0.2$

M1

Sixth residual = 0.2

A1

Answer: 0.2

3 marks

30 (7 marks)**(a)** $v = \frac{y - 10}{2}$, $u = \frac{x}{10}$: $\frac{y - 10}{2} = p + q \frac{x}{10}$

M1

 $y - 10 = 2p + \frac{q}{5}x$

M1

 $y = 10 + 2p + \frac{q}{5}x$

A1

Answer: Shown

3 marks

(b) $u = 1, 2, 3, 4, 5$ and $v = 1, 4.5, 10.5, 14, 21$

M1

 $q = 4.95$, $p = -4.65$

M1 A1

 $y = 10 + 2(-4.65) + \frac{4.95}{5}x = y = 0.7 + 0.99x$

A1

Answer: $y = 0.7 + 0.99x$

4 marks

Interpretation of Regression**1 (2 marks)**

12 is the gradient of the regression line

B1

Each extra customer is associated with an extra £12 of weekly profit

E1

Answer: £12 more profit per extra customer

2 marks

2 (2 marks)3.2 is the value of h when $t = 0$

B1

It is the estimated height of the plant at the start (3.2 cm)

E1

Answer: Estimated initial height 3.2 cm

2 marks

3 (3 marks)

The battery loses 2.5 percentage points of charge per hour

E1

 $y = 90 - 2.5 \times 8$

M1

 $= 70\%$

A1

Answer: -2.5 per hour; 70%

3 marks

4 (2 marks)

Explanatory: rainfall; response: yield

B1

Use the regression line of yield on rainfall

B1

Answer: Yield on rainfall

2 marks

5 (2 marks)

$$y = 20 + 5 \times 4$$

M1

$$= 40$$

A1

Answer: 40

2 marks

6 (2 marks)

No: correlation does not imply causation

E1

The size of the fire causes both more engines to attend and more damage

E1

Answer: No: the size of the fire is the cause of both

2 marks

7 (3 marks)

$$\text{Fitted value} = 2.143 + 1.895 \times 14 = 28.6762$$

M1 A1

$$\text{Residual} = 30 - 28.6762 = 1.324, \text{ so the point lies above the line}$$

A1

Answer: residual 1.324, above the line

3 marks

8 (3 marks)

School A has the steeper gradient: each extra revision hour is associated with 4.5 extra marks compared with 3

E1

School B has the higher intercept: 40 marks with no revision compared with 32

E1

The lines cross when $32 + 4.5x = 40 + 3x$, so at $x = 5.333$ hours

B1

Answer: A steeper, B higher start; cross at 5.33 h

3 marks

9 (2 marks)

$$0.75 \text{ kg per cm} = 750 \text{ g per 10 mm}$$

M1

$$= 75 \text{ g per mm}$$

A1

Answer: 75 g per mm

2 marks

10 (4 marks)

The line of y on x minimises vertical distances, so it is for predicting y, not x

E1

$$S_{xy} = 770, S_{yy} = 610$$

M1

$$\text{Gradient of x on y} = \frac{S_{xy}}{S_{yy}} = 1.2623$$

A1

$$\text{The rearranged line has x-gradient } \frac{1}{0.77} = 1.2987, \text{ which is different}$$

E1

Answer: Use x on y: gradient 1.262

4 marks

11 (3 marks)

$$y = -45 + 9.2 \times 17 = 111.4$$

M1 A1

Rounded to 111, because the number of visitors must be a whole number

E1

Answer: 111

3 marks

12 (2 marks)

The residuals alternate in sign with no trend or curve, and are small

B1

This suggests that the linear model is a good fit

E1

Answer: Linear model is suitable

2 marks

13 (3 marks)

$$\text{Residual} = 258 - 240 = 18$$

B1

$$\frac{18}{240} \times 100$$

M1

$$= 7.50\%$$

A1

Answer: 18; 7.5%

3 marks

14 (6 marks)

(a) $b = -4.6095$

M1 A1

$y = 400.857 - 4.61x$

A1

Answer: $y = 400.857 - 4.61x$

3 marks

(b) Each 1p rise in price is associated with about 4.61 fewer items sold

E1

Answer: About 4.61 fewer per week for each 1p

1 mark

(c) $a = 400.857$ is the predicted sales at a price of 0p

B1

A price of zero is far outside the data, so this is not meaningful

E1

Answer: Not meaningful

2 marks

15 (4 marks)

$$c = \frac{f - 32}{1.8}$$

M1

$$y = 85 - 2.4 \times \frac{f - 32}{1.8}$$

M1

$y = 85 + 42.6667 - 1.3333f = 127.6667 - 1.3333f$

A1

Gradient: usage falls by about 1.333 units for each 1 °F rise

E1

Answer: $y = 127.6667 - 1.3333f$

4 marks

16 (3 marks)

$$b = r \frac{s_h}{s_a} = 0.7 \times \frac{12}{2.5}$$

M1

$= 3.36$

A1

On average, a child's height is 3.36 cm greater for each additional year of age

E1

Answer: 3.36 cm per year

3 marks

17 (7 marks)

(a) $S_{hh} = 348\,000 - \frac{2040^2}{12} = 1200$

M1

$S_{hm} = 145\,560 - \frac{2040 \times 852}{12} = 720$

A1

Answer: 1200 and 720

2 marks

(b) $b = \frac{720}{1200} = 0.6$

M1 A1

$a = \frac{852}{12} - 0.6 \times \frac{2040}{12} = -31$, so $m = -31 + 0.6h$

A1

Answer: $m = -31 + 0.6h$

3 marks

(c) 10 cm more height gives $0.6 \times 10 = 6$ kg more mass

M1

The claim of 8 kg is too high; the data suggest about 6 kg

E1

Answer: About 6 kg, not 8 kg

2 marks

18 (4 marks)Line 1: $0.62 < 0.6215$, not significant (but close)

M1 A1

Line 2: $0.62 > 0.2638$, significant evidence of correlation

A1

Line 2 is based on more data, so is more reliable

E1

Answer: Line 2

4 marks

19 (5 marks)

(a) Least squares line: $\Sigma e^2 = 1.3977$

M1 A1

Line through origin: $\Sigma e^2 = 5.3042$

M1 A1

Answer: 1.398 and 5.304

4 marks

(b) The least squares line has the smaller sum of squares, so fits better; there is no reason for the intercept to be zero unless the context requires it

E1

Answer: Not supported

1 mark

20 (5 marks)

(a) $42 - 3.1 \times 4 = 29.6$

B1

Answer: 29.6

1 mark

(b) $42 - 3.1x < 20$	M1
$x > \frac{22}{3.1} = 7.0968$	A1
So 8 sessions	A1
Answer: 8	3 marks
(c) The model would eventually predict a negative time, which is impossible, and it is outside the range of the data	E1
Answer: Unrealistic for large x	1 mark
21 (8 marks)	
(a) $b = 1.8583$, $a = 20.05$	M1 A1
$r = 0.9914$	A1
Answer: $y = 20.05 + 1.858x$; $r = 0.991$	3 marks
(b) $0.9914 > 0.5494$	M1
Reject H_0 : evidence of positive correlation	A1
Answer: Reject H_0	2 marks
(c) Each extra year of experience is associated with about £1858 more salary	E1
The intercept £20050 is the estimated starting salary with no experience	E1
Answer: About £1858 more per year; £20050 starting salary	2 marks
(d) Other factors, such as qualifications or age, may affect both	E1
Answer: Other factors	1 mark
22 (7 marks)	
(a) With: $y = 1.467 + 4.914x$	M1 A1
Without: $y = 7.8 + 2.2x$	M1 A1
Answer: $y = 1.467 + 4.914x$; $y = 7.8 + 2.2x$	4 marks
(b) The gradient increases from 2.2 to 4.914 and the intercept falls from 7.8 to 1.467	B1 B1
Prediction at 3.5: 18.667 with, 15.5 without, so it is pulled by the outlier	E1
Answer: Gradient 2.2 to 4.914; prediction at 3.5: 15.5 to 18.667	3 marks
23 (6 marks)	
(a) $\ln y = \ln a + x \ln b$	M1
$\ln y = 1.4552 + 0.1824x$	M1 A1
$a = 4.2856$, $b = 1.2001$	A1
Answer: $a = 4.286$, $b = 1.2$	4 marks
(b) a is the value at the start ($x = 0$), about £4.286 thousand	E1
b is the yearly growth factor: growth of about 20.013% per year	E1
Answer: $a = 4.286$ thousand pounds; growth of 20.013% a year	2 marks
24 (7 marks)	
(a) y on x : $y = 2.657 + 1.149x$	M1 A1
x on y : $x = -1.794 + 0.852y$	M1 A1
Answer: $y = 2.657 + 1.149x$; $x = -1.794 + 0.852y$	4 marks
(b) Correct line (x on y): $x = -1.794 + 0.852 \times 30 = 23.7786$	M1 A1
Rearranged y on x : $x = \frac{30 - 2.657}{1.149} = 23.806$, which differs	A1
Answer: 23.7786	3 marks
25 (5 marks)	
(a) Fitted values $7.464 + 0.245x$	M1
Residuals: $-4.71, -0.85, 1.20, 2.76, 3.21, 2.77, 1.22, -0.92, -4.67$	M1 A1
Answer: $-4.71, -0.85, 1.20, 2.76, 3.21, 2.77, 1.22, -0.92, -4.67$	3 marks
(b) Residuals are negative, then positive, then negative: a curved ($\cap$ shaped) pattern	B1
A straight line is not suitable; a quadratic model would fit better	E1
Answer: Systematic pattern: not linear	2 marks

26 (9 marks)

(a) p on m: gradient $0.72 \times \frac{12}{10} = 0.864$; $p - 62 = 0.864(m - 58)$

M1 A1

m on p: gradient $0.72 \times \frac{10}{12} = 0.6$; $m - 58 = 0.6(p - 62)$

M1 A1

Answer: $p = 11.888 + 0.864m$; $m = 20.8 + 0.6p$

4 marks

(b) Physics: $62 + 0.864 \times 12 = 72.368$

M1 A1

Maths: $58 + 0.6 \times (70 - 62) = 62.8$

A1

Answer: 72.368 and 62.8

3 marks

(c) Since $r < 1$ the predicted value regresses towards the mean

E1

Each prediction uses the line for the variable being predicted

E1

Answer: Regression to the mean

2 marks

27 (13 marks)

(a) $S_{xx} = 204 - \frac{36^2}{8} = 42$, $S_{xy} = 1224 - \frac{36 \times 244}{8} = 126$

M1 A1

$b = \frac{126}{42} = 3$

A1

$a = \frac{244}{8} - 3 \times \frac{36}{8} = 17$, so $y = 17 + 3x$

A1

Answer: $y = 17 + 3x$

4 marks

(b) Predicted profit in month x is $17 + 3x$, an arithmetic sequence

M1

Months 9 to 20: 12 terms, first 44, last 77

M1

Total = $\frac{12}{2}(44 + 77) = 726$

M1 A1

Answer: £72600

4 marks

(c) Sum of first N terms = $\frac{N}{2}(2 \times 20 + (N - 1) \times 3)$

M1

= $\frac{N(3N + 37)}{2}$

A1

$N = 30$ gives 1905 and $N = 31$ gives 2015

M1

Least $N = 31$

A1

Answer: $N = 31$

4 marks

(d) Months beyond 8 are extrapolation, so the answer is unreliable

E1

Answer: Extrapolation

1 mark

28 (8 marks)

(a) Gradient = $0.5 \times \frac{6.5}{6.5} = 0.5$

M1 A1

$s - 177 = 0.5(f - 175)$, so $s = 89.5 + 0.5f$

A1

Answer: $s = 89.5 + 0.5f$

3 marks

(b) $s = 89.5 + 0.5 \times 188 = 183.5$ cm

M1 A1

$\frac{183.5 - 177}{6.5} = 1$ standard deviation

A1

Answer: 183.5 cm; 1 SD

3 marks

(c) The predicted son is closer to the mean (in standard deviations) than his father, because $r < 1$

E1

Extreme fathers tend to have less extreme sons, since height depends partly on other factors

E1

Answer: Regression to the mean

2 marks

29 (9 marks)

(a) $\ln(T - 20) = \ln A - kt$

M1

Gradient $-k$, intercept $\ln A$

A1

Answer: Gradient $-k$, intercept $\ln A$

2 marks

(b) $\ln(T - 20) = 4.317, 3.916, 3.544, 3.203, 2.907, 2.66$

M1

$\ln(T - 20) = 4.257 - 0.0666t$

M1 A1

$A = e^{4.257} = 70.599$, $k = 0.067$

A1

Answer: $A = 70.599$, $k = 0.067$

4 marks

(c) A is the initial excess temperature above the room (about 70.599°C)

E1

$$\text{Half-life} = \frac{\ln 2}{k}$$

$$= 10.408 \text{ minutes}$$

Answer: 10.408 minutes

M1

A1

3 marks

30 (10 marks)

(a) $b = 1.4531$, $a = 6.1045$

M1 A1

$$y = 6.105 + 1.453x$$

A1

Answer: $y = 6.105 + 1.453x$

3 marks

(b) Group means: (2.5, 9.55); (6.5, 15.55); (10.5, 21.55)

M1

$$\text{Line: } y = 5.8 + 1.5x$$

M1 A1

Answer: $y = 5.8 + 1.5x$

3 marks

(c) Within-group gradients: 1.04, 1.06, 1.06

M1 A1

The gradient from group means (1.5) is steeper than within groups

B1

Relationships between group averages need not hold for individuals within groups

E1

Answer: Group-level gradient differs from individual-level

4 marks

Extrapolation and Interpolation

1 (2 marks)

Interpolation: predicting y for a value of x inside the range of the data

B1

Extrapolation: predicting y for a value of x outside the range of the data

B1

Answer: Inside the data range versus outside it

2 marks

2 (4 marks)

$x = 35$: interpolation

B1

$x = 60$: extrapolation

B1

$x = 8$: extrapolation

B1

$x = 50$: interpolation (the end of the range)

B1

Answer: 35 interpolation; 60 extrapolation; 8 extrapolation; 50 interpolation

4 marks

3 (2 marks)

Interpolation

B1

The relationship is known to hold within the range of the data but may change outside it

E1

Answer: Interpolation

2 marks

4 (3 marks)

$$y = 5 + 2 \times 25 = 55$$

M1 A1

$x = 25$ is within the range (interpolation), so it is likely to be reliable

E1

Answer: 55; reliable

3 marks

5 (1 mark)

The pattern may not continue beyond the observed range of x

E1

Answer: Relationship may change outside the range

1 mark

6 (1 mark)

$$12 \leq x \leq 27$$

B1

Answer: $12 \leq x \leq 27$

1 mark

7 (6 marks)

(a) $\bar{x} = 30$, so $\bar{y} = 14 + 1.1 \times 30 = 47$

M1

$$\Sigma y = 235$$

M1

$$k = 235 - (25 + 35 + 61 + 67) = 47$$

A1

Answer: 47

3 marks

(b) $y = 14 + 1.1 \times 80 = 102$

M1 A1

$x = 80$ is outside the range 10 to 50, so this is extrapolation and may be unreliable

E1

Answer: 102; unreliable

3 marks

8 (3 marks)

$$y = 500 - 85 \times 6 = -10$$

M1 A1

A negative number of sales is impossible and £6 is outside the range, so the prediction is not valid

E1

Answer: -10: not valid

3 marks

9 (3 marks)

$$h = 80 + 6.2 \times 30 = 266 \text{ cm}$$

M1 A1

Growth stops in adulthood; this is extrapolation far outside the data and gives an impossible

E1

height

Answer: 266 cm: absurd

3 marks

10 (2 marks)

A strong correlation only shows the relationship is linear within the observed range

E1

There is no evidence the same relationship holds outside the range

E1

Answer: Not justified

2 marks

11 (3 marks)

$$2040: t = 25$$

B1

$$y = 18.143 + 3.143 \times 25 = 96.718 \text{ thousand}$$

M1

This is a long extrapolation (20 years beyond the data), so it is unreliable

E1

Answer: 96.718 thousand: unreliable

3 marks

12 (3 marks)

$$y = 52 + 3.5 \times 15 = 104.5\%$$

M1 A1

A percentage cannot exceed 100, so the model fails outside the range

E1

Answer: 104.5%: impossible

3 marks

13 (3 marks)

$$\text{Interpolation: halfway, } y = \frac{27 + 33}{2} = 30$$

M1 A1

$$\text{Regression: } y = 2 + 1.17 \times 25 = 31.25$$

A1

Answer: 30 and 31.25

3 marks

14 (7 marks)

$$(a) b = 2.0679$$

M1 A1

$$y = 2.864 + 2.068x$$

A1

Answer: $y = 2.864 + 2.068x$

3 marks

$$(b) x = 9: 21.475; x = 20: 44.2214; x = 1: 4.9321$$

M1 A1

$x = 9$ is interpolation ($4 \leq x \leq 16$): reliable

E1

$x = 20$ and $x = 1$ are extrapolation: less reliable

E1

Answer: 21.475, 44.221, 4.932; only $x = 9$ is reliable

4 marks

15 (5 marks)

(a) Temperature is to be predicted, so use x on y

E1

$$x = -16.8911 + 0.9341y$$

M1 A1

$$y = 56: x = 35.418 \text{ } ^\circ\text{C}$$

A1

Answer: 35.418 $^\circ\text{C}$

4 marks

(b) 56 g lies within the range of observed yields (28 to 70), so it is interpolation and reasonably reliable

E1

Answer: Yes, interpolation

1 mark

16 (4 marks)

$$(a) 0 = 98.6 - 4.75x$$

M1

$$x = 20.758$$

A1

Answer: 20.758%

2 marks

(b) It is well outside the data range (8 to 16), so is extrapolation

E1

The speed may not fall linearly (the car could stop at a different gradient)

E1

Answer: Unreliable

2 marks

17 (6 marks)**(a)** 2022: $t = 7$, $y = 21.34$

M1 A1

2025: $t = 10$, $y = 25.3$

A1

Answer: 21.34 and 25.3

3 marks

(b) Errors: 2.66 and 8.2

B1 B1

The error grows the further we extrapolate: growth was accelerating, not linear

E1

Answer: errors grow

3 marks

18 (6 marks)**(a)** Linear: $9.44 + 0.301 \times 80 = 33.52$

M1 A1

Exponential: $y = e^{2.3033 + 0.0198 \times 80} = 48.7961$

M1 A1

Answer: 33.52 and 48.796

4 marks

(b) The models diverge increasingly outside the data range

E1

Real populations are limited by other factors, so extrapolation is unreliable

E1

Answer: Models diverge

2 marks

19 (7 marks)**(a)** 45: $-26.2 + 1.36 \times 45 = 35$

M1 A1

120: 137

A1

Answer: 35 and 137

3 marks

(b) $k = \frac{75}{70^2} = 0.0153$

M1 A1

 $y = 0.0153 \times 120^2 = 220.4082$

M1

The straight line gives a much smaller value: extrapolating a linear model for a curved relationship under-estimates

E1

Answer: $k = 0.0153$; 220.408 m at 120 km/h

4 marks

20 (9 marks)**(a)** $\bar{x} = 6$, $S_{xx} = \sum(x - 6)^2 = 16 + 4 + 0 + 4 + 16 = 40$

M1 A1

 $S_{xy} = \sum(x - 6)y = -4 \times 9 - 2k + 2 \times 22 + 4 \times 27 = 116 - 2k$

M1

 $116 - 2k = 2.4 \times 40 = 96$, so $k = 10$

A1

Answer: 10

4 marks

(b) $\bar{y} = \frac{9 + 10 + 16 + 22 + 27}{5} = 16.8$

M1

 $y = 16.8 + 2.4(x - 6) = 2.4 + 2.4x$

A1

Answer: $y = 2.4 + 2.4x$

2 marks

(c) $x = 7$: $y = 19.2$

B1

 $x = 15$: $y = 38.4$

B1

 $x = 7$ is interpolation so reliable; $x = 15$ is extrapolation so less reliable

E1

Answer: 19.2 (reliable); 38.4 (unreliable)

3 marks

21 (8 marks)**(a)** $r = 0.9956$

B1

 $0.9956 > 0.6215$

M1

Reject H_0 : evidence of positive correlation

A1

Answer: $r = 0.996$; reject H_0

3 marks

(b) $y = 7.554 + 2.089x$

M1 A1

 $x = 12$: 32.625; $x = 30$: 70.232

A1

Answer: 32.625; 70.232

3 marks

(c) $x = 12$ is within the range (5 to 19): reliable given the strong correlation

E1

 $x = 30$ is outside the range, and a week has only 168 hours, so it is unreliable

E1

Answer: 12 reliable; 30 not

2 marks

22 (6 marks)

(a) $0 = 55.667 - 3.429x$

M1

$x = 16.236$ hours

A1

Answer: 16.236 hours

2 marks

(b) $y = 55.667$

B1

A battery level above 100% is impossible, so extrapolating back to $x = 0$ fails

E1

Answer: 55.667%: above 100%

2 marks

(c) About $1 \leq x \leq 6$ (the data range), up to where y stays between 0 and 100

B1

Outside it the model is not supported by data or physical limits

E1

Answer: $1 \leq x \leq 6$

2 marks

23 (11 marks)

(a) $S_{xx} = 480 - \frac{60^2}{10} = 120$

M1 A1

$S_{yy} = 1502.5 - \frac{120^2}{10} = 62.5$, $S_{xy} = 642 - \frac{60 \times 120}{10} = -78$

A1

Answer: 120, 62.5, -78

3 marks

(b) $r = \frac{-78}{\sqrt{120 \times 62.5}} = -0.9007$

M1 A1

 $-0.9007 < -0.7155$, so reject H_0 : significant evidence of negative correlation

E1

Answer: $r = -0.901$; reject H_0

3 marks

(c) $b = \frac{-78}{120} = -0.65$

M1

$a = 12 + 0.65 \times 6 = 15.9$, so $y = 15.9 - 0.65x$

A1

Answer: $y = 15.9 - 0.65x$

2 marks

(d) $x = 7.5$: $y = 15.9 - 0.65 \times 7.5 = 11.025$ °C

B1

$x = 0$: $y = 15.9$ °C

B1

750 m is interpolation so reliable; sea level is outside the range 200 m to 1000 m (extrapolation)

E1

so less reliable

Answer: 11.025 °C; 15.9 °C

3 marks

24 (7 marks)

(a) N: $y = -1.26 + 1.32x$

M1 A1

W: $y = -0.18 + 1.29x$

A1

Answer: N: $y = -1.26 + 1.32x$; W: $y = -0.18 + 1.29x$

3 marks

(b) True value: 48

B1

N gives 38.34, error -9.66 ; W gives 38.52, error -9.48

M1 A1

Answer: N: error -9.66 ; W: error -9.48

3 marks

(c) Using a wider range of data gives a more representative gradient, but both under-estimate here since the data are curved

E1

Answer: Wider range is more representative

1 mark

25 (5 marks)

(a) $y = 0.053 + 2.794 \times 10$

M1

$= 27.993$ °C

A1

Answer: 27.993 °C

2 marks

(b) Error = $27.993 - 11 = 16.993$ °C

M1 A1

Temperature follows a seasonal cycle, not a straight line

E1

Answer: 16.993; seasonal

3 marks

26 (7 marks)

(a) y on x : $y = 23.6 + 0.78x$

M1 A1

x on y : $x = -28.75 + 1.25y$

A1

Answer: $y = 23.6 + 0.78x$; $x = -28.75 + 1.25y$

3 marks

(b) Rearranged: $x = \frac{y - 23.6}{0.78}$

M1

$y = 47$: 30 and 30

A1

$y = 67$: 55.641 and 55; $y = 107$: 106.923 and 105

A1

Answer: 30, 55.641, 106.923 (rearranged line); 30, 55, 105 (x on y line)

3 marks

(c) They agree at the mean and diverge as y moves further away (extrapolation)

E1

Answer: Diverge

1 mark

27 (9 marks)

(a) $y = a + bx$ with $a = \bar{y} - b\bar{x}$

M1

So $y = \bar{y} + b(x - \bar{x})$

A1

$\bar{x} = 25$, $\bar{y} = 40.7143$

B1

Answer: $\bar{x} = 25$, $\bar{y} = 40.7143$

3 marks

(b) Error = $\pm 0.15 \times |x - \bar{x}|$

M1

$x = 15$: 1.5; $x = 30$: 0.75; $x = 60$: 5.25

M1 A1

Errors grow with distance from the mean point

A1

Answer: 1.5, 0.75, 5.25

4 marks

(c) The uncertainty in predictions increases with distance from $\bar{x}$

E1

Beyond the data this distance is largest, and the pattern may also change

E1

Answer: Uncertainty grows with $|x - \text{mean}|$

2 marks

28 (9 marks)

(a) Extrapolation: the data cover 20 000 to 40 000 but the claim concerns millions

E1

Causation: correlation does not prove that population growth causes crime

E1

A small sample of 5 towns, and $r = 0.78$ may not be significant

E1

Answer: Extrapolation, causation, small sample

3 marks

(b) $0.78 < 0.8054$

M1

Not significant: insufficient evidence of positive correlation

A1

Answer: Not significant

2 marks

(c) Reduction of 1 000 000 people = 1000 thousands

M1

$1000 \times 4.2 = 4200$ incidents, so the arithmetic is right but the model is invalid there

A1

Answer: 4200, but invalid

2 marks

(d) Use many more towns covering a wider range of populations, with sizes similar to the cities of interest

E1

Control for other variables such as income and age profile

E1

Answer: More data, wider range, control variables

2 marks

29 (8 marks)

(a) $b = 10$

M1 A1

$y = -5 + 10x$

A1

Answer: $y = -5 + 10x$

3 marks

(b) $100 = -5 + 10t$

M1

$t = 10.5$ years

A1

A percentage cannot exceed 100; usage will level off, so the linear model must fail

E1

Answer: 10.5 years

3 marks

(c) An S-shaped (logistic) curve levelling off below 100%

E1

Only use the linear model within or near the observed range of t (about 1 to 7)

E1

Answer: S-shaped; restrict range

2 marks

30 (10 marks)

(a) $\ln y = 0.833, 1.686, 2.485, 3.35, 4.111, 4.942$

M1

$\ln y = 0.8524 + 0.4097t$

M1 A1

$a = 2.345, b = 1.5064$

A1

Answer: $a = 2.345, b = 1.5064$

4 marks

(b) Doubling time = $\frac{\ln 2}{\ln b} = \frac{0.6931}{0.4097} = 1.692$

M1 A1

$t = 20$: $y = 8495.2946$

A1

Answer: 8495.295 million

3 marks

(c) $t = 20$ is far beyond the data (extrapolation by 10 years); physical limits may stop the

E1

exponential growth

The data are curved, so a linear model for y on t would be poor

Taking logs linearises the data, giving a much better fit within the range

Answer: Unreliable extrapolation

E1

E1

3 marks